Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

Computational estimation of cell types and their dynamics in health and disease

Cumulative dissertation

to obtain the degree of
Doctor rerum naturalium (Dr. rer. nat.) at
The Faculty of Mathematics, Informatics, and Natural Sciences
Department of Informatics
University of Hamburg

submitted by

Robin Khatri

2025

Mitglieder der Prüfungskommission

Vorsitzender: Prof. Dr. Sören Laue
Stv. Vorsitzender: Prof. Dr. Fabian Kern

Gutachter:
Prof. Dr. Stefan Bonn
Prof. Dr. Jan Baumbach
Prof. Dr. Tim Beißbarth

Datum der Disputation:
16.07.2025

ACKNOWLEDGEMENTS

I am fortunate and grateful to have Stefan as my direct supervisor at the Institute of Medical Systems Biology. I learned from him about different aspects of machine learning in practice and systems biology. It has been a pleasure and a great learning experience in a productive and fun working environment. During my doctoral work, I had the opportunity to work with collaborators from the institute and from different labs, each of whom has taught me new things. I am thankful to Prof. Dr. Ulf Panzer, Prof. Dr. Christian Krebs and PD Dr. Franz Ricklefs, Dr. Yu Zhao and Dr. Sonja Hänzelmann to name a few. I would also like to thank my co-supervisor Prof. Dr. Jan Baumbach for support with my PhD administration.

I am glad to have had many interesting collaborations and conversations with colleagues from the Institute of Medical Systems Biology. In particular, I express my gratitude to Sabine, Fabian, Cedric, Pierre, Tianran, Darius, Fabian W, Zeba, Nico, Manuela, Tiphaine, Malte, Sven, Marina, Patrick and Vadim. You made my experience so much better.

LIST OF PUBLICATIONS

PEER-REVIEWED PUBLICATIONS

sorted by date of publication - newest first.

Schaub DP*, Yousefi B*, Kaiser N, **Khatri R**, Puelles VG, Krebs CF, Panzer U, Bonn S. PCA-based spatial domain identification with state-of-the-art performance. **Bioinformatics**. 2024 Dec 26;41(1):btaf005. doi: 10.1093/bioinformatics/btaf005. PMID: 39775801; PMCID: PMC11761416.

Gies SE*, Hänzelmann S*, Kylies D, Lassé M, Lagies S, Hausmann F, **Khatri R**, Zolotarev N, Poets M, Zhang T, Demir F, Billing AM, Quaas J, Meister E, Engesser J, Mühlig AK, Lu S, Liu S, Chilla S, Edenhofer I, Czogalla J, Braun F, Kammerer B, Puelles VG, Bonn S, Rinschen MM, Lindenmeyer M[#], Huber TB[#]. Optimized protocol for the multiomics processing of cryopreserved human kidney tissue. **Am J Physiol Renal Physiol.** 2024 Nov 1;327(5):F822-F844. doi: 10.1152/ajprenal.00404.2023. Epub 2024 Oct 3. PMID: 39361723.

Engesser J*, **Khatri R***, Schaub DP*, Zhao Y, Paust HJ, Sultana Z, Asada N, Riedel JH, Sivayoganathan V, Peters A, Kaffke A, Jauch-Speer SL, Goldbeck-Strieder T, Puelles VG, Wenzel UO, Steinmetz OM, Hoxha E, Turner JE, Mittrücker HW, Wiech T, Huber TB, Bonn S[#], Krebs CF[#], Panzer Ulf[#]. Immune profiling-based targeting of pathogenic T cells with ustekinumab in ANCA-associated glomerulonephritis. **Nat Commun.** 2024 Sep 19;15(1):8220. doi: 10.1038/s41467-024-52525-w. PMID: 39300109; PMCID: PMC11413367.

Caldi Gomes L*, Hänzelmann S*, Hausmann F, **Khatri R**, Oller S, Parvaz M, Tzeplaeff L, Pasetto L, Gebelin M, Ebbing M, Holzapfel C, Columbro SF, Scozzari S, Knöferle J, Cordts I, Demleitner AF, Deschauer M, Dufke C, Sturm M, Zhou Q, Zelina P, Sudria-Lopez E, Haack TB, Streb S, Kuzma-Kozakiewicz M, Edbauer D, Pasterkamp RJ, Laczko E, Rehrauer H, Schlapbach R, Carapito C, Bonetto V, Bonn S[#], Lingor P[#]. Multiomic ALS signatures highlight subclusters and sex differences suggesting the MAPK pathway as therapeutic target. **Nat Commun.** 2024 Jun 7;15(1):4893. doi: 10.1038/s41467-024-49196-y. PMID: 38849340; PMCID: PMC11161513.

Drexler R*, **Khatri R***, Sauvigny T, Mohme M, Maire CL, Ryba A, Zghaibeh Y, Dührsen

L, Salviano-Silva A, Lamszus K, Westphal M, Gempt J, Wefers AK, Neumann JE, Bode H, Hausmann F, Huber TB, Bonn S, Jütten K, Delev D, Weber KJ, Harter PN, Onken J, Vajkoczy P, Capper D, Wiestler B, Weller M, Snijder B, Buck A, Weiss T, Göller PC, Sahm F, Menstel JA, Zimmer DN, Keough MB, Ni L, Monje M, Silverbush D, Hovestadt V, Suvà ML, Krishna S, Hervey-Jumper SL, Schüller U, Heiland DH[#], Hänzelmann S[#], Ricklefs FL[#]. A prognostic neural epigenetic signature in high-grade glioma. **Nat Med.** 2024 Jun;30(6):1622-1635. doi: 10.1038/s41591-024-02969-w. Epub 2024 May 17. PMID: 38760585; PMCID: PMC11186787.

Khatri R, Machart P, Bonn S. DISSECT: deep semi-supervised consistency regularization for accurate cell type fraction and gene expression estimation. **Genome Biol.** 2024 Apr 30;25(1):112. doi: 10.1186/s13059-024-03251-5. PMID: 38689377; PMCID: PMC11061925.

Er-Lukowiak M*, Hänzelmann S*, Rothe M, Moamenpour DT, Hausmann F, **Khatri R**, Hansen C, Boldt J, Bärreiter VA, Honecker B, Bea A, Groneberg M, Fehling H, Marggraff C, Cadar D, Bonn S, Sellau J, Lotter H. Testosterone affects type I/type II interferon response of neutrophils during hepatic amebiasis. **Front Immunol.** 2023 Dec 21;14:1279245. doi: 10.3389/fimmu.2023.1279245. PMID: 38179044; PMCID: PMC10764495.

Hausmann F*, Ergen C*, **Khatri R**, Marouf M, Hänzelmann S, Gagliani N, Huber S, Machart P, Bonn S. DISCERN: deep single-cell expression reconstruction for improved cell clustering and cell subtype and state detection.**Genome Biol.** 2023 Sep 20;24(1):212. doi: 10.1186/s13059-023-03049-x. PMID: 37730638; PMCID: PMC10510283.

Drexler R, Göttsche J, Sauvigny T, Schüller U, **Khatri R**, Hausmann F, Hänzelmann S, Huber TB, Bonn S, Heiland DH, Delev D, Venkataramani V, Winkler F, Weller J, Zeyen T, Herrlinger U, Gempt J, Ricklefs FL, Dührsen L. Targeted anticonvulsive treatment of IDH-wildtype glioblastoma based on DNA methylation subclasses. **Neuro-Oncol.** 2023 May 4;25(5):1006-1008. doi: 10.1093/neuonc/noad014. PMID: 36860191; PMCID: PMC10158071.

Drexler R, Sauvigny T, Schüller U, Eckhardt A, Maire CL, **Khatri R**, Hausmann F, Hänzelmann S, Huber TB, Bonn S, Bode H, Lamszus K, Westphal M, Dührsen L, Ricklefs FL. Epigenetic profiling reveals a strong association between lack of 5-ALA fluorescence and EGFR amplification in IDH-wildtype glioblastoma. **Neurooncol Pract.** 2023 May 2;10(5):462-471. doi: 10.1093/nop/npad025. PMID: 37720395; PMCID: PMC10502788.

Khatri, R, and Bonn, S. Uncertainty Estimation for Single-cell Label Transfer. In Conformal and Probabilistic Prediction with Applications, **PMLR** 2022; 179:109-128.

PREPRINTS

sorted by date of publication - newest first.

Sultana Z*, **Khatri R***, Yousefi B*, Shaikh N, Jauch-Speer SL, Schaub DP, Engesser J, Hellmig M, Hube AL, Sivayoganathan V, Borchers A, Peters A, Kaffke A, Paust HJ, Goldbeck-Strieder T, Wenzel UO, Puelles V, Hoxha E, Wiech T, Huber TB, Panzer U[#], Bonn S[#], Krebs CF[#]. Spatio-temporal interaction of immune and renal cells determines glomerular crescent formation in autoimmune kidney disease. **bioRxiv** 2024.12.18.629206; doi: https://doi.org/10.1101/2024.12.18.629206.

Westhaeusser F*, Fuhlert P*, Dietrich E, Lennartz M, **Khatri R**, Kaiser N, Röbeck P, Bülow R, von Stillfried S, Witte A, Ladjevardi S, Drotte A, Severgardh P, Baumbach J, Puelles VG, Häggman M, Brehler M, Boor P, Walhagen P, Dragomir A, Busch C, Graefen M, Bengtsson E, Sauter G[#], Zimmermann M[#], Bonn S[#]. Robust, credible, and interpretable AI-based histopathological prostate cancer grading. **medRxiv** 2024.07.09.24310082; doi: https://doi.org/10.1101/2024.07.09.24310082.

*,[#] equal contribution

ABBREVIATIONS

AE	Autoencoder
AI	Artificial intelligence
AMPA	α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid
ANCA	Antineutrophil cytoplasmic antibody
ANCA-GN	ANCA-associated glomerulonephritis
APC	Astrocyte precursor cell
BDNF	Brain-derived neurotrophic factor
CAR T-cell	Chimeric Antigen Receptor T-cell
cDNA	complementary DNA
cfDNA	Cell free DNA
CVAE	Conditional variational autoencoder
DNA	Deoxyribonucleic acid
DNAm	DNA methylation
EGFR	Epidermal Growth Factor Receptor
evDNA	Extracellular vesicle-bound DNA
GBM	Glioblastoma
IDH	Isocitrate Dehydrogenase
IL12	Interleukin 12
IL23	Interleukin 23
JL	Johnson-Lindenstrauss
ML	Machine learning
MLP	Multilayer perceptron
mRNA	Messenger RNA
NKT	Natural killer T
NNLS	Non-negative least squares
NPC	Neural progenitor cell
OPC	Oligodendrocyte-precursor cell
PDGFRA	Platelet derived growth Factor Receptor Alpha
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RTKI	Receptor Tyrosine Kinase I
RTKII	Receptor Tyrosine Kinase II

scRNA-seq	Single cell RNA sequencing
SVM	Support vector machine
Tfh	T follicular helper
Th1	T helper 1
Th17	T helper 17
Treg	T Regulatory
VAE	Variational autoencoder

SUMMARY

Changes in cell type composition are fundamental to understanding human disease mechanisms. While single-cell omics technologies enable unprecedented resolution in cellular profiling, their widespread clinical application is limited by technical biases and cost constraints. Measurements on bulk tissue specimens, though more robust and cost-effective, lack cell-type resolution. This creates a need for computational methods that can bridge this gap. At its core, cell type deconvolution represents a semi-blind source separation problem, where the goal is to estimate both the mixing proportions and source signals from mixture measurements, given any partial information about the sources from reference data.

This dissertation includes DISSECT, a novel deep semi-supervised learning algorithm for robust cell type deconvolution. DISSECT addresses key limitations in existing approaches by integrating information from both single-cell references and bulk data, effectively handling domain shifts between reference and target datasets. Through comprehensive benchmarking across multiple experimental settings and modalities (including bulk RNA sequencing (RNA-seq), proteomics, and spatial transcriptomics), we demonstrate DISSECT's superior performance in predicting both cell type proportions and cell type-specific expression profiles, with reduced dependency on reference selection.

We used cell type deconvolution to study two distinct diseases: antineutrophil cytoplasmic antibody-associated glomerulonephritis (ANCA-GN) and glioblastoma (GBM). In ANCA-GN, deconvolution of single-cell and spatial transcriptomics data from 34 patients and 8 controls revealed specific T helper cell accumulation patterns associated with inflammation. Computational drug prediction based on this information identified ustekinumab as a potential therapeutic agent, which showed promising results in four patients with poor prognosis under standard treatment. In glioblastoma, we used cell type deconvolution to analyze DNA methylation patterns across multiple cohorts of GBM patients, identifying two distinct and temporally stable GBM groups associated with better prognostic value than established molecular subtypes, particularly in predicting response to surgical intervention.

This dissertation makes several key contributions to bioinformatics, immunology, and neuroscience: (1) a cell type deconvolution framework that advances the state-of-the-

art in source separation for biological data, (2) an integrative analysis of immune cell type-specific signals to guide therapeutic decisions in ANCA-GN, and (3) identification of clinically relevant and stable GBM subgroups based on deconvolved cell type-specific signals.

ZUSAMMENFASSUNG

Veränderungen in der Zusammensetzung von Zelltypen sind für das Verständnis menschlicher Krankheitsmechanismen von grundlegender Bedeutung. Während Einzelzell-Omics-Technologien eine beispiellose Auflösung bei der zellulären Profilierung ermöglichen, ist ihre breite klinische Anwendung durch technische und finanzielle Limitationen begrenzt. Messungen an Bulk-Gewebeproben sind zwar robuster und kostengünstiger, weisen jedoch keine zelltypspezifische Auflösung auf. Daraus ergibt sich die Notwendigkeit computergestützter Methoden, die diese Lücke schließen können. Im Kern stellt die Zelltyp-Dekonvolution ein semi-blindes Source-Separation-Problem dar, bei dem das Ziel darin besteht, sowohl die Mischungsverhältnisse als auch die Quellsignale aus Mischungsmessungen zu schätzen, basierend auf partiellen Informationen über die Quellen aus Referenzdaten.

Diese Dissertation beinhaltet DISSECT, einen neuartigen semi-supervised Learning-Algorithmus für robuste Zelltyp-Dekonvolution. DISSECT adressiert wesentliche Limitationen bestehender Ansätze durch die Integration von Informationen sowohl aus Einzelzell-Referenzen als auch aus Bulk-Daten und bewältigt dabei effektiv Domain-Shifts zwischen Referenz- und Zieldatensätzen. Durch umfassendes Benchmarking über verschiedene experimentelle Ansätze und Modalitäten (einschließlich Bulk-RNA-Sequenzierung, Proteomik und räumliche Transkriptomik) zeigen wir die überlegene Leistung von DISSECT bei der Vorhersage sowohl von Zelltyp-Proportionen als auch zelltypspezifischen Expressionsprofilen, bei reduzierter Abhängigkeit von der Referenzauswahl.

Mit Hilfe der Zelltyp-Dekonvolution haben wir zwei unterschiedliche Krankheiten untersucht: die mit antineutrophilen zytoplasmatischen Antikörpern assoziierte Glomerulonephritis (ANCA-GN) und das Glioblastom (GBM). Bei ANCA-GN ergab die Dekonvolution von Einzelzell- und räumlichen Transkriptomikdaten von 34 Patienten und 8 Kontrollpersonen spezifische T-Helferzell-Akkumulationsmuster, die mit Entzündungen einhergehen. Eine auf diesen Informationen basierende computergestützte Arzneimittelvorhersage identifizierte Ustekinumab als potenziellen therapeutischen Wirkstoff, der bei vier Patienten mit schlechter Prognose unter Standardbehandlung vielversprechende Ergebnisse zeigte. Beim Glioblastom nutzten wir die Dekonvolution von Zelltypen zur Analyse von DNA-Methylierungsmustern in mehreren GBM-Patientenkohorten und iden-

tifizierten zwei unterschiedliche und zeitlich stabile GBM-Gruppen, die einen besseren prognostischen Wert haben als etablierte molekulare Subtypen, insbesondere bei der Vorhersage des Ansprechens auf einen chirurgischen Eingriff.

Diese Dissertation leistet mehrere wichtige Beiträge zur Bioinformatik, Immunologie und Neurowissenschaft: (1) ein Zelltyp-Dekonvolutions-Framework, das den Stand der Technik bei der Quellentrennung für biologische Daten vorantreibt, (2) eine integrative Analyse von immunzellenspezifischen Signalen, um therapeutische Entscheidungen bei ANCA-GN zu treffen, und (3) die Identifizierung klinisch relevanter und stabiler GBM-Untergruppen auf der Grundlage dekonvolvierter zelltypspezifischer Signale.

CONTENTS

PREFACE

This dissertation follows a cumulative structure and contains the following three publications appearing in peer-reviewed journals.

Publication 1: Khatri R, Machart P, Bonn S. DISSECT: deep semi-supervised consistency regularization for accurate cell type fraction and gene expression estimation. **Genome Biol.** 2024 Apr 30;25(1):112. doi: 10.1186/s13059-024-03251-5. PMID: 38689377; PMCID: PMC11061925.

Publication 2: Engesser J*, **Khatri R***, Schaub DP*, Zhao Y, Paust HJ, Sultana Z, Asada N, Riedel JH, Sivayoganathan V, Peters A, Kaffke A, Jauch-Speer SL, Goldbeck-Strieder T, Puelles VG, Wenzel UO, Steinmetz OM, Hoxha E, Turner JE, Mittrücker HW, Wiech T, Huber TB, Bonn S[#], Krebs CF[#], Panzer Ulf[#]. Immune profiling-based targeting of pathogenic T cells with ustekinumab in ANCA-associated glomerulonephritis. **Nat Commun.** 2024 Sep 19;15(1):8220. doi: 10.1038/s41467-024-52525-w. PMID: 39300109; PMCID: PMC11413367.

Publication 3: Drexler R*, **Khatri R***, Sauvigny T, Mohme M, Maire CL, Ryba A, Zghaibeh Y, Dührsen L, Salviano-Silva A, Lamszus K, Westphal M, Gempt J, Wefers AK, Neumann JE, Bode H, Hausmann F, Huber TB, Bonn S, Jütten K, Delev D, Weber KJ, Harter PN, Onken J, Vajkoczy P, Capper D, Wiestler B, Weller M, Snijder B, Buck A, Weiss T, Göller PC, Sahm F, Menstel JA, Zimmer DN, Keough MB, Ni L, Monje M, Silverbush D, Hovestadt V, Suvà ML, Krishna S, Hervey-Jumper SL, Schüller U, Heiland DH[#], Hänzelmann S[#], Ricklefs FL[#]. A prognostic neural epigenetic signature in high-grade glioma. **Nat Med.** 2024 Jun;30(6):1622-1635. doi: 10.1038/s41591-024-02969-w. Epub 2024 May 17. PMID: 38760585; PMCID: PMC11186787.

For clarity, the structure of the organization of this dissertation is presented below.

Chapter 1 provides the necessary background and introduces the main topics explored in subsequent chapters. In particular, sections 1.1-1.3 introduce the basic biological background and terms required to understand the data and publications. Sections 1.4-1.7 introduce the computational methods, including analysis pipelines, deep neural networks, cell type deconvolution state-of-the-art and the proposed novel DISSECT framework. Section 1.8 presents a background on the disease and literature corresponding to the

work herein. Section 1.9 lists the research goals this dissertation addresses. Chapters 2-4 present the aforementioned publications. Each publication includes supplementary content (tables, figures, and/or notes) which appears after the main content of the respective manuscripts in chapters 2-4. Finally, chapter 5 summarizes the findings of the works presented in this dissertation, and discusses its results, contributions, and potential impact.

In each of the publications, I am a first author and have made significant contributions to project conceptualization, analysis, drafting, editing, and revising the manuscripts. The particular contributions for each of the publications are presented as appendix A.

Robin Khatri
Hamburg, 2025

1

INTRODUCTION

1.1. CELLS AND CELL TYPES

Cells are the fundamental units of living organisms, each containing a complex array of structures and molecules that work together to maintain biological functions. Cells contain various structures including a nucleus bound by membrane (Figure 1.1a). At the heart of cellular operations lies Deoxyribonucleic acid (DNA), the genetic blueprint stored within the nucleus of cells. This double-helix molecule contains the instructions for building and operating the cell (Watson and Crick, 1953). When these instructions are carried out, a process called transcription begins which results in RNA molecules (Figure 1.1b). This process of transcription can be blocked when methyl groups (CH3) are added to a DNA molecule (Figure 1.1c). During transcription, enzymes unwind a section of DNA and use it as a template to create a complementary strand of Ribonucleic acid (RNA) (Kornberg, 2007). This RNA molecule, specifically messenger RNA (mRNA), then exits the nucleus and serves as a mobile set of instructions for protein synthesis. Other types of RNA, such as transfer RNA and ribosomal RNA, also play crucial roles in translating the genetic code into functional proteins (Crick, 1970). Between species and among organs, the cell composition and biological attributes differ, and understanding the architecture and functions of an organism requires understanding the composition and function of constituting cells (Zeng, 2022). Based on certain properties, cells can be categorized into homogenous groups termed cell types. While there is no consensus on which qualities are necessary and mandatory to define a cell type, the recent efforts focus on defining cell types with respect to shared development history and molecular features such as transcriptome and epigenome (Fleck et al., 2023). Based on the data compiled from over 1500 publications, there are an estimated >28 trillion cells in adult humans with around 500 major cell types across 60 tissue systems such as the brain and kidney (Hatton et al., 2023). The cell types undergo changes in both numbers and their molecular features in different conditions of the body such as during a disease (Jagadeesh et al., 2022; Ju et al., 2013). The primary focus, here, is to understand the changes in cell type composition and molecular features in glioblastoma (GBM) and Antineutrophil cytoplasmic antibody-associated glomerulonephritis (ANCA-GN) (introduced in Section 1.8). Major cell types of the kidney and brain in a healthy state are presented in Figure 1.1d.

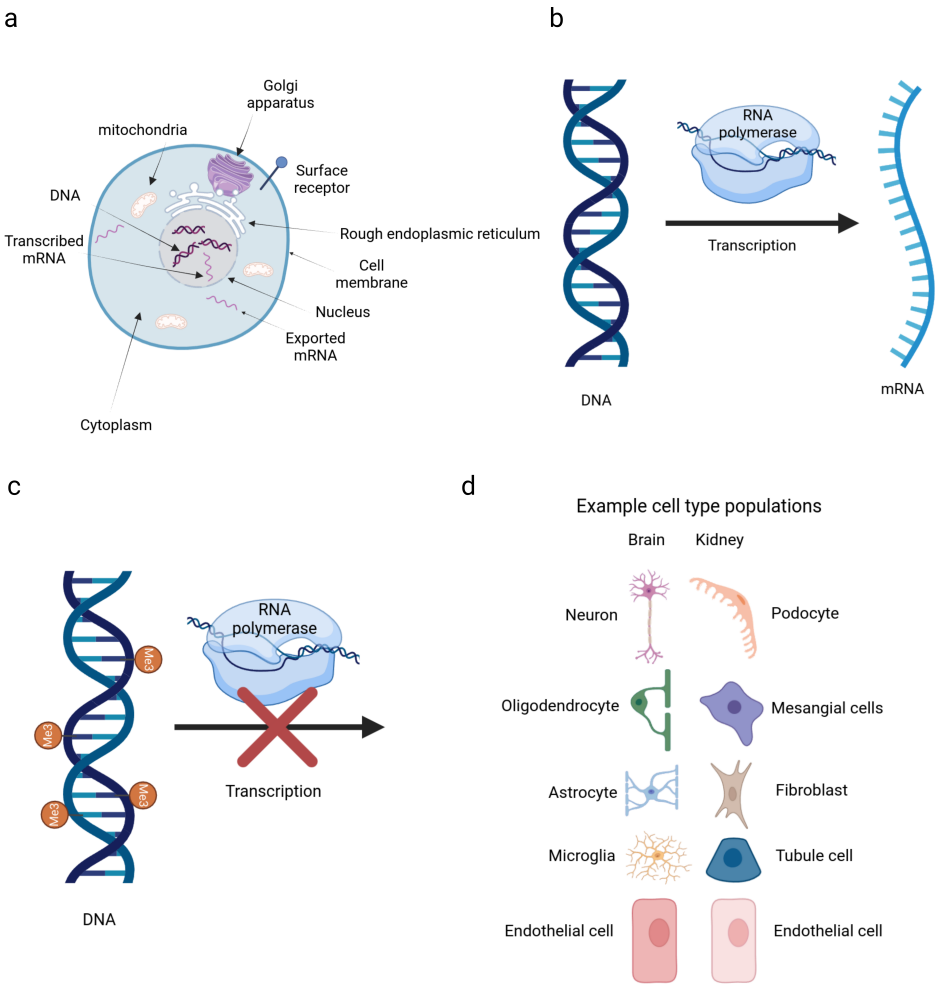


Figure 1.1 | a. Schematic illustration of a cell highlighting key organelles and processes. The cell nucleus contains DNA, which undergoes transcription to produce RNA. Mitochondria, the powerhouses of the cell, generate ATP through cellular respiration. The cytoplasm contains various organelles and newly transcribed mRNA. Exported mRNA can be seen outside the nucleus. They move to the rough endoplasmic reticulum (ER). The rough ER, studded with ribosomes, is involved in protein synthesis and modification. The Golgi apparatus further processes and packages proteins for secretion or cellular use. The cell is enclosed by a cell membrane featuring surface receptors, which are proteins that bind specific molecules and initiate cellular responses. b. Illustration of RNA transcription from a DNA template. c. Illustration of DNA methylation. DNA methylation refers to the process in which methyl groups (CH3), depicted by ME3 in the figure, are added to cytosine bases in the DNA molecule, often at CpG sites. When these methyl groups attach to a gene promoter, which is the region of DNA that initiates transcription of a particular gene, DNA methylation often restricts transcription. d. Some examples of the established cell type populations in the brain and kidney - two tissues used extensively in publications part of this dissertation. Created with BioRender.com.

1.2. RNA SEQUENCING (RNA-SEQ)

RNA-seq is a broad term for a range of technologies to measure the abundance of transcripts of each gene from a biological specimen. Multiple technologies exist for different use cases. RNA sequencing can be done either on individual cells or multiple cells such as parts or on entire organ systems. Here, to distinguish between different RNA-seq measurements, we use three terms: 1. Bulk RNA refers to RNA measurements at a tissue level, 2. Single cell RNA-seq (scRNA-seq) refers to RNA-seq at a single-cell level, and 3. Spatial transcriptomics refers to RNA-sequencing at a tissue level while preserving the locations of each measurement. Each of these are briefly described below.

1.2.1. BULK RNA-SEQ

Bulk RNA-seq is a powerful technique that uses next-generation sequencing to comprehensively profile the transcriptome. It involves extracting RNA from a sample, converting it to cDNA, fragmenting the cDNA, and sequencing the fragments using high-throughput sequencing platforms (Kukurba and Montgomery, 2015). The resulting sequencing reads are then aligned to a reference genome or transcriptome to quantify gene expression levels. Bulk RNA-seq can detect novel transcripts, splice variants, and non-coding RNAs, and has a wide dynamic range for quantifying gene expression (Kukurba and Montgomery, 2015; Rao et al., 2019).

1.2.2. SCRNA-SEQ

scRNA-seq allows for the dissection of complex tissues by profiling individual cells and providing the transcriptional states of each cell (Wolfien et al., 2021). This allows for a better understanding of cellular heterogeneity and composition (Papalexi and Satija, 2018). Several platforms exist for scRNA-seq. Two main ones with complementary benefits are introduced below.

10X CHROMIUM

The 10x Chromium platform available from 10x Genomics is a droplet-based scRNA-seq method that enables the profiling of thousands of cells in a single experiment (Wang et al., 2021). It uses microfluidics to encapsulate individual cells into nanoliter-scale droplets, along with barcoded beads for cell-specific labeling of mRNA molecules. The 10x Chromium system offers two main library preparation chemistries. 3' gene expression and 5' gene expression (Hsu et al.). Both of these capture a particular portion of

transcripts (as indicated by their names). 3' and 5' refer to the terminal regions of the RNA molecule that play crucial roles in its stability, processing, and translation (Wang et al., 2021).

SMART-SEQ2

Smart-seq2 is a plate-based scRNA-seq method from Illumina that provides full-length mRNA sequencing as opposed to specific portions (Wang et al., 2021). It involves the isolation of single cells into individual wells, followed by cell lysis and reverse transcription to generate cDNA. The cDNA is then amplified and sequenced.

Smart-seq offers several advantages, including full-length mRNA sequencing that can identify novel isoforms, and a higher sequencing depth per cell compared to 10x Chromium, allowing for the detection of lowly expressed genes. However, Smart-seq has a lower throughput compared to 10x Chromium and requires more input material per cell, and can miss rare populations from being captured (Wang et al., 2021).

1.2.3. SPATIAL TRANSCRIPTOMICS

Spatial transcriptomics is a new technology that quantifies gene expression of a tissue in a spatial context. One of the leading platforms in the field is the Visium from 10x Genomics, which allows for the comprehensive profiling of the transcriptome on circular grids of spots of around 55 µm spread across entire tissue sections. This method captures spatial information by utilizing a slide with a grid of oligonucleotide probes that hybridize to mRNA within tissue samples (Moses and Pachter, 2022). By mapping the location of gene expression, it is possible to analyze the gene expression patterns over the tissue which is critical for understanding complex biological processes such as development and disease progression (Zhou et al., 2023).

The Visium platform is particularly notable for its versatility, as it can analyze both fresh-frozen and formalin-fixed, paraffin-embedded (FFPE) tissues (Williams et al., 2022). This flexibility is essential for utilizing archival samples in research, which often contain valuable information about disease states. The technology supports whole-tissue section profiling, eliminating the need for researchers to pre-select regions of interest. This capability allows for a more comprehensive view of the tissue landscape, capturing a wide array of cellular interactions and spatial gene expression patterns. Additionally, Visium can achieve a high cellular resolution, typically averaging 1 to 10 cells per spot, depending on the tissue type, which enhances the granularity of the data obtained. However, in practice, cell type level analysis is not possible without the usage of cell type deconvolution

approaches to measure cell type fraction and gene expression estimation in each spot (Li et al., 2023). Newer technologies like Xenium from 10x genomics are rapidly developing and in the future, it may be possible to utilize them for large-scale studies, possibly in combination with Visium (Janesick et al., 2023).

1.3. DNA METHYLATION

DNA methylation, illustrated in Figure 1.1c, is an epigenetic modification that plays a vital role in regulating gene expression, genomic imprinting, X-chromosome inactivation, and maintaining genome stability. This process involves the addition of a methyl group to the 5' position of cytosine residues, predominantly occurring at CpG sites (regions of DNA where a C, cytosine nucleotide occurs before a G, guanine nucleotide). The dynamic nature of DNA methylation patterns throughout development and in response to environmental factors has made it a subject of intense research in various fields, including cancer biology, neuroscience, and developmental biology (Tucker, 2001; Jones, 2012). High-throughput technologies have revolutionized our ability to study DNA methylation patterns on a genome-wide scale. Among these technologies, DNA methylation microarrays are powerful tools for interrogating methylation status across pre-defined hundreds of thousands of CpG sites simultaneously. Two prominent platforms in this field are the Illumina Infinium HumanMethylation450 BeadChip (450k array) and its successor, the Illumina Infinium MethylationEPIC BeadChip (EPIC array).

The 450k array was designed to assess the methylation status of over 450,000 CpG sites across the human genome. This array covers 99% of RefSeq genes, with an average of 17 CpG sites per gene region distributed across the promoter, 5'UTR, first exon, gene body, and 3'UTR (Bibikova et al., 2011; Price et al., 2013). Additionally, it includes CpG sites in CpG islands, shores, and shelves, as well as miRNA promoter regions (Bibikova et al., 2011). The 450k array utilizes two types of probe designs: Infinium I and Infinium II. Infinium I probes use two bead types per CpG locus, one for methylated and one for unmethylated states, while Infinium II probes use a single bead type with two different color channels to distinguish between methylated and unmethylated states.

While 450k arrays covered a large amount of CpG sites, it still covers only a small portion of CpG sites in the human genome (28 million) (Babenko et al., 2017). EPIC array, which substantially expanded the coverage to over 850,000 CpG sites, maintains backward compatibility with the 450k array, covering more than 90% of the CpG sites present in its predecessor (Fernandez-Jimenez et al., 2019). The additional probes on the EPIC array target enhancer regions identified by the ENCODE and FANTOM5 projects, as well as open chromatin regions and DNase hypersensitive sites (Moran et al., 2016). This expanded coverage allows for a more comprehensive analysis of regulatory regions and provides greater insight into the functional relevance of DNA methylation patterns.

Both the 450k and EPIC arrays have been widely adopted in large-scale epigenome-wide association studies (EWAS), enabling researchers to identify differentially methylated re-

gions associated with various phenotypes, diseases, and environmental exposures. These arrays have contributed significantly to our understanding of the role of DNA methylation in cancer progression, neurodegenerative disorders, and aging, among other biological processes (Teschendorff and Relton, 2018).

Despite their widespread use and valuable contributions to the field, DNA methylation microarrays have some limitations. They provide a targeted approach, focusing on pre-selected CpG sites, which may miss potentially important methylation changes in unprobed regions. Additionally, the arrays are designed based on the human reference genome, which may not capture all genetic variants present in diverse populations such as murine models as a model organism (Canales and Walz, 2019). Furthermore, these arrays typically require a relatively large amount of input DNA, which can be a limiting factor when working with rare cell populations or clinical samples. To address some of these limitations and to gain insights into cellular heterogeneity, single-cell DNA methylation profiling techniques have been developed in recent years. These methods aim to capture methylation patterns at the individual cell level, providing unprecedented resolution for studying epigenetic heterogeneity within complex tissues and cell populations. Single-cell bisulfite sequencing (scBS-seq) and single-cell reduced representation bisulfite sequencing (scRRBS) are among the pioneering techniques in this field (Smallwood et al., 2014; Guo et al., 2013). However, single-cell DNA methylation profiling faces several challenges. The primary obstacle is the sparse nature of the data due to the limited amount of DNA available from a single cell and the destructive nature of bisulfite conversion. This sparsity results in low genomic coverage per cell, typically ranging from 1% to 10% of CpG sites. Another challenge is the high technical variability introduced during the amplification of small amounts of DNA, which can lead to biases and increased noise in the data. Additionally, the cost and computational resources required for single-cell methylation analysis are significantly higher compared to bulk sample analysis, limiting the number of cells that can be profiled in a given experiment. Despite these challenges, single-cell DNA methylation profiling has already provided valuable insights into cellular heterogeneity in various biological contexts, including embryonic development, tumor evolution, and brain function. As the field continues to advance, new methodologies and analytical approaches are being developed to improve genomic coverage, reduce technical biases, and enhance the integration of single-cell methylation data with other omics modalities (Luo et al., 2017).

1.4. METHODS FOR THE THE ANALYSIS OF RNA-SEQUENCING AND DNA METHYLATION DATA

1.4.1. COMPUTATIONAL ANALYSIS OF SCRNA-SEQ

This introduction outlines key methodological approaches used in the analysis of scRNA-seq data.

The first step in the analysis of scRNA-seq data is processing of the raw data obtained from sequencing (e.g. from 10x Chromium as introduced in section 1.2.2). The processing steps involve data filtering and other quality control steps to remove potential doublets (*i.e.* where more than one cell is wrongly barcoded as one cell) and cells with low quality, as measured by the number of expressed genes (Luecken and Theis, 2019), and dimensionality reduction techniques such as PCA to visualize and explore the data, resulting in a smaller set of principal components (Hwang et al., 2018). Following PCA, non-linear dimensionality reduction methods such as t-distributed Stochastic Neighbor Embedding (t-SNE) or Uniform Manifold Approximation and Projection (UMAP) are applied to further reduce the data to two or three dimensions for visualization (Van der Maaten and Hinton, 2008; McInnes et al., 2018). These techniques help identify clusters of cells with similar expression profiles, which often correspond to distinct cell types.

Clustering algorithms play a central role in identifying cell types from scRNA-seq data. Unsupervised clustering methods, such as k-means, hierarchical clustering, or graph-based approaches like Louvain or Leiden algorithms, are used to group cells based on their gene expression similarities (Kiselev et al., 2019). The choice of clustering algorithm and parameters can significantly impact the results, and we often need to experiment with different approaches to find the most biologically meaningful clustering solution.

Once clusters are identified, the next critical step is to annotate these clusters with cell type labels. This process typically involves examining the expression of known marker genes for different cell types. Differential expression analysis between clusters can help identify genes that are uniquely or preferentially expressed in each cluster, aiding in their biological interpretation (Love et al., 2014). Additionally, automated annotation tools that use existing gene expression databases and ontologies, such as SingleR or Garnett, can assist in assigning cell type labels to clusters (Aran et al., 2019; Pliner et al., 2019).

Recent advancements in scRNA-seq analysis methods have focused on addressing the challenges of batch effects and data integration. Techniques such as mutual nearest neighbors (MNN) correction, Harmony, or Seurat's integration methods allow researchers

to combine multiple scRNA-seq datasets, enabling more robust cell type identification across different experimental conditions or time points (Haghverdi et al., 2018; Korsunsky et al., 2019; Stuart et al., 2019).

1.4.2. Computational analysis of bulk RNA

The analysis of bulk RNA-seq data typically begins with quality control of raw sequencing reads, followed by alignment to a reference genome or transcriptome assembly. After quantification of gene or transcript expression levels, normalization techniques are applied to account for technical biases and enable comparisons across samples. Differential expression analysis is then performed to identify genes that are significantly up-or down-regulated between experimental conditions (Love et al., 2014). While these standard analysis steps provide valuable insights into overall gene expression changes, they do not directly address the cellular heterogeneity inherent in most biological samples. To overcome this limitation, cell type deconvolution, the estimation of cell type fractions and cell type-specific gene expression, is necessary. Cell type deconvolution is introduced in section 1.6.

1.5. DEEP NEURAL NETWORKS

In this section, we introduce the multilayer perceptrons (MLP), autoencoders (AE), and conditional variational autoencoders (CVAE) to get a background prerequisite for understanding the DISSECT framework presented in Chapter 2.

1.5.1. MULTILAYER PERCEPTRON (MLP)

Multilayer Perceptrons (MLPs) are a class of artificial neural networks that have played a role in the development of several complex deep learning architectures adapted to do different tasks. As a type of feedforward neural network, MLPs have found applications in various domains, including pattern recognition and function approximation.

The foundation of MLPs can be traced back to the simple perceptron model (Rosenblatt, 1958). However, MLPs extend this concept by introducing multiple layers of interconnected nodes (i.e. going deep), allowing them to learn and represent complex, non-linear relationships between inputs and outputs.

The fundamental structure of an MLP consists of three main components: 1) Input layer, which receives the initial data, (2) Hidden layer(s), which processes the information, and (3) Output layer(s), which produces the final results.

Each layer consists of nodes, also known as neurons or units, which are connected to nodes in adjacent layers. The strength of these connections is represented by weights, which are adjusted during the learning process.

The fundamental operation on a neuron within an MLP can be described mathematically as follows:

$$y = f(\sum_{i=1}^{n} w_i x_i + b),$$

Where:
- y is the output of the neuron
- f is the activation function
- w_i are the weights
- x_i are the inputs

- b is the bias term
- n is the number of inputs

The choice of activation function is crucial in determining the network's ability to learn non-linear relationships. Common activation functions include the sigmoid, hyperbolic tangent (tanh), and rectified linear unit (ReLU) (Ramachandran et al., 2017).

The learning in MLPs is typically achieved through backpropagation, an algorithm that calculates the gradient of the loss function with respect to the network's weights (Rumelhart et al., 1986). This gradient is then used to update the weights iteratively, minimizing the error between the network's predictions and the true values. The backpropagation algorithm can be summarized in the following steps: 1. Forward pass: Input data is propagated through the network to generate predictions. 2. Error calculation: The difference between predictions and true values is computed, such as with mean squared error. 3. Backward pass: The error is propagated backward through the network to calculate gradients. 4. Weight update: Weights are adjusted using an optimization algorithm, such as Adam (Kingma, 2014). It adapts the learning rate for each weight and incorporates concepts of momentum.

The weight update rule for Adam can be expressed as:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \cdot \hat{m}_t,$$

Where:
θ_t is the parameter (weight) at time step t
η is the learning rate
$\hat{m}_t$ is the bias-corrected first-moment estimate
$\hat{v}_t$ is the bias-corrected second-moment estimate
ϵ is a small constant to prevent division by zero

Despite their simplicity compared to more advanced neural network architectures, MLPs have demonstrated remarkable performance in various tasks. They serve as building blocks for more complex models and continue to be relevant in both research and practical applications. Further, a few cell type deconvolution methods have previously adapted MLPs (Menden et al., 2020; Yasumizu et al., 2024).

1.5.2. AUTOENCODER (AE)

Autoencoders are a class of artificial neural networks commonly used in the field of unsupervised learning and dimensionality reduction. A typical autoencoder is designed to learn efficient data representations (encoding) by training the network to ignore signal "noise". It accomplishes this by learning to reconstruct its input at the output layer. The network architecture typically consists of an encoder function $h = f(x)$ and a decoder function that produces a reconstruction $r = g(h)$ from the output of the encoder (h). The aim is to minimize the difference between the input x and its reconstruction r.

The simplest form of an autoencoder is the undercomplete autoencoder, where the hidden layer has a (much) lower dimensionality than the input and output layers. This forces the network to learn a compact representation of the input data. The learning process can be formulated as minimizing a loss function: $L(x, g(f(x)))$, where L is a loss function penalizing $g(f(x))$ for being dissimilar to x, such as the mean squared error.

Mathematically, for an input vector x, a simple encoder layer with the same activation function across all neurons produces a hidden representation h:

$$h = f(x) = s(Wx + b),$$

where W is a weight matrix, b is a bias vector, and s is an activation function applied element-wise. For multi-layer encoders, this basic transformation is applied sequentially across layers, with each layer's output serving as input to the next layer.

The decoder then attempts to reconstruct the input from this hidden representation. The output of the first decoder layer with the same activation function on all neurons can be written as,

$$r = g(h) = s'(W'h + b'),$$

where W' and b' are the weights and biases of the decoder, and s' is the decoder's activation function.

While simple autoencoders can learn to compress data efficiently, they may risk learning an identity function if the hidden layer is too large compared to an optimal one. To

address this, various regularization techniques exist. One popular approach is the sparse autoencoder, which adds a sparsity penalty to the loss function, forcing the network to activate only a small number of hidden units for each input (Ng, 2011).

Previous works have used autoencoders for cell type deconvolution, as well as for scRNA-seq data integration (Tran et al., 2021; Chen et al., 2022; Zhu et al., 2022).

1.6. CELL TYPE DEOCNVOLUTION IN BULK RNA-SEQ AND OTHER MIXED-CELL DATA MODALITIES

In this section, we first introduce a background of cell type deconvolution methods and then detail the DISSECT framework.

Cell type deconvolution methods are powerful tools to estimate the proportions of different cell types within bulk RNA-seq samples. These approaches assume that the observed gene expression in a bulk sample is the sum of the expression profiles of its constituent cell types, weighted by their relative abundances. Deconvolution algorithms aim to solve this inverse problem, inferring cell type proportions and, in some cases, cell type-specific expression profiles from the bulk data (Newman et al., 2019; Chen et al., 2022).

Deconvolution methods can be broadly categorized into reference-based and reference-free approaches. This work deals with reference-based approaches which rely on prior knowledge of cell type-specific gene expression signatures, often derived from purified cell populations or single-cell RNA-seq data. These prior signatures can be pre-computed directly and parameters (cell type fractions or gene expression) are fitted, or the prior can be learnt using machine learning (Table 1). These signatures serve as a basis for decomposing the bulk signal into its cellular components. Relying on these signatures makes cell type deconvolution a semi-blind source-separation task (Hesse and James, 2006). One of the earliest and most widely used reference-based methods is CIBERSORT (Newman et al., 2015), which employs support vector regression to estimate cell type proportions. Other popular tools in this category include xCell (Aran et al., 2019), which uses a novel gene signature-based method, and MuSiC, which uses multi-subject single-cell RNA-seq reference data to improve deconvolution accuracy (Wang et al., 2019).

Reference-free methods, on the other hand, estimate cell type proportions without relying on external gene expression signatures. These approaches often use matrix factorization techniques to decompose the bulk expression matrix into cell type-specific expression profiles and their corresponding proportions. Examples of reference-free methods include CDSeq, which employs a Bayesian hierarchical model for simultaneous deconvolution and gene expression estimation (Kang et al., 2019).

Recent advancements in deconvolution methods have focused on improving accuracy on diverse tissue types. For instance, Scaden uses deep learning (Menden et al., 2020) and SCDC combines multiple reference-based estimates to produce more reliable estimates (Dong et al., 2021). Some methods have explored the integration of deconvolution with other analysis tasks, such as differential expression analysis. Methods like csSAM and

Table 1.1 | Overview of the prominent cell type deconvolution methods.

Method name	**Algorithm type**	**Publication year**	**Journal**	**Predicts cell type fractions?**	**Predicts cell type-specific gene expression?**
TAPE	Transfer learning	2023	Nature Communications	Yes	Yes
BayesPrism	Bayesian hierarchical model	2022	Nature Cancer	Yes	Yes
Cell2Location	Probabilistic model	2022	Nature Biotechnology	Yes	No
SPOTlight	Non-negative matrix factorization	2021	Nucleic Acids Research	Yes	No
SCDC	Ensemble of models trained on different references	2020	Briefings in Bioinformatics	Yes	No
Scaden	Deep learning	2020	Science Advances	Yes	No
CIBERSORTx	Support vector machine & matrix factorization	2019	Nature Biotechnology	Yes	Yes
DWLS	Weighted least squares	2019	Nature Communications	Yes	No
MuSiC	Multi-subject deconvolution	2019	Nature Communications	Yes	No
CIBERSORT	Support vector regression	2015	Nature Methods	Yes	No

csDE enable the identification of cell type-specific differential expression from bulk RNA-seq data, providing insights into which cell types are driving observed gene expression changes (Shen-Orr et al., 2010; Chikina et al., 2015).

The choice of the deconvolution method depends on various factors, including the availability of reference data, the expected cellular composition of the samples, and the specific research questions being addressed (Avila Cobos et al., 2020). When suitable reference data are available, reference-based methods may provide more accurate and interpretable results (Avila Cobos et al., 2020; Sturm et al., 2019). However, reference-free methods can be valuable for exploring unknown cellular compositions or when reliable reference data are lacking (Zaitsev et al., 2019). It is important to note that all deconvolution methods have limitations and assumptions that should be carefully considered. Factors such as the quality and comprehensiveness of reference data, the degree of cellular heterogeneity in the samples, and the presence of unknown or rare cell types can all impact deconvolution performance (Jin and Liu, 2021; Jew et al., 2020; Patrick et al., 2020). Additionally, the resolution of cell type identification is inherently limited by the similarity of gene expression profiles between related cell types (Finotello and Trajanoski, 2018; Jin and Liu, 2021).

1.7. THE DISSECT FRAMEWORK

In this section, we introduce DISSECT framework (Khatri et al., 2024), a semi-supervised learning framework for cell type deconvolution. The complete manuscript is included as chapter 2.

1.7.1. SOURCE SEPARATION AND SEMI-SUPERVISED LEARNING

Cell deconvolution belongs to a broader class of source separation problems, where the goal is to recover individual source signals from mixed observations. In the classical blind source separation framework, this is often approached through methods like Non-negative Matrix Factorization (Lee and Seung, 2000) or Independent Component Analysis (Davies and James, 2007). However, these methods typically rely on strong assumptions about statistical independence and identical distributions.

DISSECT takes a fundamentally different approach by reformulating cell deconvolution as a semi-supervised learning problem. This reformulation is motivated by two key observations:

1. While ground truth cell type proportions for bulk RNA-seq data are generally unknown (unsupervised setting), we can generate labeled training data through single-cell RNA-seq simulations (supervised setting).

2. The physical process of cell mixing follows known constraints that can be exploited as consistency conditions.

1.7.2. CELL TYPE DECONVOLUTION AS A LEARNING PROBLEM WITH CONSISTENCY REGULARIZATION

The matrix factorization problem in cell type deconvolution is typically formulated as $\mathbf{B} = \mathbf{XS}$, where $\mathbf{B} \in \mathbb{R}^{m \times n}$ is the bulk gene expression matrix containing m samples with n genes, $\mathbf{X} \in \mathbb{R}^{m \times c}$ is a matrix containing cell type fractions of c cell types, and $\mathbf{S} \in \mathbb{R}^{c \times n}$ is a matrix consisting of cell type-specific gene expression profiles. Instead of directly solving this factorization problem, an alternative formulation is to learn a function $f : \mathbb{R}^n \to \Delta^c$ that maps bulk gene expression vectors to the probability simplex of cell type proportions.

Consistency Regularization

Given a real bulk sample B_i, a sample B^{sim} which is simulated from a scRNA-seq reference

data (refered to as simulated bulk), and a mixing coefficient $\beta \in [0,1]$, we define a mixed sample as:

$$B_i^{mix} = \beta B_i + (1-\beta) B_i^{sim}.$$

The key insight is that if f is a valid deconvolution function, it should satisfy:

$$f(B_i^{mix}) \approx \beta f(B_i) + (1-\beta) f(B_i^{sim}).$$

This consistency holds under the following assumption:

If the marker genes used for cell type identification are invariant across conditions and the mixing process is linear, then the consistency condition is exact. This can be argued as follows:

Let M be the set of marker genes, then for $g \in M$, the expression level in the mixture is:

$$B_{ig}^{mix} = \beta B_{ig} + (1-\beta) B_{ig}^{sim}, \text{and}$$

Since marker genes are invariant, their expression levels are proportional to cell type fractions, the predicted proportions must follow the same linear relationship.

Based on this validity condition, we developed the semi-supervised learning framework DISSECT that incorporates this condition into the learning objective with consistency regularization described in pseudocodes presented in algorithms 1 and 2 below. Further, the learning process of this algorithm is done in a schedule where first (in our experiments, for 2000 steps) the model is trained on purely simulated data to push model from collapsing to the solution $f(B_i^{mix}) = f(B_i) = f(B_i^{sim})$.

Input: B_{real} (bulk RNA-seq), SC (single-cell data), T (steps)
Output: f (trained fraction estimator)
$f \leftarrow$ InitializeMLP();
for $t = 1$ **to** T **do**
  $B_{sim}, X_{sim} \leftarrow$ SimulateBulk(SC);
  $\beta \leftarrow$ UniformSample(0.1, 0.9);
  $B_{mix} \leftarrow \beta \cdot B_{real} + (1-\beta) \cdot B_{sim}$;
  $X_{mix\_target} \leftarrow \beta \cdot f(B_{real}) + (1-\beta) \cdot X_{sim}$;
  if $t \leq 2000$ **then**
    $\lambda_1 \leftarrow 0$;
  else if $t \leq 4000$ **then**
    $\lambda_1 \leftarrow 15$;
  else
    $\lambda_1 \leftarrow 10$;
  end
  $L_{supervised} \leftarrow \mathrm{KL}(f(B_{sim}), X_{sim})$;
  $L_{consistency} \leftarrow \|f(B_{mix}) - X_{mix\_target}\|^2$;
  $L_{total} \leftarrow L_{supervised} + \lambda_1 \cdot L_{consistency}$;
  UpdateNetwork(f, L_{total});
end
return f

Algorithm 1: Training Fraction Estimator

Input: B_{real}, SC, f (trained fraction estimator), T_{exp} (steps)
Output: g (trained expression estimator)
$g \leftarrow$ InitializeCVAE();
for $t = 1$ **to** T_{exp} **do**
  $B_{sim}, X_{sim} \leftarrow$ SimulateBulk(SC);
  $S_{sim} \leftarrow$ ComputeSignatures(B_{sim}, X_{sim});
  $\beta \leftarrow$ UniformSample(0.1, 0.9);
  $B_{mix} \leftarrow \beta \cdot B_{real} + (1-\beta) \cdot B_{sim}$;
  $X_{real} \leftarrow f(B_{real})$;
  $X_{mix} \leftarrow \beta \cdot X_{real} + (1-\beta) \cdot X_{sim}$;
  $L_{recon} \leftarrow \|S_{sim} - g(B_{sim})\|^2$;
  $L_{consistency} \leftarrow \|X_{mix} \cdot g(B_{mix}) - \beta \cdot X_{real} \cdot g(B_{real}) - (1-\beta) \cdot X_{sim} \cdot S_{sim}\|^2$;
  $L_{KL} \leftarrow \mathrm{KL}(g_{encoder}(B_{sim}), \mathcal{N}(0,1))$;
  $L_{total} \leftarrow L_{recon} + \lambda_2 \cdot L_{consistency} + \beta_{CVAE} \cdot L_{KL}$;
  UpdateNetwork(g, L_{total});
end
return g

Algorithm 2: Training Expression Estimator

DISSECT's consistency regularization shares conceptual foundations with MixMatch (Berthelot et al., 2019) but differs in crucial ways. Firstly, MixMatch assumes unlabeled data follows the same distribution as labeled data, and DISSECT explicitly handles domain shift between simulated and real data. Secondly, MixMatch uses random convex combinations of augmented samples while DISSECT uses biologically motivated mixing based on cell type proportions. Lastly, the two differ in their loss functions. MixMatch uses cross-entropy for labeled data and L_2-norm for consistency, while DISSECT uses KL divergence for simulated data and weighted L_2-norm for consistency.

In scenarios where some bulk data with known cell type fractions is available from another experiment and its domain shift from the test bulk data is limited, the consistency framework of DISSECT offers additional theoretical advantages. One key benefit is the reduced sample complexity (*i.e.*, the required number of samples for effective training), which we examine below.

1.7.3. DISSECT REQUIRES FEWER SAMPLES TO LEARN ACCURATE REPRESENTATIONS WHEN SOME REAL BULK DATA WITH TRUE CELL TYPE PROPORTIONS IS AVAILABLE

In the classical setting without consistency regularization, learning accurate latent cell type signatures theoretically requires $\mathcal{O}(nc)$ samples, where n is the number of genes and c is the number of cell types. This complexity emerges naturally from the dimensionality of the problem: each cell type signature is represented by an n-dimensional vector, and we need to learn c such signatures. From statistical learning theory, this complexity aligns with the VC dimension of learning c hyperplanes in n dimensions (Vapnik and Chervonenkis, 2015; Blumer et al., 1989).

DISSECT's consistency regularization framework fundamentally alters this requirement by imposing constraints on the solution space. The consistency condition $f(B_i^{mix}) = \beta f(B_i) + (1-\beta) f(B_i^{sim})$ enforces strong constraints on the hypothesis space of possible deconvolution functions.

The theoretical foundation for this dimension reduction can be understood through the Johnson-Lindenstrauss (JL) lemma (Johnson, 1984), which states that N points in high-dimensional space can be projected down to $\mathcal{O}(\log N/\epsilon^2)$ dimensions while preserving all pairwise distances up to a factor of $(1 \pm \epsilon)$, for $0 < \epsilon < 1$. In the context of DISSECT, the mixing process $B_i^{mix} = \beta B_i + (1-\beta) B_i^{sim}$ with β randomly sampled from a uniform distribution between 0.1 and 0.9 creates a form of randomized projection.

DISSECT inherently satisfies several key conditions that make this theoretical reduction possible: assuming model learns well on simulated data, the mixing operation is applied to marker genes that are invariant across conditions, ensuring essential signals for cell type identification are preserved; the simulated data is generated from single-cell references containing sufficient cellular heterogeneity, which helps capture the low-dimensional manifold where cell type signatures naturally reside; and the consistency regularization enforces a constraint that the deconvolution function must be linear with respect to the mixing operation, which aligns with the linear projection properties of the JL lemma.

Under these conditions, the learning problem can be effectively reduced from $\mathcal{O}(nc)$ to approximately $\mathcal{O}(\sqrt{nc})$ complexity. The practical implications of this reduction are substantial. For a bulk RNA-seq dataset with n=5,000 genes and c = 5 cell types, traditional methods would theoretically require on the order of 25,000 samples for reliable learning (based on the $\mathcal{O}(nc)$ bound). In contrast, DISSECT could potentially achieve comparable performance with approximately 158 samples (based on the $\mathcal{O}(\sqrt{nc})$ bound). It is important to note that these are theoretical upper bounds, and in practice, most methods may require fewer samples to achieve reasonable performances.

This efficiency is particularly crucial in the context of bulk RNA-seq deconvolution, where obtaining ground truth cell type proportions is expensive and labor-intensive. Moreover, the consistency regularization not only reduces sample requirements but also improves generalization by enforcing biologically meaningful constraints on the deconvolution function.

This theoretical advantage positions DISSECT as a valuable tool for real-world applications where bulk RNA-seq data with ground truth cell type fractions is scarce, while still maintaining robust performance through its semi-supervised learning framework.

1.8. Cell type dynamics in ANCA-GN and GBM

In this section, we provide the necessary background for cell type dynamics in ANCA-GN and GBM, which are explored in chapters 3 and 4 respectively.

1.8.1. ANCA-GN

ANCA-GN is an autoimmune disease characterized by inflammation and damage to small blood vessels in the kidneys. It is a subset of ANCA-associated vasculitis (AAV), a group of systemic autoimmune diseases that can affect multiple organ systems (Jennette and Falk, 2013). The term "ANCA" refers to anti-neutrophil cytoplasmic antibodies, which play a

central role in the pathogenesis of these conditions. Below is an overview of ANCA-GN, in the context of complex dynamics of immune cells and the pathogenic mechanisms involved.

The pathogenesis of ANCA-GN is primarily driven by the production of autoantibodies targeting neutrophil proteins, specifically proteinase 3 (PR3) and myeloperoxidase (MPO) (Kitching et al., 2020). These autoantibodies interact with neutrophils and monocytes, leading to their activation and subsequent damage to blood vessel walls, particularly in the glomeruli of the kidneys. The process begins with the production of ANCA by B cells and their progenitors, followed by neutrophil priming by inflammatory stimuli, which causes neutrophils to express PR3 and MPO on their cell surface. The interaction between ANCA and these exposed autoantigens leads to neutrophil activation, degranulation, and the formation of neutrophil extracellular traps (NETs). This process, coupled with complement activation, results in endothelial injury and inflammation in the blood vessels (Radford et al., 2001; Kessenbrock et al., 2009; Kettritz, 2012; Jennette and Falk, 2013; Cornec et al., 2016).

The immune cell dynamics in ANCA-GN involve various cell types, each contributing to the disease progression and regulation in several ways. Neutrophils, as the primary effector cells, undergo degranulation, NETosis, and cytokine production upon activation by ANCA. Monocytes and macrophages also play important roles as ANCA targets, cytokine producers, and antigen-presenting cells. Dendritic cells contribute to the initiation and perpetuation of the autoimmune response through antigen presentation and cytokine production (Söderberg et al., 2015; Wilde et al., 2009).

T cells have emerged as crucial players in the pathogenesis of ANCA-GN, with various subsets contributing to disease progression. CD4+ T helper cells provide help to B cells for autoantibody production and release cytokines that promote inflammation. An expanded population of Th17 cells has been observed in ANCA-GN patients, contributing to neutrophil recruitment and activation through the production of IL-17. Regulatory T cells (Tregs) have been reported to be deficient or dysfunctional in ANCA-GN, potentially contributing to the loss of self-tolerance. CD8+ T cells have also been implicated in tissue damage and disease relapse (Abdulahad et al., 2007; Nogueira et al., 2010; Lepse et al., 2011; Chen et al., 2020).

The role of T cells in ANCA-GN is far from clearly defined. Firstly, it is not clear how different T cell subsets interact and influence each other in ANCA-GN, which specific antigens are recognized by the autoreactive T cells, and how they further contribute to the breakdown of tolerance. From a therapeutic point of view, it would be interesting

to find if there are distinct T cell signatures that could serve as biomarkers for disease activity or even predict treatment response. The potential role of tissue-resident memory T cells in maintaining long-term inflammation and potentially causing or aiding relapse remains to be fully defined. Further, the inflammatory niches and their microenvironments are not explored. These niches could serve as hotspots for sustained inflammation and autoantibody production. Understanding the composition, and heterogeneity of these niches may provide valuable insights into disease mechanisms. While ANCA-GN primarily affects glomeruli, the impact of infiltration in tubulointerstitial areas is not well understood (Boud'hors et al., 2023).

1.8.2. GBM

Glioblastoma (GBM) is the most aggressive and lethal primary brain tumor in adults, characterized by its rapid growth, invasive nature, and resistance to conventional therapies. Despite advances in neurosurgery, radiation therapy, and chemotherapy, the prognosis for GBM patients remains poor, with a median survival of approximately 15 months after diagnosis (Stupp et al., 2005). The complexity of GBM lies not only in its aggressive behavior but also in its heterogeneity, both at the cellular and molecular levels. This heterogeneity presents significant challenges for effective treatment and necessitates a deeper understanding of the disease's underlying biology.

In recent years, extensive molecular profiling efforts have identified distinct GBM subgroups based on gene expression patterns, DNA methylation profiles, and genetic alterations. The most widely recognized classification is the Verhaak subgroups, which categorizes GBM into four molecular subtypes based on transcriptomics and genetic alterations: Proneural, Neural, Classical, and Mesenchymal (MES) (Verhaak et al., 2010). The Proneural subtype is characterized by IDH1 mutations and PDGFRA amplifications, while the Classical subtype typically harbors EGFR amplifications. The Mesenchymal subtype is associated with NF1 mutations and expression of mesenchymal markers, whereas the Neural subtype shows expression of neuronal markers. However, it's important to note that individual tumors can exhibit features of multiple subtypes, and these classifications may change over time or in response to treatment, reflecting the dynamic nature of GBM (Wang et al., 2017).

More recently, DNA methylation profiling has emerged as a tool for tumor classification, leading to the identification of additional GBM subgroups. The DNA methylation-based classification system, which includes categories such as RTK I, RTK II, and MES, provides complementary information to the transcriptomic subtypes and has shown promise in refining prognostic predictions and potentially guiding treatment decisions (Capper et al.,

2018). The RTK I and RTK II subgroups are characterized by distinct patterns of receptor tyrosine kinase alterations. While the transcriptomic and DNA methylation subgroups are not always comparable, the MES subgroup aligns closely with the transcriptomic MES subtype. These methylation-based classifications offer additional granularity in understanding GBM biology and may help in identifying more homogeneous patient populations.

The cellular composition of GBM is highly complex, involving not only the malignant glioma cells but also various non-neoplastic cells within the tumor microenvironment. While the focus has traditionally been on astrocytes as the presumed cells of origin for GBM, increasing evidence suggests important roles for neurons, oligodendrocytes, and their precursors in GBM biology. Neurons, for instance, have been shown to promote the growth and progression of glioma cells through activity-dependent mechanisms, releasing factors that can enhance tumor proliferation and invasion (Venkatesh et al., 2019). This neuron-glioma interaction raises questions about the potential impact of neuronal activity on tumor behavior and whether modulating this activity could have therapeutic implications. Apart from neurons, oligodendrocytes and their precursor cells (OPCs) may also be implicated in GBM. A subset of GBMs may arise from OPCs, particularly those with IDH mutations (Lu et al., 2016). Moreover, the presence of OPCs in the tumor microenvironment may influence GBM growth and invasion. The further characterization of the potential role of oligodendrocytes in supporting or inhibiting GBM progression, especially considering their modulation by immune cells, could help find new therapeutic targets or strategies for manipulating the tumor microenvironment to enhance treatment efficacy (Kawashima et al., 2019; Harrington et al., 2020). Astrocytes, both reactive and tumor-associated, play multifaceted roles in GBM. While transformed astrocytes are often the primary malignant cell type in GBM, reactive astrocytes in the tumor microenvironment can contribute to tumor progression through the secretion of growth factors, cytokines, and extracellular matrix components (Henrik Heiland et al., 2019). The complex interplay between malignant and non-malignant astrocytes, as well as their interactions with other cell types in the tumor microenvironment, highlights the need for therapies that target not only the tumor cells but also the supporting cellular network.

The remarkable heterogeneity of GBM extends beyond cellular composition and molecular subgroups to include intratumoral heterogeneity at the genetic and epigenetic levels. Single-cell sequencing studies have identified the presence of multiple genetically distinct subclones within individual tumors, each potentially harboring different driver mutations and exhibiting varied responses to therapy (Patel et al., 2014). This genetic diversity, coupled with epigenetic plasticity, allows GBM to rapidly adapt to therapeutic

pressures, contributing to treatment resistance and recurrence. The dynamic nature of this heterogeneity poses significant challenges for developing effective targeted therapies and necessitates innovative treatment strategies that can address the evolving landscape of the tumor.

Current therapeutic approaches for GBM typically involve maximal safe surgical resection followed by radiation therapy and chemotherapy with temozolomide. However, the efficacy of these treatments is limited, and recurrence is almost inevitable. The identification of molecular subgroups and a deeper understanding of GBM biology have paved the way for more targeted therapeutic approaches. For instance, inhibitors targeting specific genetic alterations, such as EGFR mutations or IDH1 mutations, have shown promise in clinical trials for selected patient populations (Weller et al., 2017). Immunotherapies, including immune checkpoint inhibitors and CAR-T cell therapies, are also being actively investigated, with the potential to harness the immune system to combat GBM (Lim et al., 2018).

The complex tumor microenvironment and the presence of glioma stem cells contribute to treatment resistance and tumor recurrence. Novel approaches being explored include targeting the tumor microenvironment and developing combination therapies that address multiple aspects of GBM biology simultaneously. As our understanding of GBM biology continues to evolve, several key questions emerge that warrant further investigation.

In our work, we focused on capturing the heterogeneity and different cellular components of the tumor microenvironment, including neurons, oligodendrocytes, and astrocytes, that may interact to influence tumor progression and treatment response and how these interactions could be measured and therapeutically targeted.

1.9. RESEARCH OBJECTIVES

This dissertation proposes DISSECT as a deconvolution framework introduced in section 1.7, and addresses open questions pertaining to section 1.8. To this end, the following methodological and translational goals arise.

A. Evaluation of existing cell type deconvolution algorithms on RNA-seq data from various organs and diseases.

B. Development of a robust and consistent cell type deconvolution approach.

C. Study of transcriptomic changes in ANCA-GN kidney RNA-seq data using single-cell and spatial transcriptomics, and identifying targetable pathways for kidney health and better patient outcome.

D. Analysis of multi-omics data from glioblastoma to characterize cell type and state signatures, to identify biomarkers for routine neuro-oncological use. The aim here is to identify temporarily stable patient subgroups with a potentially poor prognosis and inform neurosurgical decision making.

The datasets, experiments, and associated results are presented and discussed in chapters 2-4. Objectives A and B are addressed in chapter 2. Chapter 3 addresses objective C, and Chapter 4 addresses objective D.

2

DECONVOLUTION OF BULK RNASEQ AND SPATIAL TRANSCRIPTOMICS BENEFITS FROM SEMI-SUPERVISED LEARNING

Khatri *et al. Genome Biology* (2024) 25:112
https://doi.org/10.1186/s13059-024-03251-5

Genome Biology

METHOD

Open Access

DISSECT: deep semi-supervised consistency regularization for accurate cell type fraction and gene expression estimation

Robin Khatri[1], Pierre Machart[1] and Stefan Bonn[1*]

*Correspondence:
sbonn@uke.de

[1] Institute of Medical Systems Biology, Center for Molecular Neurobiology, Center for Biomedical AI, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Abstract

Cell deconvolution is the estimation of cell type fractions and cell type-specific gene expression from mixed data. An unmet challenge in cell deconvolution is the scarcity of realistic training data and the domain shift often observed in synthetic training data. Here, we show that two novel deep neural networks with simultaneous consistency regularization of the target and training domains significantly improve deconvolution performance. Our algorithm, DISSECT, outperforms competing algorithms in cell fraction and gene expression estimation by up to 14 percentage points. DISSECT can be easily adapted to other biomedical data types, as exemplified by our proteomic deconvolution experiments.

Keywords: Cell deconvolution, Semi-supervised learning, Deep learning

Background

A prominent approach to studying tissue-specific gene expression changes in human development and disease is RNA sequencing (bulk RNA-seq). Tissues, however, usually consist of multiple cell types in different quantities and with different gene expression programs. Consequently, bulk RNA-seq from tissues measures average gene expression across the constituent cells, disregarding cell type-specific changes. The quantification of the cellular composition and cell type-specific expression that underlies bulk RNA-seq data is therefore of pivotal importance to understanding disease mechanisms and identifying potential therapeutic interventions [1].

A recent technological advancement, single-cell RNA-seq, allows for investigating gene expression in single cells for thousands of individual cells of a given tissue sample in a single experiment. However, while it provides unprecedented insights into single-cell biology, it suffers from severe technical limitations, most notably the presence of zero values in gene expression due to methodological noise, termed as "dropouts" [2]. In addition, the technology is still very costly, which essentially prohibits its application in clinical and diagnostic

settings. Bulk RNA-seq, on the other hand, can be performed for a fraction of the cost and is widely used in clinical oncology and drug discovery [3, 4].

Computational inference of cell type fraction and cell type-specific gene expression is a source-separation task, termed as "cell deconvolution" within the context of cell biology. The estimation of cell type-specific gene expression is a well established and challenging problem in the field. Prior work includes but is not limited to TAPE [5], bMIND [6], BayesPrism [7], and CibersortX (CSx) [8]. The basic aim is to provide cell type-specific gene expression information at a group or sample level. The resultant information allows deep biological insights into cell type-specific gene expression and pathway changes from bulk data. For cell deconvolution, recent computational methods utilize single-cell sequencing data to create simulated references with known fraction and expression for training [9]. While this approach achieves good deconvolution results, its performance suffers from the substantial domain shift between single-cell RNA-seq training (reference) data and the bulk RNA-seq target data. Domain refers to the statistical distribution of the source of a dataset [10]. Domain shift refers to a change in the statistical distribution of samples, which can be due to covariate shift, the presence of open sets, or both. In gene expression datasets, the covariate shift between real data and simulated datasets occurs due to changes in cell type-specific gene expression and can arise from different dropout rates and tissue conditions, for instance. When domain shifts have purely technical reasons, they are often termed batch effects. CSx [8] has previously approached the problem of batch effect removal between single cell gene expression datasets [11], using Combat [12] to remove changes in cell type-specific gene expression between a single-cell reference signature matrix and bulk RNAseq data. Open sets may occur when new cell types are encountered during test time, such as the presence of differing cell lineages [13]. Since cells go through different differentiation states, domain shift between real data and simulations may be a combination of both, the covariate shift and presence of open sets. Among many possible sources of domain variation, the most prevalent might be the presence of batch effects that refer to technological differences between two sequencing experiments and gene expression differences of biological nature.

In this work, we first formally define the task of cell deconvolution and outline the hypothesis that semi-supervised consistency regularization should improve bulk RNA-seq deconvolution when learning from single cell RNA-seq data. We then provide evidence that two novel deep learning algorithms with semi-supervised consistency regularization outperform competing state-of-the-art algorithms in deconvolution, both on a cellular and gene expression level, across a wide range of datasets. On the datasets with ground truth flow cytometry cell type proportions, DISSECT achieves consistently better Jensen-Shannon distance (*JSD*): 0.063 ± 0.015 and root mean squared error (*rmse*): 0.021 ± 0.019. In addition, DISSECT shows state-of-the-art gene expression deconvolution performance, achieving the best sample- and gene-wise correlations. Our algorithm can easily be adapted to other biomedical data types, as exemplified by our bulk proteomics and spatial expression deconvolution experiments.

Results

In this section, we first formally define the cell deconvolution task, then present our hypothesis and DISSECT deep learning models, and compare DISSECT's performance to other state-of-the-art deconvolution algorithms.

Task of cell deconvolution

Given an $m \times n$ gene expression matrix $\mathbf{B}$ consisting of m bulk gene expression vectors measuring n genes, the goal of deconvolution is to find an $m \times c$ matrix $\mathbf{X}$ of cell type fractions, where c is the number of cell types present in bulk samples such that,

$$\mathbf{B} = \mathbf{XS}, \tag{1}$$

where fractions and gene expression satisfy non-negativity $0 \leq \mathbf{X}_{ik}$, and $0 \leq \mathbf{S}_{kj}$, $\forall i \in [1, m], \forall j \in [1, n]$ and $\forall k \in [1, c]$ and sum-to-1 criterion, i.e., $\sum_{k=1}^{c} \mathbf{X}_{ik} = 1, \forall i \in [1, m]$. Here, $\mathbf{S}$ is known as the signature matrix and is unobserved. Each row $\mathbf{S}_{k\cdot}$ is a gene expression profile (or signature) of cell type k. To utilize a reference based framework, $\mathbf{S}$ can be replaced with $\mathbf{S}_{ref}$ derived from a single-cell experiment by identifying the most representative cell type specific gene expression [8].

The problem of reference-based cell deconvolution can alternatively be formulated as a learning problem, where a function f such that $f(\mathbf{B}) = \mathbf{X}$ is learnt. Since only $\mathbf{B}$ is available and $\mathbf{X}$ is generally unknown, simulations from a single-cell reference can be used to learn f. Clearly, from the above formulation of the cell deconvolution task, it is reasonable to assume linearity of deconvolution, i.e., each bulk mixture is a linear combination of expression vectors of cells spanned with corresponding cell type fractions. Thus, as defined previously in Scaden [9], multiple single cells can be combined in random proportions to generate training examples $\mathbf{B}^{\text{sim}}$ and $\mathbf{X}^{\text{sim}}$, where each row of $\mathbf{B}^{\text{sim}}$ is defined as,

$$\mathbf{B}^{\text{sim}}_{i\cdot} = \sum_{k=1}^{c} \sum_{l=1}^{\alpha_{k,i}} \mathbf{e}^{k}_{l},$$

where $\mathbf{e}^{k}_{l}$ is the expression vector of cell l belonging to cell type k, and $\alpha_{k,i}$ is the number of cells belonging to cell type k sampled to construct $\mathbf{B}^{\text{sim}}_{i\cdot}$. Correspondingly, each element of $\mathbf{X}^{\text{sim}}$ is the proportion of a cell type k in that sample i and is defined as,

$$\mathbf{X}^{\text{sim}}_{ik} = \frac{\alpha_{k,i}}{\sum_{k=1}^{c} \alpha_{k,i}},$$

In this case, since each simulated sample has a distinct signature (i.e., gene expression profile), $\mathbf{S}$ is a three dimensional matrix with each element $\mathbf{S}_{kji}$ denoting gene expression of gene j in cell type k for sample i. It is computed as following,

$$\mathbf{S}^{\text{sim}}_{k\cdot i} = \frac{\sum_{l=1}^{\alpha_{k,i}} \mathbf{e}^{k}_{l}}{\alpha_{k,i}}.$$

The predictor f, learned from a simulated dataset, can then be applied to $\mathbf{B}$ to estimate $\mathbf{X}$. Note that, the genes expressed may differ between vectors $\mathbf{e}_l$ and $\mathbf{B}$ and as such before learning function f, each $\mathbf{e}_\mathbf{l}^\mathbf{k}$ is subsetted to include genes common with $\mathbf{B}$. This is the reason why this learning problem is transductive and a separate model needs to be reconstructed for each $\mathbf{B}$.

Exploiting the linearity of deconvolution

The deconvolution task is to learn a cell type-specific gene-expression matrix (or signature matrix) $\mathbf{S}$, which serves to accurately predict cell fractions and their corresponding gene expression from a bulk gene expression matrix $\mathbf{B}$. The actual mixing process of cells to form a tissue is assumed to be linear and, as such, the relationship between $\mathbf{B}$ and $\mathbf{S}$ is linear. However, $\mathbf{S}$ is unobserved, and the deconvolution algorithm is learned using simulations. This learning process involving simulations is highly dependent on the reference being the single-cell dataset used to generate simulations, and is subjected to an inherent strong domain shift [14]. To address this, we hypothesize that a consistency-based regularization penalizing the non-linearity of mixtures of real and simulated samples would result in a mapping $\hat{f}$ that is closer to true mapping f. Non-linearity of mixtures of real and simulated samples refers to the violation of Eq. 4, defined later, for estimated $\mathbf{X}_{i\cdot}$, $\mathbf{X}_{i\cdot}^{\text{sim}}$ and $\mathbf{X}_{i\cdot}^{\text{mix}}$ using mapping f.

Consistency regularization

Consider that $\mathbf{B}$ represents gene expression matrices of real (test) bulk RNA-seq that we want to deconvolve and and $\mathbf{B}^{\text{sim}}$ represents gene expression matrix of simulated bulk samples. The number of rows (representing samples) in these two matrices may differ. To simplify the notation, we use the same index i to denote indices for real bulk samples, simulations (sim) and their mixtures (mix, defined further). Given a true bulk RNA-seq sample $\mathbf{B}_{i\cdot}$, and a simulated sample $\mathbf{B}_{i\cdot}^{\text{sim}}$ with paired proportions $\mathbf{X}_{i\cdot}^{\text{sim}}$ defined over a common set of genes, we can generate a mixture $\mathbf{B}_{i\cdot}^{\text{mix}}$ such that

$$\mathbf{B}_{i\cdot}^{\text{mix}} = \beta\mathbf{B}_{i\cdot} + (1-\beta)\mathbf{B}_{i\cdot}^{\text{sim}}, \tag{2}$$

Which gives us the relation

$$\mathbf{X}_{i\cdot}^{\text{mix}}\mathbf{S}_{\cdot i\cdot}^{\text{mix}} = \beta\mathbf{X}_{i\cdot}\mathbf{S}_{\cdot i\cdot} + (1-\beta)\mathbf{X}_{i\cdot}^{\text{sim}}\mathbf{S}_{\cdot i\cdot}^{\text{sim}}. \tag{3}$$

where $\mathbf{X}_{i\cdot}$ represents cell fractions of sample i and where $\beta \in [0,1]$. Cell types are characterized by a few marker genes that are invariant across cell states and even across tissues [15]. A network that accurately predicts cell type fractions based on gene expression of simulated or real bulk RNA-seq data would thus have to learn them. In the estimation of cell type fractions, we therefore assume that the expression of these marker genes should be identical in signatures $\mathbf{S}_{\cdot i\cdot}^{\text{mix}}$, $\mathbf{S}_{\cdot i\cdot}$ and $\mathbf{S}_{\cdot i\cdot}^{\mathbf{sim}}$. Hence,

$$\mathbf{X}_{i\cdot}^{\text{mix}} = \beta\mathbf{X}_{i\cdot} + (1-\beta)\mathbf{X}_{i\cdot}^{\text{sim}}, \tag{4}$$

Equation 4 serves as the formulation to generate pseudo ground-truths for these mixtures during learning, and it enables the use of consistency regularization without having to explicitly estimate signatures. In an iterative learning process $\mathbf{X}_{i\cdot}$ can be replaced

with predictions of the algorithm from the previous iteration. Naturally, it is also possible to only mix real samples with each other. The number of bulk RNA-seq samples is, however, considerably lower (tens to hundreds) than the amount of single-cells present in a single-cell experiment (thousands or more). Equation 4 allows to generate pseudo ground truth proportions for mixtures $\mathbf{B}_{i\cdot}^{\text{mix}}$ at each step of learning cell type fractions, while Eq. 3 allows to generate pseudo ground truth signatures at each step of learning gene expression profiles.

Network architecture and learning procedure

We approach the two tasks, estimation of cell type fractions and estimation of gene expression profiles per cell type as two different tasks because of their differing assumptions. For the estimation of cell type fractions, we assume that signatures are identical for each sample, both simulated and bulk, while to estimate gene expression, we relax this condition and involve complete consistency regularization (Eq. 3). An illustration of the method is presented in Fig. 1.

Estimation of cell type fractions

The underlying algorithm of the first part of our deconvolution method is an average ensemble of multilayered perceptrons (MLPs). The ensembling is performed to reduce the variance by averaging different runs [16]. Each MLP consists of the same architecture initialized with different weights. Each MLP has an architecture: Input (# genes) - ReLU6 (512) - ReLU6 (256) - ReLU6 (128) - ReLU6 (64) - Linear (# cell types) - Softmax. ReLU6 (output of ReLU activation clipped by a maximum value of 6) [17, 18] was chosen out of tested activations over grid search on (Linear, ReLU, ReLU6, Swish [19]). The final

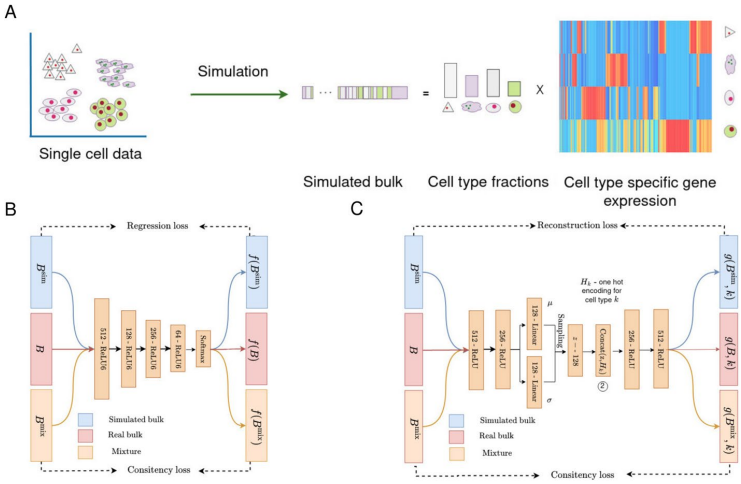


Fig. 1 **A** Illustration of the simulation procedure using reference single-cell data. The figure shows the simulation of one sample which consists of cell type fractions, simulated gene expression and cell type specific gene expression profiles (i.e., signature matrix). **B** Detailed overview of an MLP used to estimate cell type fractions. **C** Overview of an autoencoder used to estimate cell type specific gene expression profiles

application of a softmax activation function allows to achieve the non-negativity and sum to 1 criteria of deconvolution. We train the network with batch size 64 to minimize the loss function per batch defined below with an Adam Optimizer with initial learning rate of $1e-5$.

$$\begin{aligned}\mathcal{L}_{\text{total}}\left(\mathbf{X}_{i\cdot}^{\text{sim}}, f\left(\mathbf{B}_{i\cdot}^{\text{sim}}\right), \mathbf{X}_{i\cdot}^{\text{mix}}, f\left(\mathbf{B}_{i\cdot}^{\text{mix}}\right)\right) &= \mathcal{L}_{\text{KLdivergence}}\left(\mathbf{X}_{i\cdot}^{\text{sim}}, f\left(\mathbf{B}_{i\cdot}^{\text{sim}}\right)\right) \\ &\quad + \lambda_1 * \mathcal{L}_{\text{cons}}\left(\mathbf{X}_{i\cdot}^{\text{mix}}, f\left(\mathbf{B}_{i\cdot}^{\text{mix}}\right)\right),\end{aligned} \tag{5}$$

where $\mathcal{L}_{\text{KLdivergence}}(\cdot,\cdot)$ is the Kullback-Leibler divergence and $\mathcal{L}_{\text{cons}}(\cdot,\cdot)$ is the consistency loss defined as:

$$\mathcal{L}_{\text{cons}}\left(\mathbf{X}_{i\cdot}^{\text{mix}}, f\left(\mathbf{B}_{i\cdot}^{\text{mix}}\right)\right) = \left\|\mathbf{X}_{i\cdot}^{\text{mix}} - f\left(\mathbf{B}_{i\cdot}^{\text{mix}}\right)\right\|_2^2, \text{ and}$$

$$\mathbf{X}_{i\cdot}^{\text{mix}} = \beta f(\mathbf{B}_{i\cdot}) + (1-\beta)\mathbf{X}_{i\cdot}^{\text{sim}}.$$

To generate mixtures, for each batch, we sample β uniformly at random for Eq. 4. The interval [0.1, 0.9] was chosen for the uniform distribution to allow for at least some real and some simulated gene expression in the mixture. Since the number of simulations is generally larger (in our experiments, set to 1,000 times the number of cell types) than that of real data, we sample real data to create additional bulk samples, $\mathbf{B}_{i\cdot}$, until the size equals that of the simulated data, $\mathbf{B}_{i\cdot}^{\text{sim}}$. This pair of data together with simulated proportions, $\mathbf{X}_{i\cdot}^{\text{sim}}$, is then used to create training batches of size 64. For every batch, we generate mixtures according to Eq. 2.

Our loss is inspired by MixMatch [20], which uses unlabelled samples to mix up and match sample predictions. Our adaptation in Eq. 5 addresses the limited samples available from true bulk RNA-seq, unavailability of sample fractions and is derived from the definition of the task itself. In essence, Eq. 5 integrates domain knowledge into the objective.

To avoid a scenario where the network does not learn and outputs predictions such that $f(\mathbf{B}_{i\cdot}^{\text{mix}}) = f(\mathbf{B}_{i\cdot}^{\text{sim}}) = f(\mathbf{B}_{i\cdot})$, which is a solution to Eq. 4, we first let the model learn purely from simulated examples. This allows the model to learn meaningful expression profiles to achieve accurate results on simulated examples. We selected λ_1 based on a grid search over constant and step-wise functions. We adopt a step-wise function for λ_1, given as:

$$\lambda_1 = \begin{cases} 0 & \text{if step} \leq 2000, \\ 15 & \text{elif } 2000 \leq \text{step} \leq 4000, \\ 10 & \text{else.} \end{cases}$$

We train the network for a predefined number of steps as opposed to epochs, since it is possible to generate infinitely many simulated samples without increasing the intrinsic dimensionality of the data. In our experiments, we limit the number of steps to 5000 as found optimal in Scaden [9].

Estimation of per sample cell type specific gene expression profiles Estimation of cell type fractions from bulk RNA-seq requires an assumption that signatures of cell types are

shared across single cell and bulk RNA-seq. However, cell type gene expression profiles (at least for genes that are not invariant across tissue states) may differ between samples. Previously, works such as CSx [8] and TAPE [5] have explored utilizing cell type fractions to estimate gene expression per sample. Here, we make use of a β-variational autoencoder with standard normal distribution as prior to estimate average gene expression of the different cell types from bulk RNA-seq expression levels. To jointly train the network on all cell types, we condition the decoder (at its input layer) with cell type labels. This allows for training a single model to estimate gene expression of each cell type for a sample. To make use of bulk RNA seq during the training, we regularize the reconstruction loss with a consistency loss defined over per cell type signature. Denoting f as before and $g(\cdot, k)$ as the output of the autoencoder with condition k (corresponding to cell type label) on the decoder input, this consistency loss is defined as:

$$\mathcal{L}_{\text{cons}}^{\text{VAE}}\left(f, g, \mathbf{B}_{i\cdot}^{\text{mix}}, \mathbf{B}_{i\cdot}, \mathbf{X}_{i\cdot}^{\text{sim}}, \mathbf{S}_{ki\cdot}^{\text{sim}}\right) = \Big\| f\left(\mathbf{B}_{i\cdot}^{\text{mix}}\right)_k g\left(\mathbf{B}_{i\cdot}^{\text{mix}}, k\right) - \beta f(\mathbf{B}_{i\cdot})_k g(\mathbf{B}_{i\cdot}, k) - (1-\beta)\mathbf{X}_{i\cdot}^{\text{sim}}\mathbf{S}_{ki\cdot}^{\text{sim}}\Big\|_2^2,$$

where $\mathbf{B}_{i}^{\text{mix}}$ is given by Eq. 2, and $f(\mathbf{B}_{i\cdot}^{\text{mix}})_k$ is the proportion of cell type k in sample i as estimated during cell type fraction estimation and is fixed during training. In implementation, we replace $f(\mathbf{B}_{i\cdot}^{\text{mix}})_k$ with $\beta f(\mathbf{B}_{i\cdot})_k + (1-\beta)\mathbf{X}_{i\cdot}^{\text{sim}}$. Thus, this loss forces the learned signature for cell type k, $g(\mathbf{B}_{i\cdot}^{\text{mix}}, k)$, to be closer to signatures for both real and simulated bulk samples. This loss function makes the assumption that mixing two bulk samples is similar to mixing individual cell type specific signatures that constitute those bulks. We added this loss function with a regularization parameter λ_2 (with default value 0.1) to the loss of the standard β-variational autoencoder (the weight on the KL divergence, denoted as β^{VAE}, is set to 0.1 by default). The total loss function sums up to:

$$\begin{aligned}\mathcal{L}_{\text{total}}^{\text{VAE}}\left(f, g, \mathbf{B}_{i\cdot}^{\text{sim}}, \mathbf{B}_{i\cdot}^{\text{mix}}, \mathbf{B}_{i\cdot}, \mathbf{X}_{i\cdot}^{\text{sim}}, \mathbf{S}_{ki\cdot}^{\text{sim}}\right) = &\left\| \mathbf{S}_{ki\cdot}^{\text{sim}} - g\left(\mathbf{B}_{i\cdot}^{\text{sim}}, k\right)\right\|_2^2 \\ &+ \lambda_2 \mathcal{L}_{\text{cons}}^{\text{VAE}}\left(f, g, \mathbf{B}_{i\cdot}^{\text{mix}}, \mathbf{B}_{i\cdot}, \mathbf{X}_{i\cdot}^{\text{sim}}, \mathbf{S}_{ki\cdot}^{\text{sim}}\right) \\ &+ \beta^{VAE}\mathcal{L}_{\text{KLdivergence}}(\mathcal{N}(\mu,\sigma), \mathcal{N}(0,1)),\end{aligned}$$

where $\mathcal{N}(0,1)$ is standard normal distribution, and μ and σ are the empirical mean and standard deviation estimated from the output of the encoder. Both the encoder and decoder consist of two hidden layers. Under default settings used throughout this work, we train the network to minimize the loss function with an Adam optimizer with initial learning rate of $1e-3$, and the values for hyperparameters λ_2 and β^{VAE} are respectively 0.1 and $1e-2$. The network is trained for $5000 \times k$, k being the number of cell types.

Estimation of cell type fractions and comparison with flow cytometry

To quantitatively assess the deconvolution algorithm, we first deconvolve six different peripheral blood mononuclear cells (PBMC) bulk datasets for which cell type proportions have already been quantified using flow cytometry (Additional file 1: Table S1). To evaluate deconvolution performance, we utilize root-mean-squared error (*rmse*) and Pearson correlation (r) for cell type-wise comparisons and Jensen-Shannon distance

(*JSD*) for sample-wise comparisons between estimated fractions and ground truth proportions. The evaluation metrics are defined in the "Evaluation metrics" section. To evaluate our approach, we compared it to state-of-the-art deconvolution methods, MuSiC [21], CSx [8], Scaden [9] and TAPE (TAPE-O and TAPE-A) [5], BayesPrism and BayesPrism-M [7], and bMIND [6]. MuSiC and CSx were chosen for their best performances in benchmarking studies [22, 23]. Scaden and TAPE are selected as both are deep learning-based deconvolution approaches, the latter of which, TAPE-A, performs an adaptation of the network weights for test samples. Since deconvolution is linear, we also considered linear MLPs as a deconvolution algorithm. Further details can be found under the "State of the art" section.

We utilize the *PBMC8k* single cell RNA-seq dataset as reference (Additional file 1: Table S2) for all methods. The first two principal components of combined simulated and real PBMC datasets are visualized in Additional file 2: Fig. S1A, illustrating a domain shift between datasets.

For each dataset, DISSECT always obtained the best *JSD* across all datasets (Fig. 2A), leading to an average improvement over the second-placed algorithms of 6 percentage points. On the *GSE65133* dataset, for instance, DISSECT outperforms second-paced Scaden by 8 percentage points (DISSECT: $JSD = 0.145$, Scaden: $JSD = 0.222$). Similarly, DISSECT always obtains the best *rmse* across all datasets and improves over

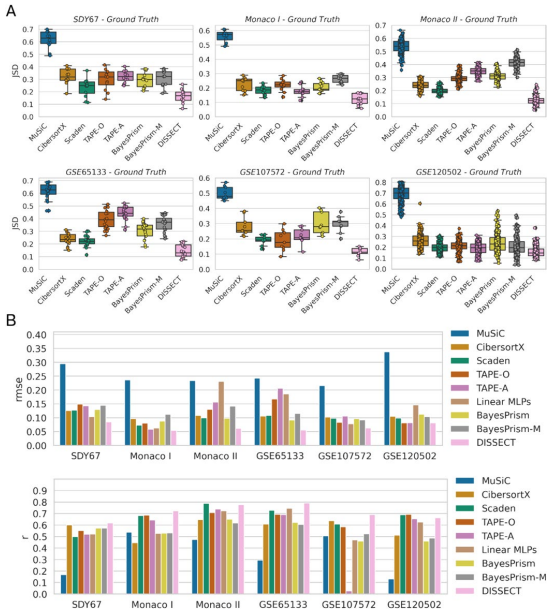


Fig. 2 Evaluation of deconvolution algorithm on six datasets with ground truth information. **A** Per-sample Jensen-Shannon divergence (*JSD*). Each plot corresponds to a dataset. From left to right and top to bottom: *SDY67*, *Monaco I*, *Monaco II*, *GSE65133*, *GSE107572*, and *GSE120502*. **B** Root mean-squared-error (*rmse*, top) averaged over cell types for each of the dataset. Datasets are listed on *x*-axis. Pearson's correlation (*r*, bottom) averaged over cell types

second-placed algorithms by 2 percentage points, on average (Fig. 2B). In addition, it achieved the best *r* on 4 out of 6 datasets (Fig. 2B).

Furthermore, we computed *macro-* level *r* and *rmse* by computing the metrics without making a distinction of cell types as performed previously in [9]. Note that in this setting, *JSD* remains unaffected as it is a sample-level metric and is therefore excluded. We observe that DISSECT achieves consistently best *rmse* across all datasets while achieving best *r* on 5 out of the 6 datasets (Additional file 2: Fig. S1).

Since MuSiC can take advantage of multi-sample references, we also evaluated MuSiC using blood data from the Immune Cell Atlas (ICA) (Additional file 1: Table S2). We also evaluated MuSiC with pre-selected marker genes (MuSiC-M) that were selected by CSx. MuSiC-M showed increased performance in 4 out of 6 datasets (Additional file 2: Fig. S2A-B). MuSiC also shows improved performance in the multi-sample setting in both rmse (Additional file 2: Fig. S2A) and r (Additional file 2: Fig. S2B). DISSECT still reaches best performance in *rmse* (on average 8 percentage points better) and *r* (on average 13 percentage points better) across all datasets.

Next, we evaluated the cell fraction deconvolution performance on the *Monaco I* (Additional file 1: Table S1) dataset, which contains several closely related and rare cell types and constitutes a relatively hard cell deconvolution task, using *Ota* dataset (Additional file 1: Table S1). With a correlation of 0.6, DISSECT's average performance is 14 percentage points better than the second placed Scaden (Additional file 1: Table S3), while Scaden's average RMSE was marginally (1 percentage point) better than second placed DISSECT (Additional file 1: Table S4). To validate that the performance improvement in DISSECT is due to the semi-supervised learning and consistency loss, we performed an ablation study on data *SDY67* by successively and cumulatively removing components of the algorithm and testing it again. The following components were removed successively: consistency regularization, KL Divergence loss (mean squared error instead), and the nonlinear activation function (identity function instead). The ablation results are shown in Additional file 1: Table S5.

In summary, these results provide strong evidence that DISSECT consistently outperforms current state-of-the-art cell type deconvolution algorithms across six different datasets with ground truth information.

Consistency of predictions and relationship between cell type fractions and biological phenotypes

To further corroborate the above results, we evaluate DISSECT's performance on three datasets that do not have paired flow cytometry data. In this section, we compare to other established biological facts as well as divergences over different reference single-cell datasets. The bulk datasets together with literature-based expected biological relationships of cell types are listed in Additional file 1: Table S1.

Brain

The *ROSMAP* dataset consists of 508 bulk RNA-seq samples from the dorsolateral prefrontal cortex (DLPFC) of patients with Alzheimer's disease (AD) as well as non-AD samples (Additional file 1: Table S1). For 463 of these samples, Braak stages of disease severity have been quantified. Correspondingly, single-nuclei RNA-seq (snRNA-seq)

for 48 individuals from the same cohort is available [24]. For 41 of these samples, cell type fractions based on immunohistochemistry (IHC) from a previous work exist [25]. It should be noted that IHC was performed for all neurons and as a result, comparison with respect to excitatory vs inhibitory neurons was not possible. Here, we consider two biological ground truths: first is the ratio of excitatory neurons to inhibitory neurons (Additional file 1: Table S1), and second is the neurodegeneration, or the loss of neurons with increasing Braak Stages [26]. We deconvolved *ROSMAP* using the *Allen Brain Atlas* reference (Additional file 1: Table S2).

We computed the *JSD* between the estimated fractions and IHC cell type proportions. DISSECT estimated fractions had the best average *JSD*s and provides the expected excitatory-inhibitory neuron ratio of (3:1–9:1), while other methods generally underestimated this ratio (Fig. 3A). All methods recover a negative correlation between increasing Braak stages and the fraction of neurons (Additional file 2: Fig. S3).

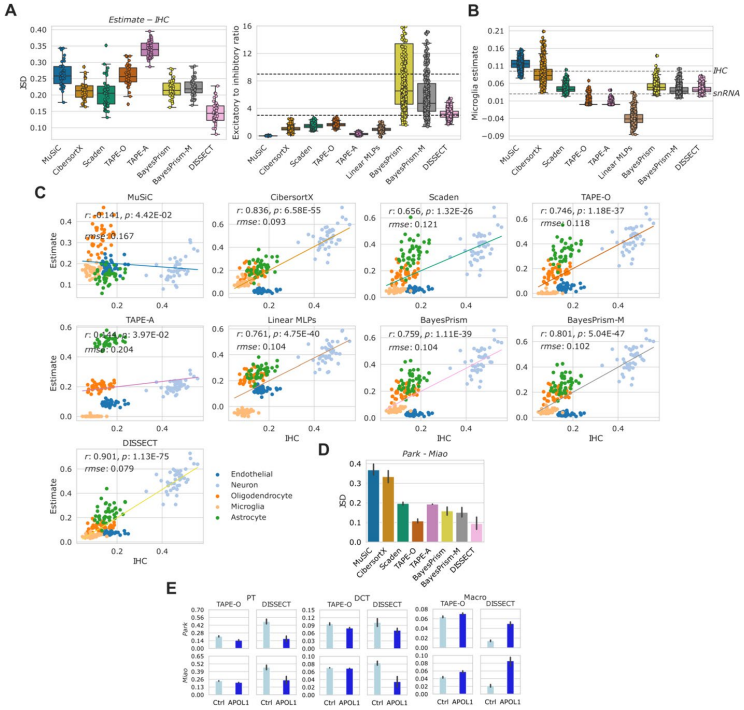


Fig. 3 **A** Left: Box-plots showing *JSD* between estimated fractions and IHC based cell type proportions from 41 individuals from *ROSMAP*. Right: Ratio of excitatory to inhibitory neurons computed from *ROSMAP*. Expected ratios lie between 3:1 and 9:1 as indicated by dashed lines. **B** Boxplots showing microglia proportion as estimated by different methods. Median proportions of microglia estimated using snRNA-seq and IHC are labeled. **C** Correlations between estimates (*y-axis*) and IHC cell type proportions (*x-axis*). **D** *JSD* between predicted proportions from *Kidney* between experiments with *Miao* and *Park* as references. **E** Predictions from TAPE-O and DISSECT from *Kidney*. From left to right: Proximal tubule (PT), ductal convoluted tubules (DCT), and macrophages (Macro). Each row indicates a reference. Error bars show standard deviations, while height of the bars shown mean prediction

Previously, it has been noted that snRNA-seq and IHC data provide different estimates for some cell types, notably microglia and endothelial cells [25]. It is interesting to observe that DISSECT and Scaden were the only methods where the estimates of microglia resembled closely those obtained from snRNA-seq and IHC data (Fig. 3B). We also computed *r* and *rmse* between the IHC cell type proportions and estimated fractions (Fig. 3C). With a correlation *r* of 0.901 DISSECT proved to be 14 percentage points better than the second-placed linear MLP. DISSECT also displayed the best *rmse* at 0.079.

Overall, the comparison to IHC and snRNA-seq ground truth information for the ROSMAP data further strengthens our claim that consistency regularization with DISSECT robustly improves cell deconvolution.

Pancreas

The *GSE50244* bulk RNAseq dataset consists of 89 pancreas samples from healthy and type 2 diabetes (T2D) individuals (Additional file 1: Table S1). For 77 of these samples, hemoglobic 1C levels are available as ground truth information. We performed the deconvolution using three single-cell reference datasets *Baron*, *Segerstolpe*, and *Xin* (Additional file 1: Table S2). Both *Baron* and *Segerstolpe* datasets contain alpha, beta, gamma, delta, acinar, and ductal cell types. While only alpha, beta, gamma, and delta cell types were present in the *Segerstolpe* dataset. To measure the consistency of deconvolution algorithms, we measured *JSD*s between estimated fractions using each of the three references (Additional file 2: Fig. S4A). While several methods showed considerable divergences, indicating reference-dependent deconvolution results, DISSECT displayed the most consistent results with a *JSD* of ~0.1–0.2 across the three pairs. In terms of recovery of significant negative correlations between the estimated fractions of beta cells and hemoglobin 1C (hba1c) levels, DISSECT provided highly significant correlations of between -0.45 and -0.47 across the three references (Additional file 2: Fig. S4B). These results further suggest that DISSECT is both precise and robust in cell type deconvolution on real data and is comparatively less affected by the choice of single-cell reference.

Kidney

The *GSE81492* dataset consists of 10 kidney samples of APOL1 mutant mice, which is a mouse model of chronic kidney disease (CKD) (Additional file 1: Table S1). We deconvolved the dataset using two single cell references: *Miao* and *Park* (Additional file 1: Table S2). Similar to our experiments on the pancreas tissue, we computed *JSD* between the estimated cell type fractions from the two references. DISSECT provided the best average *JSD* (0.09) out of all considered methods (Fig. 3D). We further compare the methods on the recovery of expected relation of cell type fractions with the biological phenotype (Additional file 1: Table S1). Figure 3E compares two best methods on *JSD*, DISSECT, and TAPE-O, while Additional file 2: Fig. S5 presents these results on all cell types for all methods. It is known that CKD results in the decrease in proximal tubule cells (PT) and distal convoluted tubules (DCT). Cell type fractions estimated with DISSECT showed a significant loss of PTs and DCTs and a corresponding increase in macrophages, while TAPE-O provided much smaller differences between the control and CKD model (Fig. 3E). PTs are the most abundant cell type in kidney making up around

50% of a mouse kidney [27]. DISSECT correctly estimates the high abundance of PTs in healthy kidney, while TAPE-O underestimates them (Fig. 3E).

In summary, it is noteworthy that DISSECT shows state-of-the-art precision and robustness in cell type deconvolution across various ground truth information and 9 datasets, including PBMC, brain, pancreas, and kidney bulk RNA-seq samples. DISSECT also shows superior robustness to the choice of single cell reference.

Application to proteomics and spatial transcriptomics

It is conceivable that DISSECT's consistency regularization for bulk RNA-seq cell type deconvolution should also lend itself to other biomedical datatypes in which domain shifts might be a problem. Applications might include, for example, the deconvolution of spatial transcritomic (ST) and bulk proteomic data with supra-cellular resolution. In order to evaluate these potential use-cases, we performed deconvolution of spatial transcriptomics and proteomics samples. Here, our aim is to test the hypothesis of applicability of DISSECT on these data modalities and we do not intend to perform an exhaustive comparison to multiple methods developed for these modalities. For comparisons on spatial transcriptomics, we consider four state-of-the-art spatial deconvolution methods, RCTD [28], Cell2location (C2L) [29] as shown to perform among the best in the benchmarking study [30]. We also include SONAR [31] and CARD [32], both of which can utilize spatial information. For comparisons on proteomic deconvolution, we consider the tested bulk deconvolution methods.

Spatial transcriptomics

We evaluated DISSECT on the task of spatial deconvolution using mouse brain and human lymph node samples (Additional file 1: Table S1). As a ground truth, we considered relationships with biological phenotypes in line with our application of kidney and pancreas datasets (Additional file 1: Table S1). Due to the spatial nature of the ST, we could verify the recovery of neuronal layers in brain (Additional file 2: Fig. S6) and discernment of germinal centers in lymph node (Additional file 2: Fig. S7). DISSECT performs on par with C2L and RCTD on both datasets. The results are provided and discussed in detail in the Additional file 2: Supplementary Note.

Proteomics

To compare the ability of the tested deconvolution methods to recover cell type proportions from proteomics mixtures, we utilized 50 human brain samples (Additional file 1: Table S1). We applied each deconvolution method on these samples using the Allen Brain Atlas reference (Additional file 1: Table S2). Compared to other methods, DISSECT recovered excitatory neurons to be the expected majority population in both datasets while maintaining the excitatory to inhibitory neuron ratio to be around expected range of (3:1–9:1) (Additional file 2: Fig. S8). These results strongly suggest that DISSECT reaches state-of-the-art performance on proteomic cell type deconvolution and might be applicable to other biomedical data types.

Evaluation of DISSECT under domain shifts

To assess the impact of consistency regularization on the performance of DISSECT and other algorithms, we used *Ota* dataset (Additional file 1: Table S1). Using this dataset in a dynamic domain shift setup (see the "Domain shift experimental setup" section), we evaluated the performance of deconvolution methods. We also included DISSECT without consistency (DISSECT w/o consistency) to asses the impact of semi-supervised learning under varying shifts. The performance of all methods dropped significantly for test sets with domain shifts (Additional file 2: Fig. S9). However, the drop in performance was much lower for DISSECT than other methods. Furthermore, a clear advantage of semi-supervised learning with consistency regularization is observed in comparison to DISSECT without consistency, especially in terms of *rmse*.

Estimation of cell type-specific gene expression

So far, we have shown that DISSECT can reliably deconvolve cell fractions. In this section, we focus on the deconvolution and inference of cell type-specific gene expression from bulk RNA-seq mixtures using our novel conditional autoencoder based algorithm (Fig. 1). While we were able to use ground truth flow cytometry data for the evaluation of cell fractions, no such gold-standard is available for cell type-specific gene expression information. In consequence, we measure DISSECT's gene expression inference performance on simulated bulk RNA-seq data. To maintain a domain shift between the training and test datasets, we simulated data for training and testing using different single-cell datasets. We compared the performance of DISSECT with that of TAPE-A, bMIND, and BayesPrism, all of which can infer cell type-specific gene expression per sample. We simulated bulk samples from one of the four reference single-cell PBMC datasets listed in Additional file 1: Table S2 and created training simulations from the remaining three. Simulations from each single-cell dataset consisted of 6000 samples. To evaluate the performance of DISSECT and other methods, we compared the true and estimated gene expression profiles of each cell type for each simulated sample (sample-wise) and for each gene (gene-wise) using Spearman correlation. These sets of results were aggregated across cell types and averaged. DISSECT displays the best sample- and gene-wise correlations in 6 out of 8 experiments, outperforming TAPE-A by 0.025 ± 0.023 in the sample-wise comparisons and by 0.012 ± 0.029 in the gene-wise comparisons (Table 1). Moreover, DISSECT exhibited an improvement in both sample and gene-wise metrics, exemplifying its advantage.

These results indicate that DISSECT's consistency regularization robustly performs state-of-the-art cell type-specific gene expression deconvolution.

Discussion

In this work, we first formally define the task of cell deconvolution and outline the hypothesis that semi-supervised consistency regularization should improve bulk RNA-seq deconvolution when learning from single cell RNA-seq data. We then provide evidence that our novel deep learning-based algorithm, DISSECT, outperforms competing state-of-the-art algorithms in deconvolution, both on a cellular and gene expression level, across many different datasets. This included 6 PBMC datasets with

Table 1 Spearman correlation between ground truth and estimated gene expression profiles on simulated datasets averaged over samples. The column *Dataset* indicates the single-cell dataset used to create simulations for the test set

Dataset	TAPE-A	bMIND	BayesPrism	DISSECT
sample-wise r				
PBMC6k	**0.83±0.09**	0.80 ± 0.07	**0.83 ± 0.11**	0.82 ± 0.08
PBMC8k	0.79 ± 0.09	0.80 ± 0.08	0.81 ± 0.09	**0.84±0.11**
DonorA	0.85 ± 0.11	0.84 ± 0.09	0.80 ± 0.09	**0.89±0.10**
DonorC	0.81 ± 0.12	**0.83±0.11**	0.80 ± 0.08	**0.83±0.08**
gene-wise r				
PBMC6k	0.42 ± 0.14	**0.46±0.14**	0.41 ± 0.14	**0.46±0.15**
PBMC8k	**0.51±0.12**	0.44 ± 0.18	0.45 ± 0.12	0.48 ± 0.14
DonorA	**0.48 ± 0.20**	0.45 ± 0.16	0.46 ± 0.18	**0.48±0.18**
DonorC	0.45 ± 0.11	0.43 ± 0.15	0.45 ± 0.12	**0.49±0.12**

For each dataset, values with the highest mean correlation are displayed in bold font

ground truth flow cytometry information and 3 datasets (brain, pancreas, and kidney) with other established biological facts as ground truth information. Across the board, DISSECT provided the best cell type deconvolution results when compared to four state-of-the-art methods, while also being comparatively robust to the choice of single-cell reference. We follow a two-step procedure because the assumptions for each of the algorithms differ, and we do not foresee any significant benefit from iteratively deconvolving cell type fractions and gene expression. In a case study, we also show how our algorithm can easily be adapted to deconvolve cell types of proteomic and spatial expression data. For the spatial transcriptomics data, DISSECT estimates cell type fractions per spot, which are constrained to sum to 1. To be able to estimate the number of cells per cell type for each spot, and to map single cells, DISSECT estimates can be used as a prior for algorithms such as CytoSpace [33]. CytoSpace infers both the number of cells in each spot and solves an optimization problem to map single cells to their spatial locations. To estimate only the number of cells per cell type for each spot, the total number of cells as estimated by CytoSpace can be multiplied with the output of DISSECT. While these results are not exhaustive, they nevertheless show the applicability of DISSECT on other biomedical data types, a research avenue we might pursue in more depth in the future. In addition to DISSECT's state-of-the-art cell type fraction deconvolution (an average improvement of 0.063 in *JSD* and 0.021 in *rmse* over the state of the art on the datasets with ground truth cell type fractions), it achieved best cell type-specific gene expression deconvolution results in 6 out of 8 comparisons across four simulated datasets with an average improvement of 0.025 in the sample-wise and 0.012 in the gene-wise comparisons.

While we focused on MLPs for the estimation of cell type fractions and an autoencoder for gene expression estimation in this work, consistency regularization might also improve other deconvolution algorithms.

No gold standard ground truth exists for quantitative assessment of estimated cell type-specific gene expression between two conditions for real bulk RNA-seq datasets. This is a limitation of the experimental setup presented for cell type-specific gene

expression estimation. A potential solution will be to develop biologically valid benchmark datasets that can be evaluated at scale.

While DISSECT outperforms competing algorithms in cell type fraction and cell type-specific gene expression deconvolution, some results leave room for further improvement. DISSECT accurately distinguishes cell types where the transcriptional difference reflects cell subtypes, for instance PBMCs (CD4 T cells and CD8 T cells), pancreas (pancreatic islets), kidney (tubular epithelial cells), and brain (OPC and oligodendrocytes). However, when estimating granular cell type proportions in the Monaco I dataset, error rates exceeded the ground truth proportions (*rmse*>0.01 for cell subsets present at less than 1%). Therefore, for cell types that make up less than 1% of all cells and cells with very similar gene expression, for instance CD4 T and activated CD4 T cells, deconvolution algorithms should be used with caution. Future research into semi-supervised and contrastive algorithms as well as data augmentation and integration techniques should further enhance DISSECT's performance on hard deconvolution tasks.

Conclusions

In conclusion, DISSECT provides a semi-supervised deep learning framework to estimate cell type proportions and per-sample cell type-specific gene expression, is robust across datasets and tissues, and is easily applicable to other data modalities. DISSECT delivers state-of-the-art deconvolution performance, as long as cell types are not too closely related and make up more than 1% of all cells.

Methods

Evaluation metrics

To quantitively evaluate estimated cell type fractions across samples, we used two metrics, namely Pearson's correlation (*r*) and root-mean-squared error (*rmse*). Given x and y as estimated fractions and ground truth respectively,

$$r = \frac{cov(x, y)}{\sigma_x, \sigma_y} \tag{6}$$

$$rmse = \sqrt{Avg(x - y)^2} \tag{7}$$

To compute sample-wise divergences two list of fractions x_i and y_i for the same sample i, we used Jensen-Shannon distance (JSD) which is the square root of Jensen-Shannon divergence. *JSD* is given as

$$JSD(x \| y) = \sqrt{\frac{D(x_i \| m_i) + D(y_i \| m_i)}{2}}, \tag{8}$$

where $m_i = \frac{(x_i + y_i)}{2}$ and D is the Kullback-Leibler divergence.

State of the art

Here, we briefly detail the state-of-the-art deconvolution approaches. Out of these methods, CSx, TAPE, BayesPrism, and bMIND can also estimate per sample cell type-specific gene expression signatures.

MuSiC

MuSiC [21] uses weighted non-negative least squares. MuSiC maintains cross-cell and cross-sample consistencies by appropriately weighting genes based on their informativity during an iterative procedure. We used MuSiC R package (version 1.0.0). Deconvolution using MuSiC was performed according to the authors recommendations. Since MuSiC is a method that utilizes multi-subject scRNA-seq datasets, when available, we used cells from multiple subjects in deconvolution with MuSiC. We used the default hyperparameters to execute MuSiC. For single-cell datasets with multiple donors (Additional file 1: Table S2), we ran MuSiC with single-cell data from all available donors.

CSx

CSx [8] is a deconvolution method that addresses domain gap problems with scRNA-seq and bulk samples by aiming to correct batch effects. It uses scRNA-seq to generate a cell type specific signature matrix and uses ν-support vector regression as the underlying algorithm. To construct the signature matrix, we used the following hyperparameters for CSx as recommended by the authors: kappa = 999, q-value = 0.01 and number of genes within a range of 300 and 500. The quantile normalization was also disabled. CSx comprises two modes, S- and B-modes, to address the domain gap. S-mode is used when deconvolving with a signature matrix constructed using a scRNA-seq dataset, while B-mode is used when deconvolving with a signature matrix constructed using purified samples. We followed the documentation provided by the authors to run CSx and used the S-mode. CSx can also predict gene expression signatures for each sample for which it uses a non-negative matrix factorization based iterative algorithm. However, CSx only estimates genes likely to be differentially expressed in one of the bulk samples and as such the evaluations for simulations from healthy PBMC single-cells are not possible. We ran CSx through docker container obtained from [34].

Scaden

Scaden [9] is an average ensemble of three deep neural networks with different architectures that was developed for cell fraction deconvolution. Each network is trained only on simulated pseudo bulk data generated from an scRNA-seq reference similar to described above. Scaden is provided as a Python package. We used the official Scaden package (version 1.1.2) with the instructions provided by the authors to train the networks.

TAPE

TAPE [5] is a fully connected autoencoder where the bottleneck consists of cell type fractions. The architecture of the encoder is similar to the archictecture of Scaden but with CeLU activations. The decoder consists of linear activations and outputs gene expression of the input vector. The adaptive mode of TAPE (TAPE-A) aims at optimizing the network for bulk samples, while the overall mode trains for fractions with an added loss function that reconstructs input bulk expression from fractions. Since TAPE-A reconstructs gene expression from fractions (bottleneck), the signature matrix is visible in the (linear) decoder. To estimate gene expression signatures for each bulk sample, decoder weights are optimized per-sample using an iterative optimization strategy. Network weights are changed during the two modes, we compare with both and refer to TAPE in overall mode as TAPE-O and in adaptive mode as TAPE-A. We used the official scTAPE package (version 1.1.2) implemented in Python.

Linear MLPs

The solution to the deconvolution problem could be, in principle, a linear function. For this reason, we also compared to an MLP ensemble that has similar architecture to DISSECT, but in which we replaced all non-linear activations with an identity function and removed the consistency loss.

BayesPrism and BayesPrism-M

Primarily a method developed for oncology bulk datasets, BayesPrism [7] is a Bayesian framework to infer cell type fraction and cell type specific per-sample gene expression. It models gene expression as multinonmial distribution and calculates the cumulative posterior across cell states to derive the statistics for individual cell types. To evaluate BayesPrism with preselected marker genes using *select.marker* function. We utilize official implementation of BayesPrism in R (version 2.1.2).

bMIND

bMIND [6] is a Bayesian method to infer cell type specific gene expression per sample based on single-cell gene expression for given cell types. Using the prior from single-cell gene expression, bMIND models bulk gene expression as the product of gene expression and cell type fractions as a Bayesian mixed-effects model. bMIND uses cell type fractions as estimated by other deconvolution methods as its input. We used default settings of bMIND in our experiemnts with its R implementation (version 0.3.3).

Pre-processing and simulations

Quality control

Before simulating from reference datasets, we remove cells with less than 200 expressed genes and genes which are expressed in less than 3 cells. Furthermore, we also remove cells expressing more than 4% mitochondrial genes. Thereafter, before each deconvolution, we subset reference and bulk datasets to include only the common genes between the two. This quality control step was identical for all methods.

Simulations for deconvolution of bulk RNA-seq samples and proteomics

For deep learning methods, we sampled $\alpha_{k,i}$ uniformly to generate simulations *s.t.* $\sum_{k=1}^{c} \alpha_{k,i} = 100, \forall i$ if the dataset is single-cell. For experiments on granular level cell types where simulations are done from purified cell samples, we modified the simulation procedure to reflect this. In this case, a simulated sample is given by $\mathbf{B}_{i\cdot}^{\text{sim}} = \sum_{k=1}^{c} \mathbf{X}_{ik}^{\text{sim}} \mathbf{b}_{\mathbf{l}}^{\mathbf{k}}$, where $\mathbf{b}_{\mathbf{l}}^{\mathbf{k}}$ is the expression vector of purified sample l belonging to cell type k. For all experiments, we simulated total $1000 \times c$ simulations where c is number of cell types in the reference dataset.

Simulations for deconvolution of 10x Visium ST samples

We adjusted simulation procedure to mimic ST datasets. 10x Visium (one of the technologies to generate ST samples) consists of around 10 cells per spot. To reflect this, we simulated between 5 and 12 cells to generate one spot (i.e., $\sum_{k=1}^{c} \alpha_{ki} \sim [5, 12]$). Since ST is much sparser, to generate one spot, we kept between 2 and 6 cell types. Due to sparsity of spots, not all cell types are present in a given spot. To account for this and to make comparison across spots possible, we utilized the outputs of the last layer (before performing softmax operation) and set negative predictions to zero. Thereafter, we re-normalized these absolute scores by such that each prediction sum to one. For all experiments, we simulated total $1000 \times c$ simulations where c is number of cell types in the reference dataset.

Deconvolution of proteomics data

For deconvolution of proteomics data, it is not valid to mix protein intensities and gene expression due to different normalizations. Instead of mixing simulated samples with real samples, proteomics samples were mixed with each other, i.e., at each training step, $\mathbf{B}_{i\cdot}^{\text{mix}} = \beta \mathbf{B}_{r_1\cdot} + (1-\beta)\mathbf{B}_{r_2\cdot}$, where r_1 and r_2 are two randomly selected proteomics samples at the training step.

Pre-processing for estimation of cell type fractions

For Scaden, TAPE, linear MLPs, and DISSECT, before passing simulated and real bulk samples to the network, we normalize samples to sum to a million counts (counts per million (CPM)) and log scale them with base 2 after adding 1. CPM normalization was performed to maintain total mRNA expressed per gene to be out of a fixed total gene expression, and CPM is widely used in computational genomics. During training, for each batch, we normalize each sample by *MinMax* scaling. These are standard preprocessing steps [9].

For MuSiC and CSx (under S-mode), data was supplied on a linear scale as suggested in their respective publications and no change was made to the default normalization methods of both [8, 21].

To estimate cell type specific gene expression profiles, we need to maintain relationship between gene expression of individual cell types and simulated bulks, which would

be lost if we perform CPM normalization of both simulated samples and corresponding cell type specific gene expression profiles. Hence, instead of performing CPM normalization of simulated bulks, we normalize each test bulk sample to sum to the mean of sums of simulated samples. Furthermore, for estimating cell type specific gene expression, we want to maintain gene level information across samples. To achieve this, instead of normalizing each sample using *MinMax* scaling, we perform *MinMax* scaling globally over all samples.

For TAPE, since the signature matrix is observed in decoder (see the "State of the art" section), preprocessing step is similar to the preprocessing done in estimating cell type fractions.

Hyperparameters and fine-tuning

We fine tuned the network for activation functions, learning rate, and batch size using randomized search with hyperopt [35] with the root mean squared error as the objective function. The following grids were used for the optimization: activations = [linear, ReLU, ReLU6, Swish], learning rate = [5e−3, 1e−3, 5e−4, 1e−4, 5e−5, 1e−5], λ_1 = [0,1,5,10,15] with or without scheduled change at every 2000 steps and batch sizes = [32, 64, 128, 256] with 50 iterations on Ascites bulk dataset as used in Scaden [9]. Other hyperparameters were fixed to the default hyperparameters of Scaden. The optimal hyperparameters were fixed for all experiments, with batch size = 64, learning rate = 1e−5, activation function = ReLU6, λ_1 according to schedule [0,15,10] at steps [0,2000,4000], and number of steps = 5000.

Domain shift experimental setup

Using the *Ota* dataset (Additional file 1: Table S1) that contains 9852 purified samples belonging to immune cell subsets including several B cell and T cell subsets as shown in Additional file 1: Table S3, we created an experimental setup with domain shifts involving the following 4 scenarios. *20% split*: We randomly split the dataset into training (80%) and test sets (20%). *Activated 1*: We used the same split as in *20% split*. We removed certain CD4 and CD8 T cell subsets, namely, CD4 T memory, CD8 TEM, and CD8 TE from the training split while they were kept in the test set. In the test set, on the other hand, other subsets (CD4 T naive, CD8 T naive, and CD8 TCM) were removed. *Activated 2*: We followed the same procedure as in *Activated 1* except we removed certain B cell subsets, namely, B NSM, BEx, and BSM from the training set while they were kept in the test set. B naive subset was removed from the test set. Finally, for a model-based domain shift, we used DISCERN [36] to project the test set of *20% split* to the dataset simulated from *pbmc8k* and used in deconvolving PBMC bulk RNAseq. The CD4 T cell, CD8 T cell, and B cell subsets, regardless of their subtype identity, were labeled as CD4Tcells, CD8Tcells, and Bcells to allow comparisons. In each scenario, 6000 samples were simulated.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03251-5.

Additional file 1. Supplementary tables. The file contains supplementary tables [47-74].

Additional file 2. Supplementary figures. The file contains supplementary figures and supplementary note [75, 76].

Additional file 3. Review history. The file contains the peer review history.

Acknowledgements
We would like to thank Fabian Hausmann for helpful discussions and the MAXOMOD consortium for providing the proteomic data.

Review history
The review history is available as Additional file 3.

Peer review information
Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
SB initiated, and SB, PM, and RK conceptualized the study. RK wrote the code and implemented DISSECT. RK analyzed the data with help from SB. SB and RK interpreted the results and wrote the manuscript. PM provided ideas and reviewed the manuscript. All authors read and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. R.K. was supported by FOR5068 P9 and the 3R initiative of the UKE, P. M. by SFB1286 SP02, and S.B. by EU E-rare MAXOMOD, the M3I excellence initiative of the UKE, and SFB1192 B8 and C3.

Availability of data and materials
The datasets analyzed in this work are publicly available. Summary of the datasets are provided in Additional file 1: Tables S1 (bulk) and S2 (single-cell). PBMC single-cell RNA seq was obtained from 10x Genomics [37]. Single-cell RNA-seq dataset of the brain was obtained from Allen Brain Map [38]. Single-cell RNA-seq datasets from kidneys and pancreas can be accessed using Gene Expression Omnibus using corresponding accession codes: GSE157079 (***Miao***) and GSE107585 (***Park***), GSE81608 (***Xin***), GSE84133 (***Baron***). The raw single-cell data for ***Segertolpe*** is available at ArrayExpress (EBI) with accession code E-MTAB-5061. Cross-tissue Immune Cell Atlas (ICA) is available from [39]. Bulk RNA-seq datasets titled ***Monaco I***, ***Monaco II***, ***GSE120502***, ***GSE107572***, ***GSE50244***, and ***GSE81492*** are available from Gene Expression Omnibus [40] with following accession codes: GSE107011, GSE106898, GSE120502, GSE107572, GSE50244, and GSE81492. Bulk RNA-seq dataset ***SDY67*** was obtained from data resource provided in [9]. The original source for ***SDY67*** is ImmPort with accession code SDY67. ***ROSMAP*** cohort dataset is available from Synapse [41] with accession code syn3219045. The pre-processed data was obtained from [42]. The bulk proteomics data of post-mortem human brain samples was obtained from MAXOMOD consortium [43]. Allen brain reference with cortex annotations was obtained from [44].

Code availability
DISSECT is implemented in Python using tensorflow and keras (both versions 2.7.0) frameworks. The code is available at GitHub [45] with installation and usage instructions and has been deposited to Zenodo [46] under MIT license.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 10 July 2023 Accepted: 17 April 2024
Published online: 30 April 2024

References

1. Li X, Wang CY. From bulk, single-cell to spatial RNA sequencing. Int J Oral Sci. 2021;13(1):1–6.
2. Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, et al. Eleven grand challenges in single-cell data science. Genome Biol. 2020;21(1):1–35.
3. Zhou JG, Liang B, Jin SH, Liao HL, Du GB, Cheng L, et al. Development and validation of an RNA-seq-based prognostic signature in neuroblastoma. Front Oncol. 2019;9:1361.
4. Roberts KG, Li Y, Payne-Turner D, Harvey RC, Yang YL, Pei D, et al. Targetable kinase-activating lesions in Ph-like acute lymphoblastic leukemia. N Engl J Med. 2014;371(11):1005–15.

5. Chen Y, Wang Y, Chen Y, Cheng Y, Wei Y, Li Y, et al. Deep autoencoder for interpretable tissue-adaptive deconvolution and cell-type-specific gene analysis. Nat Commun. 2022;13(1):6735.
6. Wang J, Roeder K, Devlin B. Bayesian estimation of cell type-specific gene expression with prior derived from single-cell data. Genome Res. 2021;31(10):1807–18.
7. Chu T, Wang Z, Pe'er D, Danko CG. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. Nat Cancer. 2022;3(4):505–17.
8. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol. 2019;37(7):773–82.
9. Menden K, Marouf M, Oller S, Dalmia A, Magruder DS, Kloiber K, et al. Deep learning–based cell composition analysis from tissue expression profiles. Sci Adv. 2020;6(30):eaba2619.
10. Long M, Cao Y, Wang J, Jordan M. Learning transferable features with deep adaptation networks. In: Bach F, Blei D, editors. Proceedings of the 32nd International Conference on Machine Learning. vol. 37 of Proceedings of Machine Learning Research. Lille: PMLR; 2015. pp. 97–105. https://proceedings.mlr.press/v37/long15.html.
11. Büttner M, Miao Z, Wolf FA, Teichmann SA, Theis FJ. A test metric for assessing single-cell RNA-seq batch correction. Nat Methods. 2019;16(1):43–9.
12. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118–27.
13. Wang S, Pisco AO, McGeever A, Brbic M, Zitnik M, Darmanis S, et al. Leveraging the Cell Ontology to classify unseen cell types. Nat Commun. 2021;12(1):5556.
14. Maden SK, Kwon SH, Huuki-Myers LA, Collado-Torres L, Hicks SC, Maynard KR. Challenges and opportunities to computationally deconvolve heterogeneous tissue with varying cell sizes using single-cell RNA-sequencing datasets. Genome Biol. 2023;24(1):288.
15. Domínguez Conde C, Xu C, Jarvis L, Rainbow D, Wells S, Gomes T, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science. 2022;376(6594):eabl5197.
16. Ju C, Bibaut A, van der Laan M. The relative performance of ensemble methods with deep convolutional neural networks for image classification. J Appl Stat. 2018;45(15):2800–18.
17. Hannun A, Case C, Casper J, Catanzaro B, Diamos G, Elsen E, et al. Deep speech: scaling up end-to-end speech recognition. 2014. arXiv preprint arXiv:14125567. https://arxiv.org/abs/1412.5567.
18. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA. 2018. pp. 4510–20. https://ieeexplore.ieee.org/document/8578572.
19. Ramachandran P, Zoph B, Le QV. Searching for activation functions. 2017. arXiv preprint arXiv:171005941. https://arxiv.org/abs/1710.05941.
20. Berthelot D, Carlini N, Goodfellow I, Papernot N, Oliver A, Raffel CA. Mixmatch: a holistic approach to semi-supervised learning. Adv Neural Inf Process Syst. 2019;32:5050–60.
21. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat Commun. 2019;10(1):1–9.
22. Avila Cobos F, Alquicira-Hernandez J, Powell JE, Mestdagh P, De Preter K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. Nat Commun. 2020;11(1):1–14.
23. Jin H, Liu Z. A benchmark for RNA-seq deconvolution analysis under dynamic testing environments. Genome Biol. 2021;22(1):1–23.
24. Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, et al. Single-cell transcriptomic analysis of Alzheimer's disease. Nature. 2019;570(7761):332–7.
25. Patrick E, Taga M, Ergun A, Ng B, Casazza W, Cimpean M, et al. Deconvolving the contributions of cell-type heterogeneity on cortical gene expression. PLoS Comput Biol. 2020;16(8):e1008120.
26. Braak H, Del Tredici K, Rüb U, De Vos RA, Steur ENJ, Braak E. Staging of brain pathology related to sporadic Parkinson's disease. Neurobiol Aging. 2003;24(2):197–211.
27. Clark JZ, Chen L, Chou CL, Jung HJ, Lee JW, Knepper MA. Representation and relative abundance of cell-type selective markers in whole-kidney RNA-Seq data. Kidney Int. 2019;95(4):787–96.
28. Cable DM, Murray E, Zou LS, Goeva A, Macosko EZ, Chen F, et al. Robust decomposition of cell type mixtures in spatial transcriptomics. Nat Biotechnol. 2022;40(4):517–26.
29. Kleshchevnikov V, Shmatko A, Dann E, Aivazidis A, King HW, Li T, et al. Cell 2location maps fine-grained cell types in spatial transcriptomics. Nat Biotechnol. 2022;40(5):661–71.
30. Li B, Zhang W, Guo C, Xu H, Li L, Fang M, Hu Y, Zhang X, Yao X, Tang M, Liu K. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. Nat Methods. 2022;19(6):662–70.
31. Liu Z, Wu D, Zhai W, Ma L. SONAR enables cell type deconvolution with spatially weighted Poisson-Gamma model for spatial transcriptomics. Nat Commun. 2023;14(1). https://doi.org/10.1038/s41467-023-40458-9.
32. Ma Y, Zhou X. Spatially informed cell-type deconvolution for spatial transcriptomics. Nat Biotechnol. 2022;40(9):1349–59. https://doi.org/10.1038/s41587-022-01273-7.
33. Vahid MR, Brown EL, Steen CB, Zhang W, Jeon HS, Kang M, Gentles AJ, Newman AM. High-resolution alignment of single-cell and spatial transcriptomes with CytoSPACE. Nat Biotechnol. 2023;41(11):1543–8.
34. CIBERSORTx. https://cibersortx.stanford.edu/. Accessed 30 Jan 2024.
35. Bergstra J, Yamins D, Cox D. Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. In: International conference on machine learning. PMLR; 2013. pp. 115–23.
36. Hausmann F, Ergen C, Khatri R, Marouf M, Hänzelmann S, Gagliani N, et al. DISCERN: deep single-cell expression reconstruction for improved cell clustering and cell subtype and state detection. Genome Biol. 2023;24(1):212.
37. 10x Genomics. https://www.10xgenomics.com. Accessed 30 Jan 2024.
38. Allen Brain Map. https://portal.brain-map.org. Accessed 30 Jan 2024.

39. Cross-tissue Immune Cell Atlas. https://www.tissueimmunecellatlas.org. Accessed 30 Jan 2024.
40. Gene Expression Omnibus (GEO). https://www.ncbi.nlm.nih.gov/geo/. Accessed 30 Jan 2024.
41. Synapse. https://www.synapse.org. Accessed 30 Jan 2024.
42. Deconvolution of cellular heterogeneity in brain transcriptomes. https://github.com/ellispatrick/CortexCellDeconv. Accessed 30 Jan 2024.
43. Caldi Gomes L, Hänzelmann S, Hausmann F, Khatri R, Oller S, Parvaz M, et al. Multiomic ALS signatures highlight sex differences and molecular subclusters and identify the MAPK pathway as therapeutic target. bioRxiv. 2023;2023–08.
44. Reference Atlas :: Allen Brain Atlas: Mouse Brain. https://mouse.brain-map.org/static/atlas. Accessed 30 Jan 2024.
45. Khatri R, Machart P, Bonn S. Deep semi-supervised consistency regularization for accurate cell type fraction and gene expression estimation. https://github.com/imsb-uke/DISSECT. Accessed 30 Jan 2024.
46. Khatri R, Machart P, Bonn S. Deep semi-supervised consistency regularization for accurate cell type fraction and gene expression estimation. Zenodo. 2024. https://doi.org/10.5281/zenodo.10570404.
47. Zimmermann MT, Oberg AL, Grill DE, Ovsyannikova IG, Haralambieva IH, Kennedy RB, et al. System-wide associations between DNA-methylation, gene expression, and humoral immune response to influenza vaccination. PloS ONE. 2016;11(3):e0152034.
48. Monaco G, Lee B, Xu W, Mustafah S, Hwang YY, Carre C, et al. RNA-Seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. Cell Rep. 2019;26(6):1627–40.
49. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12(5):453–7.
50. Finotello F, Mayer C, Plattner C, Laschober G, Rieder D, Hackl H, et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. Genome Med. 2019;11(1):1–20.
51. Harrison GF, Sanz J, Boulais J, Mina MJ, Grenier JC, Leng Y, et al. Natural selection contributed to immunological differences between hunter-gatherers and agriculturalists. Nat Ecol Evol. 2019;3(8):1253–64.
52. Ota M, Nagafuchi Y, Hatano H, Ishigaki K, Terao C, Takeshima Y, et al. Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases. Cell. 2021;184(11):3006–21.
53. Alejandro EU, Gregg B, Blandino-Rosano M, Cras-Méneur C, Bernal-Mizrachi E. Natural history of β-cell adaptation and failure in type 2 diabetes. Mol Asp Med. 2015;42:19–41.
54. Saisho Y. β-cell dysfunction: its critical role in prevention and management of type 2 diabetes. World J Diabetes. 2015;6(1):109.
55. Wang X, Misawa R, Zielinski MC, Cowen P, Jo J, Periwal V, et al. Regional differences in islet distribution in the human pancreas-preferential beta-cell loss in the head region in patients with type 2 diabetes. PLoS ONE. 2013;8(6):e67454.
56. Fadista J, Vikman P, Laakso EO, Mollet IG, Esguerra JL, Taneera J, et al. Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. Proc Natl Acad Sci. 2014;111(38):13924–9.
57. Venkatachalam MA, Weinberg JM, Kriz W, Bidani AK. Failed tubule recovery, AKI-CKD transition, and kidney disease progression. J Am Soc Nephrol. 2015;26(8):1765–76.
58. Liu BC, Tang TT, Lv LL, Lan HY. Renal tubule injury: a driving force toward chronic kidney disease. Kidney Int. 2018;93(3):568–79.
59. Malhotra R, Craven T, Ambrosius WT, Killeen AA, Haley WE, Cheung AK, et al. Effects of intensive blood pressure lowering on kidney tubule injury in CKD: a longitudinal subgroup analysis in SPRINT. Am J Kidney Dis. 2019;73(1):21–30.
60. Beckerman P, Bi-Karchin J, Park ASD, Qiu C, Dummer PD, Soomro I, et al. Transgenic expression of human APOL1 risk variants in podocytes induces kidney disease in mice. Nat Med. 2017;23(4):429–38.
61. Streit WJ, Braak H, Xue QS, Bechmann I. Dystrophic (senescent) rather than activated microglial cells are associated with tau pathology and likely precede neurodegeneration in Alzheimer's disease. Acta Neuropathol. 2009;118(4):475–85.
62. Hindle JV. Ageing, neurodegeneration and Parkinson's disease. Age Ageing. 2010;39(2):156–61.
63. Fu H, Possenti A, Freer R, Nakano Y, Hernandez Villegas NC, Tang M, et al. A tau homeostasis signature is linked with the cellular and regional vulnerability of excitatory neurons to tau pathology. Nat Neurosci. 2019;22(1):47–56.
64. Alreja A, Nemenman I, Rozell CJ. Constrained brain volume in an efficient coding model explains the fraction of excitatory and inhibitory neurons in sensory cortices. PLoS Comput Biol. 2022;18(1):e1009642.
65. Winer J, Larue D. Populations of GABAergic neurons and axons in layer I of rat auditory cortex. Neuroscience. 1989;33(3):499–515.
66. Ouellet L, de Villers-Sidani E. Trajectory of the main GABAergic interneuron populations from early development to old age in the rat primary auditory cortex. Front Neuroanat. 2014;8:40.
67. Braitenberg V, Schüz A. Cortex: Statistics and geometry of neuronal connectivity. 2nd thoroughly revised edition of: Anatomy of the cortex. Statistics and geometry (1991), 249. Springer Verlag Tiergarten. 1998;17:69121.
68. Beaulieu C. Numerical data on neocortical neurons in adult rat, with special reference to the GABA population. Brain Res. 1993;609(1–2):284–92.
69. Mostafavi S, Gaiteri C, Sullivan SE, White CC, Tasaki S, Xu J, et al. A molecular network of the aging human brain provides insights into the pathology and cognitive decline of Alzheimer's disease. Nat Neurosci. 2018;21(6):811–9.
70. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. Cell Syst. 2016;3(4):346–60.
71. Segerstolpe Å, Palasantza A, Eliasson P, Andersson EM, Andréasson AC, Sun X, et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. Cell Metab. 2016;24(4):593–607.
72. Xin Y, Kim J, Okamoto H, Ni M, Wei Y, Adler C, et al. RNA sequencing of single human islet cells reveals type 2 diabetes genes. Cell Metab. 2016;24(4):608–15.
73. Park J, Shrestha R, Qiu C, Kondo A, Huang S, Werth M, et al. Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. Science. 2018;360(6390):758–63.

74. Miao Z, Balzer MS, Ma Z, Liu H, Wu J, Shrestha R, et al. Single cell regulatory landscape of the mouse kidney highlights cellular differentiation programs and disease targets. Nat Commun. 2021;12(1):1–17.
75. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. In: 2010 20th international conference on pattern recognition. IEEE; 2010. pp. 3121–4.
76. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.

Publisher's Note

DISSECT: deep semi-supervised consistency regularization for accurate cell type fraction and gene expression estimation
– Supplementary tables –

Robin Khatri, Pierre Machart, Stefan Bonn*

Institute of Medical Systems Biology, Center for Molecular Neurobiology
Center for Biomedical AI
University Medical Center Hamburg-Eppendorf, Hamburg, Germany

*To whom correspondence should be addressed; E-mail address: sbonn@uke.de.

Datasets

Table S1: Details on bulk datasets used to evaluate deconvolution methods. For six datasets, the ground truth proportions were available while for others, relationship with the biological phenotypes was considered. Biological hypotheses based on literature serve as proxy ground truths. These are listed in "Biological hypothesis based on literature".

Tissue	**Dataset**	**# samples**	**# Type**	**Flow cytometry**	**Biological hypothesis based on literature**	**Original Source**
PBMC	SDY67	12	RNA-seq	Yes	-	[47]
PBMC	Monaco I	12	RNA-seq	Yes	-	[48]
PBMC	Monaco II	164	Microarray	Yes	-	[48]
PBMC	GSE65133	20	Microarray	Yes	-	[49]
PBMC	GSE107572	9	RNA-seq	Yes	-	[50]
PBMC	GSE120502	250	RNA-seq	Yes	-	[51]
PBMC	Ota	9852	RNA-seq	-	-	[52]
Pancreas	GSE50244	89 (77 with information on hemoglobic 1C levels)	RNA-seq	No	Fraction of beta cells are negatively associated with severity of type 2 diabetes indicated by hemoglobin A1c (hba1C) level [53]-[55].	[56]
Kidney	GSE81492	10	RNA-seq	No	Tubule cells diminish with chronic kidney disease (CKD) [57]-[59].	[60]
Brain	ROSMAP	508 (463 with corresponding annotation of Braak stages)	RNA-seq	No	1. Neurodegeneration with advanced Braak stage [61]-[63], and 2. Between 3:1 and 9:1 ratio of excitatory and inhibitory neurons [64]-[68].	[69]
Brain	PFC proteomics	50	Mass spec.	No	Between 3:1 and 9:1 ratio of excitatory and inhibitory neurons [64]-[68].	[43]
Lymph node	Lymph node	4,035 spots	10x Visium	No	Identification of germinal centers (GC) by co-localization of GC associated cell types	10x Genomics
Brain	Anterior sagittal mouse brain	2,695 spots	10x Visium	No	Identification of excitatory neuronal layers	10x Genomics

Single-cell datasets

Table S2: Single cell datasets used as reference. To deconvolve PBMC datasets in Table S1, single-cell datasets from the corresponding tissues were considered. We used the *PBMC8k* as a reference single-cell dataset from a healthy donor for all methods considered here. To maintain same genes between the single-cell data and bulk RNA-seq, we subset both datasets over common gene-set. For the multi-sample setting to use with MuSiC, we considered *Immune Cell Atlas (ICA)*. The atlas was restricted to blood to match the bulk tissue with donor: 621B (103 cells), 637C (760 cells), A35 (1,368 cells), A36 (3,124 cells), D496 (9,065 cells), D503 (12,208 cells). Cell types that were present in less than 5 samples were dropped. All PBMC datasets were annotated with B cells, CD4 T cells, CD8 T cells, Monocytes and NK cells while the remaining cells were mixed to form an unknown cluster. To deconvolve pancreas, kidney, brain and lymph node samples in Table S1, single-cell dataset from the corresponding tissues were considered.

Tissue	**Dataset**	**# cells**	**Multi-sample**	**Original source**
PBMC	PBMC8k	8,381	no	10x Genomics (8k PBMCs from a Healthy Donor)
PBMC	PBMC6k	5,419	no	10x Genomics (6k PBMCs from a Healthy Donor)
PBMC	DonorA	2,900	no	10x Genomics (Frozen PBMCs Donor A)
PBMC	DonorC	9,519	no	10x Genomics (Frozen PBMCs Donor C)
PBMC	Immune Cell Atlas (ICA)	26,628	yes	[15]
Pancreas	Baron	8,569	yes	[70]
Pancreas	Segerstolpe	3,514	yes	[71]
Pancreas	Xin	1,492	yes	[72]
Kidney	Park	43,745	yes	[73]
Kidney	Miao	16,887	no (only adult)	[74]
Brain	Allen Brain Atlas	49,418	yes	[38]
Lymph node, spleen, tonsil	lymph node reference	73,620	yes	[29]

Table S3: Pearson correlation coefficient (r) between estimates from different methods and flow cytometry for granular cell type fractions in *Monaco I*.

Dataset	**MuSiC**	**CSx**	**Scaden**	**TAPE-O**	**TAPE-A**	**Linear MLPs**	**DISSECT**
B Ex	nan	**0.43**	0.18	0.22	0.030	0.10	0.32
B NSM	nan	-0.08	**0.12**	0.10	-0.15	-0.22	0.09
B Naive	nan	0.95	0.87	0.8	0.43	0.71	**0.96**
B SM	**0.85**	nan	0.57	0.45	0.15	0.26	0.63
Monocytes C	0.30	0.29	**0.63**	0.57	0.52	0.11	0.62
Monocytes I	0.41	0.36	0.90	0.87	0.81	0.54	**0.93**
Monocytes NC	0.25	0.09	0.31	0.35	0.48	0.19	**0.66**
NK	0.80	**0.82**	0.58	0.59	0.65	0.49	**0.82**
Neutrophils LD	0.2	nan	**0.89**	0.48	0.57	0.03	0.56
Plasmablasts	0.62	0.85	0.86	0.65	0.66	0.42	**0.92**
CD4 T Naive	0.66	0.47	0.68	0.70	0.34	0.14	**0.76**
CD4 T Memory	**0.47**	-0.15	0.27	0.27	0.12	0.08	0.24
CD8 T Naive	0.52	**0.7**	0.36	0.38	0.32	0.27	0.49
CD8 T CM	nan	-0.65	0.19	0.13	**0.21**	0.01	0.12
CD8 T EM	nan	nan	0.02	0.47	0.45	0.11	**0.62**
CD8 T TE	0.25	**0.9**	0.28	0.35	0.36	0.35	0.86
mDC	nan	0.46	0.47	0.39	0.40	0.05	**0.68**
pDC	0.55	**0.57**	0.19	0.42	0.31	0.3	0.55
Average	0.49	0.40	0.46	0.45	0.37	0.22	**0.60**

Table S4: *rmse* between estimates from different methods and flow cytometry for granular cell type fractions in *Monaco I*.

Dataset	**MuSiC**	**CSx**	**Scaden**	**TAPE-O**	**TAPE-A**	**Linear MLPs**	**DISSECT**
B Ex	**0.01**	0.05	0.04	0.04	**0.01**	0.02	0.02
B NSM	0.02	**0.01**	0.02	0.03	0.05	0.04	0.02
B Naive	**0.01**	0.03	0.03	0.03	0.04	0.06	0.03
B SM	0.05	**0.01**	**0.01**	0.02	0.02	0.03	**0.01**
Monocytes C	0.05	**0.02**	0.04	**0.02**	**0.02**	0.06	0.06
Monocytes I	0.12	0.15	**0.03**	0.06	0.10	0.12	0.04
Monocytes NC	0.20	0.09	**0.02**	0.07	0.05	0.10	**0.02**
NK	**0.05**	0.08	0.08	0.11	0.11	0.05	0.08
Neutrophils LD	0.02	0.03	**0.01**	**0.01**	**0.01**	0.01	0.02
Plasmablasts	**0.01**	**0.01**	0.02	**0.01**	0.04	0.01	0.04
CD4 T Naive	**0.02**	0.03	0.05	0.05	**0.02**	0.02	0.05
CD4 T Memory	0.10	0.07	**0.03**	0.12	0.15	0.21	**0.03**
CD8 T Naive	0.21	0.07	**0.04**	0.05	0.05	0.01	**0.04**
CD8 T CM	**0.01**	0.08	0.02	0.05	0.12	0.01	0.03
CD8 T EM	0.02	**0.01**	0.02	0.02	**0.01**	0.03	0.02
CD8 T TE	**0.01**	0.07	0.09	0.08	0.11	0.16	0.08
mDC	**0.01**	0.04	0.04	0.08	0.09	0.03	0.03
pDC	0.02	**0.00**	0.02	0.01	0.01	0.01	0.02
Average	0.06	0.05	**0.03**	0.05	0.06	0.05	0.04

Table S5: Average performance over five random experiments for SDY67 (Table S1.) Each column indicates the additional part.

Metric	Linear MLP	Activations	KL Divergence	KL Divergence + Consistency
r	0.51 ± 0.018	0.55 ± 0.016	0.54 ± 0.006	0.63 ± 0.005
rmse	0.13 ± 0.008	0.13 ± 0.006	0.11 ± 0.004	0.09 ± 0.002

DISSECT: deep semi-supervised consistency regularization for accurate cell type fraction and gene expression estimation
– Supplementary figures and Supplementary Note –

Robin Khatri, Pierre Machart, Stefan Bonn*

Institute of Medical Systems Biology, Center for Molecular Neurobiology
Center for Biomedical AI
University Medical Center Hamburg-Eppendorf, Hamburg, Germany

*To whom correspondence should be addressed; E-mail address: sbonn@uke.de.

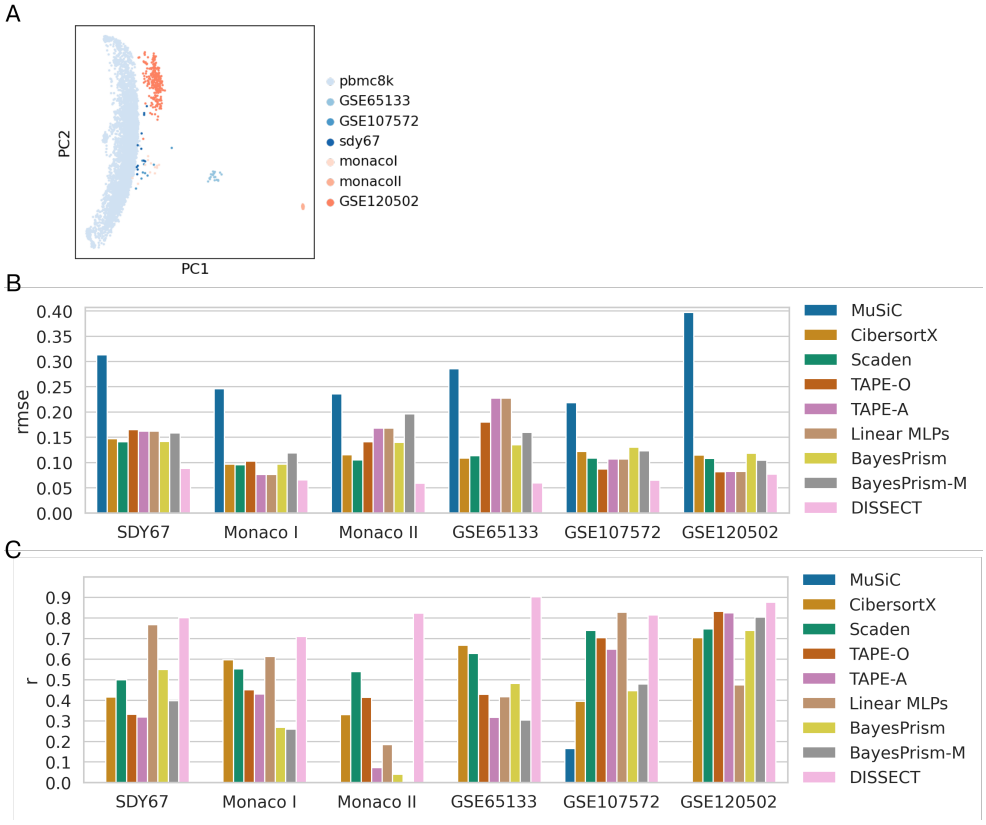


Fig. S1: **A.** PC (Principal component) embeddings of simulated and real PBMC datasets computed using the union of the top 2,000 highly variable genes per dataset. **B.** Overall Pearson's correlation (*r*) and **C.** root-mean-squared-error (*rmse*) for each of the dataset. Datasets are listed on x-axis.

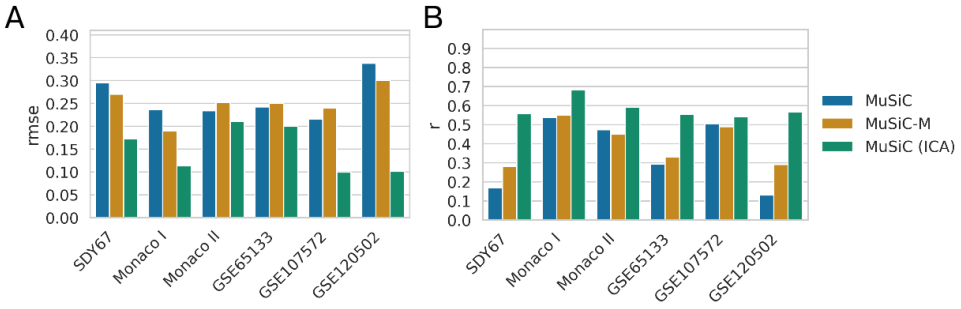


Fig. S2: Comparison of MuSiC (Fig. 3D), MuSiC with marker genes (MuSiC-M) and MuSiC with blood data from Immune Cell Atlas *ICA* (MuSiC-ICA). **A**. root mean-squared-error (*rmse*) and **B**. Pearson's correlation (*r*).

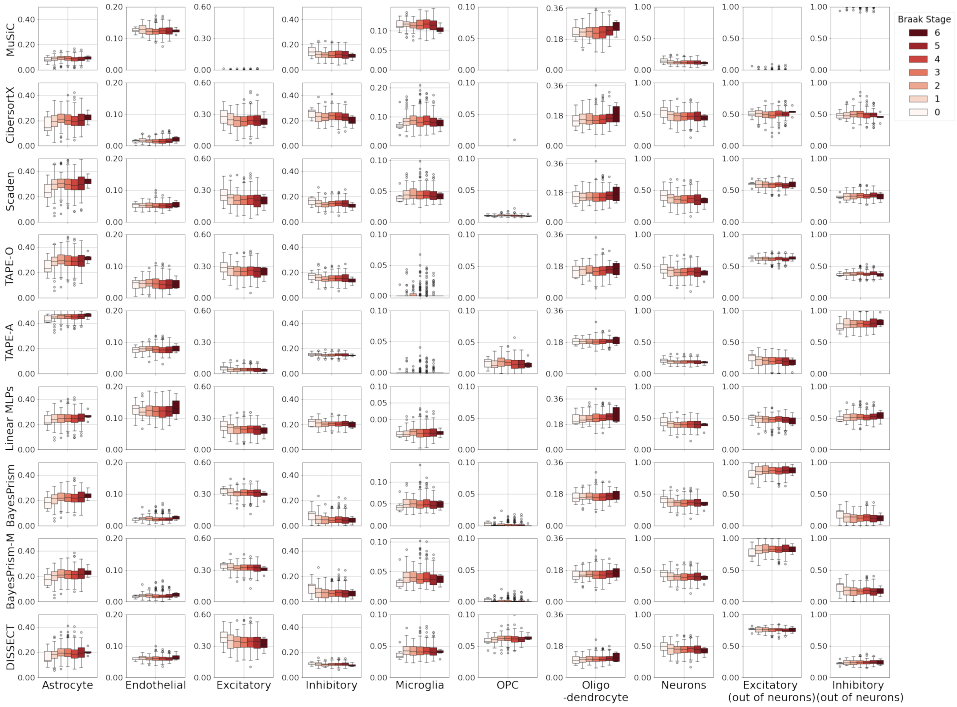


Fig. S3: Estimated cell type proportions on *ROSMAP* separated by Braak Stage. Rows indicate methods, columns indicate cell type.

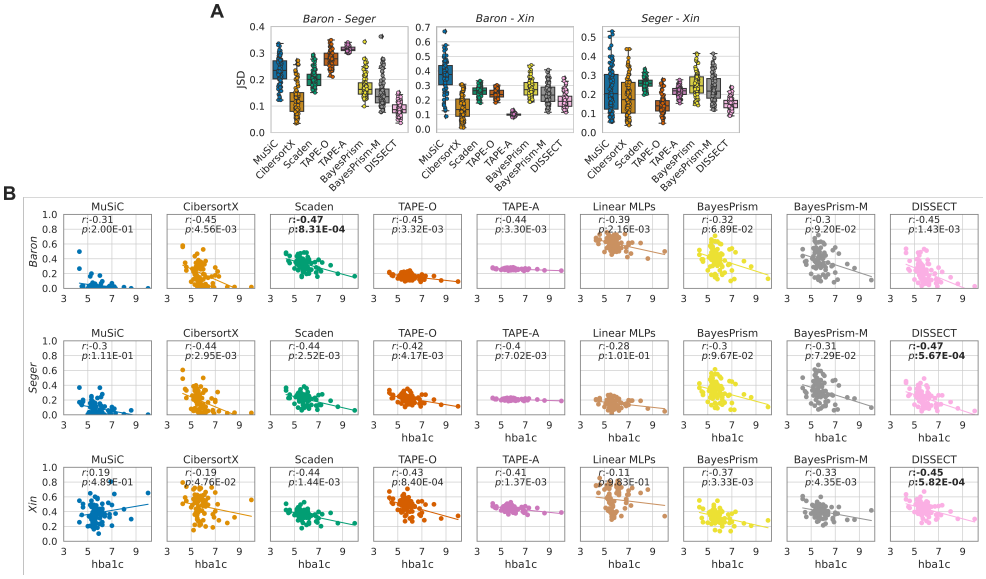


Fig. S4: **A**. Box-plots showing *JSD*s between predicted proportions from *Pancreas* using different single-cell references. Each plot shows *JSD*s between two references. From left to right: *Baron* and *Seger*, *Baron* and *Xin*, and *Seger* and *Xin*. **B**. Associations between predicted beta proportions and hba1c levels assessed through multiple linear regression with hba1c as dependent variable and beta estimates, age, BMI and gender as independent variables. *P*-values correspond to 2-tailed Student's t-test for significance of coefficients for beta estimates. *r* is the Pearson correlation coefficient between beta estimates and hba1c. Each column indicates a method and each row indicates a reference.

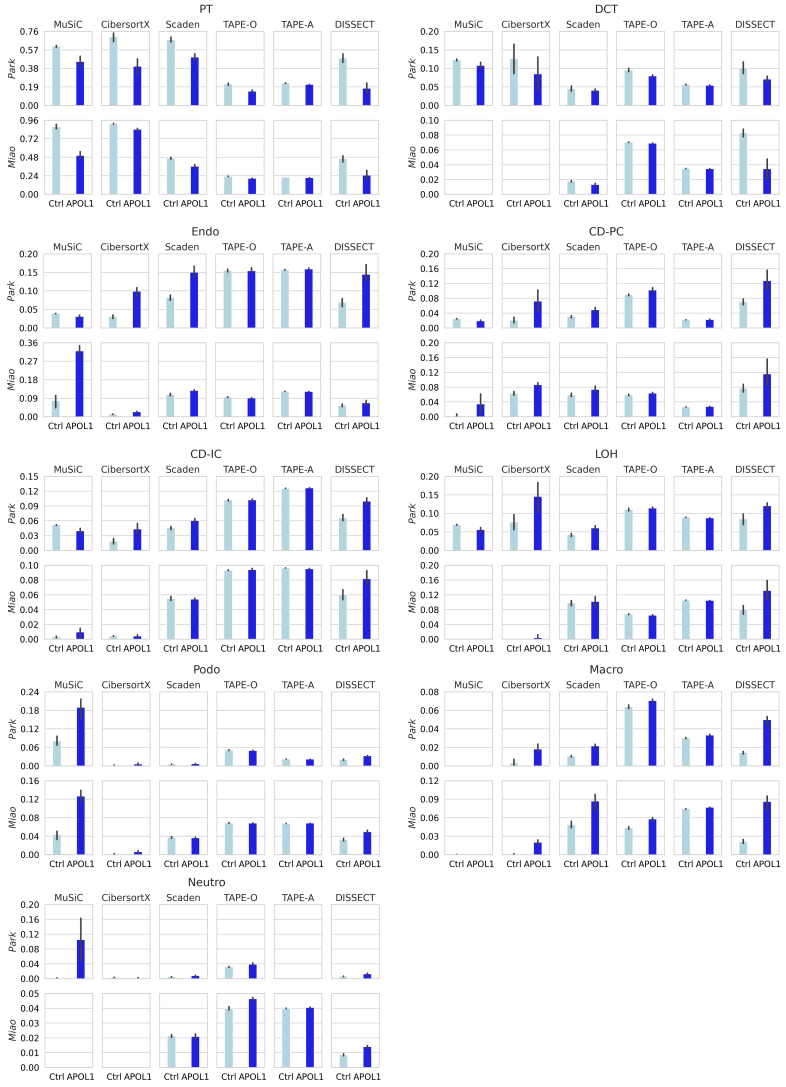


Fig. S5: The figure extends Fig. 3D to include all methods and cell type. From left to right and top to bottom: Proximal tubule (PT), ductal convoluted tubules (DCT), endothelial cells (Endo), collecting duct principal cells (CD-PC), collecting duct intercalated cells (CD-IC), loop of henle (LOH), podocytes (Podo), macrophages (Macro) and neutrophils (Neutro). Each row indicates a reference (*Miao, Park*).

Supplementary Note

To exemplify the applicability of DISSECT on other data types. We aimed to deconvolve spatial transcriptomics datasets. The 10x Genomics Visium™ platform, for instance, delivers spatial gene expression information with a spot diameter of 55 μm. This resolution is not enough to capture single cells and spot-based gene expression on Visium™ is therefore a mixture of its constituent cells. In this section we measure DISSECT's deconvolution performance on Visium™ ST data.

We performed deconvolution on two ST samples obtained from 10x Genomics website corresponding to Lymph node and brain and compared against RCTD and Lymph node. Both datasets are accompanied by H&E images of the underlying tissue.

Anterior sagittal mouse brain - Brain is a highly structured organ with information about the structures of neurons in the cortex. We utilized the reference data from Allen Brain Atlas, as previously used in deconvolution of *ROSMAP* data. Using the ST adapted simulations, we computed proportions of different cell types, including different layers of neurons and visualized them on top of a corresponding hematoxylin and eosin (H&E) stained image (Fig. S6A). DISSECT faithfully captured the spatial layering of the cortical areas of the brain, as well as known 'hot-spots' of neurons, oligodendrocytes, astrocytes, and inhibitory neurons such as somatostatin- and parvalbumin-positive neurons.

To identify cortical layers, we performed louvain clustering (resolution 1) on the estimated cell type fractions per spot, and labelled the clusters enriched for different layers of neurons. Layers L2/3 IT, L4, L5 IT, L5 PT, L6 CT, L6 IT, and L6b were mapped respectively to clusters 1, 7, 12, 8, 3, 14 and 15 (Fig. S6B,C). The identified cortical layers corresponded with the cortical layers annotated in the Allen Reference Atlas (Fig. S6D). We applied RCTD, C2L, SONAR and CARD using the same setting. Compared to RCTD and C2L, DISSECT achieves better separation of excitatory neuronal layers. CARD slightly outperforms DISSECT in this task with a 0.02 increase in schilloute scores. The quantification was made using silhouette score with euclidean metric. (Fig. S6E).

Lymph node - Next, we evaluated DISSECT on the spatial deconvolution of lymph node tissue. Lymph nodes consist of various immune cell subsets and localized germinal centers (GCs). We used a lymph node single-cell reference and used the manually annotated germinal centers that were provided with the study as ground truth [29]. To verify whether DISSECT estimated and localized cell fractions per spot correctly, we visualized GC-related cell types, namely Cycling B cells, Germinal center B cells and follicular dendritic cells (FDCs) (Fig. S7A). DISSECT also identified T cells associated regions in and around these GCs (Fig. S7A). To obtain binary GC predictions to compare with the ground truth GC annotations, we computed louvain clusters (resolution 1) on the cell type proportions, and labelled spots with cluster 1 as GCs based on enrichment of GC associated cell types (Fig. S7B). Since the number of GC spots (378) is considerably lower than non-GC spots (3,657), balanced accuracy as implemented in Scikit-learn was used to account for this imbalance [75]-[76]. Comparison with the ground truth revealed a balanced accuracy of 0.94, indicating that DISSECT deconvolved GCs with high accuracy and

on par performance with C2L and RCTD (Fig. S7C). For CARD and SONAR, the balanced accuracy were 0.93 and 0.91 respectively.

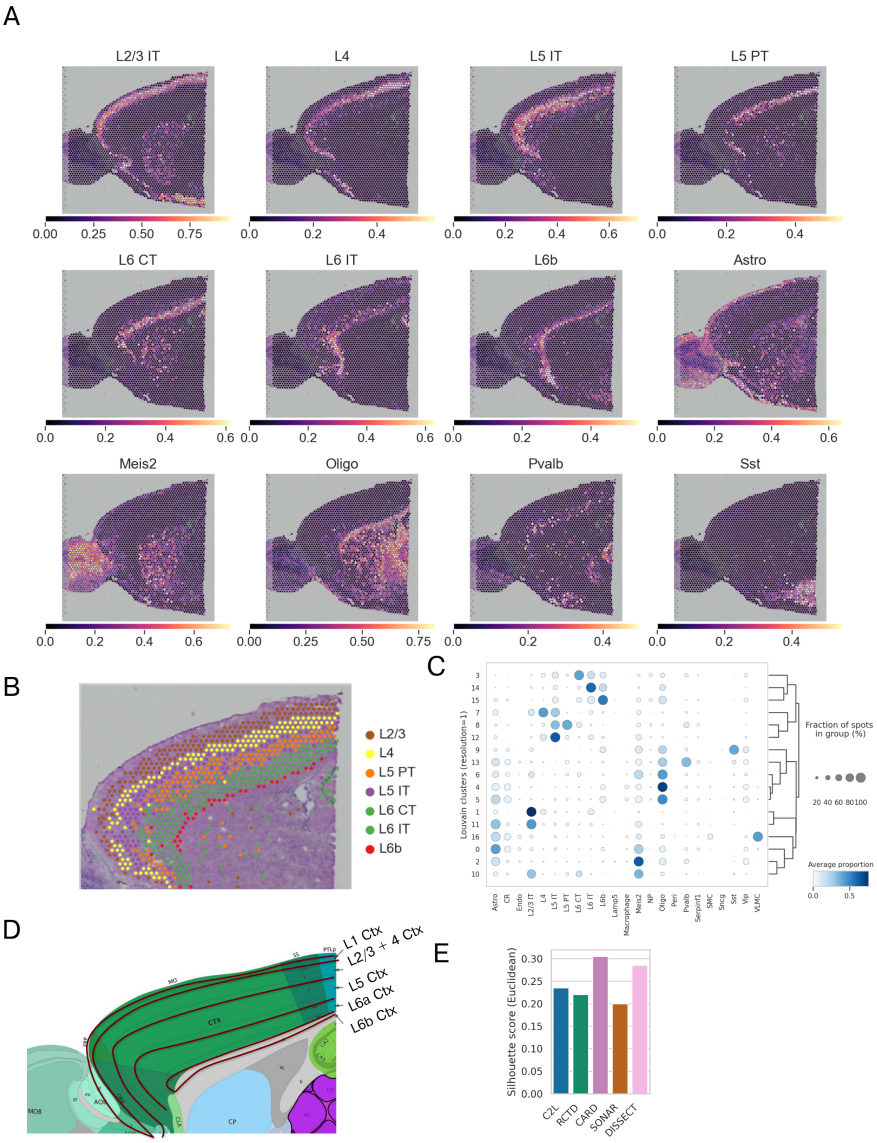


Fig. S6. **A**. Estimated cell type proportions from mouse brain tissue visualized over H&E image of the corresponding tissue. L2/3 - L6b indicate different layers of neurons. *Astro:* Astrocytes, *Oligo:* Oligodendrocytes. **B**. Estimated cortical layers using enrichment of cell type proportions in louvain clusters presented in **C**. **D**. Annotations of cortical layers from Allen Brain Reference obtained from [38]. For visibility, cortex boundaries were highlighted. **E**. Shilloute scores for C2L, RCTD, SONAR, CARD and DISSECT.

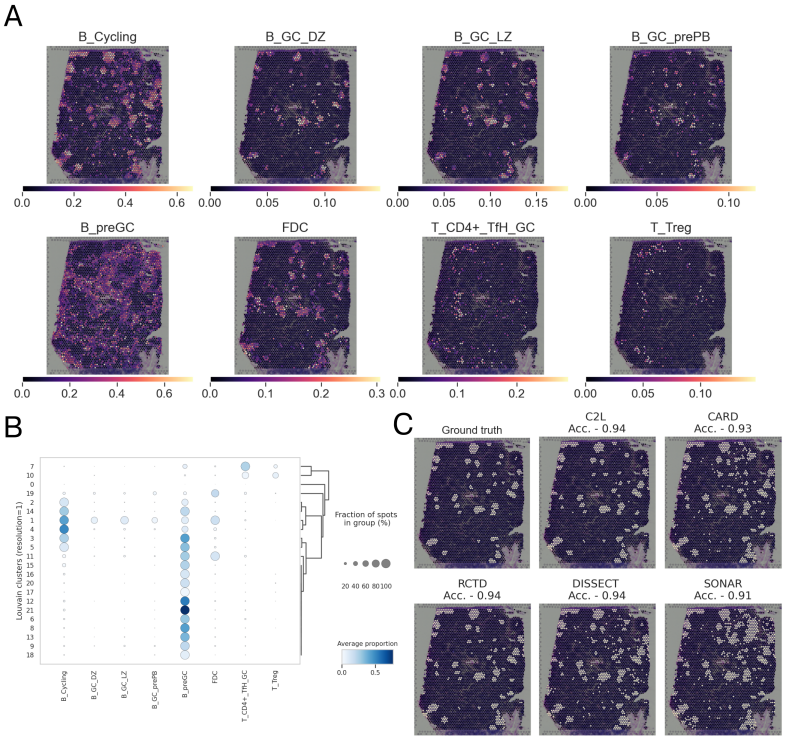


Fig. S7: **A**. Estimated cell type proportions visualized over H&E image of the corresponding lymph node tissue. *B_cycling*: Cycling B cells, *B_DC_DZ*: Dark zone germinal center B cells, *B_GC_LZ*: Light zone germinal center B cells, *B_GC_PrePB*: germinal center Pre-plasmablast/plasma cells, *B_preGC*: pre-germinal center B cells, *FDC*: Follicular dendritic cells, *T_CD4*$^{+}$*TfH_GC*: Germinal center follicular helper CD4^{+} T cells, *T_Treg*: Regulatory T cells. **B**. Louvain clustering on cell type proportions. y-axis lists cluster numbers and x-axis lists cell types. **C** Comparison of identified clusters, from left to right and top to bottom: Ground truth GC-spots, predicted annotations for C2L, RCTD, DISSECT, CARD and SONAR.

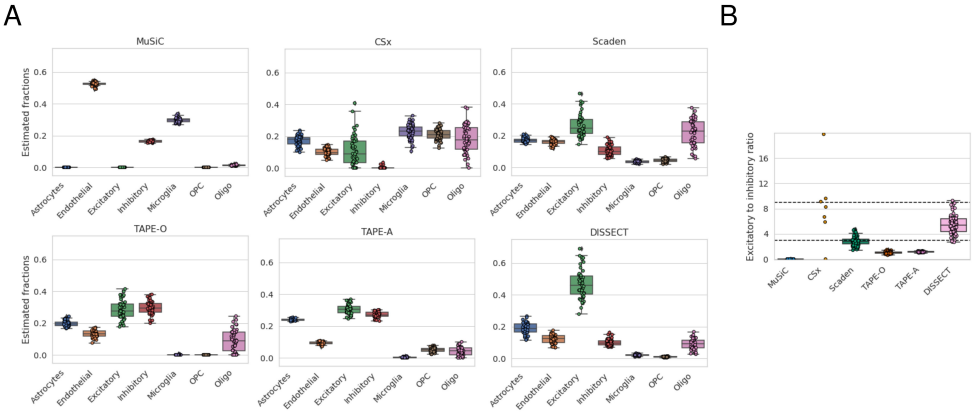


Fig. S8: **A**. Boxplots showing estimated cell type proportions from human proteomics samples. Title of each plot is indicated at the top of the plot and x-axis list cell types. **B**. Boxplots showing predicted excitatory to inhibitory neuron ratios for each method for the proteome samples. Expected ratios lie between 3:1 and 9:1 as indicated by dashed lines. To make the plot discernible possible, in **B**, the y-axis was limited to a value of a maximum value of 17.

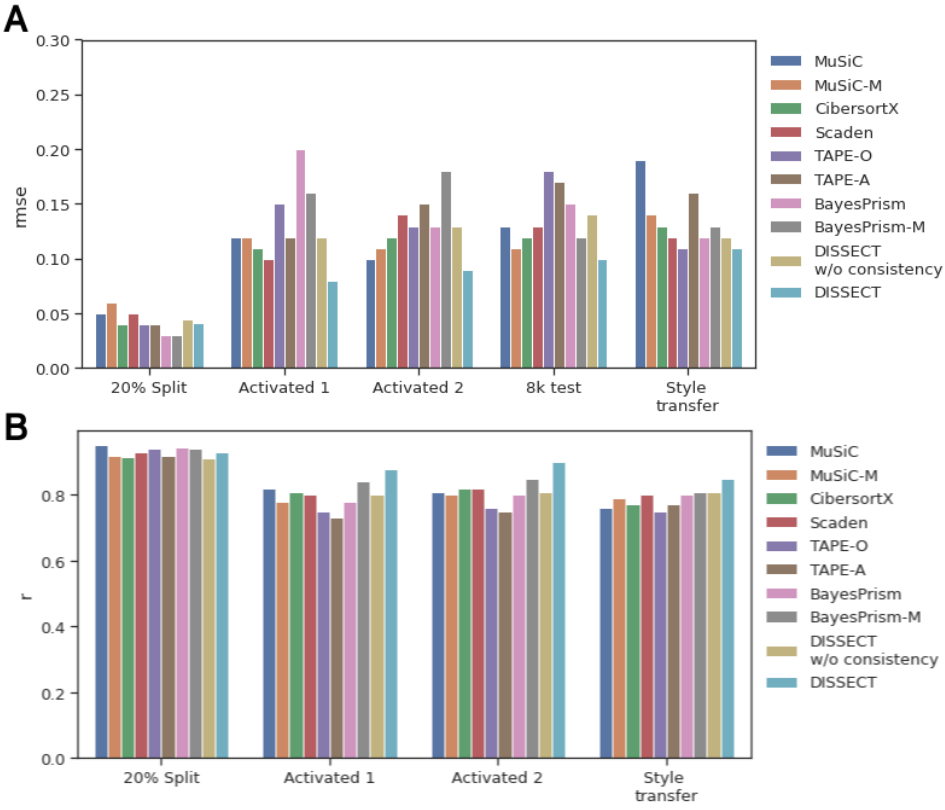


Fig. S9: Performance of deconvolution algorithms in estimating cell types fractions under domain shift.

3

Treatment with a monoclonal antibody, ustekinumab, to target pathogenic T cells in ANCA-GN improves patient outcome

nature communications

Article

https://doi.org/10.1038/s41467-024-52525-w

Immune profiling-based targeting of pathogenic T cells with ustekinumab in ANCA-associated glomerulonephritis

Received: 14 May 2024

Accepted: 11 September 2024

Published online: 19 September 2024

Check for updates

Jonas Engesser[1,2,7], Robin Khatri[2,3,7], Darius P. Schaub[2,3,7], Yu Zhao[1,2,3], Hans-Joachim Paust[1,2], Zeba Sultana[1,2,3], Nariaki Asada[1,2], Jan-Hendrik Riedel[1,2], Varshi Sivayoganathan[1,2], Anett Peters[1], Anna Kaffke[1], Saskia-Larissa Jauch-Speer[1], Thiago Goldbeck-Strieder[1], Victor G. Puelles[1,4], Ulrich O. Wenzel[1], Oliver M. Steinmetz[1], Elion Hoxha[1,4], Jan-Eric Turner[1,2], Hans-Willi Mittrücker[2,5], Thorsten Wiech[4,6], Tobias B. Huber[1,2,4], Stefan Bonn[2,3,4,8] ✉, Christian F. Krebs[1,2,4,8] ✉ & Ulf Panzer[1,2,4,8] ✉

Antineutrophil cytoplasmic antibody (ANCA)–associated vasculitis is a life-threatening autoimmune disease that often results in kidney failure caused by crescentic glomerulonephritis (GN). To date, treatment of most patients with ANCA-GN relies on non-specific immunosuppressive agents, which may have serious adverse effects and be only partially effective. Here, using spatial and single-cell transcriptome analysis, we characterize inflammatory niches in kidney samples from 34 patients with ANCA-GN and identify proinflammatory, cytokine-producing $CD4^+$ and $CD8^+$ T cells as a pathogenic signature. We then utilize these transcriptomic profiles for digital pharmacology and identify ustekinumab, a monoclonal antibody targeting IL-12 and IL-23, as the strongest therapeutic drug to use. Moreover, four patients with relapsing ANCA-GN are treated with ustekinumab in combination with low-dose cyclophosphamide and steroids, with ustekinumab given subcutaneously (90 mg) at weeks 0, 4, 12, and 24. Patients are followed up for 26 weeks to find this treatment well-tolerated and inducing clinical responses, including improved kidney function and Birmingham Vasculitis Activity Score, in all ANCA-GN patients. Our findings thus suggest that targeting of pathogenic T cells in ANCA-GN patients with ustekinumab might represent a potential approach and warrants further investigation in clinical trials.

Antineutrophil cytoplasmic antibody (ANCA)-associated vasculitis is a group of systemic autoimmune diseases characterized by inflamed and necrotic small to medium-sized blood vessels[1]. Kidney involvement is common and is associated with a substantial risk of end-stage renal disease and death. Renal involvement typically manifests as rapidly progressive crescentic glomerulonephritis (ANCA-GN) with a fast decline in kidney function[2–4]. Despite recent advances in the treatment and management of ANCA-GN, such as

[1]Department of Medicine III, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [2]Hamburg Center for Translational Immunology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [3]Institute of Medical Systems Biology, Center for Biomedical AI, Center for Molecular Neurobiology Hamburg, Hamburg, Germany. [4]Hamburg Center for Kidney Health (HCKH), University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [5]Institute for Immunology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [6]Institute of Pathology, Division of Nephropathology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [7]These authors contributed equally: Jonas Engesser, Robin Khatri, Darius P. Schaub. [8]These authors jointly supervised this work: Stefan Bonn, Christian F. Krebs, Ulf Panzer. ✉e-mail: s.bonn@uke.de; c.krebs@uke.de; panzer@uke.de

B cell depletion using rituximab[5–7] and complement C5a receptor blockade with avacopan[8], the rate of end-stage kidney disease and side effects remains high, emphasizing the unmet need for more effective and immunopathogenesis-based treatment strategies in ANCA-GN.

Several studies investigated the gene expression profiles of blood samples from patients with ANCA-associated vasculitis (AAV) showing distinct endotypes and potential prognostic biomarkers[9–11]. Moreover, recent flow cytometric, single-cell RNA sequencing (scRNA-seq), and immunohistochemical analyses of kidney biopsy samples have provided deeper insights into the pathological mechanisms mediated by immune cells in ANCA-GN[12–15]. However, the relevant specific spatial localization of immune cells and their cellular interactions are largely unknown. Decoding the localization and function of immune cells in the kidney is highly relevant because the local immune responses could be the drivers of renal injury and disease progression, offering unique opportunities for the identification and characterization of treatment targets for ANCA-GN.

ANCA-GN patients remain at significant risk of renal failure and increased mortality, highlighting the need to develop more effective and safer therapies. Here, we combine spatial transcriptomics and single-cell RNAseq and identify Th1 and Th17 cells as major contributors to immune-mediated renal injury in ANCA-GN. Based on these results we treat four ANCA-GN patients with ustekinumab, which specifically targets Th1 and Th17 cells, as add-on therapy. The rapid clinical response in all four patients suggests that ustekinumab could be a promising therapy and should be further investigated in clinical trials. Our approach to combining high-dimensional single-cell and spatial immune profiling with clinical and histopathological data facilitates personalized pathogenesis-based treatments and could be a promising strategy for other autoimmune diseases.

Results

Study cohort and experimental overview

We included two independent patient groups with biopsy-confirmed ANCA-GN from the Hamburg GN Registry[14,16] in our study (Fig. 1). The exploratory group consists of 34 ANCA-GN patients. From each of these patients, two renal biopsy cores were taken. One was used for routine pathological evaluation and the other sample was used for spatial ($n = 28$) and single-cell ($n = 27$) transcriptomic analysis (Fig. 1 and Table 1). The treatment group consists of four-patients with relapsing ANCA-GN that were treated with ustekinumab, steroids, and low-dose cyclophosphamide and underwent single cell, and flow cytometry immune profiling, as well as pathological examination and clinical follow-up analysis for 26 weeks (Fig. 1).

Spatial transcriptomics reveals inflammatory glomerular and tubulointerstitial niches linked with T cell activation in ANCA-GN

Kidney inflammation is a hallmark of ANCA-GN but the underlying immunopathology is not well understood. To characterize the inflammatory niches, pathological cell-cell interactions, and key molecular pathways that drive kidney inflammation in ANCA-GN, we generated spatial transcriptome (ST) sequencing data from 28 renal biopsies of the exploratory cohort using the Visium platform (Fig. 2a and Supplementary Data 1). By unsupervised clustering of the spatial data, we were able to define 12 tissue compartments. Based on marker gene expression, we identified normal glomeruli, inflamed glomeruli, tubulointerstitium, inflamed tubulointerstitium, vasculature, and several tubular compartments. The latter includes proximal tubules (PT), connecting tubules (CNT), distal convoluted tubules (DCT), collecting duct (CD), and loop of Henle (LOH) (Fig. 2b,c and Supplementary Fig. 1a–c)[17]. We verified our clustering-based annotations by comparison to expert annotations of glomerular compartments on H&E-stained images, exhibiting an annotation concordance of over 90% on normal glomerular regions (Supplementary Fig. 1d). To understand the compositional difference of the identified compartments between ANCA-GN and healthy controls, we included 8 healthy control samples in our analysis (Supplementary Fig. 1e). ANCA-GN samples were enriched for inflamed glomeruli and inflamed tubulointerstitium as compared to healthy control samples (Fig. 2c and Supplementary Data 2).

Next, we aimed to identify key molecular pathways and cell subtypes involved in immunopathology in the inflammatory niches of the kidney. An unsupervised analysis of cluster-defining genes from the two inflamed compartments identified T cell activation as the most differentially expressed gene ontology term (Fig. 2d and Supplementary Data 3–5). By co-analyzing the neighborhood composition of compartments and their enrichment for T cell-specific pathways, we found that gene sets for Th1 and Th17 cell differentiation as well as T cell-mediated cytotoxicity were enriched in inflamed compartments (Fig. 2e, f and Supplementary Data 4, 5). Further gene sets upregulated in inflamed glomerular compartments indicate increased interleukin-1, extracellular matrix organization, and regulation of fibroblast proliferation (Supplementary Data 5).

Enrichment of proinflammatory cytokine-producing Th1/Tc1 and Th17/Tc17-like effector T cells in the kidney of ANCA-GN patients

To further clarify and define the role of specific T cell subtypes and their signaling cascades in ANCA-GN, we generated a single cell transcriptome and epitope atlas of T cells, encompassing 72,416T cells from renal biopsy and blood samples of 27 ANCA-GN patients of the exploratory cohort (Fig. 3a and Supplementary Fig. 2a and Supplementary Data 6). Unsupervised clustering identified 15 T cell clusters,

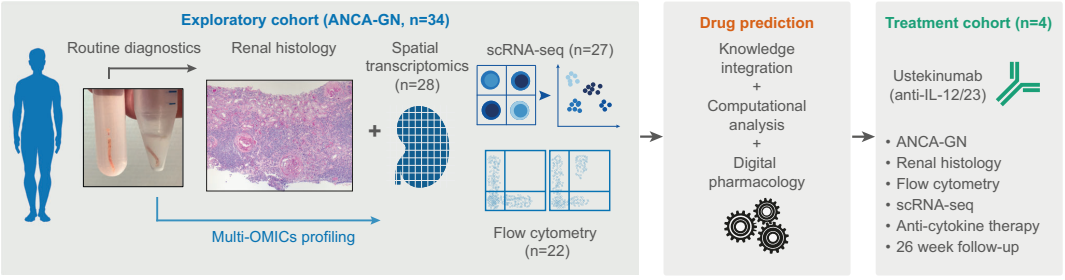


Fig. 1 | Study overview. 34 patients from the Hamburg GN Registry underwent diagnostic kidney biopsy and multi-OMIC high dimensional immune profiling. Based on these results drug prediction revealed ustekinumab as the strongest candidate for treatment of ANCA-GN. Subsequently, 4 patients with severely relapsing ANCA-GN were treated with ustekinumab and followed up for 26 weeks.

Table 1 | Basic and Clinical Characteristics Exploratory cohort

	N = 34
Demographics	
Age—years, median (IQR)	64.5 (57.75–74.25)
Sex, *n* (%)	
Female	14 (41.18)
Male	20 (58.82)
BMI, median (IQR)[a]	24.65 (22.90–27.36)
ANCA status, *n* (%)	
MPO	22 (64.71)
PR3	12 (35.29)
Initial organ involvement, *n* (%)	
General	22 (64.71)
Renal	34 (100)
ENT	6 (17.65)
Lung (DAH)	11 (32.35)
Nervous system	3 (8.82)
Cutaneous	2 (5.88)
Abdominal	2 (5.88)
Eye	2 (5.88)
Heart	1 (2.94)
Histological ANCA renal risk score[16], *n* (%)	
Low	12 (35.29)
Medium	15 (44.12)
High	7 (20.59)
Laboratory values, median (IQR)	
Creatinine (mg/dl)	2.12 (1.66–4.55)
eGFR (ml/min)	23.5 (11.75–42.75)
ACR (mg/g)	720 (328.5–1500)
Immunosuppressant induction treatment, *n* (%)	
Glucocorticoids	34 (100)
Rituximab	9 (26.47)
Cyclophosphamide	22 (64.71)
Cyclophosphamide and Rituximab	3 (8.82)
PLEX	4 (11.76)

Source data are provided as a Source Data file.
IQR interquartile range, *ANCA* antineutrophile cytoplasmatic antibody, *MPO* myeloperoxidase, *PR3* proteinase 3, *ENT* ear nose throat, *DAH* diffuse alveolar hemorrhage, *PLEX* therapeutic plasma exchange, *eGFR* estimated glomerular filtration rate, *ACR* albumin-creatinine-ratio.
[a]*n* = 33.

containing $CD4^+$ T effector cells ($CD4^+$ Teff), $CD8^+$ T effector cells ($CD8^+$ Teff), $CD4^+$ naïve T cells, $CD8^+$ naïve T cells, $CD8^+$ T effector memory cells (Teff/em), $CD4^+$ central memory T cells (Tcm), stressed T cells, regulatory T cells (Treg), γδ T cells, mucosal-associated invariant T cells (MAIT), natural killer T cells (NKT), $CD4^+$ cytotoxic T cells (CTL), natural killer cells (NK cells) and proliferating T cells (Fig. 3a and Supplementary Fig. 2b). Cytokine expression analysis revealed the highest cytokine scores in $CD4^+$ and $CD8^+$ T effector cells (clusters 1 and 2) (Fig. 3b and Supplementary Fig. 3a). Interestingly, these effector $CD4^+$ and $CD8^+$ T cells were enriched in the inflamed kidney but not in the peripheral blood, highlighting their relevance in renal inflammation (Fig. 3c). Further analyses showed that the $CD4^+$ T effector cell cluster had a high proportion of Th1, Th1-like, and Th17 cells and $CD8^+$ T effector cell cluster of Tc1, and Tc17-like cells (Fig. 3d and Supplementary Fig. 3b, c). Subgroup analysis of proteinase 3 (PR3) ANCA versus myeloperoxidase (MPO) ANCA patients showed no differences in composition of T effector cells (Supplementary Fig. 3d).

To understand the spatial location of these pathogenic $CD4^+$ and $CD8^+$ effector T cells in the inflamed kidney, we next used single cell information to deconvolve the spatial transcriptomic data. Consistent with the up-regulation of T cell activation markers, $CD4^+$ Th1 and Th17 as well as $CD8^+$ Tc1 cells were exclusively localized to inflammatory glomerular and tubulointerstitial niches (Fig. 3e and Supplementary Fig. 3e).

Digital pharmacology identifies ustekinumab as drug candidate

Based on the combined analysis of spatial and single cell transcriptome, type 1 and 3 cytokine producing T cells constitute a potential immunopathogenesis-based therapeutic target in ANCA-GN. We employed digital pharmacology, the mapping of drugs to cells based on their molecular interaction, to search for approved drugs that specifically target these pathogenic T cells in the kidney. To narrow our search space to immunomodulating drugs, we preselected 277 drugs consisting of antineoplastic agents, endocrine therapy drugs, and immunosuppressants (Anatomical Therapeutic Chemical (ATC) codes L01, L02, L04, respectively) that could potentially interact with $CD4^+$ and $CD8^+$ Teff subsets in the inflamed glomerular and inflamed tubulointerstitial compartments. To prioritize these drugs, we constructed a dictionary of drug-gene interactions based on the spatial and single cell transcriptome information and subsequently filtered drugs for chemical viability and FDA-approval (Fig. 3f and Supplementary Data 7). Among the drugs with high differential interaction scores in the inflamed renal compartments, we identified ustekinumab as the drug exhibiting the highest specificity for $CD4^+$ and $CD8^+$ effector T cells. Ustekinumab is a human monoclonal antibody directed against the p40 subunit of both IL-12 and IL-23, which has the potential to inhibit Th1/Tc1 and Th17/Tc17 cell responses.

Rapid biopsy immune profiling

ScRNA-sequencing is a time consuming and expensive technology. To enable rapid and cost-effective screening of pathogenic immune cell infiltrates in a clinical setting, we performed flow cytometry-based single cell immune biopsy profiling of the exploratory cohort (Fig. 4a, b and Supplementary Fig. 4). This approach delivers patient-specific immune profiles within hours after biopsy and might be instrumental in establishing immunopathogenesis-driven targeted biological therapies. Based on our results of the single cell and spatial transcriptomic data, we focused on the identification of pathogenic Th1/Tc1 and Th17/Tc17 cells.

Rapid single cell immune biopsy profiling showed that Th1/Tc1($CXCR3^+$, $CCR6^-$) and Th17/ Tc17-like cells ($CCR6^+$, $CCR4^+$)[18] were the dominant T cell subsets in the inflamed kidney of ANCA-GN patients (Fig. 4b). Subsequent multiplex immunofluorescence staining showed that these pathogenic T cells were mainly localized to glomerular and interstitial inflammatory areas (Fig. 4c), further supporting an anti-T cell-cytokine treatment with ustekinumab.

Demographic, clinical, and immune characteristics of the ustekinumab treatment group

Based on our findings from the exploratory ANCA-GN cohort and results from preclinical GN models, we decided to use ustekinumab in combination with low-dose cyclophosphamide in patients with AAV that had relapsing disease and a relative contraindication or incomplete response to current standard therapies (ustekinumab treatment cohort). The basic and clinical characteristics of the treatment group, encompassing four ANCA-GN patients are briefly summarized below and are shown in more detail in Table 2 and Supplementary Fig. 5.

Patient 1: A 73-year-old male, with known MPO-ANCA positive vasculitis under remission maintenance therapy with rituximab, presented with fatigue, dizziness, lower limb edema, gross hematuria and acute kidney injury to our nephrology clinic. Kidney biopsy was performed and revealed active, crescentic ANCA-GN. No other organ manifestation was noted. At time of relapse, the patient received rituximab remission maintenance therapy. Despite appropriate rituximab dosing and intervals, full B cell depletion was not achieved. With known urinary flow obstruction and the need for urinary diversion, the

https://doi.org/10.1038/s41467-024-52525-w

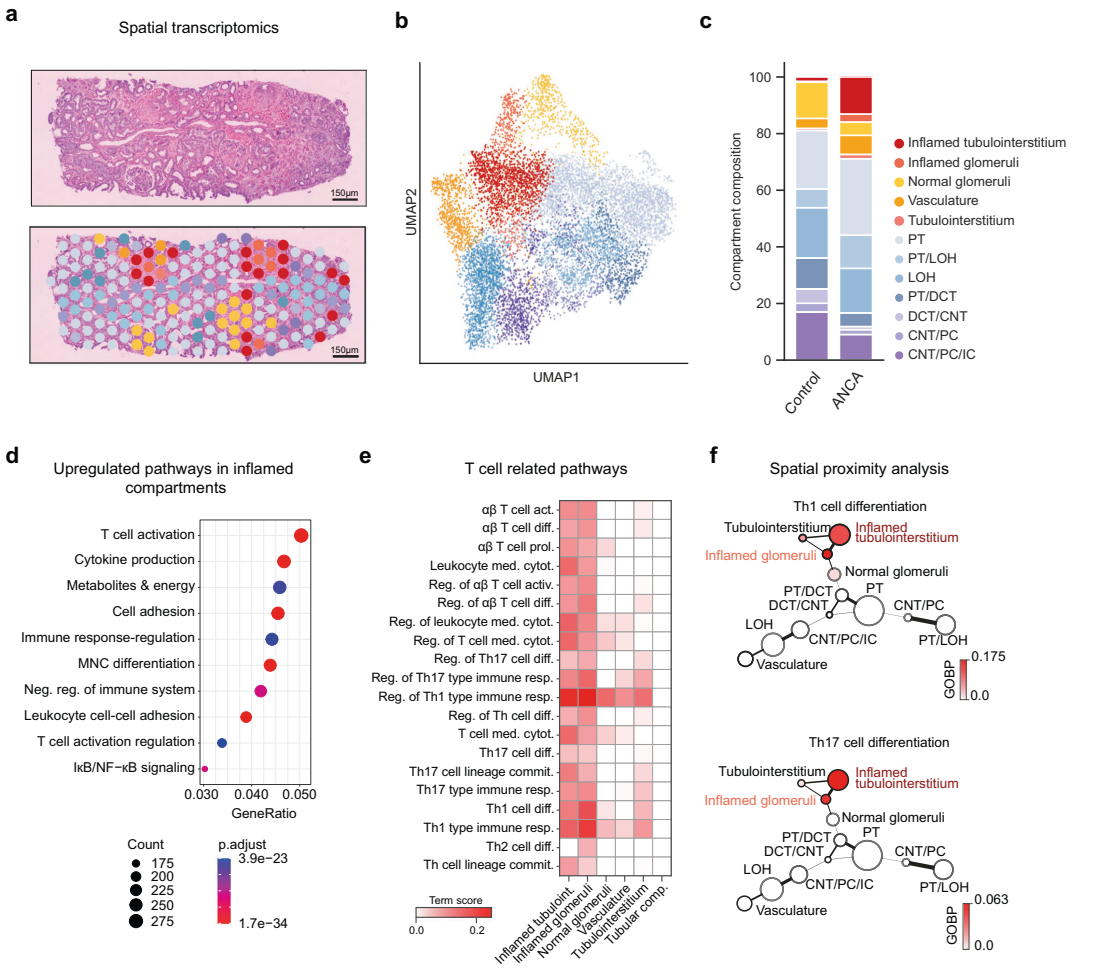


Fig. 2 | Spatial transcriptome analysis of the ANCA-GN exploratory group. **a** Left, Representative section of an H&E-stained kidney biopsy. Right, Spatial distribution of renal compartments overlaid on the representative section. **b** UMAP embedding displaying annotated renal compartments across 10,763 spots from all ST slides. See Supplementary Fig. 1c for the expression of cell type markers in annotated compartments. **c** Barplots showing the composition of renal compartments in the control (21,420 spots) and ANCA-GN exploratory group. The significance of the difference in composition was assessed with differential population analysis and is presented in Supplementary Data 2. **d** Top 10 enriched gene ontology terms in inflamed compartments. Count: number of DE genes in the term. Gene ratios: ratio of the number of DE genes in the term to the number of all DE genes. The colors show adjusted *p*-values (*p*.adjust) computed using the *enrichGO* function from R package clusterProfiler with right-tailed Fisher's exact test and Benjamini–Hochberg multiple test correction. **e** Scores of alpha-beta T cell-related gene ontology terms computed using GSVA in different renal compartments. **f** Graph showing the spatial proximity of renal compartments and the enrichment of T cell activation. The node sizes and edge widths are proportional to compartment size and spatial proximity, respectively. The nodes are colored by an increasing Th1 and Th17 cell differentiation term scores computed using GSVA. PT, proximal tubules. LOH, loop of Henle. DCT distal convoluted tubules, CNT connecting tubules, PC principal cells, IC intercalated cells. Source data are provided as a Source Data file.

patient was hesitant for full dose cyclophosphamide therapy. (Supplementary Fig. 5a).

Patient 2: A 52-year-old male patient was admitted to our clinic with a creatinine increase as well as active urinary sediment after six pulses of i.v. cyclophosphamide, because of recently diagnosed MPO-ANCA positive vasculitis with extensive organ manifestations. Kidney biopsy was performed and revealed active, crescentic ANCA-GN. The early phase of the COVID-19 pandemic raised substantial concerns over B cell depleting therapies, thus prompted us to decide against re-induction therapy containing rituximab (Supplementary Fig. 5b).

Patient 3: A 32-year-old male was admitted to our nephrology ward with fever, night sweats, weight loss, progressive dyspnea and hemoptysis. Four weeks earlier the patient received rituximab and steroids because of a pulmonary and ENT relapse of known PR3-ANCA positive vasculitis. Chest imaging and bronchoscopy showed progressive DAH (diffuse alveolar hemorrhage). Urinary analysis displayed glomerular hematuria and laboratory testing showed acute kidney injury with pronounced elevation of CRP and ANCA-level. Kidney biopsy was initiated and revealed active, crescentic glomerulonephritis. Because of concomitant severe

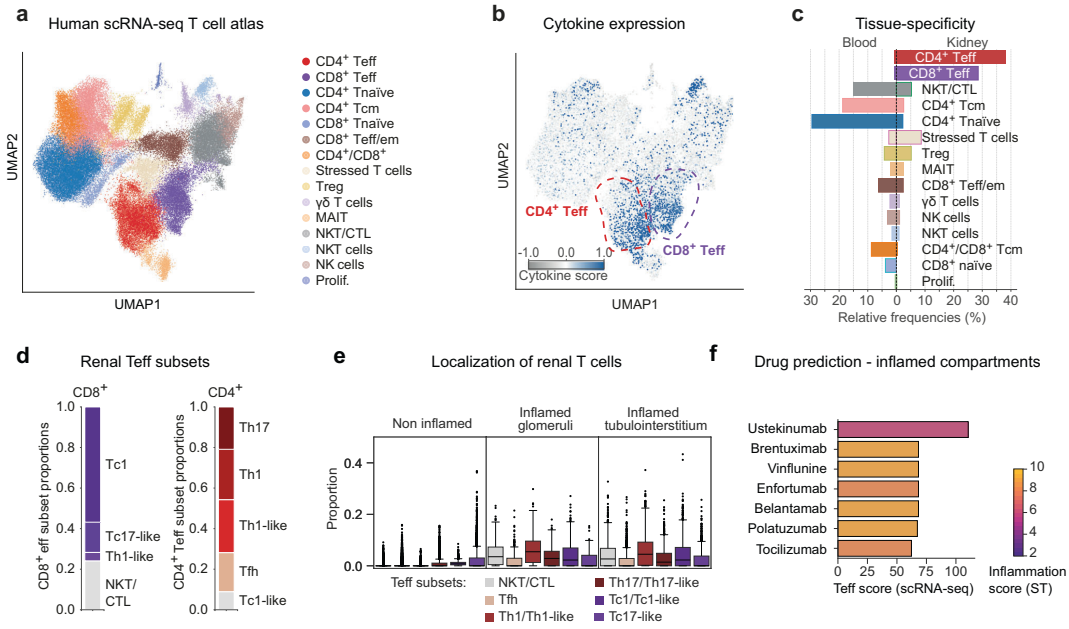


Fig. 3 | Single T cell transcriptome analysis and drug prediction. a UMAP projection and cluster annotations of the combined human single cell atlas of renal and blood T cells. In total 72,416 cells are shown of which 22,187 stem from the kidney and 50,229 from the blood. See Supplementary Fig. 2b for the expression of cell type-specific marker genes. **b** Combined type 1-3 cytokine expression score (type 1: *IFNG, TNF, IL2, IL18, LTA, CSF2*; type 2: *IL4, IL5, IL9, IL13*; type 3: *IL17A, IL17F, IL22, IL26*) per cell overlayed on the UMAP projection. The positions of $CD4^+$ and $CD8^+$ T effector cell clusters are highlighted manually. The detailed expression per cytokine type and cell type is shown in Supplementary Fig. 3a. **c** Relative tissue composition per cell type. The frequencies are computed separately for blood and kidney cells, i.e., both sides add up to 100%. **d** $CD4^+$ and $CD8^+$ T effector cell subsets and their relative proportions. Both T effector cell subsets show some overlap due to the proximity of their respective expression profiles, e.g., the $CD4^+$ Teff cluster contains a small proportion of $CD8^+$ Tc1-like cells. The marker gene expression is detailed in Supplementary Fig. 3b, c. **e** Distribution of Teff cell subsets within the non-inflamed and inflamed compartments. For detailed proportions in individual non-inflamed compartments, see Supplementary Fig. 3e. Boxplots show the median (middle horizontal line), interquartile range (box), Tukey-style whiskers (lines beyond the box), outliers (data points beyond 1.5*interquartile or below−1.5*interquartile) for proportion of Teff subsets in 10,763 spots from all ANCA ST slides. **f** Ranking of drugs based on their interaction scores within Teff single cells, with colors representing their interaction scores specifically within inflamed renal compartments. Source data are provided as a Source Data file.

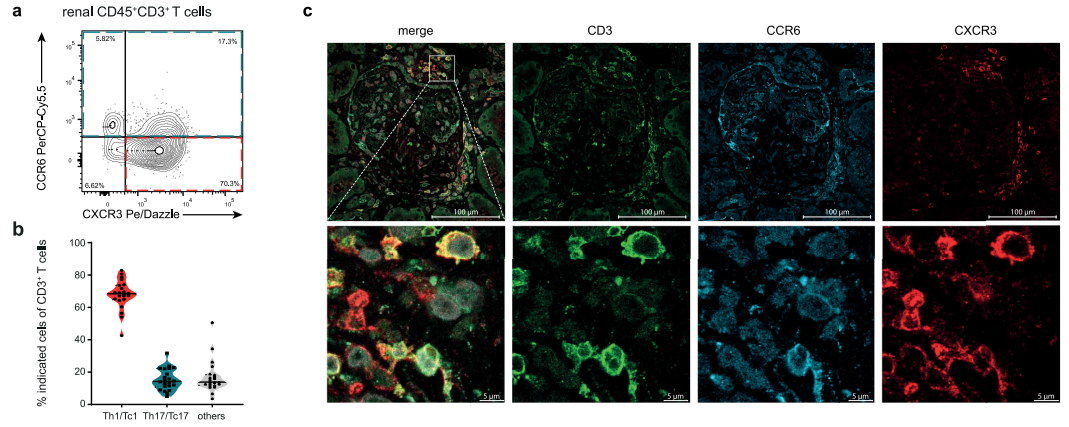


Fig. 4 | Immune profiling of renal T cells. a Representative flow cytometry plot showing the identification of chemokine receptor expression from cells isolated from biopsy samples of patients with ANCA-GN (exploratory cohort, $n = 22$). **b** Quantification of chemokine receptor expression CXCR3 (Th1/Tc1) and CCR6 (Th17/Tc17) from renal $CD3^+$ T cells. Violin plots show mean, symbols represent individual data points. ($n = 22$). **c** Representative immunofluorescence staining of chemokine receptors CXCR3 and CCR6 on $CD3^+$ T cells in human kidney tissue of ANCA-GN. Lower row zoomed in areas. Source data are provided as a Source Data file.

77
https://doi.org/10.1038/s41467-024-52525-w

Table 2 | Basic and Clinical Characteristics treatment cohort

	Patient 1	Patient 2	Patient 3	Patient 4
Demographics				
Age—years	73	52	32	69
Sex (female/male)	M	M	M	F
BMI	34.68	25.90	24.38	27.06
Time until relapse (weeks)	143	16	168	138
ANCA status				
MPO	+	+	–	+
PR3	–	–	+	–
Organ involvement				
General	+	+	+	+
Renal	+	+	+	+
ENT	–	+	+	–
Lung (DAH)	–	+	+	–
Nervous system	–	+	–	–
Cutaneous	+	+	–	–
Abdominal	–	–	–	–
Eye	–	–	–	–
Heart	–	+	–	–
Histological ANCA renal risk score[16]				
Low	–	+	+	–
Medium	–	–	–	–
High	+	–	–	+
Laboratory values at relapse				
Creatinine (mg/dl)	4.25	2.27	1.88	3.42
eGFR (ml/min)	13	32	46	13
ACR (mg/g)	3345.7	590.0	582.3	1134.9
Previous immunosuppressant treatment				
Glucocorticoids	+	+	+	+
Cyclophosphamide	+	+	–	+
PLEX	+	–	–	–
Rituximab	+	–	+	+
Azathioprine	+	–	–	+

ANCA anti-neutrophile cytoplasmatic antibody, *MPO* myeloperoxidase, *PR3* proteinase 3, *ENT* ear nose throat, *DAH* diffuse alveolar hemorrhage, *eGFR* estimated glomerular filtration rate, *ACR* albumin-creatinine ratio, *PLEX* therapeutic plasma exchange.

leukopenia, cyclophosphamide could not be given at full dose (Supplementary Fig. 5c).

Patient 4: A 72-year-old female patient, with known MPO-ANCA positive vasculitis and remission maintenance therapy with rituximab, was sent to our nephrology ward with acute kidney injury and reduced general condition. Chest imaging ruled out relevant thoracic pathologies. Urinary analysis showed glomerular hematuria and kidney biopsy was issued. Here, active crescentic ANCA-GN was seen and diagnosis of relapsing ANCA-GN was made. Because the patient suffered a relapse while being on remission maintenance with rituximab, she was deemed a poor responder to rituximab. Furthermore, she suffered from myelodysplastic syndrome with bicytopenia (leukopenia and anemia), thus full dose cyclophosphamide was deemed unsuitable, because of increased risk for myelotoxicity. (Supplementary Fig. 5d).

Flow cytometry-based rapid immune biopsy profiling in each of the four ANCA-GN patients demonstrated a strong infiltration of Th1/Tc1 and Th17/Tc17-like cells into the inflamed kidney (Supplementary Fig. 6a, b). Additional single cell transcriptome sequencing of the four patients provided a more comprehensive renal T cell profile and confirmed the observed Th1/Tc1 and Th17/Tc17-cell responses (Supplementary Figs. 7a–e, 8a-d and Supplementary Data 8).

Clinical response of the ustekinumab ANCA-GN treatment group

The four ANCA-GN patients of the treatment cohort were given ustekinumab s.c. (90 mg) in combination with low dose cyclophosphamide and steroids, following the RITUXVAS trial approach[19], as a re-induction therapy. All patients received ustekinumab at weeks 0, 4, 12, and 24 in combination with two to three low doses of cyclophosphamide (cumulative dose 1.5–2.0 g) and glucocorticoids according to the PEXIVAS trial reduced dose regimen[20]. Starting at week 16, patients 3 and 4 (patient 2 at week 22) received a low dose remission maintenance therapy with either azathioprine or mycophenolate mofetil (MMF) (Fig. 5a). At six months, the prednisolone dose was tapered to 5 mg daily in all four patients.

All patients showed a rapid clinical and serological response to this re-induction treatment protocol. Mean serum creatinine levels decreased from a median of 2.8 (2.0–4.0) mg/dl to 1.6 (1.3–1.7) mg/dl at 6 months. According to the albumin-creatinine-ratio, median albuminuria decreased from 862.5 (584–2793) mg/g at the time of relapse to 604 (359–1402) mg/g at 6 months (Fig. 5a). ANCA serum levels decreased from 65 (24–126) U/ml to 32 (9–57) U/ml. The Birmingham vasculitis activity score (BVAS) declined from a median of 12.5 (9.75–13) at the beginning of ustekinumab treatment to 2.5 (2–4.5) at 6 months (Fig. 5b). C-reactive protein (CRP) levels rapidly improved throughout the 6-month treatment period (ranging from 5 to 190 mg/l at the beginning of treatment to < 4–36 mg/l at 6 months) (Fig. 5b). The treatment with ustekinumab was well tolerated. No serious adverse effects were observed during the 6-month treatment period.

Discussion

Despite numerous advances in therapy for ANCA-GN, these patients still have a substantial risk of kidney failure and increased mortality[21,22]. Today, the leading causes of death in ANCA-GN are infections, followed by cardiovascular disease and malignancies[23], all associated with immunosuppressive therapy. This highlights the unmet need to balance disease control against the risk of side effects. However, the limited understanding of the underlying immunopathology, particularly within the inflamed kidney, impedes the implementation of tailored and effective treatment options. Therefore, in this study we sought to tackle this issue with an approach consisting of four distinct steps.

First, using unsupervised analyses of spatial transcriptomics data derived from the kidneys of patients with ANCA-GN, we identified a distinct enrichment of inflammatory glomerular and tubulointerstitial niches, linked to T cell activation, as the key molecular pathways. Evidence for a pathogenic role of T cells in ANCA-GN patients, is derived from genetic studies showing a significant association with distinct human leukocyte antigen (HLA) class II haplotypes[24–26], an unbalanced activation state of blood and kidney T cells[27–29] and a therapeutic T cell-depletion study in refractory AAV patients[30]. The T cell subsets and cytokine networks that promote tissue injury and loss of renal function, however, remain to be fully elucidated. Thus, secondly, we performed unsupervised single cell transcriptomics and epitope mapping of renal T cells from our exploratory cohort, revealing the dominance of proinflammatory cytokine-producing Th1/Tc1 and Th17/Tc17-like effector T cells in the kidneys of ANCA-GN patients. Thirdly, based on the dominant effector T cell subsets in inflammatory glomerular and tubulointerstitial niches of nephritic kidneys we used digital pharmacology to investigate relevant drugs targeting pathways expressed in these T cells, and could thereby identify ustekinumab as a potential treatment approach. Ustekinumab, a human monoclonal antibody directed against the p40 subunit of both IL-12 and IL-23 thereby targeting the Th1/Tc1 and Th17/Tc17 immune responses, is approved for the treatment of psoriasis[31], psoriatic arthritis[32], and inflammatory bowel disease[33,34]. Several studies highlighted its efficacy and good tolerability in these patients[35–40]. To date, there are no data available for using ustekinumab in ANCA-GN, despite the fact that experimental GN models, including

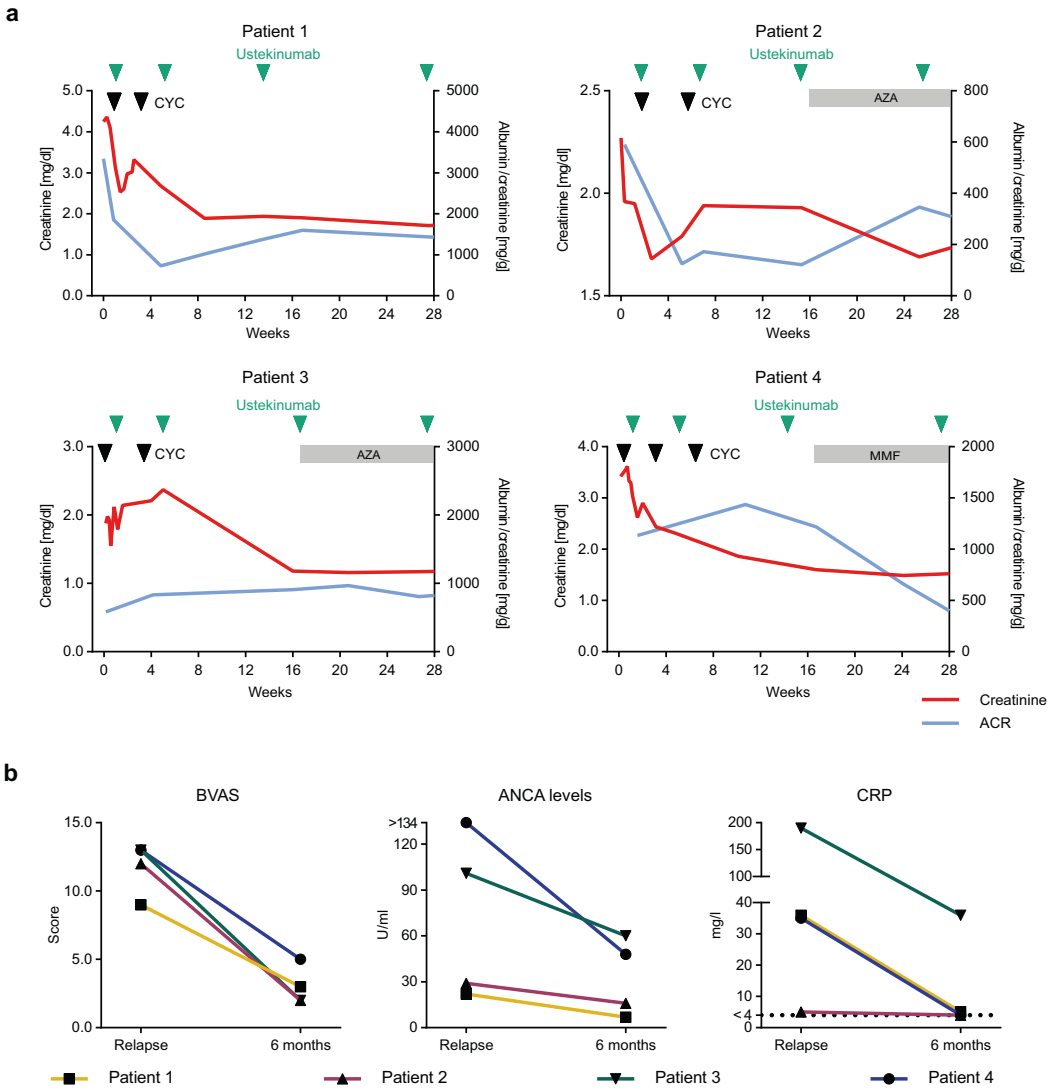


Fig. 5 | Clinical outcome of the ustekinumab treatment cohort. a Course of serum creatinine and albuminuria during ustekinumab treatment. Black arrowheads indicate cyclophosphamide and green arrowheads ustekinumab administration. Gray bands indicate low dose remission maintenance therapy with either AZA or MMF. **b** BVAS, ANCA levels measured via ELISA and CRP levels at baseline and 6 months after initiation ustekinumab treatment ($n = 4$). (CYC cyclophosphamide; AZA azathioprine; MMF mycophenolate mofetil; BVAS Birmingham Vasculitis Activity Score; CRP C-reactive protein; ACR albumin creatinine ratio). Source data are provided as a Source Data file.

preclinical ANCA-GN models[41–45], provide a clear rationale for targeting the IL-12/IL-23 axis in immune mediated kidney disease.

Fourthly and finally, given our findings in the exploratory cohort, we assessed the efficacy of ustekinumab in four ANCA-GN patients with relapsing disease. Our treatment protocol was designed as an add-on therapy of ustekinumab at weeks 0, 4, 12, and 24 with up to three low-dose pulses of cyclophosphamide, similar to recent trials establishing rituximab and the complement inhibitor avacopan as add-on treatments for ANCA vasculitis (RITUXVAS and ADVOCATE trials)[8,19]. The immunosuppression was complemented by an intravenous prednisolone and oral glucocorticoid therapy according to the PEXIVAS study reduced dose regimen[20]. All four patients responded rapidly to therapy, with improvements in kidney function, ANCA levels, CRP, and BVAS. Importantly, all patients tolerated the treatment well and no adverse ustekinumab-related effects were observed.

This report pioneers an immunopathology-based anti-T cell-cytokine therapy for immune-mediated kidney diseases. Our data further suggests that combining high-dimensional single cell immune profiling with clinical and histopathological information facilitates personalized pathogenesis-based treatments. Moreover, this study indicates that rapid single cell immune biopsy profiling by flow cytometry is a feasible approach that might be routinely applied in patients

with ANCA-GN and could prove to be a potential strategy for other organ-specific autoimmune and inflammatory diseases.

Although our results provide an immunopathogenesis-based rationale for targeting Th1/Tc1 and Th17/Tc17 responses with ustekinumab in ANCA-GN, our study has several limitations. The treatment protocol was designed as an add-on therapy, making it more difficult to assess the intrinsic efficacy of ustekinumab, and it is likely that part of the observed response to treatment is due to the concomitant use of low-dose cyclophosphamide and steroids. In addition, our study focused on the renal manifestation of AAV, and it is unclear whether these results also apply to the involvement of other organs.

In addition, long-term data need to be acquired to confirm the overall safety profile of ustekinumab in ANCA-GN. Furthermore, the data generated from this case series is based on a low number of patients without a control group. Therefore, these results should be interpreted with caution and need to be confirmed in adequately designed clinical trials for which the appropriate treatment protocol and patient subgroups remain to be determined. Taken together, our study suggests that ustekinumab is a well-tolerated therapeutic option for the treatment of ANCA-GN, which should be further investigated in clinical trials.

Methods

Patients

We included two independent ANCA-GN patient groups from the Hamburg GN Registry[14,16] in our study. The exploratory cohorts consist of 34 patients and the ustekinumab treatment group of four patients. For the spatial transcriptomic analysis of control samples, the healthy parts of the kidney, which was removed due to tumor nephrectomy, were used. Informed consent was obtained from all participating patients in accordance to the CARE guidelines and in accordance with the ethical principles stated in the Declaration of Helsinki. All four patients in the ustekinumab treatment group also provided written informed consent before receiving ustekinumab as an off-label treatment. Detailed information on the patient cohorts and the performed analysis are provided in Tables 1, 2 and Supplementary Data 10.

Sex- and gender-based analyses were not performed. Information about the sex of the patients is provided in Table 1 for the exploratory cohort and in Table 2 for the treatment cohort. Biological sex and self-reported sex were identical in both the exploratory and treatment cohort.

These studies were approved by the Institutional Reviewing Board (IRB) of the University Medical Center Hamburg-Eppendorf and Ethik-Kommission der Ärztekammer Hamburg (local ethics committee of the chamber of physicians in Hamburg), and covered by the licenses PV4806, PV5026, and PV5822.

Spatial transcriptomics

Preprocessing of the spatial transcriptomics slides. For spatial transcriptomics, formalin-fixed paraffin-embedded (FFPE) tissue sections from patients with ANCA-associated glomerulonephritis and controls (healthy tissue from tumornephrectomies) were transferred on Visium (10x Genomics) slides (spatial for FFPE gene expression human transcriptome) and processed according to the manufacturer's instructions. Next-generation sequencing was performed on an Illumina NovaSeq 6000 aiming at 25,000 reads per spot (PE150).

For alignment to the genome of ST slides ($n = 20$) from 30 patients, the human genome assembly GRCh38-2020-A was used. Mapping to the genome was performed using 10x Genomics Space Ranger (v2.0.1). Alignment metrics from spaceranger are provided in Supplementary Data 1. The same alignment method and libraries as used for the exploratory group were used to align ST slides of the internal controls ($n = 3$).

Quality control

After alignment of the ST slides to the genome, 1 slide was excluded from analysis due to low gene counts (280 median genes per spot compared to 3499.58 ± 972.97, Supplementary Data 1). Data analysis of the ST gene expression data was performed using Scanpy[46] (v1.9.3) in Python (v3.9.7). The following parameters in Scanpy's preprocessing pipeline were used to filter poor-quality spots: *min_genes* = 100, *min_spots* = 3, *min_counts* = 2000, *max_counts* = 35000. The filtered ST data consisted of 10,763 spots and 17,847 genes. The filtered spot counts were normalized to sum to 10,000, and data was $\log_2$-transformed with a pseudo-count of 1.

Clustering and annotation

Principal components (*n_comps* = 50) were computed on the highly variable genes (*highly_variable_genes* in Scanpy with default settings and slide-name as *batch_key*). The batch effect corresponding to the slide was removed using harmony[47] (v0.1.0) in R (v4.1.1). To identify clusters, Leiden clustering (*scanpy.tl.leiden*) was performed on Uniform Manifold Approximation and Projection (UMAP) data projections with a resolution of 1.2. The UMAP projections were generated on a neighborhood graph constructed using *scanpy.pp.neighbors* with *n_neighbors* = 10. Cluster annotations were performed using the following cell type specific markers from a reference kidney single cell dataset[17] - Proximal tubules (PT): *LRP2, CUBN, SLC13A1*, Distal convoluted tubules (DCT): *SLC12A3, CNNM2, FGF13, KLHL3, LHX1, TRPM6*, Connecting tubules (CNT): *SLC8A1, SCN2A, HSD11B2, CALB1*, Principal cells (PC): *GATA3, AQP2, AQP3*, Intercalated cells (IC): *ATP6VOD2, ATP6V1C2, TMEM213, CLNK*, Ascending thin loop of Henle (Thin limb): *CRYAB, TACSTD2, SLC44A5, KLRG2, COL26A1, BOC*, Thick ascending loop of Henle (TAL): *CASR, SLC12A1, UMOD*. Endothelial cells (Endo): *CD34, PECAM1, PTPRB, MEIS2, EMCN*, vascular smooth muscle cells (vSMC)/Pericyte: *NOTCH3, PDGFRB, ITGA8*, Fibroblasts: *COL1A1, COL1A2, C7, NEGR1, FBLN5, DCN, CDH11*, Podocytes: *PTPRQ, WT1, NTNG1, NPHS1, NPHS2, CLIC5, PODXL*, Immune cells: *PTPRC, CD3D, CD14, CD19*. The expression of these marker genes for each annotated renal compartment is shown in the Supplementary Fig. 1c. The distribution of total gene counts, number of spots across slides, and annotated compartments are presented in the Supplementary Fig. 1a.

Quantification of spatial proximity

The spatial neighborhood enrichment was performed with Squidpy[48] (v1.2.2) over all slides. The underlying multi-sample spatial graph was constructed by merging all sample-specific spatial graphs into a single graph, resulting in one connected component per sample. The sample-specific graphs were constructed by connecting each spot to its nearest neighbors. To visualize the neighborhood enrichment matrix, the compartments were considered as nodes with the number of spots in a compartment as node sizes, z-scores as edge weights, and *neato* as layout engine from the library Pygraphviz (v1.11). Negative z-scores were set to 0, essentially removing spatially distant compartments. For visualization, the resulting weights were downscaled by 0.25.

Integration of control samples

ST data from kidney nephrectomies ($n = 8$) was integrated with the previously generated embedding of ST data from the ANCA-GN exploratory group. In total, we used 3 slides generated at the UKE Hamburg (Supplementary Data 9) and 5 slides previously generated by Lake et al. [17]., totaling 21,420 spots. The integration was performed with Symphony[49] (v0.1.0) with highly variable genes computed over the ANCA-GN exploratory group.

Differential population analysis

To identify the renal compartments differentially abundant between the control and ANCA-GN samples, differential population analysis was applied using scCODA[50] (v0.1.9) with CNT/PC as the reference cell type

and a false discovery rate of 0.05. We identified inflamed glomerular, inflamed tubulointerstitial, PT, PT/LOH, and Tubulointerstitial/Vessels to be differentially abundant between control and ANCA-GN (Supplementary Data 2).

Gene set enrichment analysis

First, we identified differentially expressed (DE) genes in inflamed glomerular and inflamed interstitial compartments using a Wilcoxon test (adjusted p-value cutoff of 0.05 and $\log_2$-fold change cutoff of 0.25) through *scanpy.tl.rank_genes_groups*. We then performed gene set enrichment analysis on a functional level with the differentially expressed (DE) genes as input, using the *enrichGO* function from *clusterProfiler* R package[51] (*v4.2.2*) and *biological processes* as gene ontology (GO). The function *simplify* was used to remove redundant GO terms. Gene-set variation analysis (GSVA)[52] was used to compute the scores of gene set ontology terms.

Annotation of H&E slides

The same biopsy samples as used in 10x Visium were manually annotated by an expert into three categories: normal, crescentic, and uncertain. The third group contained the tissue that could not be confidently assigned to either normal or crescentic categories. The original images were exported to TIF and processed using ImageJ (v1.54f). The manual annotations were performed using Napari (v0.5.0a2.dev171+gf2d7d437).

Single cell RNA-sequencing: preprocessing and quality control

The Cell Ranger software (v5.0.1 and v7.1.0, 10x Genomics) was used to demultiplex cellular barcodes and map reads to the reference genome GRCh38-3.0.0 and GRCh38-2020-A. All quality control and preprocessing steps were performed in Seurat[53] (v4.0.4) and R (v4.1.1). The Seurat demultiplexing function *HTODemux* was used to demultiplex the hashtag samples. We removed the cells in which less than 500 or more than 5000 expressed genes were detected. We further filtered out low-quality cells with more than 10% mitochondrial genes. Subsequently, raw counts were normalized to 10,000 and log1p transformed, batch corrected and integrated with harmony[47] using the 2000 most highly variable genes, and clustered using the Louvain algorithm with resolution 0.1. T cells were isolated by removal of all cell clusters with low CD3 expression. We merged the tissue-specific datasets for each cohort by keeping the union of all genes for blood and kidney samples. Subsequently, we removed all cells belonging to the top 0.1% total counts quantile or expressing less than 200 genes as well as any genes that were expressed in less than 10 cells. We further removed cells with more than 12,000 detected surface proteins and proteins that were present in less than 10 cells. In the case of the transcriptome information we normalized the raw counts to sum up to 10,000 and log1p transformed them. For the raw protein counts we performed centered log-ratio normalization. The filtered, processed, and combined single-cell data for the exploratory cohort contains 72,416 cells (22,187 kidneys, 50,229 blood) and 21,419 genes (Supplementary Fig. 2a and Supplementary Data 6). For the treatment group, the combined single-cell data contains 34,810 cells (15,372 kidney, 19,438 blood) and 38,224 genes (Supplementary Fig. 7e and Supplementary Data 8).

Clustering and cell type identification

In the following, we provide full information only about the analysis workflows for the exploratory group but mention differences to the analogous workflow for the treatment group. All analyses were performed either in R (v4.1.1) using Seurat (v4.0.4) or in Python (v3.9.17) using Scanpy[46] (v1.9.1).

For the exploratory group, we combined two integration workflows to identify and annotate cell types: one to identify broad T cell clusters and another to annotate specific $CD4^+$ and $CD8^+$ T cell subsets. We first performed a principal component analysis on the top 2000 highly variable genes and then applied harmony[47] on the first 30 principal components to correct batch effects between patients. After computing the nearest neighbor graph, we clustered the data using the Louvain algorithm with a resolution of 0.1. We annotated the cell clusters using canonical cell type markers for broad T cell subsets (Fig. 3a and Supplementary Fig. 2b). Type 1-3 cytokine scores (type 1: *IFNG, TNF, IL2, IL18, LTA, CSF2*; type 2: *IL4, IL5, IL9, IL13*; type 3: *IL17A, IL17F, IL22, IL26*) were calculated using the Scanpy function *score_genes* (Fig. 3b and Supplementary Fig. 3a).

In a second step, we isolated the $CD4^+$ Teff and $CD8^+$ Teff cells, mostly composed of kidney cells, and removed patients containing less than 2 total cells (blood and kidney) and genes present in less than 10 cells. We reintegrated the remaining cells with totalVI[54] based on the top 4000 highly variable genes and all surface proteins, treating the patient ID as a categorical covariate. After Leiden clustering with a resolution of 0.8 and 1.0, respectively, we annotated the cell clusters based on canonical markers for $CD4^+$ and $CD8^+$ T cell subsets (Fig. 3d and Supplementary Figs. 2b and 3b,c).

For the ustekinumab group, we followed an analogous integration, clustering, and annotation workflow (Supplementary Figs. 7a–e and 8a–d) with the following differences. First, we regressed out the counts based on the number of genes, the total number of counts, and the fraction of mitochondrial genes per cell before integration of the full dataset and the T effector subsets. Secondly, we performed the reintegration of both T effector subsets with harmony instead of totalVI.

Cell type deconvolution of spatial transcriptomics

In combination with the ANCA-GN renal T cell single cell atlas described in this study, the single cell atlas from Stewart et al. [55]. was integrated to estimate cell type proportions, resulting in total 24 cell types: LOH, B cell, CD, $CD4^+$ T naive, $CD4^+$ Tcm, $CD8^+$ T naive, CNT, DC, Endo, Fib, Macrophage, Mast cell, Monocyte, Myofib, Neutrophil, NKT, Podo, PT, Tfh, Treg, Th1 and Th17, Tc1 and Tc17. Cell-type deconvolution was performed using a reference-based algorithm, DISSECT[56], with default training parameters. To generate simulated data from training using the PropsSimulator module of DISSECT, the following changes were made: *n_samples* = 8000 and *downsample* = 0.1.

Drug prediction

Drugs from ATC/DDD classification L (Antineoplastic and immunomodulating drugs), excluding immunostimulants from ATC/DDD classification L03, were used as potential candidate drugs. The targets of these drugs were extracted from ChEMBL[57] and putative drug interactions from the drug-target interaction database DGIdb[58], resulting in a total of 277 drugs. To assess the targets of the drugs, we used drug2cell (v0.1.0)[59] with inflamed glomerular and interstitium renal compartments as clusters of interest. Default parameters were used in drug2cell in addition to a cutoff of 0.25 for $\log_2$-fold change in the drug-target expression. This resulted in 14 drugs (Supplementary Data 7) whose targets were differentially enriched in the two compartments. To prioritize drugs, we looked for drugs that affect primarily the inflamed compartments and are less enriched in others. Only drugs that were enriched at least 75% in the compartments of interest and less than 75% in all other compartments were selected. This criterion resulted in 7 potential drugs targeting the inflamed compartments: belantamab mafodotin, brentuximab vedotin, vinflunine, ustekinumab, enfortumab vedotin, tocilizumab, and polatuzumab vedotin. To select the final target out of these drugs and to integrate T cell specificities of the drugs, we computed scores of their targets in the $CD4^+$ Teff and $CD8^+$ Teff clusters using *scanpy.tl.score_genes* function and ordered them in decreasing order.

Isolation and flow cytometry of human biopsy leukocytes

Single-cell suspensions were obtained from human kidney biopsies by enzymatic digestion in RPMI 1640 medium with collagenase D at 0.4 mg/ml (Roche, 11088858001) and deoxyribonuclease I (DNase I;

10 µg/ml; Sigma-Aldrich, 10104159001) at 37 °C for 30 min followed by dissociation with gentle MACS (Miltenyi Biotec). Leukocytes from blood samples were separated using Leucosep tubes (Greiner Bio-One, 10349081). Cells were stained with fluorochrome-conjugated antibodies from BioLegend and BD Biosciences, CD45 BV510 (BioLegend, clone HI30, catalog number 304036, dilution 1:100), CD3 BV785 (BioLegend, clone OKT3, catalog number 317330, dilution 1:200), CD4 BV650 (BioLegend, clone RPA-T4, catalog number 300536, dilution 1:200), CD8 APC-R700 (BD Bioscoences, clone RPA-T8, catalog number 565165, dilution 1:100), CXCR3 Pe/Dazzle (BioLegend, clone G025H7, catalog number 353736, dilution 1:100), CCR6 PerCP-Cy5-5 (BioLegend, clone G034E3, catalog number 353406, dilution 1:100). Cells were also stained with a dead cell stain (Molecular Probes, L10119) to exclude dead cells from analysis. Electronic compensation was performed with antibody (Ab) capture beads stained separately with individual monoclonal antibodies (MABs) used in the experimental panel. FACS was performed on a FACSAria Fusion cell sorter (BD Biosciences). Data analysis was performed using FlowJo software (Treestar) or FACSDiva software (BD Biosciences).

FACS and scRNA-seq processing of human leukocytes

Single-cell suspension of human leukocytes was prepared as described in the section Isolation and Flow cytometry of human biopsy leukocytes. ScRNA-seq of human samples from the kidney and peripheral blood was performed from FACS-sorted CD3 positive T cells using the Chromium Next GEM Single Cell 5′ Kit v2 (10x Genomics) according to manufacturer's instructions. The gating strategy, shown in Supplementary Fig. 4, is identical for the flow cytometry analysis as well as the FACS sorting. It is based on leukocytes, singlets, the living cells, CD45, and for sorting the CD3 population was collected. Libraries were sequenced aiming at 50,000 reads per cell on an Illumina NovaSeq (P150), using the CG00330 protocol from 10X Genomics.

Immunofluorescence staining

For immunofluorescence staining, paraffin-embedded kidney sections (2 µm) from ANCA-GN patients were stained with primary antibodies against CD3 (Abcam, ab11089, dilution 1:100), CCR6 (Sigma, HPA014488/Origene TA316610, dilution 1:100), and CXCR3 (BD Biosciences, 557183, dilution 1:100) after dewaxing and antigen retrieval (pH6 for 15 min). Following washing in phosphate-buffered saline, fluorochrome-labeled secondary antibodies were applied. Staining was visualized using an LSM800 with Airyscan and the ZenBlue software (all Carl Zeiss, Jena, Germany).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All gene expression data used in this manuscript are publicly available via the NCBI Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/). The newly generated data for this study is accessible under GSE253633 and GSE250138. The accession codes for all other gene expression data used are listed in the Supplementary Data 10. Source data are provided with this paper.

Code availability

The code to process and analyze the single cell sequencing and ST data is available at https://github.com/imsb-uke/ANCA-GN_transcriptomics. The source code is also deposited at Zenodo (https://doi.org/10.5281/zenodo.13208437).

References

1. Kitching, A. R. et al. ANCA-associated vasculitis. *Nat. Rev. Dis. Prim.* **6**, 71 (2020).
2. Anders, H.-J., Kitching, A. R., Leung, N. & Romagnani, P. Glomerulonephritis. Immunopathogenesis and immunotherapy. *Nat. Rev. Immunol.* **23**, 453–471 (2023).
3. Kurts, C., Panzer, U., Anders, H.-J. & Rees, A. J. The immune system and kidney disease. Basic concepts and clinical implications. *Nat. Rev. Immunol.* **13**, 738–753 (2013).
4. Wilde, B., van Paassen, P., Witzke, O. & Tervaert, J. W. C. New pathophysiological insights and treatment of ANCA-associated vasculitis. *Kidney Int.* **79**, 599–612 (2011).
5. Guillevin, L. et al. Rituximab versus azathioprine for maintenance in ANCA-associated vasculitis. *N. Engl. J. Med.* **371**, 1771–1780 (2014).
6. Specks, U. et al. Efficacy of remission-induction regimens for ANCA-associated vasculitis. *N. Engl. J. Med.* **369**, 417–427 (2013).
7. Stone, J. H. et al. Rituximab versus cyclophosphamide for ANCA-associated vasculitis. *N. Engl. J. Med.* **363**, 221–232 (2010).
8. Jayne, D. R. W., Merkel, P. A., Schall, T. J. & Bekker, P. Avacopan for the treatment of ANCA-associated vasculitis. *N. Engl. J. Med.* **384**, 599–609 (2021).
9. Lyons, P. A. et al. Novel expression signatures identified by transcriptional analysis of separated leucocyte subsets in systemic lupus erythematosus and vasculitis. *Ann. Rheum. Dis.* **69**, 1208–1213 (2010).
10. Grayson, P. C. et al. Neutrophil-related gene expression and low-density granulocytes associated with disease activity and response to treatment in antineutrophil cytoplasmic antibody-associated vasculitis. *Arthritis Rheumatol.* **67**, 1922–1932 (2015).
11. Banos, A. et al. The genomic landscape of ANCA-associated vasculitis. Distinct transcriptional signatures, molecular endotypes and comparison with systemic lupus erythematosus. *Front. Immunol.* **14**, 1072598 (2023).
12. Nishide, M. et al. Single-cell multi-omics analysis identifies two distinct phenotypes of newly-onset microscopic polyangiitis. *Nat. Commun.* **14**, 5789 (2023).
13. O'Reilly, V. P. et al. Urinary soluble CD163 in active renal vasculitis. *J. Am. Soc. Nephrol.* **27**, 2906–2916 (2016).
14. Krebs, C. F. et al. Pathogen-induced tissue-resident memory TH17 (TRM17) cells amplify autoimmune kidney disease. *Sci. Immunol.* **5**, eaba4163 (2020).
15. Paust, H.-J. et al. CD4+ T cells produce GM-CSF and drive immune-mediated glomerular disease by licensing monocyte-derived cells to produce MMP12. *Sci. Transl. Med.* **15**, eadd6137 (2023).
16. Brix, S. R. et al. Development and validation of a renal risk score in ANCA-associated glomerulonephritis. *Kidney Int.* **94**, 1177–1188 (2018).
17. Lake, B. B. et al. An atlas of healthy and injured cell states and niches in the human kidney. *Nature* **619**, 585–594 (2023).
18. Zielinski, C. E. et al. Pathogen-induced human TH17 cells produce IFN-γ or IL-10 and are regulated by IL-1β. *Nature* **484**, 514–518 (2012).
19. Jones, R. B. et al. Rituximab versus cyclophosphamide in ANCA-associated renal vasculitis. *N. Engl. J. Med.* **363**, 211–220 (2010).
20. Walsh, M. et al. Plasma exchange and glucocorticoids in severe ANCA-associated vasculitis. *N. Engl. J. Med.* **382**, 622–631 (2020).
21. Tan, J. A. et al. Mortality in ANCA-associated vasculitis. A meta-analysis of observational studies. *Ann. Rheum. Dis.* **76**, 1566–1574 (2017).
22. Flossmann, O. et al. Long-term patient survival in ANCA-associated vasculitis. *Ann. Rheum. Dis.* **70**, 488–494 (2011).
23. Sánchez Álamo, B. et al. Long-term outcomes and prognostic factors for survival of patients with ANCA-associated vasculitis. *Nephrol. Dial. Transplant.* **38**, 1655–1665 (2023).
24. Heckmann, M. et al. The Wegener's granulomatosis quantitative trait locus on chromosome 6p21.3 as characterised by tagSNP genotyping. *Ann. Rheum. Dis.* **67**, 972–979 (2008).

25. Lyons, P. A. et al. Genetically distinct subsets within ANCA-associated vasculitis. *N. Engl. J. Med.* **367**, 214–223 (2012).
26. Wang, H.-Y. et al. Risk HLA class II alleles and amino acid residues in myeloperoxidase-ANCA-associated vasculitis. *Kidney Int.* **96**, 1010–1019 (2019).
27. McKinney, E. F. et al. A CD8+ T cell transcription signature predicts prognosis in autoimmune disease. *Nat. Med.* **16**, 586 (2010).
28. Abdulahad, W. H., Kallenberg, C. G. M., Limburg, P. C. & Stegeman, C. A. Urinary CD4+ effector memory T cells reflect renal disease activity in antineutrophil cytoplasmic antibody-associated vasculitis. *Arthritis Rheum.* **60**, 2830–2838 (2009).
29. Nogueira, E. et al. Serum IL-17 and IL-23 levels and autoantigen-specific Th17 cells are elevated in patients with ANCA-associated vasculitis. *Nephrol. Dial. Transplant.* **25**, 2209–2217 (2010).
30. Schmitt, W. H. et al. Treatment of refractory Wegener's granulomatosis with antithymocyte globulin (ATG). An open study in 15 patients. *Kidney Int.* **65**, 1440–1448 (2004).
31. Griffiths, C. E. M. et al. Comparison of ustekinumab and etanercept for moderate-to-severe psoriasis. *N. Engl. J. Med.* **362**, 118–128 (2010).
32. Gottlieb, A. et al. Ustekinumab, a human interleukin 12/23 monoclonal antibody, for psoriatic arthritis. Randomised, double-blind, placebo-controlled, crossover trial. *Lancet* **373**, 633–640 (2009).
33. Feagan, B. G. et al. Ustekinumab as induction and maintenance therapy for Crohn's disease. *N. Engl. J. Med.* **375**, 1946–1960 (2016).
34. Sands, B. E. et al. Ustekinumab as induction and maintenance therapy for ulcerative colitis. *N. Engl. J. Med.* **381**, 1201–1214 (2019).
35. Jin, Y. et al. Risk of hospitalization for serious infection after initiation of ustekinumab or other biologics in patients with psoriasis or psoriatic arthritis. *Arthritis Care Res.* **74**, 1792–1805 (2022).
36. Papp, K. et al. Safety surveillance for ustekinumab and other psoriasis treatments from the psoriasis longitudinal assessment and registry (PSOLAR). *J. Drugs Dermatol.* **14**, 706–714 (2015).
37. Ghosh, S. et al. Ustekinumab safety in psoriasis, psoriatic arthritis, and Crohn's disease. An integrated analysis of phase II/III clinical development programs. *Drug Saf.* **42**, 751–768 (2019).
38. Ritchlin, C. et al. Efficacy and safety of the anti-IL-12/23 p40 monoclonal antibody, ustekinumab, in patients with active psoriatic arthritis despite conventional non-biological and biological anti-tumour necrosis factor therapy. 6-month and 1-year results of the phase 3, multicentre, double-blind, placebo-controlled, randomised PSUMMIT 2 trial. *Ann. Rheum. Dis.* **73**, 990–999 (2014).
39. McInnes, I. B. et al. Efficacy and safety of ustekinumab in patients with active psoriatic arthritis. 1 year results of the phase 3, multicentre, double-blind, placebo-controlled PSUMMIT 1 trial. *Lancet* **382**, 780–789 (2013).
40. van Vollenhoven, R. F. et al. Efficacy and safety of ustekinumab, an IL-12 and IL-23 inhibitor, in patients with active systemic lupus erythematosus. Results of a multicentre, double-blind, phase 2, randomised, controlled study. *Lancet* **392**, 1330–1339 (2018).
41. Li, H. et al. IL-23 reshapes kidney resident cell metabolism and promotes local kidney inflammation. *J. Clin. Investig.* **131**, e142428 (2021).
42. Gan, P.-Y. et al. Biologicals targeting T helper cell subset differentiating cytokines are effective in the treatment of murine anti-myeloperoxidase glomerulonephritis. *Kidney Int.* **96**, 1121–1133 (2019).
43. Kitching, A. R., Holdsworth, S. R. & Tipping, P. G. IFN-gamma mediates crescent formation and cell-mediated immune injury in murine glomerulonephritis. *J. Am. Soc. Nephrol.* **10**, 752–759 (1999).
44. Paust, H.-J. et al. The IL-23/Th17 axis contributes to renal injury in experimental glomerulonephritis. *J. Am. Soc. Nephrol.* **20**, 969–979 (2009).
45. Schreiber, A. et al. Neutrophil gelatinase-associated lipocalin protects from ANCA-induced GN by inhibiting TH17 immunity. *J. Am. Soc. Nephrol.* **31**, 1569–1584 (2020).
46. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY. Large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
47. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
48. Palla, G. et al. Squidpy. A scalable framework for spatial omics analysis. *Nat. Methods* **19**, 171–178 (2022).
49. Kang, J. B. et al. Efficient and precise single-cell reference atlas mapping with Symphony. *Nat. Commun.* **12**, 5890 (2021).
50. Büttner, M., Ostner, J., Müller, C. L., Theis, F. J. & Schubert, B. scCODA is a Bayesian model for compositional single-cell data analysis. *Nat. Commun.* **12**, 6876 (2021).
51. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler. An R package for comparing biological themes among gene clusters. *Omics J. Integr. Biol.* **16**, 284–287 (2012).
52. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA. Gene set variation analysis for microarray and RNA-seq data. *BMC Bioinforma.* **14**, 7 (2013).
53. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
54. Gayoso, A. et al. Joint probabilistic modeling of single-cell multiomic data with totalVI. *Nat. Methods* **18**, 272–282 (2021).
55. Stewart, B. J. et al. Spatiotemporal immune zonation of the human kidney. *Science* **365**, 1461–1466 (2019).
56. Khatri, R., Machart, P. & Bonn, S. DISSECT. Deep semi-supervised consistency regularization for accurate cell type fraction and gene expression estimation. *Genome Biol.* **25**, 112 (2024).
57. Mendez, D. et al. ChEMBL. Towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940 (2019).
58. Freshour, S. L. et al. Integration of the drug-gene interaction database (DGIdb 4.0) with open crowdsource efforts. *Nucleic Acids Res.* **49**, D1144–D1151 (2021).
59. Kanemaru, K. et al. Spatially resolved multiomics of human cardiac niches. *Nature* **619**, 801–810 (2023).

Acknowledgements

This study was supported by grants from the Deutsche Forschungsgemeinschaft (DFG) to U.P. (SFB 1192 A1 and C3), C.F.K. (SFB 1192 A5 and C3; KR 3483/3-1) and S.B. (SFB 1192 A2, B8, and C3). R.K. was additionally supported by the 3R (Replace, Reduce, Refine) funding of the UKE. FACS was performed at the UKE FACS sorting core facility. Single-cell RNA sequencing was performed at the UKE Single Cell Core Facility.

Author contributions

Conceptualization: S.B., C.F.K., and U.P. Methodology: J.E., R.K., D.P.S., Y.Z., H.J.P., Z.S., N.A., Formal analysis: J.E., R.K., D.P.S., Y.Z., H.J.P., N.A., S.B., C.F.K., and U.P. Spatial transcriptomics: R.K., D.P.S., Z.S., V.S., A.P., A.K., S.L.J.S., T.G.S., V.G.P., T.W., S.B., C.F.K., and U.P., scRNA sequencing: R.K., D.P.S., Y.Z., and C.F.K. Data analysis: J.E., R.K., D.P.S., Y.Z., H.J.P., Renal histology: T.W., and U.P. Patient cohorts: J.E., J.H.R., U.O.W., O.M.S., E.H., T.W., T.B.H., C.F.K., and U.P. Writing original draft: S.B., C.F.K., and U.P. Writing review and editing: J.E., R.K., D.P.S., Y.Z., J.H.R, S.B., J.E.T., H.W.M., C.F.K., and U.P. Visualization: J.E., R.K., D.P.S., S.B., C.F.K., and U.P. Supervision: S.B., C.F.K. and U.P. Funding acquisition: T.B.H., S.B., C.F.K., U.P.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-52525-w.

Correspondence and requests for materials should be addressed to Stefan Bonn, Christian F. Krebs or Ulf Panzer.

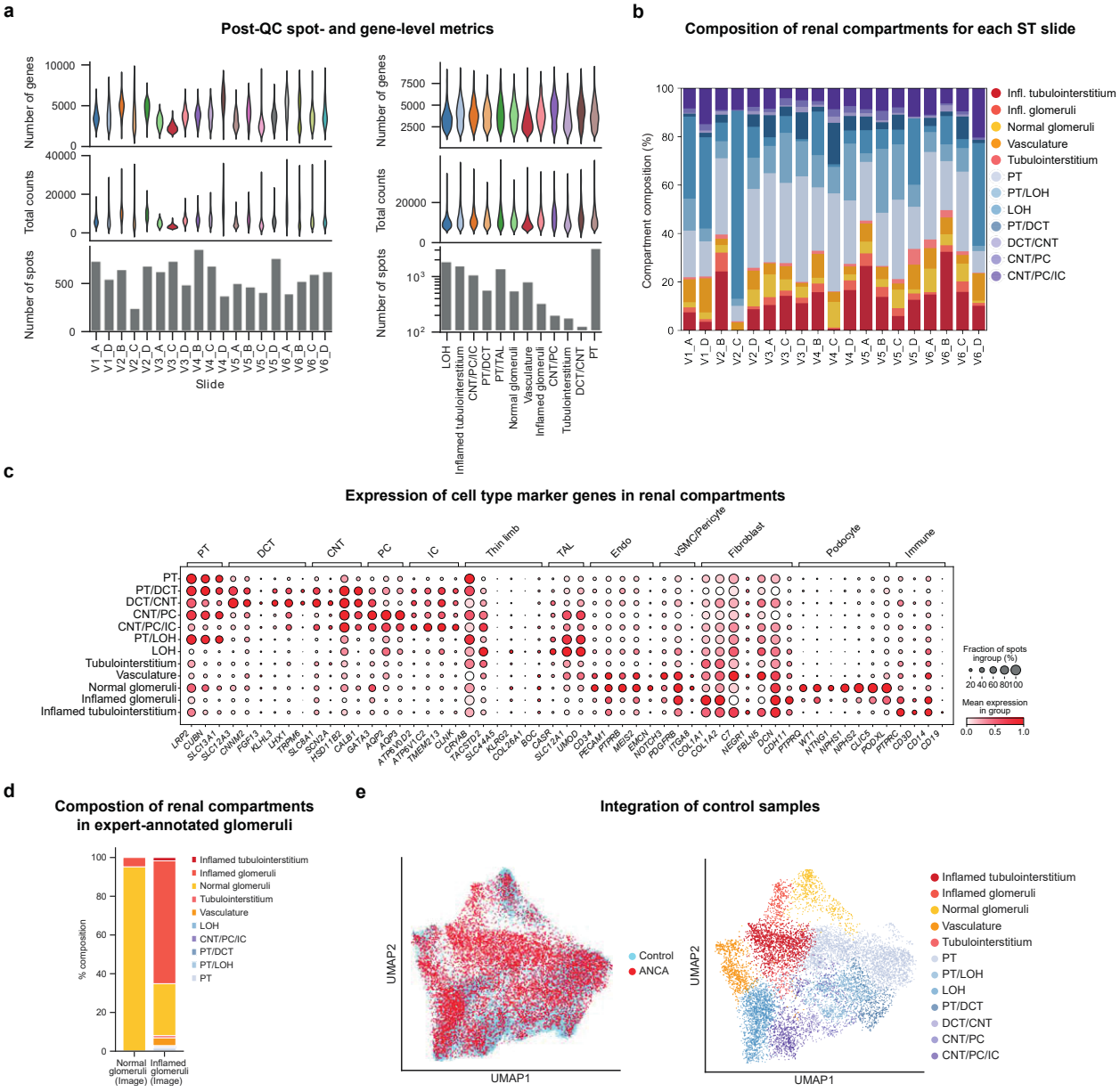
Peer review information *Nature Communications* thanks Michael Eadon, Ulrich Specks and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Immune profiling-based targeting of pathogenic T cells with ustekinumab in ANCA-associated glomerulonephritis

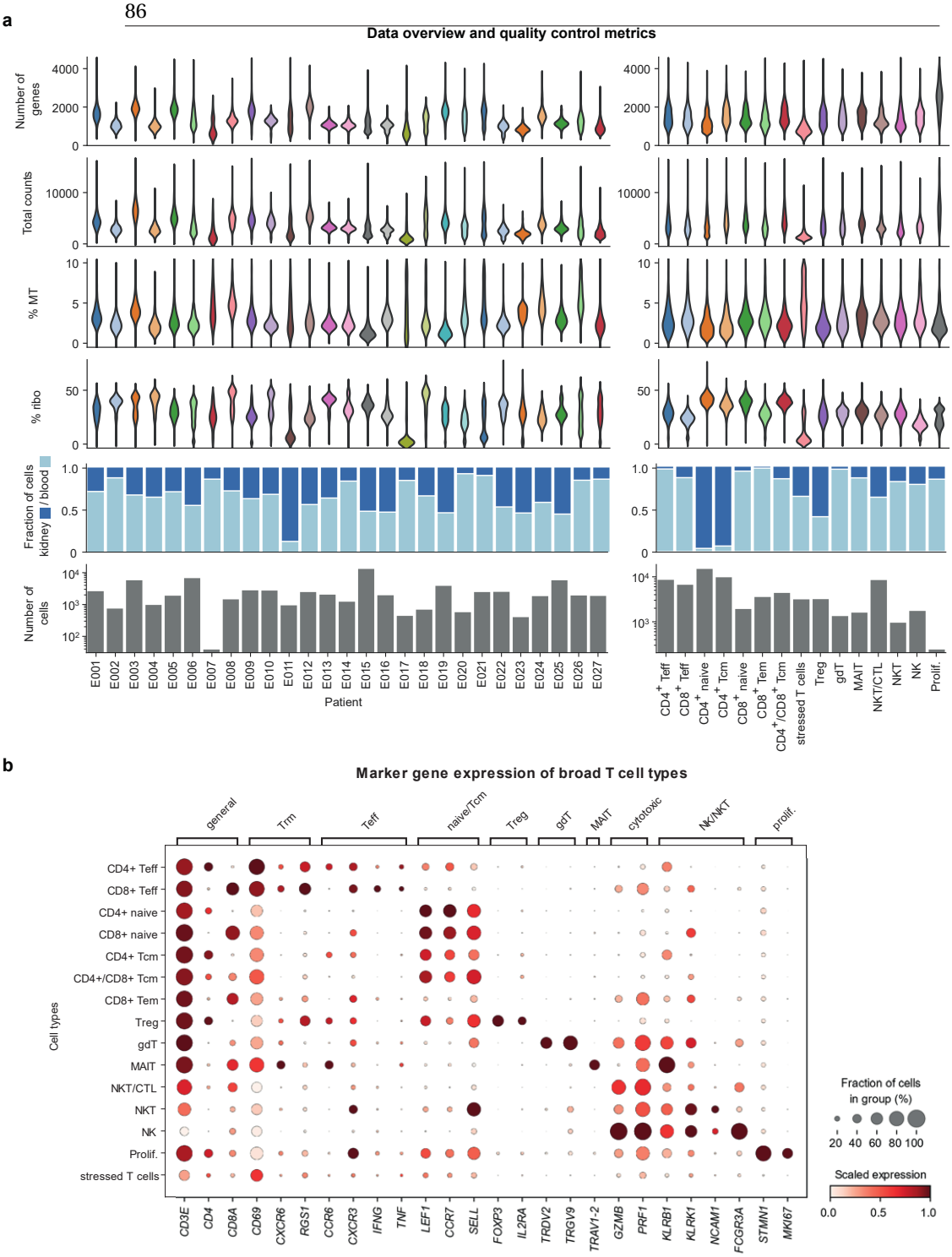
Jonas Engesser[1,2#], Robin Khatri[2,3#], Darius P. Schaub[2,3#] , Yu Zhao[1,2,3], Hans-Joachim Paust[1,2], Zeba Sultana[1,2,3], Nariaki Asada[1,2], Jan-Hendrik Riedel[1,2], Varshi Sivayoganathan[1,2], Anett Peters[1], Anna Kaffke[1], Saskia-Larissa Jauch-Speer[1], Thiago Goldbeck-Strieder[1], Victor G. Puelles[1,4], Ulrich O. Wenzel[1], Oliver M. Steinmetz[1], Elion Hoxha[1,4], Jan-Eric Turner[1,2], Hans-Willi Mittrücker[2,5], Thorsten Wiech[4,6], Tobias B. Huber[1,2,4], Stefan Bonn[2,3,4*], Christian F. Krebs[1,2,4*], Ulf Panzer[1,2,4*]

Affiliations:
[1]III. Department of Medicine, University Medical Center Hamburg-Eppendorf; Hamburg 20251, Germany
[2]Hamburg Center for Translational Immunology, University Medical Center Hamburg-Eppendorf; Hamburg 20251, Germany
[3]Institute of Medical Systems Biology, Center for Biomedical AI, Center for Molecular Neurobiology Hamburg; Hamburg 20251, Germany
[4]Hamburg Center for Kidney Health (HCKH), University Medical Center Hamburg-Eppendorf, Hamburg 20251, Germany
[5]Institute for Immunology, University Medical Center Hamburg-Eppendorf, Hamburg 20251, Germany
[6]Institute of Pathology, Division of Nephropathology, University Medical Center Hamburg-Eppendorf; Hamburg 20251, Germany

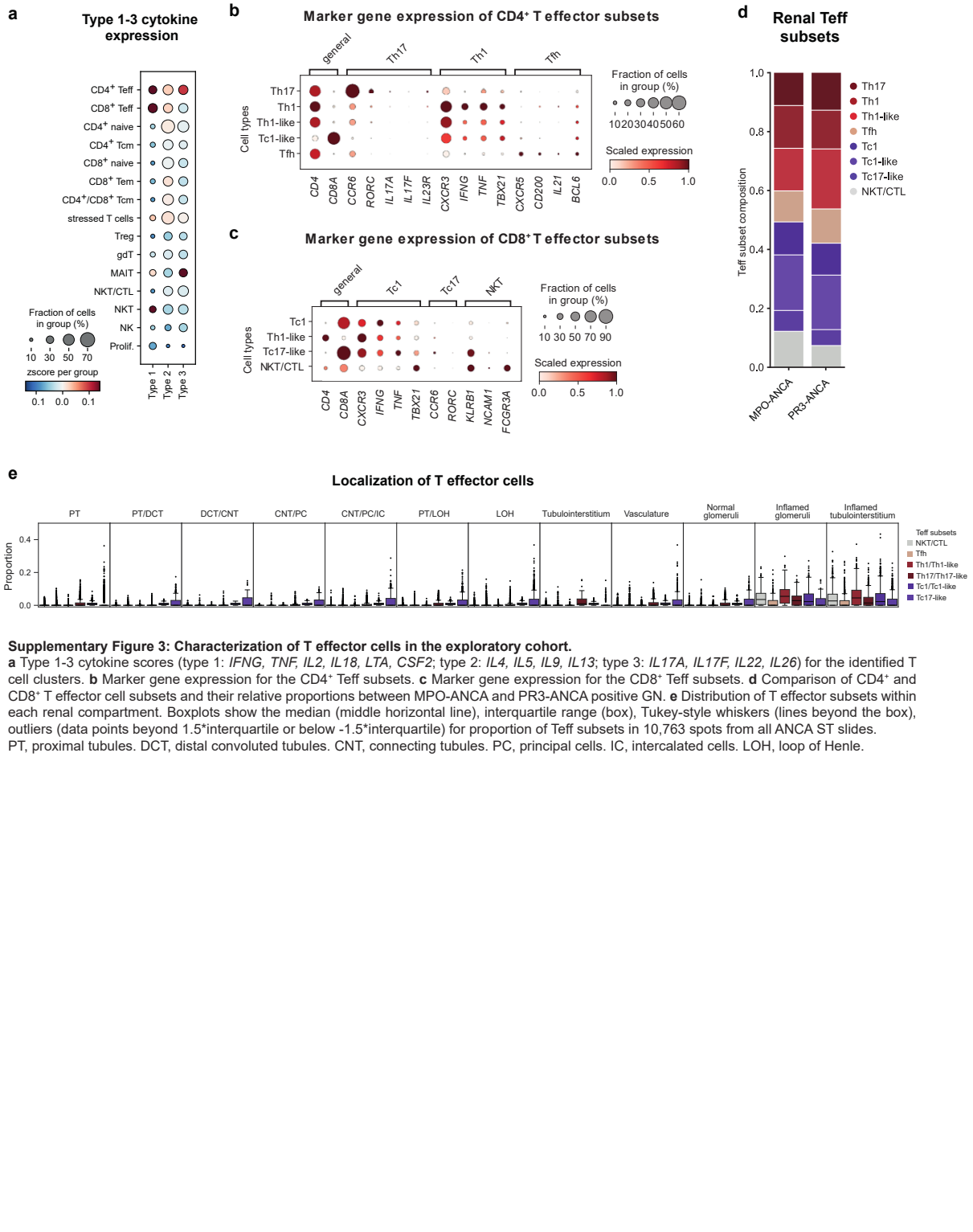
[#]These authors contributed equally: Jonas Engesser, Robin Khatri and Darius P. Schaub
*These authors jointly supervised this work: Stefan Bonn, Christian F. Krebs and Ulf Panzer



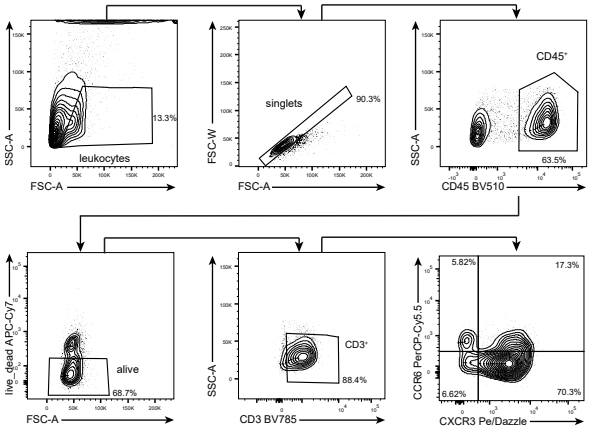
Supplementary Figure 1: Overview of spatial transcriptomics data from the exploratory group.
a Post-quality control (QC) distributions of number of genes, total UMI counts per spot, and number of spots per Visium ST slide (left) and renal compartment (right). **b** Barplots showing composition of each ST slide. **c** Dot plot showing the expression of marker genes across the renal compartments. **d** Barplots showing the distribution of gene-expression-based annotations compared to expert-based image annotations of normal and inflamed glomerular compartments (x-axis). **e** Joint UMAP-embedding of control and ANCA-GN exploratory group showing condition and renal compartments after the integration of ST slides containing control samples (8 slides and 21,420 spots).
LOH, loop of Henle. CNT, connecting tubules. PC, principal cells. IC, intercalated cells. PT, proximal tubules. DCT, distal convoluted tubules. TAL, thick ascending loop of Henle. vSMC, vascular smooth muscle cells.



Supplementary Figure 2: Exploratory cohort single cell T cell atlas.
a Quality control metrics and tissue composition across patients and cell type clusters. Violin plots show distributions of the number of genes, total counts, percentage of mitochondrial counts, and percentage of ribosomal counts. Barplots visualize the relative tissue composition and the total number of cells on a log scale. **b** Marker gene expression for the broad T cell annotations.



Supplementary Figure 3: Characterization of T effector cells in the exploratory cohort.
a Type 1-3 cytokine scores (type 1: *IFNG, TNF, IL2, IL18, LTA, CSF2*; type 2: *IL4, IL5, IL9, IL13*; type 3: *IL17A, IL17F, IL22, IL26*) for the identified T cell clusters. **b** Marker gene expression for the CD4+ Teff subsets. **c** Marker gene expression for the CD8+ Teff subsets. **d** Comparison of CD4+ and CD8+ T effector cell subsets and their relative proportions between MPO-ANCA and PR3-ANCA positive GN. **e** Distribution of T effector subsets within each renal compartment. Boxplots show the median (middle horizontal line), interquartile range (box), Tukey-style whiskers (lines beyond the box), outliers (data points beyond 1.5*interquartile or below -1.5*interquartile) for proportion of Teff subsets in 10,763 spots from all ANCA ST slides.
PT, proximal tubules. DCT, distal convoluted tubules. CNT, connecting tubules. PC, principal cells. IC, intercalated cells. LOH, loop of Henle.



Supplementary Figure 4: Gating strategy rapid biopsy immune profiling. Gating strategy for the identification of the chemokine receptors CXCR3 (Th1/Tc1) and CCR6 (Th17/Tc17) of renal $CD45^+CD3^+$ T cells used in Figure 4 and Supplemental Figure 6. Bottom right panel coressponding to Figure 4a

a

Patient 1

Case

~~A 79 year old male patient with a history of MPO-ANCA positive vasculitis was referred to a regional hospital in~~ august 2020 with an episode of macro-hematuria, as well as fatigue, dizziness and progressive lower limb edema. Laboratory assessment showed acute kidney injury and elevated CRP levels. Chest imaging excluded relevant pathologies. Urinary analysis showed gross hematuria and macro-albuminuria. BVAS was elevated to 9 points. Suspecting a relapse of AAV, the patient was transferred to university hospital Hamburg-Eppendorf.

AAV history

Initial diagnosis of AAV was made in 2017, with general symptoms, skin and kidney involvement. Kidney biopsy showed pauci-immune crescentic glomerulonephritis with small amounts of mesangial IgA deposition. Induction treatment consisted of PLEX, steroids and cyclophosphamide (cumulative CYC dose 5.0 g). Azathioprine was given for remission maintenance. After 7 months, azathioprine was discontinued due to severe side effects, and rituximab was started. Rituximab was well tolerated, and stable remission achieved, although full B cell depletion was not achieved during maintenance therapy with rituximab. Last dose RTX was given 6 months before re-admission with suspected AAV relapse.

Diagnostic workup at relapse

Hemoglobin [g/dl]	12.0
Leukocytes [10^9/l]	7.8
Platelets [10^9/l]	271
Creatinie [mg/dl]	4.25
eGFR [ml/min]	13
CRP [mg/l]	36
MPO-ANCA [U/ml]	22
B cells [%]	6 (34/µl

Kidney biopsy was performed, displaying a relapse of ANCA-GN with a high ARRS. Rapid biopsy immune profiling of the kidney biopsy revealed an enrichment of Th1/Tc1 an Th17/Tc17 cells in the kidney tissue.

Treatment considerations

Patient tolerated rituximab well, but full B cell depletion could not be achieved, despite appropriate RTX-dosage. Furthermore, the patient suffered a relapse while on RTX treatment. Urinary flow obstruction and the need of urinary diversion led the patient away from full dose CYC, while still giving consent to two doses of i.v. CYC in the inpatient setting. Rapid immune profiling showed an increase in Th1/Tc1 and Th17/Tc17 cells, suggesting suitability for an anti-IL12/IL-23 treatment.

Treatment

- Pulsed i.v. steroids (3 x 500 mg prednisolone), followed by an oral tapering, according to the PEXIVAS trial reduced dose regimen. By 6 months prednisolone was tapered to 5mg daily (cumulative prednisolone dose: 3900 mg).
- Two i.v. pulses of cyclophosphamide (cumulative dosage 2g) were given in the inpatient setting.

Outcome

Combined treatment of ustekinumab, steroids and low dose CYC resulted in a rapid clinical improvement with resolution of fatigue, edema and dizziness. Creatinine rapidly improved and uACR declined, as well as ANCA-levels, CRP, and BVAS. At 6 months the patient had an elevated BVAS of 3 because of grade 1 hypertension and persistent albuminuria, mostly attributed to glomerular scarring.

organ involvement

medical history
- monoclonal B cell lymphocytosis
- exogen-allergic asthma
- OSA
- obesity

b

Patient 2

Case

A 52 year old male patient with known history of MPO-ANCA positive vasculitis with extensive organ manifestation was admitted to university medicine hospital Hamburg-Eppendorf, because of increasing creatinine and albuminuria after induction treatment with steroids (3 x 500 mg prednisolone and oral taper) and six pulses of i.v. cyclophosphamide (cumulative dose of 6.0 g; last dose given 4 weeks prior to admission). Initial manifestations (ENT, pulmonary, nervous and cutaneous) had improved since starting of induction treatment with steroids and CYC. BVAS was increased to 12 at admission. With suspected worsening ANCA-GN kidney biopsy was initiated.

AAV history

Initial diagnosis of AAV was made 02/2021 with general, renal, ENT, lung, nervous system, cutaneous, renal and heart manifestation. Treatment with i.v. prednisolone followed by an oral taper (according to the PEXIVAS reduced dose regimen), as well as pulsed i.v. cyclophosphamide (according to EUVAS scheme, cumulative dose 6.0 g) was initiated.

Diagnostic workup at relapse

Hemoglobin [g/dl]	10.7
Leukocytes [10^9/l]	6.5
Platelets [10^9/l]	282
Creatinie [mg/dl]	2.27
eGFR [ml/min]	32
CRP [mg/l]	5
MPO-ANCA [U/ml]	29

Urinary sediment showed dysmorphic erythrocytes (acanthocytes). Kidney biopsy demonstrated pauci-immune crescentic glomerulonephritis with chronic and active lesions, compatible with active ANCA-GN with a low ARRS. Rapid immune-profiling of the kidney biopsy revealed an enrichment of Th1/Tc1 and Th17/Tc17 cells in the kidney tissue.

Treatment considerations

Cyclophosphamide treatment along with steroids could not achieve remission and renal manifestation showed a progressive disease with active glomerular lesions. COVID-19 pandemic was still on the rise, raising concerns over rituximab treatment. Rapid immune profiling of the kidney showed an enrichment of Th1/Tc1 and Th17/Tc17 cells.

Treatment

- Pulsed steroids (3x500mg prednisolone), followed by an oral tapering. By 6 months prednisolone was tapered to 5mg daily (cumulative prednisolone dose: 3755 mg).
- Two i.v. pulses of cyclophosphamide (cumulative dosage 2.0 g) were given.
- Ustekinumab 90mg s.c. was administered at weeks 0, 4, 12, and 24.
- Azathioprine (1 mg/kg) was started at week 22 after beginning of induction therapy as additional maintenance therapy.

Outcome

Combined treatment of ustekinumab, steroids and low dose CYC resulted in a rapid clinical improvement with reduction in creatinine, uACR and ANCA-levels (16 U/ml). At 6 months the patient had an elevated BVAS of 2, because of persistent microalbuminuria, undulating around 200 mg/day.

organ involvement

medical history
- hyperlipidemia
- postherpetic neuralgia

c

Patient 3

Case

~~A 92 year old male with a history of PR3-ANCA positive vasculitis was re-admitted to~~ university hospital Hamburg-Eppendorf, because of fever, night sweats, weight loss, and progressive dyspnea with hemoptysis. Four weeks before admission the patient was diagnosed with an AAV relapse- rituximab (two pulses of 1g) and pulsed steroids as well as an oral tapering regimen were started. Clinical examination at re-admission showed moist crackles in chest auscultation. BVAS was 13 at that time.

AAV history

Initial diagnosis of AAV was made 2019 and treatment with i.v. prednisolone followed by an oral taper (according to the PEXIVAS reduced dose regimen) as well as rituximab was initiated. After two doses (1g each) of rituximab, the patient was lost to follow up. In 09/2022 aforementioned relapse with pulmonary and ENT manifestation occurred and pulsed prednisolone as well as re-induction with rituximab (two times 1g) was started. Initially the patient's symptoms improved and he was discharged.

Diagnostic workup at relapse

Hemoglobin [g/dl]	8.2
Leukocytes [10^9/l]	1.7
Platelets [10^9/l]	435
Creatinie [mg/dl]	1.88
eGFR [ml/min]	46
CRP [mg/l]	190
PR3-ANCA [U/ml]	101
B cells [%]	0 (1/µl)

CT scans and bronchoscopy showed progressive DAH. Urinary sediment showed extensive acanthocyturia. Kidney biopsy was performed and displayed pauci-immune crescentic glomerulonephritis with active lesions, compatible with ANCA-GN, with a low ARRS. Rapid immune-profiling of the kidney biopsy revealed an enrichment of Th1/Tc1 and Th17/Tc17 cells in the kidney tissue.

Treatment considerations

Severe leukopenia contradicted full dose cyclophosphamide re-induction treatment. Patient was also concerned of cyclophosphamide treatment-related adverse effects, and initially rejected an i.v. treatment. B cells were fully depleted, as rituximab was given four weeks prior. Rapid immune profiling of the kidney showed an enrichment of Th1/Tc1 and Th17/Tc17 cells.

Treatment

- Pulsed steroids (3x250mg prednisolone), followed by an oral tapering. By 6 months prednisolone was tapered to 5mg daily (cumulative prednisolone doses 3520 mg).
- Oral cyclophosphamide was given as the patient was initially rejecting i.v. therapy. After repeated shared decision talks, the patient gave consent to one more i.v pulse under close monitoring of WBC (cumulative dose: 500 mg p.o. and 1000mg i.v.).
- Ustekinumab 90mg s.c. was administered at weeks 0, 4, 12, and 24.
- Starting at week 16 after re-induction therapy AZA was started for additional maintenance therapy and eventually switched to low dose MMF (2x500 mg) because of AZA side effects.

Outcome

Combined treatment of ustekinumab, steroids and low dose CYC resulted in a rapid clinical improvement with suspended night sweats, fever and hemoptysis, a reduction in creatinine, ANCA-levels, and BVAS. At 6 months the patient had an elevated BVAS of 2, because of persisting albuminuria. Antiproteinuric treatment with ACEi was not well tolerated and could only be given at the lowest dose.

organ involvement

medical history
- no chronic diseases

d

Patient 4

Case

A 72 year old female with a history of MPO-ANCA positive vasculitis (renal limited disease) and myelodysplastic syndrome was admitted to university medicine hospital Hamburg-Eppendorf, because of acute kidney injury and reduced general condition. The patient was receiving rituximab as remission maintenance, with the last dose given 5 months prior to hospital admission. BVAS was 13 at admission. With suspected relapse of ANCA-GN diagnostic workup was initiated.

AAV history

Initial diagnosis of AAV was made in august 2020 with a renal limited phenotype. Treatment with i.v. prednisolone followed by an oral taper (according to the PEXIVAS reduced dose regimen) as well as pulsed i.v. cyclophosphamide (cumulative dose 3600 mg) was initiated. After remission induction azathioprine was given for remission maintenance. Eventually, azathioprine had to be discontinued because of side effects and rituximab at doses of 500 mg/6 months was initiated in 05/2022.

Diagnostic workup at relapse

Hemoglobin [g/dl]	8.9
Leukocytes [10^9/l]	3.0
Platelets [10^9/l]	264
Creatinie [mg/dl]	3.42
eGFR [ml/min]	13
CRP [mg/l]	28
MPO-ANCA [U/ml]	>134
B cells [%]	2 (20/µl)

Chest X-ray showed no signs of pulmonary involvement. Urinary analysis showed extensive acanthocyturia and macro-albuminuria. Kidney biopsy was performed and showed pauci-immune crescentic glomerulonephritis with active lesions, compatible with ANCA-GN, with a high ARRS. Rapid immune-profiling of the kidney biopsy revealed an enrichment of Th1/Tc1 and Th17/Tc17 cells in the kidney tissue.

Treatment considerations

Leukopenia of 3.0 x10^9/l with known myelodysplastic syndrome under high dose erythropoietin therapy contradicted full dose cyclophosphamide re-induction therapy. Retrospective the patient showed a rise in MPO-ANCAs in 01/2023 while B cells were depleted (0% / 1/µl). Rapid immune profiling of the kidney showed an enrichment of Th1/Tc1 and Th17/Tc17 cells.

Treatment

- Pulsed steroids (3 x 500 mg prednisolone), followed by an oral tapering regimen according to the PEXIVAS study reduced dose arm. By 6 months prednisolone was tapered to 5mg daily (cumulative prednisolone dose: 4285 mg).
- Three pulses of i.v. cyclophosphamide were given (cumulative dose of 1800mg).
- Ustekinumab 90mg s.c. was administered at weeks 0, 4, 12, and 24.
- Low dose MMF (2 x 500 mg) was stared at week 16 after beginning of induction therapy as additional maintenance therapy.

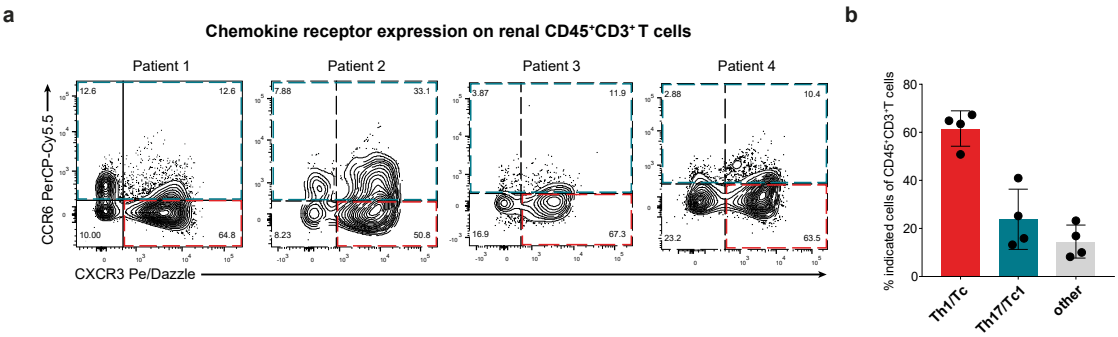
Outcome

Combined treatment of ustekinumab, steroids and low dose CYC resulted in a rapid clinical improvement with substantial improvement of general condition, reduction in creatinine, uACR, ANCA-levels, and BVAS reduction to 5. At 6 months the patient had an elevated BVAS of 5, because of albuminuria and eumorphic hematuria.

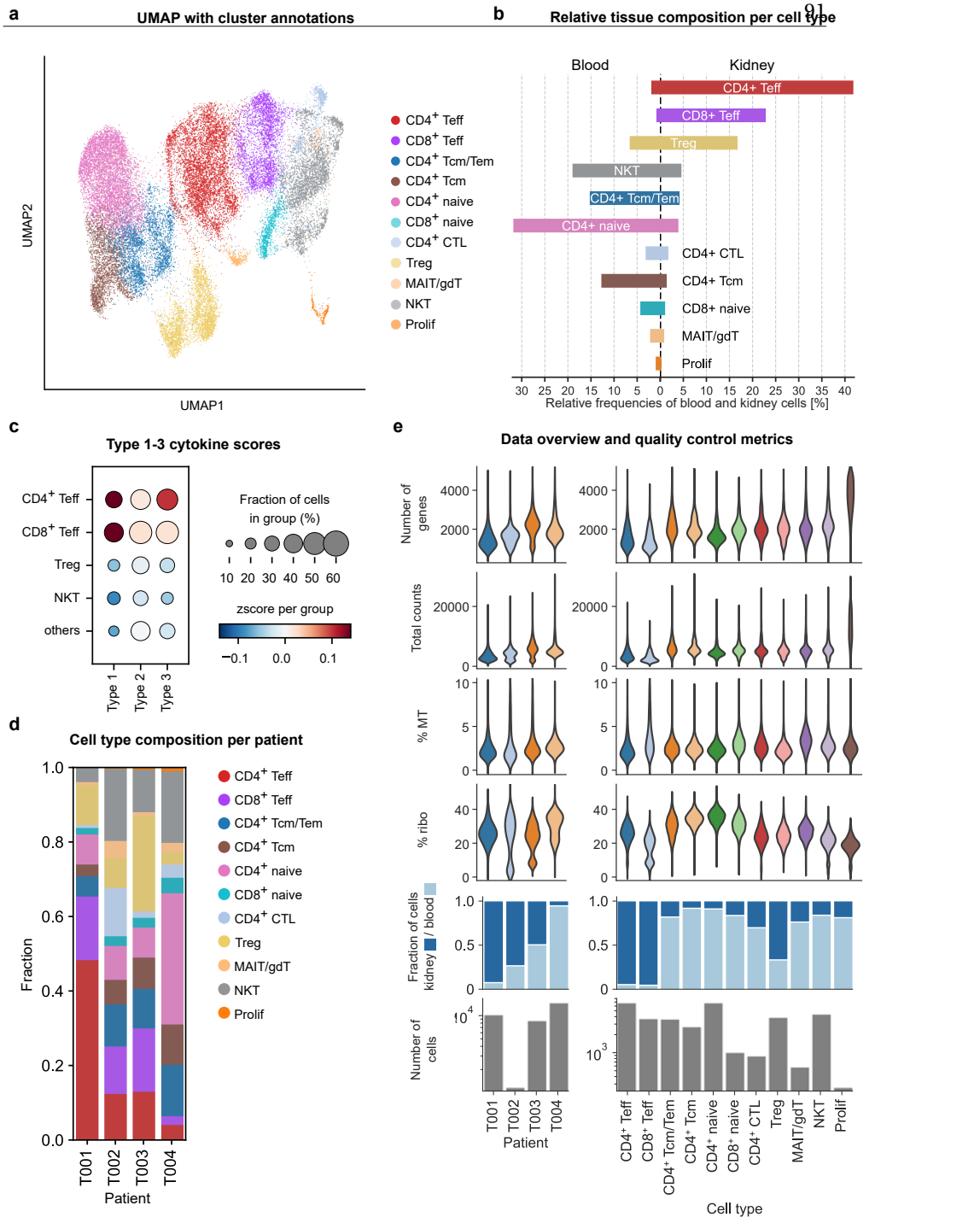
organ involvement

medical history
- myelodysplastic syndrome
- osteoporosis
- Hashimoto's thyroiditis

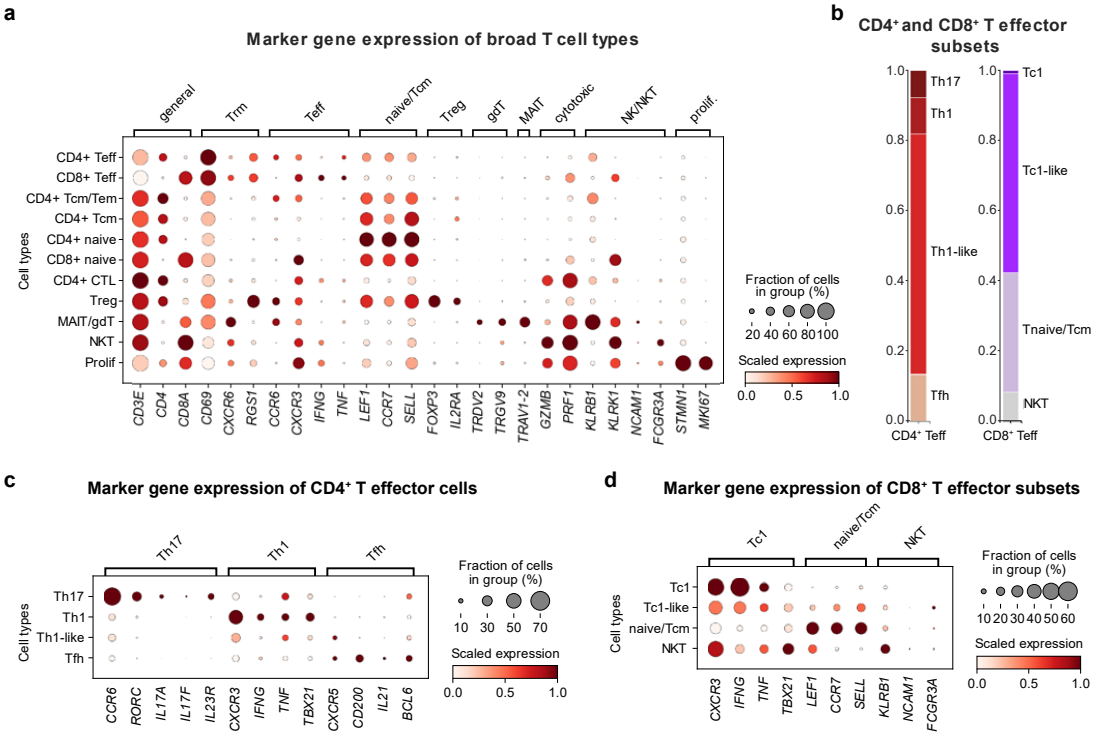
Supplementary Figure 5: Detailed case vignettes for each patient of the ustekinumab treatment cohort.
Case vignettes illustrating a brief case description, history, diagnostic workup, as well as treatment considerations, treatment and outcomes for each patient of the ustekinumab treatment cohort. **a** Patient 1. **b** Patient 2. **c** Patient 3. **d** Patient 4. BVAS, Birmingham Vasculitis Activity Score. AAV, ANCA-associated vasculitis. PLEX, therapeutic plasma exchange. CYC, cyclophosphamide. ARRS, ANCA renal risk score (21). RTX, rituximab. uACR, urinary albumin to creatinine ratio. OSA, obstructive sleep apnoea. ENT, ear nose throat. DAH, diffuse alveolar hemorrhage. MMF, mycophenolate mofetil. ACEi, angiotensin-converting enzyme inhibitor.



Supplementary Figure 6: Rapid Immunoprofiling of T cells in the kidneys of the treatment cohort.
a Flow cytometry-based identification of chemokine receptor expression from T cells isolated from biopsy samples of patients with ANCA-GN (n=4). **b** Quantification of chemokine receptor expression CXCR3 (Th1/Tc1) and CCR6 (Th17/Tc17) from renal $CD45^+CD3^+$T cells. Bar graphs show mean with SD, symbols represent individual data points.



Supplementary Figure 7: Ustekinumab treatment group single-cell T cell atlas.
a UMAP projection of the integrated single-cell embeddings with corresponding cluster annotations. **b** Tissue composition for the different T cell clusters ordered by descending kidney enrichment. **c** Type 1-3 cytokine scores (type 1: *IFNG, TNF, IL2, IL18, LTA, CSF2*; type 2: *IL4, IL5, IL9, IL13*; type 3: *IL17A, IL17F, IL22, IL26*) for the T cell clusters most enriched in the kidney and the other clusters combined. **d** Cell type composition per patient. **e** Quality control metrics and tissue composition across patients and cell type clusters. Violin plots show distributions of the number of genes, total counts, percentage of mitochondrial counts, and percentage of ribosomal counts. Barplots visualize the relative tissue composition and the total number of cells on a log scale.



Supplementary Figure 8: Ustekinumab treatment group single cell T cell atlas marker genes and T effector cell type composition. **a** Marker gene expression for the broadly annotated T cell clusters. **b** CD4+ and CD8+ T effector subset cell type composition. **c** Marker gene expression for the CD4+ Teff subsets. **d** Marker gene expression for the CD8+ Teff subsets.

Description of Additional Supplementary Files

Supplementary Data 1: Alignment metrics of ST data from the ANCA-GN exploratory group.

Supplementary Data 2: Summary of differential population analysis in the ST data between the control samples and the ANCA-GN exploratory group. The analysis was performed using the standard scCODA model with CNT/PC cluster as the reference cell type.

Supplementary Data 3: Differential gene expression of the renal compartments in the ST data from the ANCA-GN exploratory group. Differential gene expression was performed using two-sided Wilcoxon rank-sum tests with Benjamini-Hochberg multiple test correction.

Supplementary Data 4: Gene set enrichment analysis of the inflamed interstitial compartment in the ANCA-GN exploratory group. Enrichment was performed using clusterProfiler enrichGO function that used right-tailed Fisher's Exact test with Benjamini-Hochberg multiple test correction. Statistical significance was tested using the *enrichGO* function from R package clusterProfiler with right-tailed Fisher's exact test and Benjamini-Hochberg multiple test correction.

Supplementary Data 5: Gene set enrichment analysis of the inflamed glomerular compartment in the ANCA-GN exploratory group. Statistical significance was tested using the *enrichGO* function from R package clusterProfiler with right-tailed Fisher's exact test and Benjamini-Hochberg multiple test correction.

Supplementary Data 6: Post QC quality metrics for the single cell sequencing data of the exploratory cohort.

Supplementary Data 7: Final drug candidates and their scores in ANCA-GN exploratory group and Teff cells from the ANCA-GN T cell atlas. Statistical significance was tested using the *enrichGO* function from R package clusterProfiler with right-tailed Fisher's exact test and Benjamini-Hochberg multiple test correction.

Supplementary Data 8: Post QC quality metrics for the single cell sequencing data of the ustekinumab treatment cohort.

Supplementary Data 9: Alignment metrics of the ST data from the internal control group.

Supplementary Data 10: Data overview including publication status for the CITE-/scRNA-seq datasets per patient and tissue.
Detailed information about performed analysis of each patient, as well as information about the immunosuppressive treatment up to 7 days prior to kidney biopsy. Furthermore, all accession codes to the gene expression data listed.
(B, blood: k, kidney; MPO, myeloperoxidase; PR3, proteinase 3; RTX, rituximab; CYC, cyclophosphamide; AZA, azathioprine)

4

Neural glioblastoma integrates into neuron-glioma-networks and predicts therapeutic vulnerability

nature medicine

Article

https://doi.org/10.1038/s41591-024-02969-w

A prognostic neural epigenetic signature in high-grade glioma

Received: 7 August 2023

Accepted: 3 April 2024

Published online: 17 May 2024

Check for updates

A list of authors and their affiliations appears at the end of the paper

Neural–tumor interactions drive glioma growth as evidenced in preclinical models, but clinical validation is limited. We present an epigenetically defined neural signature of glioblastoma that independently predicts patients' survival. We use reference signatures of neural cells to deconvolve tumor DNA and classify samples into low- or high-neural tumors. High-neural glioblastomas exhibit hypomethylated CpG sites and upregulation of genes associated with synaptic integration. Single-cell transcriptomic analysis reveals a high abundance of malignant stemcell-like cells in high-neural glioblastoma, primarily of the neural lineage. These cells are further classified as neural-progenitor-cell-like, astrocyte-like and oligodendrocyte-progenitor-like, alongside oligodendrocytes and excitatory neurons. In line with these findings, high-neural glioblastoma cells engender neuron-to-glioma synapse formation in vitro and in vivo and show an unfavorable survival after xenografting. In patients, a high-neural signature is associated with decreased overall and progression-free survival. High-neural tumors also exhibit increased functional connectivity in magnetencephalography and resting-state magnet resonance imaging and can be detected via DNA analytes and brain-derived neurotrophic factor in patients' plasma. The prognostic importance of the neural signature was further validated in patients diagnosed with diffuse midline glioma. Our study presents an epigenetically defined malignant neural signature in high-grade gliomas that is prognostically relevant. High-neural gliomas likely require a maximized surgical resection approach for improved outcomes.

The importance of the nervous system as a regulator of brain tumors has been repeatedly highlighted but has not yet been translated into a therapeutically relevant setting[1–5]. Particularly in gliomas, studies have demonstrated that the activity-driven formation of malignant neuron-to-glioma networks is critical for cancer progression[4,6–8], and that glioma cells remodel neuronal circuits by increasing neuronal hyperexcitability[4,9–12]. Further insight into molecular mechanisms identified connected and unconnected glioblastoma cells that form distinct cell states and differ in their gene signatures as well as functions within neuron-to-glioma networks[13]. Additionally, glioblastomas exhibiting high functional connectivity have been shown to be associated with poorer survival[12]. Moreover, callosal projection neurons were shown to promote glioma progression and widespread infiltration underpinning the importance of the central nervous system as a critical regulator[14].

High-grade glioma consists of both malignant and nonmalignant cells[15,16]. Therefore, their cell-type composition can be determined through epigenetic bulk DNA analysis, which allows for the identification of molecular differences. Here, we aimed to use brain tumor-related epigenetic signatures to understand isocitrate dehydrogenase (IDH)-wild-type high-grade gliomas, suggesting that certain epigenetic subclasses may be more likely to be integrated into neuron-to-glioma networks with clinical relevance. We analyzed the epigenetic neural signature of central nervous system (CNS) tumors,

e-mail: f.ricklefs@uke.de

categorizing glioblastoma and H3K27-altered diffuse midline glioma (DMG) into low- and high-neural subgroups, which were characterized molecularly, functionally and clinically.

Results

Epigenetic neural signature predicts patients outcome

To address our hypotheses, we applied the epigenetic neural signature of Moss et al.[17] to estimate cellular composition (Fig. 1a) of a combined dataset of epigenetically profiled CNS tumors of Capper et al.[18] and our institutional cohorts (Fig. 1b) as well as healthy tissue (Extended Data Fig. 1a). Using this combined dataset, glioblastoma samples ($n = 1{,}058$) were dichotomized for defining a cutoff separating low- and high-neural tumors (cutoff based on median neural proportion 0.41; Fig. 1c,d). We demonstrate that more than two clusters did not show significant separability of survival among the resulting clusters (Extended Data Fig. 1b,c). The reproducibility of the cutoff (0.41) was validated across multiple cohorts (Extended Data Fig. 1d–f). The cutoff was applied to 363 patients with glioblastoma from our clinical cohort who received surgical treatment followed by standard-of-care combined chemoradiotherapy. Survival analysis revealed a significantly shorter overall survival ($P < 0.0001$, median overall survival 14.2 versus 21.2 months; Fig. 1e) and progression-free survival (PFS) ($P = 0.02$, median PFS 6.2 versus 10.0 months; Fig. 1f) for patients with a high-neural glioblastoma (Extended Data Table 1). This finding was replicated in an external cohort with 187 patients from The Cancer Genome Atlas (TCGA)-GBM database[19] ($P < 0.01$, median overall survival 12.0 versus 17.1 months; Fig. 1g). The neural classification was identified as an independent prognostic factor for overall survival (odds ratio (OR) 1.96; 95% confidence interval (CI) 1.45–2.64, $P < 0.01$; Fig. 1h) and PFS (OR 1.51; 95% CI 1.13–2.02, $P < 0.01$; Fig. 1i). Other infiltrating brain tumor cell types of the lymphoid or myeloid lineage did not show an association with patient survival (Extended Data Fig. 1g–j).

High-neural glioblastomas exhibit a synaptic character

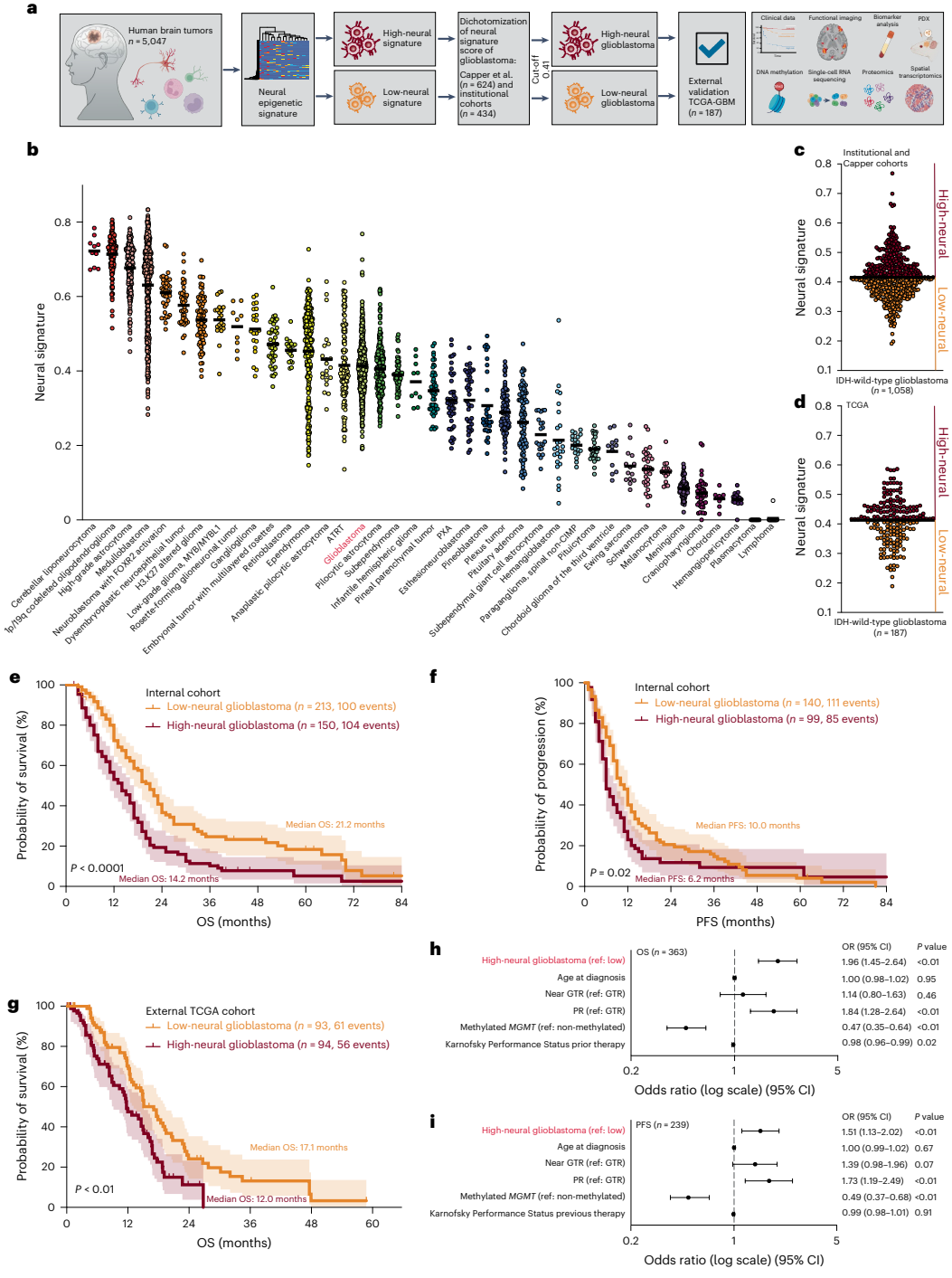
To discern epigenetic differences in low- and high-neural glioblastomas, we applied the 'invasivity signature'[13] (172 genes linked to neural features, migration and invasion) to the DNA methylation data of our clinical cohort (Supplementary Table). High-neural tumors were hypomethylated at CpG sites within gene loci of the invasivity signature compared to low-neural tumors (Extended Data Fig. 2a). In addition, two gene sets that are either associated with neuron-to-glioma synapse formation[20] ('neuronal signature genes'; Supplementary Table) or trans-synaptic signaling[21] ('trans-synaptic signaling genes'; Supplementary Table) were hypomethylated in high-neural glioblastomas (Extended Data Fig. 2a), whereas synapse-related genes were upregulated in high-neural glioblastoma (Extended Data Fig. 2b).

Next, we used an integrative analysis of paired epigenetic and transcriptomic datasets of glioblastoma samples ($n = 86$). First, we computed a scale-free gene expression network (weighted correlation network analysis; WGCNA[22]) resulting in gene expression modules, which were further correlated to the neural signature through module significance measurement by quantifying the absolute correlation between the epigenetic signature and the individual module-derived gene expression profiles (Fig. 2a,b). We identified three expression modules significantly correlated with the epigenetic status of high-neural glioblastomas: module green ($R^2 = 0.55$, $P = 3.5 \times 10^{-6}$), module cyan ($R^2 = 0.67$, $P < 2.2 \times 10^{-22}$) and module midnight blue ($R^2 = 0.41$, $P = 9.3 \times 10^{-5}$) (Fig. 2c,d). Gene Ontology analysis revealed that these modules were associated with synaptic functions (*GRIN3A*, *SYT4* and *SNAP25*), regulating the expression of genes involved in neuronal differentiation (*NEUROD2*) and calcium-dependent cell adhesion (*CDH22*, *CNTNAP5* and *CNTN3*) (Fig. 2e,f).

We projected module eigengene signatures onto an integrated single-cell dataset of malignant (GBMap[23]) and healthy brain cells from the motor cortex (Allen Brain Institute). This analysis revealed a significant enrichment of the corresponding expression modules clustering to cells of the neural lineage such as healthy neurons along with malignant neural-progenitor-like cells (NPCs) and oligodendrocyte-progenitor-like cells (OPCs) (module green and cyan, $P < 0.01$), as well as nonmalignant oligodendrocytes (module midnight blue, $P < 0.01$) (Fig. 2g–i and Extended Data Fig. 3a). This correlation with the signature, dominated by typical neuronal marker genes, was anticipated. To assess whether the neural signature in our samples reflects malignant cell properties or merely the presence of neurons, we analyzed the relationship between DNA purity and the neural signature, finding a notable positive correlation ($P < 0.001$, $R^2 = 0.19$; Extended Data Fig. 3b), whereas microglia ($P < 0.001$, $R^2 = 0.35$; Extended Data Fig. 3c) and immune cell signatures ($P < 0.001$, $R^2 = 0.67$; Extended Data Fig. 3d) showed a negative correlation. Our study, using only glioblastoma samples with a reliable diagnostic output from the DKFZ methylation classifier (Methods) showed that the calibrated score for 'IDH-wild-type glioblastoma' was unaffected by the epigenetic neural signature, nor vice versa ($P = 0.39$, $R^2 = 0.003$; Extended Data Fig. 3e). Additionally, a non-reference-based multi-dimensional single-cell deconvolution algorithm[24] was used to differentiate the neural signature in tumor cells from neuronal contamination. The analysis, which included glioblastoma tissue, matching tumor monocultures ($n = 17$), healthy cortex ($n = 9$) and sorted NeuN$^+$ cells ($n = 5$), confirmed a higher stem-cell-like signature in glioblastoma tissue and cell cultures (Extended Data Fig. 3f) and the distinct neuronal signature in NeuN$^+$ cells and healthy cortex (Extended Data Fig. 3g). Integrating RNA sequencing (RNA-seq) data, we observed 64 out of 67 samples (95.52%; Extended Data Fig. 3h) clustered into the established Verhaak transcriptomic glioblastoma subtypes (classical, mesenchymal and proneural)[25]. Ultimately, we analyzed the neural signature in cell cultures from 17 freshly resected patients with glioblastoma and observed a well-preserved neural signature (Extended Data Fig. 3i), which remained stable even in long-term cultures (Extended Data Fig. 3j) without the presence of NeuN$^+$ cells (Extended Data Fig. 3k).

The synaptic character of high-neural glioblastoma was further validated in the tumor proteome (Extended Data Fig. 4a–f), showing an increase in proteins related to synaptic transmission (Extended Data Fig. 4a–d) and characteristics of malignant OPC-like, astrocyte-like and NPC-like cells (Extended Data Fig. 4e,f). Histopathological

Fig. 1 | Epigenetic neural classification predicts outcome of patients with glioblastoma. a, Schematic of the study workflow. In humans ($n = 5{,}047$) diagnosed with a CNS tumor we performed deconvolution using DNA methylation arrays (850k or 450k) for determining the neural signature. IDH-wild-type glioblastomas were stratified into subgroups with a low- or high-neural signature for further analyses. **b**, Epigenetic neural signature in all CNS tumor entities ($n = 5{,}047$). **c**, Dichotomization of the combined dataset from Capper et al.[18] and three institutional cohorts (Hamburg, Berlin and Frankfurt, all Germany) into low- and high-neural glioblastomas. The black line indicates a median neural score of all included patients with glioblastoma ($n = 1{,}058$) and represents the cutoff (0.41) for stratification into low- and high-neural glioblastoma. **d,** External validation of the cutoff value using the TCGA-GBM dataset ($n = 187$). The black line indicates the median neural score. **e**–**i**, Survival analysis of patients with low- and high-neural glioblastoma treated by radiochemotherapy after surgery. **e**, Overall survival (OS) of 363 patients with glioblastoma of the internal clinical cohort. log-rank test, $P = 0.000005$. Error bands represent 95% CI. **f**, PFS of 226 patients with glioblastoma of the internal clinical cohort. log-rank test, $P = 0.0233$. Error bands represent 95% CI. **g**, Overall survival of 187 patients with glioblastoma of the TCGA-GBM cohort. log-rank test, $P = 0.0017$. Error bands represent 95% CI. **h**,**i**, Forest plots illustrating multivariate analysis of patients with glioblastoma from the internal clinical cohort. Means are shown by closed circles and whiskers represent 95% CI. GTR, gross total resection; PR, partial resection; MGMT, O^6-methylguanine-DNA-methyltransferase.


a
Human brain tumors n = 5,047
Neural epigenetic signature
High-neural signature
Low-neural signature
Dichotomization of neural signature score of glioblastoma: Capper et al. (n = 624) and institutional cohorts (n = 434)
Cut-off 0.41
High-neural glioblastoma
Low-neural glioblastoma
External validation TCGA-GBM (n = 187)
Clinical data
Functional imaging
Biomarker analysis
PDX
DNA methylation
Single-cell RNA sequencing
Proteomics
Spatial transcriptomics
b
Neural signature
Cerebellar liponeurocytoma
1p/19q codeleted oligodendroglioma
High-grade astrocytoma
Medulloblastoma
Neuroblastoma with FOXR2 activation
Dysembryoplastic neuroepithelial tumor
H3K27 altered glioma
Low-grade glioma, MYB/MYBL1
Rosette-forming glioneuronal tumor
Embryonal tumor with multilayered rosettes
Ganglioglioma
Retinoblastoma
Ependymoma
Anaplastic pilocytic astrocytoma
ATRT
Glioblastoma
Pilocytic astrocytoma
Subependymoma
Infantile hemispheric glioma
Pineal parenchymal tumor
PXA
Esthesioneuroblastoma
Pineoblastoma
Plexus tumor
Pituitary adenoma
Subependymal giant cell astrocytoma
Hemangioblastoma
Paraganglioma, spinal, non-CIMP
Chordoid glioma of the third ventricle
Pituicytoma
Ewing sarcoma
Schwannoma
Melanocytoma
Meningioma
Craniopharyngioma
Chordoma
Hemangiopericytoma
Plasmacytoma
Lymphoma
c
Institutional and Capper cohorts
Neural signature
High-neural
Low-neural
IDH-wild-type glioblastoma (n = 1,058)
d
TCGA
Neural signature
High-neural
Low-neural
IDH-wild-type glioblastoma (n = 187)
e
Internal cohort
Low-neural glioblastoma (n = 213, 100 events)
High-neural glioblastoma (n = 150, 104 events)
Probability of survival (%)
Median OS: 21.2 months
P < 0.0001
Median OS: 14.2 months
OS (months)
f
Internal cohort
Low-neural glioblastoma (n = 140, 111 events)
High-neural glioblastoma (n = 99, 85 events)
Probability of progression (%)
Median PFS: 10.0 months
P = 0.02
Median PFS: 6.2 months
PFS (months)
g
External TCGA cohort
Low-neural glioblastoma (n = 93, 61 events)
High-neural glioblastoma (n = 94, 56 events)
Probability of survival (%)
Median OS: 17.1 months
P < 0.01
Median OS: 12.0 months
OS (months)
h
OS (n = 363)
High-neural glioblastoma (ref: low) 1.96 (1.45–2.64) <0.01
Age at diagnosis 1.00 (0.98–1.02) 0.95
Near GTR (ref: GTR) 1.14 (0.80–1.63) 0.46
PR (ref: GTR) 1.84 (1.28–2.64) <0.01
Methylated MGMT (ref: non-methylated) 0.47 (0.35–0.64) <0.01
Karnofsky Performance Status prior therapy 0.98 (0.96–0.99) 0.02
OR (95% CI) P value
Odds ratio (log scale) (95% CI)
i
PFS (n = 239)
High-neural glioblastoma (ref: low) 1.51 (1.13–2.02) <0.01
Age at diagnosis 1.00 (0.99–1.02) 0.67
Near GTR (ref: GTR) 1.39 (0.98–1.96) 0.07
PR (ref: GTR) 1.73 (1.19–2.49) <0.01
Methylated MGMT (ref: non-methylated) 0.49 (0.37–0.68) <0.01
Karnofsky Performance Status previous therapy 0.99 (0.98–1.01) 0.91
OR (95% CI) P value
Odds ratio (log scale) (95% CI)

https://doi.org/10.1038/s41591-024-02969-w

staining demonstrated a higher fraction of OLIG2-positive tumor cells in high-neural glioblastoma samples but comparable sparse infiltration of NeuN$^+$ cells within the tumor samples (Extended Data Fig. 4g,h).

Next, we leveraged spatially resolved transcriptomic data with paired methylation profiling ($n = 24$) to examine the molecular architecture and cell-type distribution in low- and high-neural glioblastoma

Fig. 2 | Integrated epigenetic and transcriptomic analysis reveals synaptic functions and a malignant NPC/OPC-like character in high-neural glioblastoma. a, Illustration of the workflow to integrate epigenetic and transcriptional data. Gene co-regulation networks are correlated to the epigenetic deconvolution signature. **b**, Hierarchical dendrogram of the gene expression modules derived from the weighted correlation network analysis. Dot-plot of the neural signature with gene expression models using Pearson correlation (bottom). Size and color indicate the correlation coefficient, nonsignificant correlation is marked. **c**, Bar-plot of the differential gene expression of module eigengenes (log_2-transformed fold change) in low- and high-neural glioblastoma (cutoff 0.41). **d**, Dimensional reduction (UMAP) of the gene expression modules (named by colors). **e**, A detailed visualization of the modules: green, cyan and midnight blue (significantly associated with high-neural tumors). **f**, Gene Ontology analysis of gene expression modules in low- and high-neural tumors. **g**, UMAP dimensional reduction of the GBMap reference dataset. Colors indicate the different cell types. **h**, Module eigengene expression of low- and high-neural glioblastoma in the GBMap reference dataset. **i**, Gene expression enrichment of low- and high-neural-associated module eigengenes across glioblastoma cell states. AC, astrocytes; DC, dendritic cells; GBM, glioblastoma; NK, natural killer; OGD, oligodendrocytes; TAM, tumor-associated macrophages.

samples (Fig. 3). We hypothesized that these tumors have distinct architectures, reflected by a unique spatial arrangement of transcripts that predict their epigenetic neural subgroup.

To this end, we trained a graph-neural network (GNN) using 1,000 randomly chosen microenvironments within the samples. Each microenvironment was centered on a 55-µm spot and extended up to 450 µm. These subgraphs were representative of the broader sample and were instrumental for the GNN training, achieving an R^2 of 0.99 and an F1 score of 0.98, indicating that the neural score can be reliably predicted from the transcriptional landscape (Fig. 3a,b).

We applied our neural score threshold of 0.41 to categorize microenvironments as 'neural high' or 'neural low'. Of note, 41.2% of the samples exhibited a blend of both categories, including those at the threshold and those with the most elevated neural scores (Fig. 3c). For instance, a sample with a neural score of 0.58 showed two prominent peaks at 0.38 and 0.58, suggesting a diverse microenvironmental composition (Fig. 3d); however, a pure or predominant neural type was present in all but one of the 24 samples (95.8%). Further analysis revealed that high-neural score microenvironments typically encompass NPC-like and astrocyte-like tumor cells (Fig. 3e), alongside a significant presence of oligodendrocytes and OPC-like cells, painting a picture of the tumor microenvironment's unique architecture associated with the high-neural phenotype.

In conclusion, single-cell and spatially resolved transcriptomic analyses decipher that the neural signature in glioblastomas predominantly originates from cells of the neural lineage exhibiting an OPC/NPC/astrocyte-like phenotype and is characterized by a distinct tumor microenvironment.

High-neural glioblastomas resemble a malignant stem cell-like state

Using a nonreference-based multi-dimensional single-cell deconvolution algorithm, we observed a higher stem/progenitor cell-like state but lower immune component in high-neural glioblastoma (28.05%) compared to all newly diagnosed glioblastoma (17.31%) and low-neural glioblastoma (14.14%) (Extended Data Fig. 4i). Both components were significantly correlated with the neural signature (Extended Data Fig. 4j,k).

No significant copy-number variations were observed between low- and high-neural subgroups (conumee R package v.1.28.0)[26,27] (Extended Data Fig. 5a). Next-generation sequencing (NGS) of 201 genes showed a higher frequency of *PIK3CA* (0 out of 65 (0.0%) versus 9 out of 60 (15.0%)) and *TP53* (6 out of 65 (9.23%) versus 19 out of 60 (31.67%)) mutations in high-neural tumors (Extended Data Fig. 5b,c). These findings were confirmed by an analysis of paired epigenetic and sequencing data of the TCGA dataset (Extended Data Fig. 5d,e).

High-neural glioblastomas integrate into neuron-to-glioma networks

The transcriptional and proteomic analysis revealed an increased synaptogenic character in high-neural glioblastomas. This led us to explore their integration into neuron-to-glioma networks. After xenografting, an increased colocalization of neuron-to-glioma synapse puncta ($P < 0.01$; Fig. 4a–c) was observed in high-neural glioblastoma which was proven using electron microscopy ($P = 0.008$; Fig. 4d). An increase of colocalization of synapse puncta in high-neural glioblastoma cells after co-culturing with cortical neurons was found ($P < 0.001$; Fig. 4e).

For clinical translation, we assessed functional tumor connectivity using magnetoencephalography ($n = 38$; Fig. 4f,g) and resting-state functional magnetic resonance imaging ($n = 44$; Fig. 4h–k) in patients with glioblastoma. Both modalities showed a significantly higher peritumoral connectivity within the high-neural subgroup ($P < 0.01$; Fig. 4f–i). This aligns with recent studies on cellular states in regions of HFC-glioblastoma[12]. Comparing the connectivity phenotype[12] to our neural classification showed high concordance (Fig. 4g); however, no increased connectivity was seen between the tumor region and the contralateral hemisphere (Fig. 4j). Volumetric analysis showed significantly smaller volumes of contrast enhancement ($P = 0.03$; Extended Data Fig. 6a) in high-neural glioblastoma, but no association with fluid-attenuated inversion recovery (FLAIR) ($P = 0.18$; Extended Data Fig. 6b) and necrotic volume ($P = 0.78$; Extended Data Fig. 6c). These findings indicate that high-neural glioblastomas engender neuron-to-glioma synaptogenesis and have a distinct role within neuron-to-glioma networks exhibiting functional connectivity.

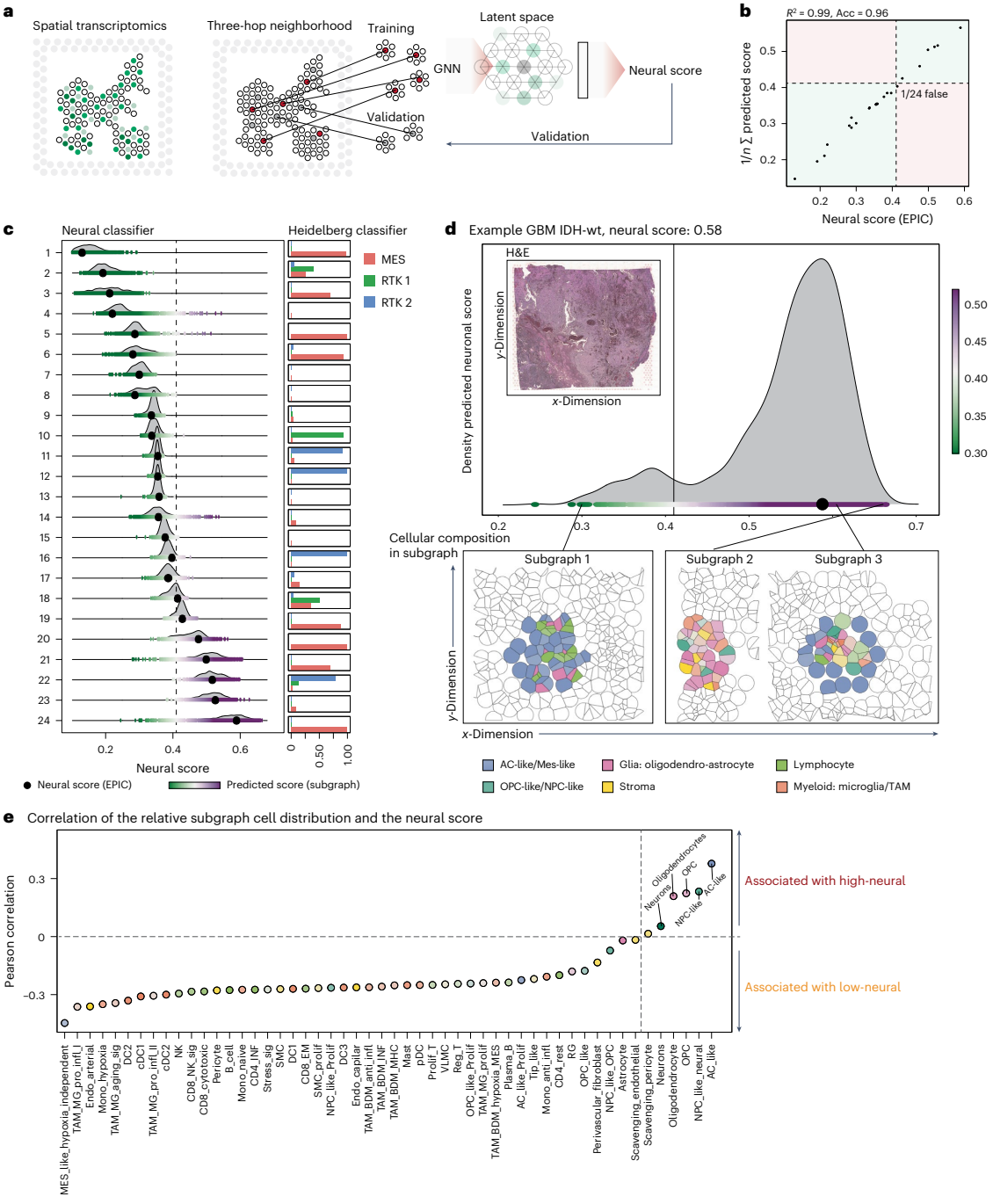
Epigenetic neural signature is transferable to in vivo and in vitro models

Most studies elucidating the biology of cancer neuroscience in high-grade glioma were performed in preclinical models. Therefore, we examined the translatability of our epigenetic neural signature in cell cultures and patient-derived xenograft (PDX) models. We observed a well-preserved neural signature in 82.3% of our cell cultures compared to the original tumor samples (Fig. 5a), confirming that our preclinical models sufficiently reflect the characteristics of the original tumor. Comparison of low- and high-neural glioblastoma in PDX models of an internal cohort ($n = 30$ mice of seven patient-derived glioblastoma cell cultures; Fig. 5b) and two publicly available cohorts[28,29] ($n = 96$

Fig. 3 | Spatially resolved architecture of low- and high-neural glioblastoma. **a**, Illustration of the workflow. Spatial transcriptomic data were used to identify neighborhoods defined as subgraphs. A GNN was trained to predict the neural score based on the spatial arrangements of transcripts. **b**, Scatter-plot of the mean sample predictions and the ground truth values. **c**, Illustration of the variance of neural score (predictions) compared to the threshold of 0.41. Bar plot indicates the Heidelberg classifier values of the glioblastoma subclasses ($n = 24$) (right). The dashed black line indicates the neural score threshold of 0.41. **d**, Example of a high-neural glioblastoma sample with a large blend of low- and high-neural predicted scores. The hematoxylin and eosin (H&E) image demonstrate the histology of the sample. Spatial neighborhoods derived from subgraphs with high- and low-neural scores are demonstrated (bottom). The single-cell maps are generated through single-cell deconvolution (Cell2Location) and CytoSpace spatial deconvolution. wt, wild type. **e**, Overview of the cell-type abundance correlated with the neural score.

https://doi.org/10.1038/s41591-024-02969-w

patient-derived glioblastoma cell cultures; Fig. 5c) showed a significantly shorter survival of mice bearing high-neural tumors (internal cohort, $P = 0.0009$; external cohort, $P = 0.001$). Additionally, an increased proliferation index was seen in high-neural glioblastoma in vivo using immunodeficient mice ($P < 0.01$; Fig. 5d–f) as well as in co-cultures with cortical neurons ($P < 0.001$; Fig. 5g,h). In accordance with current literature describing neuronal activity-driven widespread infiltration of glioblastoma cells[14], we observed a significantly wider



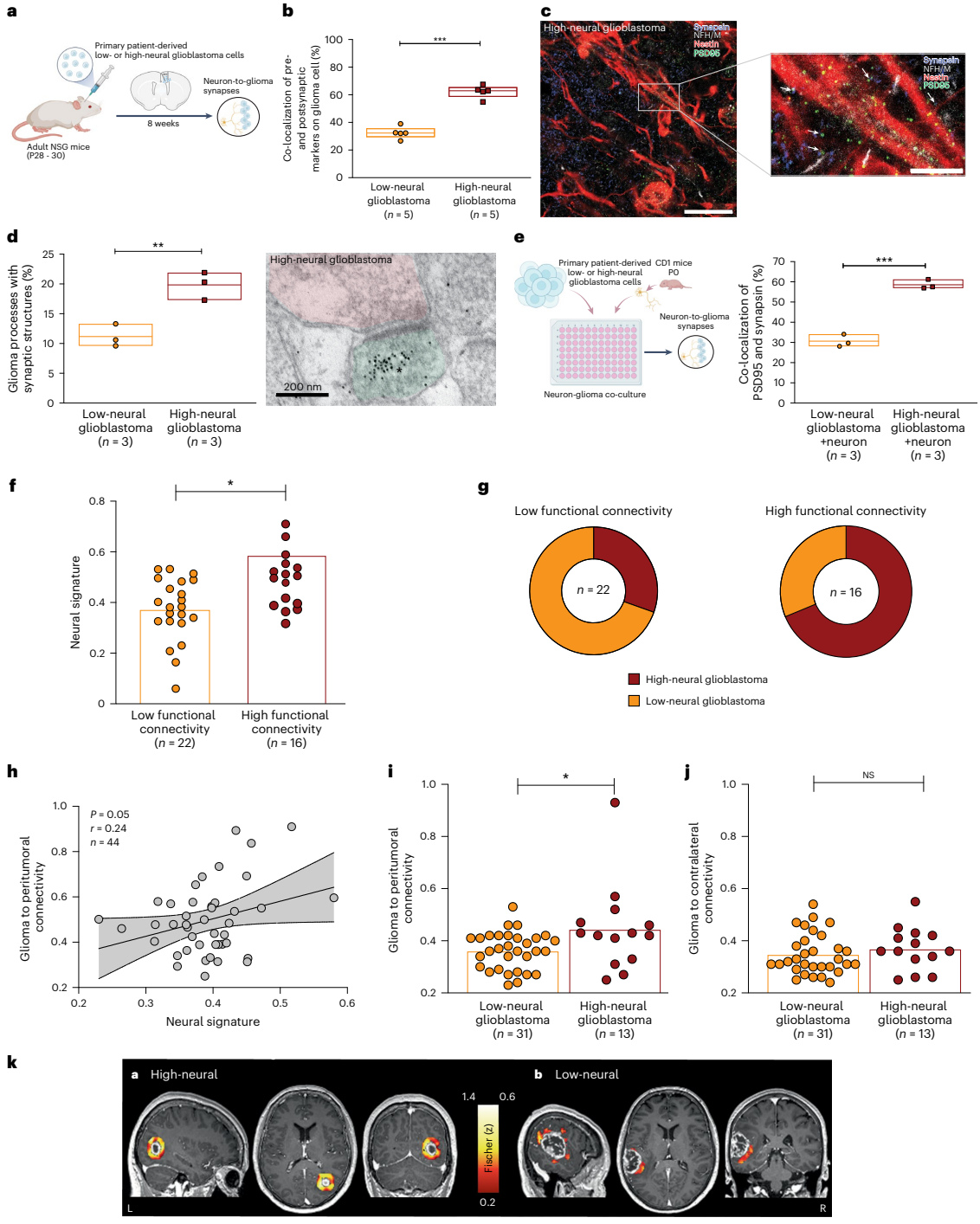

a
Primary patient-derived low- or high-neural glioblastoma cells
Adult NSG mice (P28 - 30)
8 weeks
Neuron-to-glioma synapses
b
Co-localization of pre- and postsynaptic markers on glioma cell (%)
***
Low-neural glioblastoma (n = 5)
High-neural glioblastoma (n = 5)
c
High-neural glioblastoma
Synapsin
NFH/M
Nestin
PSD95
d
Glioma processes with synaptic structures (%)
**
High-neural glioblastoma
200 nm
Low-neural glioblastoma (n = 3)
High-neural glioblastoma (n = 3)
e
Primary patient-derived low- or high-neural glioblastoma cells
CD1 mice P0
Neuron-to-glioma synapses
Neuron-glioma co-culture
Co-localization of PSD95 and synapsin (%)
***
Low-neural glioblastoma +neuron (n = 3)
High-neural glioblastoma +neuron (n = 3)
f
Neural signature
*
Low functional connectivity (n = 22)
High functional connectivity (n = 16)
g
Low functional connectivity
High functional connectivity
n = 22
n = 16
High-neural glioblastoma
Low-neural glioblastoma
h
P = 0.05
r = 0.24
n = 44
Glioma to peritumoral connectivity
Neural signature
i
Glioma to peritumoral connectivity
*
Low-neural glioblastoma (n = 31)
High-neural glioblastoma (n = 13)
j
Glioma to contralateral connectivity
NS
Low-neural glioblastoma (n = 31)
High-neural glioblastoma (n = 13)
k
a High-neural
b Low-neural
Fischer (z)
1.4
0.6
0.2
L
R

Fig. 4 | High-neural glioblastomas are integrated into neuron-to-glioma networks. a, Experimental workflow. **b**, Quantification of the colocalization of presynaptic and postsynaptic markers in low-neural ($n = 22$ regions, five mice) and high-neural ($n = 21$ regions, five mice) glioblastoma xenografts. $P = 0.0008$, two-tailed Student's t-test. Data are mean ± s.e.m. **c**, Confocal image of infiltrated whiter matter of high-neural glioblastoma xenograft. White box and arrowheads highlight magnified view of synaptic puncta colocalization. Blue, synapsin-1 (presynaptic puncta); white, neurofilament heavy and medium (axon); red, nestin (glioma cell processes); green, PSD95 (postsynaptic puncta). Scale bars, 500 µm (top) and 250 µm (bottom). **d**, Electron microscopy of red fluorescent protein (RFP)-labeled glioblastoma cells. Quantification of neuron-to-glioma synaptic structures as a percentage of all visualized glioma cell processes (left) and image of neuron-to-glioma process in a high-neural glioblastoma xenograft (right). Asterix denotes immunogold particle labeling of RFP. Postsynaptic density in RFP$^+$ tumor cell (green), synaptic cleft and vesicles in presynaptic neuron (red) identify synapses. $^{**}P < 0.01$, two-tailed Student's t-test. Scale bar, 200 nm. Data are mean ± s.e.m. $n = 3$ biological replicates. **e**, Colocalization of PSD95 and synapsin-1 in low- and high-neural glioblastoma cells in co-cultures with neurons. $P = 0.0007$, not significant (NS), $P > 0.05$, two-tailed Student's t-test, $n = 3$ biological replicates. Data are mean ± s.e.m. **f**, Neural signature categorized into low functional connectivity (LFC) and high functional connectivity (HFC) as defined by magnetoencephalography. $P = 0.0327$, two-tailed Student's t-test. **g**, Overlap between samples classified to the functional connectivity by Krishna et al.[12] and the epigenetic-based neural classification of our study. **h**, Correlation of neural signature with degree of peritumoral connectivity as defined by resting-state functional magnetic resonance imaging (rs-fMRI). Simple linear regression $P = 0.05$, error bands representing the 95% CI. **i**, Peritumoral functional connectivity (defined by rs-fMRI) in low- and high-neural glioblastoma. $P = 0.0416$, two-sided Mann–Whitney U-test. **j**, Functional connectivity to the contralateral hemisphere (defined by rs-fMRI) in low- and high-neural glioblastoma groups. NS, $P > 0.05$, two-sided Mann–Whitney U-test. **k**, Examples showing the region of interest (ROI)-to-voxel functional connectivity of the contrast-enhancing area to its peritumoral surrounding. Peritumoral connectivity of a high-neural glioblastoma (0.457) and mean functional connectivity to its peritumoral area of 0.837 (left). By contrast, a low-neural glioblastoma (0.347) is shown with mean functional connectivity to its peritumoral area of 0.294 (right).

migration of high-neural glioblastoma cells in vitro ($P < 0.05$; Fig. 5i,j) and in vivo ($P < 0.001$; Fig. 5k). These findings demonstrate the robustness of the epigenetic neural signature in vitro and in vivo and indicate higher proliferation when receiving neuronal input.

Epigenetic neural classification remains spatiotemporally stable

As heterogeneity is a hallmark of glioblastoma, we investigated the spatiotemporal heterogeneity of the epigenetic neural signature. First, we analyzed 143 spatially collected biopsies from 34 patients (3–7 samples per patient). Among them, 23 patients (67.6%) demonstrated a pure low- or high-neural signature, while ten patients (29.4%) exhibited a predominant signature (Extended Data Fig. 6d). Temporal stability was assessed in 39 patients with matched tissue from both initial and recurrence surgery (Extended Data Fig. 6e). Here, 31 out of 39 patients (79.5%) remained in the same neural subgroup at recurrence (Extended Data Fig. 6f). Overall, the neural subgroup seemed to be spatiotemporally stable in contrast to transcriptional states that change in a larger proportion of patients[30,31].

Drug sensitivity analysis of neural glioblastoma cells

Patients with glioblastoma routinely undergo combined radiochemotherapy after surgical resection[32]. We evaluated 27 different agents for their efficacy in the treatment of low- and high-neural glioblastoma cells (Extended Data Fig. 7a). We observed a trend for increased cleaved caspase 3 (Extended Data Fig. 7b) and reduced tumor cell size (Extended Data Fig. 7c) after treatment with lomustine (CCNU), JNJ10198400 and cyclosporine-treated high-neural glioblastoma cells, whereas talazoparib showed a trend for greater sensitivity in low-neural glioblastoma cells; however, none of these compounds reached statistical significance (Extended Data Fig. 7d). Therefore, we wondered about the prognostic impact of surgical resection as we previously demonstrated survival differences for other methylation-based glioblastoma subclasses[33].

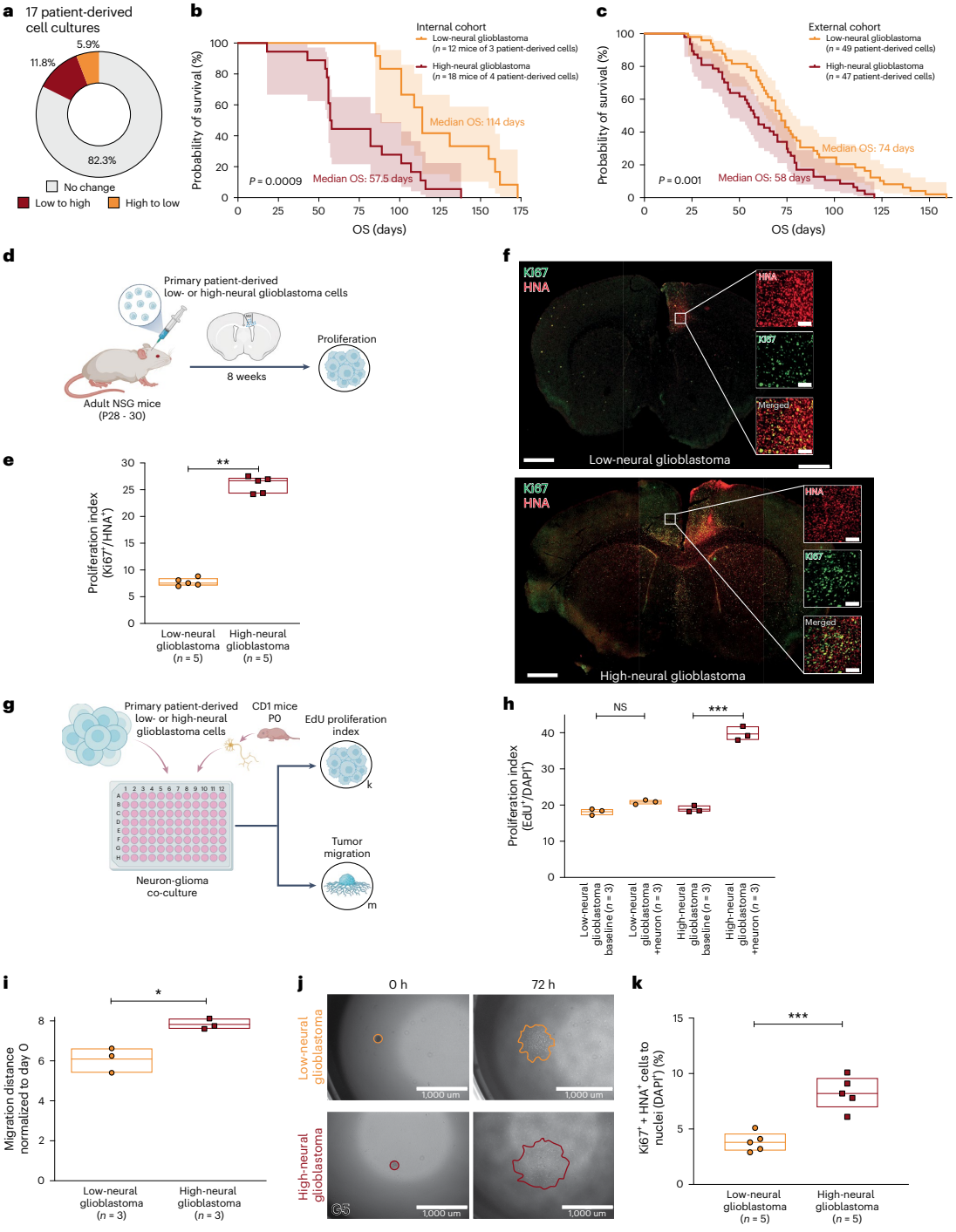
Neural classification predicts benefit of resection

Glioblastomas are epigenetically assigned to different subclasses[34]. Here, RTK I and RTK II (receptor tyrosine kinase I and II subtypes) tumors showed a comparable high-neural signature, whereas mesenchymal (MES) tumors had the lowest neural signature (Extended Data Fig. 7a). Given the different neural signatures between methylation-based subclasses, we hypothesized that the neural signature might constitute a factor for determining benefit from extent of resection (EOR). In low-neural glioblastoma, a significant survival benefit of gross total resection (GTR) (100% CE resection) and near GTR (≥90% CE resection) was observed compared to partial resection (<90% CE resection) ($P < 0.001$; Fig. 6a). By contrast, the survival benefit of a near GTR was not seen in high-neural glioblastoma (Fig. 6b). These findings held true in multivariate analyses (Extended Data Fig. 8b,c) and after applying the current criteria of the RANO (Response Assessment in Neuro-Oncology) resect group[35] (Extended Data Fig. 8d,e). A methylated *MGMT* promoter showed a survival benefit in both neural subgroups, but a striking difference in low-neural glioblastoma with a median overall survival difference of 12.0 months depending on the *MGMT* promoter methylation status ($P < 0.0001$; Fig. 6c). Our combined survival data demonstrate that high-neural glioblastomas have an unfavorable outcome and a greater resection may be required to achieve a survival benefit in this distinct subclass.

Fig. 5 | Neural classification is conserved in cell culture and correlates with survival as well as proliferation. a, Comparison of neural signature between patient's tumor tissue and cell culture in 17 glioblastomas. **b**,**c**, Survival after xenografting of patient-derived low- and high-neural glioblastoma cells in our internal cohort (**b**) and two combined external cohorts (**c**). log-rank test, $P = 0.0009$ (**b**), $P = 0.001$ (**c**). Error bands represent 95% CI. **d**, Primary patient-derived low- and high-neural glioblastoma cell suspensions ($n = 1$ per group) were implanted into premotor cortex (M2) of adult NSG mice ($n = 5$ mice per group). Mice were perfused after 8 weeks of tumor growth and brains sectioned in the coronal plane for further immunofluorescence analyses. **e**, Proliferation index (measured by total number of HNA$^+$ cells co-labeled with Ki67 divided by the total number of HNA$^+$ tumor cells counted across all areas quantified) in low- and high-neural glioblastoma-bearing mice ($n = 5$ mice per group). $P = 0.00819$, two-tailed Student's t-test. Data are mean ± s.e.m. **f**, Representative confocal images of proliferation index in low-neural (top) and high-neural glioblastoma (bottom) xenografts. Human nuclear antigen (HNA), red; Ki67, green. Scale bars, 1 µm (overview images) and 200 µm (magnified images). **g**, Experimental workflow. **h**, EdU proliferation index (measured by total number of DAPI$^+$ cells co-labeled with EdU divided by the total number of DAPI$^+$ tumor cells counted across all areas quantified) in low-neural ($P = 0.418$) and high-neural ($P = 0.0000172$) glioblastoma as monocultures and co-cultured with neurons. Two-tailed Student's t-test, $n = 3$ biological replicates. Data are mean ± s.e.m. **i**,**j**, 3D migration assay analysis comparing distance of migration 72 h after seeding (**i**) and representative images at time 0 h (left) and 72 h (right) of low- and high-neural glioblastoma cells (**j**). $P = 0.0115$, two-tailed Student's t-test, $n = 3$ biological replicates. Scale bars, 1 µm. Data are mean ± s.e.m. **k**, In vivo spread of tumor cells into corpus callosum in low- and high-neural glioblastoma. $P < 0.0004$, two-tailed Student's t-test. Data are mean ± s.e.m. EdU, 5-ethynyl-2′-deoxyuridine; DAPI, 4,6-diamidino-2-phenylindole.

Serum biomarkers of neural glioblastoma

Next, we examined the feasibility of preoperatively determining the epigenetic neural subclassification in the blood of patients with glioblastoma to further reach clinical translation. By analyzing serum levels of brain-derived neurotrophic factor (BDNF) in 94 patients at diagnosis, we found higher BDNF levels in high-neural glioblastoma



compared to low-neural glioblastoma, patients with meningioma ($n = 13$) and healthy individuals ($n = 19$) (Fig. 6d). The serum BDNF levels positively correlated with the epigenetic neural signature ($P < 0.01$, $R^2 = 0.28$; Fig. 6e). Conversely, glioblastomas with higher BDNF serum levels had a decreased immune cell signature (Fig. 6f), consistent with the lower immune cell signature of high-neural tissue samples. We observed elevated BDNF levels in patients with glioma-associated seizures at the time of diagnosis ($P = 0.02$; Fig. 6g) and during follow-up ($P < 0.001$; Fig. 6h), which aligns with the known activity-regulated release of BDNF, most likely from healthy neurons (Fig. 6i,j) within high-neural glioblastoma networks.

Furthermore, we identified the neural signature in circulating extracellular vesicle-associated DNA (EV-DNA) and cell-free DNA (cfDNA) in patients' plasma (Extended Data Fig. 8f–i). Circulating extracellular vesicles, a surrogate marker for glioblastoma[36,37] and involved in neuronal synchronization[38], correlated with the neural signature (Extended Data Fig. 8f). Epigenetic profiling of EV-DNA in plasma revealed a neural signature that was absent in cfDNA (Extended Data Fig. 8g). The neural signature detected in EV-DNA exhibited a significant increase in glioblastoma compared to samples from healthy donors and patients with meningioma (Extended Data Fig. 8g). Notably, high-neural tumors showed a higher incidence of a detectable neural signature in circulating EV-DNA (Extended Data Fig. 8h). While plasma-derived EV-DNA displayed markedly lower levels of neural signatures, cerebrospinal fluid EV-DNA exhibited lower but more comparable levels to tissue scores (Extended Data Fig. 8i).

Our findings suggest that BDNF could assist in stratifying patients with glioblastoma based on their neural subgroup, potentially facilitating targeted therapy in the future and that the neural signature is detectable in circulating extracellular vesicles.

Epigenetic neural classification informs survival in diffuse midline glioma

Besides glioblastoma, previous studies have highlighted the importance of neuronal activity-driven proliferation in DMG[6,7]. We identified the epigenetic neural signature in a cohort of H3 K27-altered DMG consisting of pediatric and adolescent patients from our institutional cohort ($n = 21$), Chen et al.[39] ($n = 24$) and Sturm et al.[34] ($n = 10$). The neural signature was evenly distributed among tumors in the thalamus, pons and medulla (Extended Data Fig. 9a). Similar to glioblastomas, areas in genes related to trans-synaptic signaling were mainly hypomethylated in high-neural DMGs (Extended Data Fig. 9b). A notable association with stem and glial cell states (Extended Data Fig. 9c) and increased synaptic gene expression[4] ($P = 0.01$; Extended Data Fig. 9d) was observed in high-neural DMGs. Survival analysis of 72 patients showed an unfavorable outcome for high-neural DMG ($P < 0.01$; Extended Data Fig. 9e–g). These results confirm the relevance of the neural signature in an additional type of IDH-wild-type high-grade glioma.

Discussion

In recent years, the bidirectional interaction between glioma cells and neural cells, with their ability to form synapses and integrate into neuronal networks, has been identified as a major factor in tumor progression[4,6,13,40]. In this study, we identified an epigenetically defined malignant neural signature as a potential marker for neural-to-glioma interactions and present the following findings: (1) A malignant neural signature is increased in glioblastoma and DMG, compared to nonmalignant brain tumors. (2) High-neural glioblastoma confers an unfavorable survival in humans and mice, and in addition, the neural signature is associated with higher functional connectivity in patients with glioblastoma. (3) High-neural glioblastoma shows an increased malignant stem cell and neural lineage character but decreased immune infiltration. (4) The neural signature remains robust in vitro and in vivo and high-neural glioblastoma-bearing mice show higher proliferation when receiving neuronal input as well as increased neuron-to-glioma synapse formation. (5) High-neural tumors benefit from a maximized resection. (6) Elevated BDNF serum levels are present in patients with high-neural glioblastoma. (7) The prognostic value can also be seen in H3K27-altered DMG.

Gliomas encompass a variety of cellular components of the tumor microenvironment and subgroups can be described according to distinct cellular states[15]. Epigenome profiling and deconvolution have been effective in characterizing these glioma subclasses[41,42]. A recent study highlighted the importance of epigenetic regulation across various cancer types and demonstrated a close epigenomic relationship between glioblastoma cells and OPCs[43]. Our determination of an epigenetic neural signature revealed an increase in glioblastoma and DMG, echoing findings of previous studies in preclinical models[4,7]. Nonetheless, it is essential to note that the neural signature was derived from a single cortical neuron reference generated from three IDAT files, and while we integrated DNA methylation data from healthy brain regions for comparison, a larger sample size might have provided clearer differentiation between low- and high-neural tumors.

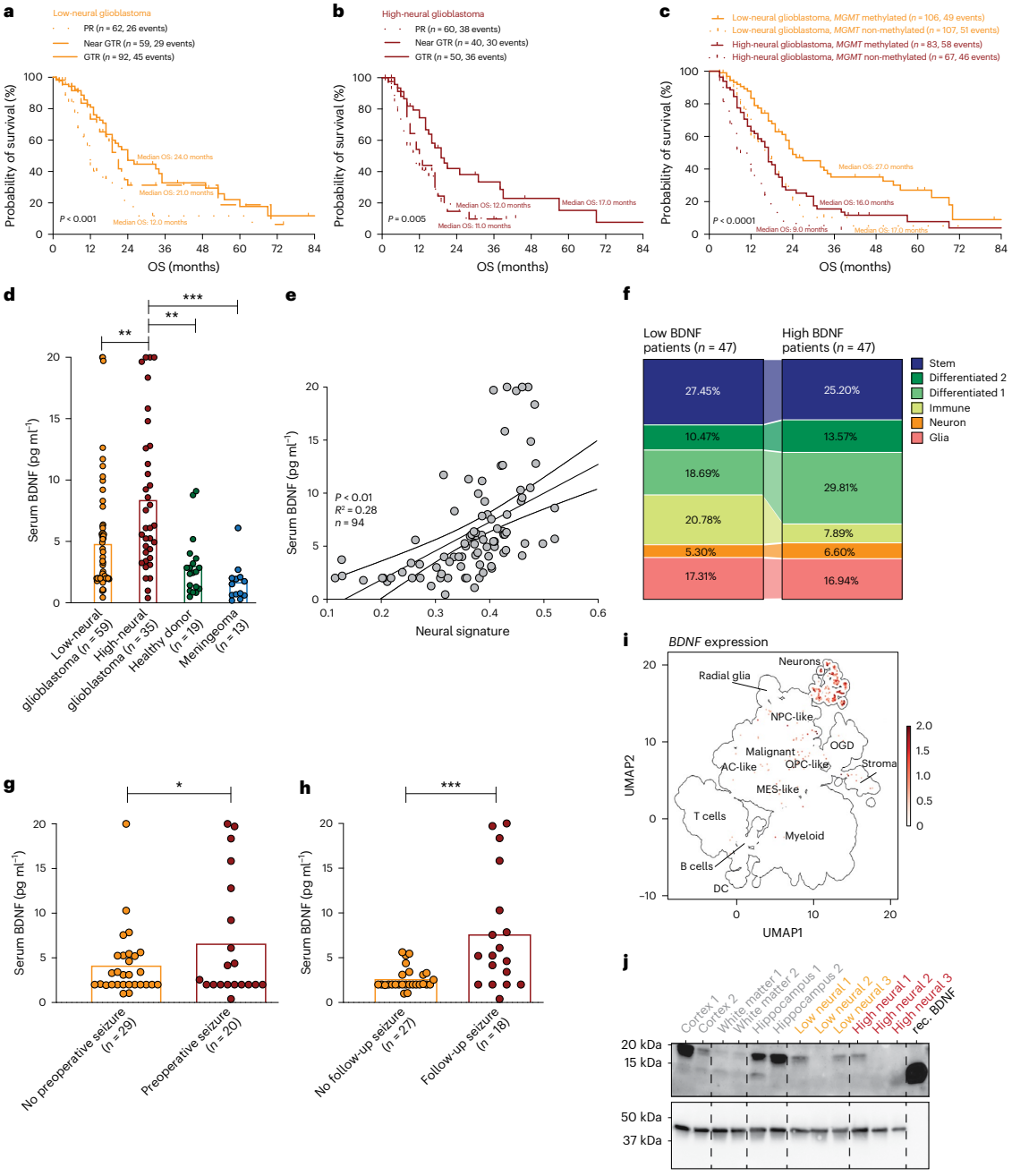
High-neural glioblastoma showed gene upregulation and hypomethylation associated with invasiveness and neuro-glioma synapse formation. Glioma growth is known to involve paracrine signaling and glutamatergic synaptic input[4–8], and recently a study subdivided glioblastoma cells into unconnected and connected cells with unique cell states, explaining brain infiltration through hijacking of neuronal mechanisms[13]. Our spatial transcriptomic analysis has unveiled the malignant stem-cell-like characteristics of high-neural glioblastoma, primarily clustering with cells of the neural lineage, such as OPC/NPC/astrocyte-like cells, alongside healthy oligodendrocytes and neurons. These findings align with the previously described unconnected glioblastoma cells that hijack neuronal mechanisms and drive brain invasion. While tumors with an OPC/NPC-like cellular state have been shown to overlap with the classical and proneural TCGA subtypes[15], which have been assumed as having a better prognosis[25], our identified high-neural glioblastoma demonstrated a poor patient outcome. This possible discrepancy may be explained by our integrated RNA-seq analysis, which revealed a wide heterogeneity of the transcriptomic TCGA subtypes in our epigenetic low- and high-neural tumors. In addition, this difference can largely be attributed to the noted transcriptional heterogeneity and plasticity within tumor populations[15,44]. Our study posits that the epigenetic signature offers a more stable marker than purely transcriptional profiles. Unlike the transient nature of transcriptional states, epigenetic signatures encompass not only the cells in OPC/NPC/astrocyte-like states but also reflect broader dependencies and interactions within the tumor microenvironment. Therefore, we argue that our high-neural phenotype should be interpreted as

Fig. 6 | Neural classification predicts benefit of EOR and *MGMT* promoter methylation status and can be detected in serum of patients with glioblastoma. a,**b**, Survival outcome categorized after EOR in patients with glioblastoma treated by radiochemotherapy with a low-neural (**a**) and high-neural (**b**) tumor. log-rank test, $P = 0.0003$ (**a**), $P = 0.005$ (**b**). Error bands represent 95% CI. **c**, Survival outcome categorized by *MGMT* promoter methylation status in patients with glioblastoma treated by radiochemotherapy with a low- and high-neural tumor. log-rank test, $P = 2.719 \times 10^{-11}$. Error bands represent 95% CI. **d**,**e**, Immunoassay quantification of serum BDNF concentration of 94 patients with glioblastoma and healthy donors as well as patients with meningioma as control groups at the time of diagnosis. $^{**}P < 0.01$, $^{***}P < 0.001$, two-tailed Student's *t*-test; error bands represent 95% CI. **f**, Cell composition analysis in glioblastoma with low and high BDNF serum levels. **g**,**h**, Seizure outcome of patients with glioblastoma considering BDNF serum levels at the time of surgery (**g**) and during follow-up (**h**). $^{*}P < 0.05$, $^{***}P < 0.001$, two-tailed Student's *t*-test. **i**, Transcriptomic analysis of BDNF expression. **j**, Western blotting of BDNF in various healthy brain tissue samples and low- as well as high-neural glioblastoma. $n = 3$ biological replicates.

being driven by epigenetic factors that incline cells toward OPC/NPC/astrocyte-like states, rather than solely being a direct consequence of transcriptional variability.

Of note, the observed diploid oligodendrocyte transcriptomic module may represent a tumor cell population of primary near-diploid state as glioblastomas are karyotypically heterogeneous tumors[45–47]. Alternatively, it might be possible that surrounding healthy oligodendrocytes are affecting the neuronal activity-driven mechanisms on glioma cells[2].

The clinical relevance of our findings is supported by the observation that patients suffering from high-neural glioblastoma or DMG had an unfavorable outcome. A greater EOR must be achieved to have



prognostic improvement in high-neural glioblastoma, which may explain the results of our previous study examining the impact of DNA methylation subclasses[33]. Our findings are in line with a recent study by Krishna et al.[12] demonstrating poorer survival in patients with glioblastoma exhibiting high functional connectivity. Integrating connectivity data from resting-state functional MRI and magnetoencephalography (MEG) linked an increased functional connectivity to its peritumoral surrounding with a higher neural signature in our patients. If a reliable stratification of the neural classification by MEG or MRI is predictable remains to be discussed in further studies. The synaptogenic character with increased connectivity of high-neural glioblastomas could be replicated with in vivo and in vitro experiments. Collectively, these data underscore the tremendous importance of the synaptic integration of gliomas into neuronal circuits and targeting these neuron-to-glioma networks seems to be a promising therapeutic approach[1,48].

One factor drawing attention is BDNF, a neuronal activity-regulated neurotrophin, which has been found to promote glioma growth[6,49] and interrupting BDNF–TrkB signaling has been shown to confer survival benefit in mice[5]. We found elevated serum BDNF levels in patients with high-neural glioblastoma and further correlation with increased seizure frequency. Potential sources of elevated BDNF include neurons in a glioma-induced state of hyperexcitability[4], given the known activity regulation of BDNF secretion[50–52] or possibly from glioblastoma cells[53]. In brief, neuronal activity arising from glioma-to-neuron interactions during tumor growth or seizure initiation seems to be a pivotal driver for BDNF release and identifies a potential biomarker of high-neural glioblastoma.

While the BDNF–TrkB axis may represent a therapeutic target for high-neural glioblastoma, we further identified low-neural tumors as immune-enriched based on transcriptomic and cell state composition analysis. Consequently, one could hypothesize that two opposing glioblastoma subtypes seem to be differentiated here and will need to be pursued in future studies and therapeutic avenues. The identification of an immunosuppressive state in high-neural glioblastoma is concordant with recent findings which described immunosuppressive mechanisms in thrombospondin-1-upregulated glioma samples[54]. This stratification of IDH-wild-type gliomas based on their epigenetic neural signature may provide a potential tool for predicting response to neuroscience-guided therapies.

Conclusion

Overall, the definition of a high-neural signature in IDH-wild-type glioma revealed a malignant NPC/OPC/astrocyte-like character that affects patient survival, remains stable during therapy and is conserved in preclinical models. This knowledge supports clinicians in stratifying patients with glioma according to their prognosis and determining the surgical and neuro-oncological benefit for current standard of care. Last, the here-presented clinical translation in the field of glioma neuroscience using an epigenetic neural signature may advance the development of trials with neuroscience-guided therapies.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-024-02969-w.

References

1. Winkler, F. et al. Cancer neuroscience: state of the field, emerging directions. *Cell* **186**, 1689–1707 (2023).
2. Taylor, K. R. & Monje, M. Neuron-oligodendroglial interactions in health and malignant disease. *Nat. Rev. Neurosci.* **4**, 733–746 (2023).
3. Monje, M. Synaptic communication in brain cancer. *Cancer Res.* **80**, 2979–2982 (2020).
4. Venkatesh, H. S. et al. Electrical and synaptic integration of glioma into neural circuits. *Nature* **573**, 539–545 (2019).
5. Taylor, K. R. et al. Glioma synapses recruit mechanisms of adaptive plasticity. *Nature* **623**, 366–374 (2023).
6. Venkatesh, H. S. et al. Neuronal activity promotes glioma growth through neuroligin-3 secretion. *Cell* **161**, 803–816 (2015).
7. Venkatesh, H. S. et al. Targeting neuronal activity-regulated neuroligin-3 dependency in high-grade glioma. *Nature* **549**, 533–537 (2017).
8. Venkataramani, V. et al. Glutamatergic synaptic input to glioma cells drives brain tumour progression. *Nature* **573**, 532–538 (2019).
9. Campbell, S. L., Buckingham, S. C. & Sontheimer, H. Human glioma cells induce hyperexcitability in cortical networks. *Epilepsia* **53**, 1360–1370 (2012).
10. Campbell, S. L. et al. GABAergic disinhibition and impaired KCC2 cotransporter activity underlie tumor-associated epilepsy. *Glia* **63**, 23–36 (2015).
11. Buckingham, S. C. et al. Glutamate release by primary brain tumors induces epileptic activity. *Nat. Med.* **17**, 1269–1274 (2011).
12. Krishna, S. et al. Glioblastoma remodelling of human neural circuits decreases survival. *Nature* **617**, 599–607 (2023).
13. Venkataramani, V. et al. Glioblastoma hijacks neuronal mechanisms for brain invasion. *Cell* **185**, 2899–2917 (2022).
14. Huang-Hobbs, E. et al. Remote neuronal activity drives glioma progression through SEMA4F. *Nature* **619**, 844–850 (2023).
15. Neftel, C. et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. *Cell* **178**, 835–849 (2019).
16. Mathur, R. et al. Glioblastoma evolution and heterogeneity from a 3D whole-tumor perspective. *Cell* **187**, 446–463 (2024).
17. Moss, J. et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat. Commun.* **9**, 5068 (2018).
18. Capper, D. et al. DNA methylation-based classification of central nervous system tumours. *Nature* **555**, 469–474 (2018).
19. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
20. Venkataramani, V. et al. Disconnecting multicellular networks in brain tumours. *Nat. Rev. Cancer* **22**, 481–491 (2022).
21. Südhof, T. C. Towards an understanding of synapse formation. *Neuron* **100**, 276–293 (2018).
22. Pruim, R. H. R. et al. ICA-AROMA: a robust ICA-based strategy for removing motion artifacts from fMRI data. *NeuroImage* **112**, 267–277 (2015).
23. Ruiz-Moreno, C. et al. Harmonized single-cell landscape, intercellular crosstalk and tumor architecture of glioblastoma. Preprint at *bioRxiv* https://doi.org/10.1101/2022.08.27.505439 (2022).
24. Silverbush, D., Suva, M. & Hovestadt, V. LTBK-08. Inferring cell type and cell state composition in glioblastoma from bulk DNA methylation profiles using multi-omic single-cell analyses. *Neuro-Oncol.* **24**, vii300 (2022).
25. Wang, Q. et al. Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment. *Cancer Cell* **32**, 42–56 (2017).
26. Hovestadt, V. & Zapatka, M. conumee. Enhanced copy-number variation analysis using Illumina DNA methylation arrays. *Bioconductor* https://doi.org/10.18129/b9.bioc.conumee (2017).
27. Verhaak, R. G. W. et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**, 98–110 (2010).

28. Vaubel, R. A. et al. Genomic and phenotypic characterization of a broad panel of patient-derived xenografts reflects the diversity of glioblastoma. *Clin. Cancer Res.* **26**, 1094–1104 (2020).
29. Golebiewska, A. et al. Patient-derived organoids and orthotopic xenografts of primary and recurrent gliomas represent relevant patient avatars for precision oncology. *Acta Neuropathol.* **140**, 919–949 (2020).
30. Varn, F. S. et al. Glioma progression is shaped by genetic evolution and microenvironment interactions. *Cell* **185**, 2184–2199 (2022).
31. Barthel, F. P. et al. Longitudinal molecular trajectories of diffuse glioma in adults. *Nature* **576**, 112–120 (2019).
32. Stupp, R. et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N. Engl. J. Med.* **352**, 987–996 (2005).
33. Drexler, R. et al. DNA methylation subclasses predict the benefit from gross total tumor resection in IDH-wildtype glioblastoma patients. *Neuro-Oncol.* **25**, 315–325 (2022).
34. Sturm, D. et al. Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma. *Cancer Cell* **22**, 425–437 (2012).
35. Karschnia, P. et al. Prognostic validation of a new classification system for extent of resection in glioblastoma: a report of the RANO resect group. *Neuro-Oncol.* **25**, 940–954 (2022).
36. Ricklefs, F. L. et al. Circulating extracellular vesicles as biomarker for diagnosis, prognosis and monitoring in glioblastoma patients. *Neuro. Oncol.* https://doi.org/10.1093/neuonc/noae068 (2024).
37. Ricklefs, F. L. et al. Imaging flow cytometry facilitates multiparametric characterization of extracellular vesicles in malignant brain tumours. *J. Extracell. Vesicles* **8**, 1588555 (2019).
38. Spelat, R. et al. The dual action of glioma-derived exosomes on neuronal activity: synchronization and disruption of synchrony. *Cell Death Dis.* **13**, 705 (2022).
39. Chen, L. H. et al. The integrated genomic and epigenomic landscape of brainstem glioma. *Nat. Commun.* **11**, 3077 (2020).
40. Mancusi, R. & Monje, M. The neuroscience of cancer. *Nature* **618**, 467–479 (2023).
41. Singh, O., Pratt, D. & Aldape, K. Immune cell deconvolution of bulk DNA methylation data reveals an association with methylation class, key somatic alterations, and cell state in glial/glioneuronal tumors. *Acta Neuropathol. Commun.* **9**, 148 (2021).
42. Wu, Y. et al. Glioblastoma epigenome profiling identifies SOX10 as a master regulator of molecular tumour subtype. *Nat. Commun.* **11**, 6434 (2020).
43. Terekhanova, N. V. et al. Epigenetic regulation during cancer transitions across 11 tumour types. *Nature* **623**, 432–441 (2023).
44. Ravi, V. M. et al. Spatially resolved multi-omics deciphers bidirectional tumor-host interdependence in glioblastoma. *Cancer Cell* **40**, 639–655 (2022).
45. Shapiro, J. R. & Shapiro, W. R. The subpopulations and isolated cell types of freshly resected high grade human gliomas: their influence on the tumor's evolutionin vivo and behavior and therapyin vitro. *Cancer Metastasis Rev.* **4**, 107–124 (1985).
46. Hu, L. S. et al. Integrated molecular and multiparametric MRI mapping of high-grade glioma identifies regional biologic signatures. *Nat. Commun.* **14**, 6066 (2023).
47. Gill, B. J. et al. MRI-localized biopsies reveal subtype-specific differences in molecular and cellular composition at the margins of glioblastoma. *Proc. Natl Acad. Sci. USA* **111**, 12550–12555 (2014).
48. Shi, D. D. et al. Therapeutic avenues for cancer neuroscience: translational frontiers and clinical opportunities. *Lancet Oncol.* **23**, e62–e74 (2022).
49. van Kessel, E. et al. Tumor-related molecular determinants of neurocognitive deficits in patients with diffuse glioma. *Neuro-Oncol.* **24**, 1660–1670 (2022).
50. Greenberg, M. E., Xu, B., Lu, B. & Hempstead, B. L. New insights in the biology of BDNF synthesis and release: implications in CNS function. *J. Neurosci.* **29**, 12764–12767 (2009).
51. Tao, X., Finkbeiner, S., Arnold, D. B., Shaywitz, A. J. & Greenberg, M. E. Ca2+ influx regulates BDNF transcription by a CREB family transcription factor-dependent mechanism. *Neuron* **20**, 709–726 (1998).
52. Wrann, C. D. et al. Exercise induces hippocampal BDNF through a PGC-1α/FNDC5 pathway. *Cell Metab.* **18**, 649–659 (2013).
53. Wang, X. et al. Reciprocal signaling between glioblastoma stem cells and differentiated tumor cells promotes malignant progression. *Cell Stem Cell* **22**, 514–528 (2018).
54. Nejo, T. et al. Glioma-neuronal circuit remodeling induces regional immunosuppression. Preprint at *bioRxiv* https://doi.org/10.1101/2023.08.04.548295 (2023).

Richard Drexler[1,2,37], Robin Khatri[3,4,37], Thomas Sauvigny[1], Malte Mohme[1], Cecile L. Maire[1], Alice Ryba[1], Yahya Zghaibeh[1], Lasse Dührsen[1], Amanda Salviano-Silva[1], Katrin Lamszus[1], Manfred Westphal[1], Jens Gempt[1], Annika K. Wefers[5,6], Julia E. Neumann[5,7], Helena Bode[8], Fabian Hausmann[3,4], Tobias B. Huber[9,10], Stefan Bonn[3,4], Kerstin Jütten[11], Daniel Delev[11,12], Katharina J. Weber[13,14,15,16], Patrick N. Harter[13,17], Julia Onken[18], Peter Vajkoczy[18], David Capper[19], Benedikt Wiestler[20], Michael Weller[21,22], Berend Snijder[23], Alicia Buck[21,22], Tobias Weiss[21,22], Pauline C. Göller[24,25], Felix Sahm[24,25], Joelle Aline Menstel[26], David Niklas Zimmer[26], Michael B. Keough[2], Lijun Ni[2], Michelle Monje[2], Dana Silverbush[27,28], Volker Hovestadt[29,30], Mario L. Suvà[29,31], Saritha Krishna[32], Shawn L. Hervey-Jumper[32], Ulrich Schüller[5,8,33], Dieter H. Heiland[12,26,34,35,36,38], Sonja Hänzelmann[3,4,9,10,38] & Franz L. Ricklefs[1,38] ✉

[1]Department of Neurosurgery, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [2]Department of Neurology, Stanford University, Stanford, CA, USA. [3]Institute of Medical Systems Biology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [4]Center for Biomedical AI, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [5]Institute of Neuropathology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [6]Mildred Scheel Cancer Career Center HaTriCS4, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [7]Center for Molecular Neurobiology Hamburg (ZMNH), University Hospital Hamburg Eppendorf, Hamburg, Germany. [8]Research Institute Children's Cancer Center Hamburg, Hamburg, Germany. [9]III. Department of Medicine, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [10]Hamburg Center for Kidney Health (HCKH), University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [11]Department of Neurosurgery, University Hospital Aachen, Aachen, Germany. [12]Department of Neurosurgery, University Clinic Erlangen, Friedrich-Alexander Universität Erlangen-Nürnberg, Erlangen, Germany. [13]Neurological Institute (Edinger Institute), University Hospital Frankfurt, Frankfurt am Main, Germany. [14]German Cancer Consortium (DKTK), Heidelberg, Germany and German Cancer Research Center (DKFZ), Heidelberg, Germany. [15]Frankfurt Cancer Institute (FCI), Frankfurt am Main, Germany. [16]University Cancer Center (UCT) Frankfurt, Frankfurt am Main, Germany. [17]Institute of Neuropathology, Faculty of Medicine, LMU Munich, Munich, Germany. [18]Department of Neurosurgery, Charité - Universitätsmedizin Berlin, Berlin, Germany. [19]Department of Neuropathology, Charité - Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany. [20]Department of Neuroradiology, Klinikum rechts der Isar, School of Medicine, Technical University Munich, Munich, Germany. [21]Department of Neurology, Clinical Neuroscience Center, University Hospital Zurich, Zurich, Switzerland. [22]Department of Neurology, University of Zürich, Zurich, Switzerland. [23]Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland. [24]Hopp Children's Cancer Center Heidelberg (KiTZ), Heidelberg, Germany. [25]Department of Neuropathology, University Hospital Heidelberg, Heidelberg, Germany. [26]Department of Neurosurgery, Medical Center University of Freiburg, Freiburg, Germany. [27]Department of Cancer Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [28]Abramson Family Cancer Research Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [29]Broad Institute of Harvard and MIT, Cambridge, MA, USA. [30]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [31]Department of Pathology and Center for Cancer Research, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [32]Department of Neurological Surgery, University of California, San Francisco, CA, USA. [33]Department of Pediatric Hematology and Oncology, Research Institute Children's Cancer Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [34]Translational Neurosurgery, Friedrich-Alexander Universität Erlangen-Nürnberg, Erlangen, Germany. [35]Department of Neurological Surgery, Northwestern University Feinberg School of Medicine, Chicago, IL, USA. [36]German Cancer Consortium (DKTK), Partner Site Freiburg, Freiburg, Germany. [37]These authors contributed equally: Richard Drexler, Robin Khatri. [38]These authors jointly supervised this work: Dieter H. Heiland, Sonja Hänzelmann, Franz L. Ricklefs. ✉e-mail: f.ricklefs@uke.de

Methods

Patient cohorts

Several patient cohorts were analyzed based on the glioma subclass. A clinical cohort of 363 patients who underwent IDH-wild-type glioblastoma resection at University Medical Center Hamburg-Eppendorf, University Hospital Frankfurt or Charité University Hospital Berlin was analyzed. Informed written consent was obtained from all patients and experiments were approved by the medical ethics committee of the Hamburg chamber of physicians (PV4904). The TCGA-GBM cohort was included for external validation[19]. A clinical cohort of pediatric and adolescent patients who underwent surgery for H3K27-altered DMG at University Medical Center Hamburg-Eppendorf was established and extended with cohorts from Sturm et al.[34] and Chen et al.[39]. The reference and diagnostic set (n = 3,905) from Capper et al.[18] was utilized.

Clinical definitions

Diagnosis for the clinical cohort followed World Health Organization (WHO) classification guidelines[55]. The EOR of contrast-enhancing parts was stratified into GTR, complete removal, near GTR, >90% removal and partial resection, <90% removal. Overall survival refers to diagnosis until death or last follow-up and PFS from diagnosis until progression according to RANO criteria based on local assessment[56]. Seizures and antiepileptic medication use were defined by the current International League Against Epilepsy guidelines[57]. T1-weighted and T2-weighted FLAIR MRI images were analyzed using the Brainlab program. The volume of contrast enhancement, FLAIR hyperintensity and necrotic volume was assessed in cm^3 obtained via multiplanar 3D reconstruction of the tumor ROI, enabled by delineating with the tool 'Smart Brush' manually in every slice.

Stereotactic biopsies for spatial sample collection

Biopsies were obtained using a cranial navigation system (Brainlab v.13.0) and intraoperative neuronavigation. To limit the influence of brain shift, biopsies were obtained before tumor removal at the beginning of surgery with minimal dural opening. Tissue samples were then transferred to 10% buffered formalin and sent to the Department of Neuropathology for further processing and histopathological evaluation.

Measurement of functional connectivity using magnetoencephalography

Tumor tissues with HFC and LFC sampled during surgery based on preoperative MEG were obtained from patients with IDH-wild-type glioblastoma operated on in the Department of Neurosurgery, University of California, San Francisco[12]. From each formalin-fixed paraffin-embedded (FFPE) tissue block, four serial sections at a thickness of ~10 µm each were used for DNA extraction. DNA was extracted with the QIAamp DNA FFPE kit (QIAGEN). DNA was quantified using the Nanodrop Spectrophotometer (Thermo Scientific). The ratio of optical density at 260 nm to 280 nm was calculated and served as the criterion for DNA quality.

Functional connectivity by rs-fMRI

Forty-four treatment-naive patients with glioblastoma (mean age 65 ± 9 years) underwent rs-fMRI before surgery, with tumor tissues subsequently analyzed for genome-wide DNA methylation patterns using the Illumina EPIC (850k) array. Functional data preprocessing followed a standardized protocol implemented in SPM12 (ref. 58) within MATLAB (v.9.5)[59,60]. In brief, functional images were realigned, unwarped and coregistered to the structural image. Segmentation, bias correction and spatial normalization were conducted, with functional images smoothed using a 5-mm FWHM Gaussian kernel. Further preprocessing steps included slice-time correction, regression of movement-related time series using ICA-AROMA[24] and high-pass filtering (>0.01 Hz). Tumor lesions were segmented using ITK-SNAP[61] software and utilized as regions of interest for seed-based correlation analysis to compute voxel-based tumor-to-peritumoral connectivity (Fisher z transformation). A 10-mm peritumoral distance mask was created, and mean functional connectivity between the tumor and its peritumoral surrounding was computed using a ROI-to-voxel approach.

Immunoblotting

Frozen tissue samples were lysed using RIPA buffer, containing 50 mM Tris-HCl (pH 7.5), 150 mM NaCl, aprotinin (10 mg ml^{-1}), 1 mM phenylmethylsulfonyl fluoride, leupeptin (10 mg ml^{-1}), 2 mM Na_3VO_4, 4 mM EDTA, 10 mM NaF, 10 mM sodium pyrophosphate, 1% NP-40, 0.1% sodium deoxycholate and 1% protease inhibitor (Merck). Total protein concentration was measured by the bicinchoninic acid (BCA) assay (Pierce). Proteins were separated using Tris-glycine gels, blotted into nitrocellulose membrane and probed with antibodies anti-BDNF (1:1,000 dilution, Cell Signaling, 47808) and anti-β-actin (1:1,000 dilution, Sigma-Aldrich A2228).

Immunohistochemistry

Tissue samples were fixed in 4% formaldehyde, dehydrated, embedded in paraffin and sectioned at 2 µm following standard laboratory protocols. Immunohistochemical staining for NeuN (Chemico, MAB377, 1:200 dilution), Sox2 (Abcam, AB79351, 1:200 dilution), OLIG2 (R&D Systems, AF2418, 1:50 dilution) and GFAP (DAKO, M0761, 1:200 dilution) was conducted using an automated staining machine (Ventana BenchMark TX, Roche Diagnostics). Detection was achieved using diamino-benzidine as a chromogen, with counter-staining performed using Mayer's Solution (Sigma-Aldrich).

Drug sensitivity analysis

Patient-derived glioblastoma cell lines (GS-11, GS-73, GS-84, GS-110, GS-13, GS-74, GS-80, GS-90 and GS-101) were dissociated into single cells and seeded into a 384-well plate at a density of 1,250–7,500 cells per well in neurobasal medium supplemented with B27, glutamine, pen/strep, heparin and human FGF and EGF. Cells were treated with 27 drugs and dimethylsulfoxide as a control in triplicate for 48 h at 37 °C and 5% CO_2. After treatment, cells were fixed, blocked and stained with antibodies against vimentin, cleaved caspase 3 and TUBB3. Imaging was performed using an Opera Phenix automated confocal microscope and z-stacks were segmented based on DAPI staining using CellProfiler (v.2.2.0)[62]. Downstream analysis was conducted in MATLAB v.9.13.0, where marker-positive cells/spheroids were identified using linear thresholds. Cell counts and average cell/spheroid areas were averaged per condition and compared between drug treatment and control groups.

Spatially resolved transcriptomics

Quality assessment RNA. RNA extraction from FFPE tissue sections was conducted following the 'Purification of Total RNA from FFPE tissue sections' protocol (July 2021 version). Two 10-µm sections per tissue block were processed and RNA was eluted using 14 µl RNase-free water. Subsequently, 2 µl of the eluted RNA was subjected to both the Qubit RNA High-Sensitivity Assay and the DNF-471 Standard Sensitivity RNA Protocol using the Fragment Analyzer, following the respective manufacturer's instructions. RNA quality was assessed by computing the Distribution Value 200 (DV200) using Agilent's ProSize Data Analysis Software. The DV200 represents the percentage of RNA fragments longer than 200 nucleotides within a range of 200–10,000 bp. A DV200 ≥ 50% is considered desirable according to 10x Genomics guidelines. Additionally, the software provided the RNA integrity number to supplement the quality assessment.

Tissue preprocessing. To prepare FFPE tissue for spatial transcriptomics, sections of 5-µm thickness were sliced using a microtome, floated in a 42 °C water bath and transferred onto glass slides. Following H&E staining, tissue examination under the EVOS microscope facilitated the selection of the area of interest. The 'Visium Spatial Gene Expression

for FFPE – Tissue Preparation Guide' (CG000408, Rev A) guided the initial steps of tissue preprocessing. Modifications to these steps are detailed explicitly in subsequent descriptions. For hydration and trimming, without conducting a tissue adhesion test due to intact tissue adhesion on glass slides, FFPE tissue blocks underwent hydration in an ice water bath for 20 min, followed by trimming and cutting into 4-µm thick sections using the Thermo Fisher Scientific HM355 S automatic microtome. Trimming excess paraffin and tissue parts on a standard glass slide was performed, followed by floating the section in a 42 °C water bath for extension and smoothing. Sections were then fit onto Visium slides and dried using a thermocycler at 42 °C for 3 h, before being stored in a desiccator at room temperature overnight. After heating the Visium slides at 60 °C for 2 h, they underwent two 15-min immersions in xylene, followed by serial dilutions in 100%, 96%, 85% and 70% ethanol for 3 min each. The slides were finally rinsed in Milli-Q water for 20 s. The slides were stained with 1 ml hematoxylin for 3 min, washed in two successive Milli-Q water baths, treated with 1 ml bluing buffer for 1 min, washed again and then stained with 1 ml alcoholic eosin for 1 min, followed by another wash. Imaging was carried out with an EVOS M7000 microscope from Thermo Fisher Scientific at ×20 magnification in the brightfield setting, utilizing auto-focus for the first image of each capture area. Following imaging, the slide was placed into a Visium slide cassette (PN2000282) with an alignment tool (PN3000433). Pipetting was performed carefully to prevent disturbing the tissue, ensuring full coverage of the capture area and complete removal of leftover fluids. Each well of the cassette was treated twice with 100 µl 0.1 N HCl, then rinsed with 150 µl, pH 9.0 TE buffer, followed by another TE buffer application and incubation at 70 °C for 1 h on a thermal cycler. This initiated the library construction's hybridization stage.

Library preparation. Fo the pre-hybridization mix application, each well received pre-hybridization mix, followed by a 30-min incubation at 37 °C. This was succeeded by an overnight incubation of probe hybridization mix at 50 °C, centrifugation, multiple washes and application of probe ligation mix for 1 h at 37 °C. Post-ligation wash buffer was applied, followed by several washes. For the RNase and permeabilization mix application, the RNase mix and permeabilization mix were each applied and incubated for 30 min and 1 h, respectively at 37 °C, followed by washing and probe extension mix application. For probe elution and PCR, 0.08 M KOH was utilized to elute the probe. After transferring the solution to an eight-tube-strip, 1 M, pH 7.0 Tris-HCl was added. Cycle numbers for PCR were determined using a qPCR mix and performed with a StepOnePlus Real-Time PCR System. Sample Index PCR followed, with cleanup using SPRIselect and transfer of 25 µl to a new tube strip. A second qPCR was performed with the NEBNext Library Quant kit for Illumina to determine library molarities, ensuring successful library construction and cDNA presence.

Sequencing. Sequencing of the libraries was conducted using the NextSeq 500/550 device from Illumina. Libraries were normalized to the same molarity before being combined. Denaturation and dilution of libraries were performed following the 'NextSeq System – Denature and Dilute Libraries Guide' protocol. The combined library was denatured with 0.2 N NaOH, neutralized and diluted to a loading concentration using High Output kits. PhiX control was denatured, diluted and mixed with the library. The final mix underwent sequencing with the NextSeq 500/550 High Output kit v.2.5 (75 cycles).

Isolation and analysis of extracellular vesicles

Extracellular vesicles were isolated from plasma or cerebrospinal fluid of patients with glioblastoma by differential centrifugation[37,63]. After initial centrifugation steps to eliminate cells, platelets and large vesicles, extracellular vesicle pellets were obtained through ultracentrifugation. These pellets were resuspended with filtered PBS and analyzed for concentration and size using nanoparticle tracking analysis. Extracellular vesicle-enriched samples were diluted before nanoparticle tracking analysis and the analysis was conducted using appropriate parameters. Additionally, extracellular vesicles were characterized by electron microscopy for size and morphology and by imaging flow cytometry for extracellular vesicle markers (CD9, CD63 and CD81). DNA extraction from extracellular vesicles was performed using a purification kit. For comparison, bulk cfDNA was isolated from plasma using a commercial kit.

Detection of BDNF serum levels

Plasma from patients with glioblastoma was isolated by double spin centrifugation of whole blood. Samples were aliquoted and stored at −80 °C before use. BDNF plasma levels were detected using the LEGENDplex Neuroinflammation Panel 1 (BioLegend). Data were acquired using the BD LSR Fortessa and Beckman Coulter Cytoflex LX flow cytometer and analyzed with the BioLegend LEGENDplex software.

Proteomic processing of human glioblastoma samples

FFPE samples of tumors were obtained from tissue archives from the neuropathology unit in Hamburg. Tumor samples were fixed in 4% paraformaldehyde, dehydrated, embedded in paraffin and sectioned at 10 µm for microdissection using standard laboratory protocols. For paraffin removal, FFPE tissue sections were incubated in 0.5 ml n-heptane at room temperature for 30 min, using a ThermoMixer (ThermoMixer 5436, Eppendorf). Samples were centrifuged at 14,000*g* for 5 min and the supernatant was discarded. Samples were reconditioned with 70% ethanol and centrifuged at 14,000*g* for 5 min. The supernatant was discarded. The procedure was repeated twice. Pellets were dissolved in 150 µl 1% w/v sodium deoxycholate in 0.1 M triethylammonium bicarbonate buffer and incubated for 1 h at 95 °C for reverse formalin fixation. Samples were sonicated for 5 s at an energy of 25% to destroy interfering DNA. A BCA assay was performed (Pierce BCA Protein Assay kit, Thermo Scientific) to determine the protein concentration, following the manufacturer's instructions. Tryptic digestion was performed for 20 µg protein, using the single-pot, solid-phase-enhanced sample preparation (SP3) protocol[64]. Eluted peptides were dried in a Savant SpeedVac Vacuum Concentrator (Thermo Fisher Scientific) and stored at −20 °C until further use. Directly before measurement, dried peptides were resolved in 0.1% formic acid to a final concentration of 1 µg µl^{-1}. In total 1 µg was subjected to mass spectrometric analysis.

Liquid chromatography–tandem mass spectrometer parameters

LC–MS/MS measurements were performed using a QExactive mass spectrometer (Thermo Fisher Scientific) coupled with a Dionex Ultimate 3000 UPLC system (Thermo Fisher Scientific). Tryptic peptides were injected via an autosampler, purified, and desalted using a reversed-phase trapping column (Acclaim PepMap 100 C18 trap) before separation on a reversed-phase column (Acclaim PepMap 100 C18). Trapping occurred for 5 min at a flow rate of 5 µl min^{-1}, followed by separation using a linear gradient from 2% to 30% solvent B over 65 min at 0.3 µl min^{-1}. Peptides were ionized using nano-electrospray ionization (nano-ESI) with a spray voltage of 1,800 V and analyzed in data-dependent acquisition mode. During MS1 scans, ions were accumulated for a maximum of 240 ms or until reaching a charge density of 1×10^6 ions (AGC target), with mass analysis performed at a resolution of 70,000 at $m/z = 200$ over a mass range of 400–1,200 m/z. Peptides with charge states between 2+ and 5+ and intensities above 5,000 were isolated within a 2.0 m/z isolation window in top-speed mode for 3 s from each precursor scan and fragmented using higher energy collisional dissociation with a normalized collision energy of 25%. MS2 scanning, conducted using an orbitrap mass analyzer, had a starting mass of 100 m/z with a resolution of 17,500 at $m/z = 200$ and was accumulated for 50 ms or until reaching an AGC target of 1×10^5. Peptides that were already fragmented were excluded for 20 s.

NGS of low- and high-neural glioblastoma samples

Tumor mutational profiling was conducted at the Department of Neuropathology, University Hospital Heidelberg, using a custom CNS tumor-specific NGS gene panel (Agilent, SureSelect Custom Tier2, 1,235 Mb). Library preparation followed manufacturer recommendations with the SureSelect XT HS2 DNA kit (Agilent, 5191-5688). Prepared libraries were pooled and sequenced on the Illumina Novseq6000 platform (Novaseq v.1.5 200 cycles S1 Reagent kit, 20028318). The NGS panel covers the entire coding region, along with selected intronic and promoter regions of 201 genes relevant to CNS tumors. It detects single-nucleotide variants, small insertions/deletions (indels), exonic rearrangements and recurrent fusion events. Sequenced reads were mapped to GRCh38 using the nf-core/sarek (v.3.3.2) pipeline[65–67], with single-nucleotide variant and structural variant calling performed using Strelka (v.4.4.0.0)[68] and Manta (v.1.6.0)[23]. Variant annotation was performed using SNPeff (v.5.1d)[69]. Variants were filtered based on several criteria, including mapping to exonic regions, QUAL > 20, MQ > 30, DP > 15, high/moderate impact and a population frequency <0.001 from the 1000 Genomes project. Additionally, variants with high population frequencies in the Genome Aggregation Database (gnomAD), such as SETD2 c.5885C>T and KMT2C c.2447dupA, were filtered out.

Mice housing

In vivo experiments were conducted following approved protocols from the Stanford University Institutional Animal Care and Use Committee and the University Medical Center Hamburg-Eppendorf, adhering to institutional guidelines and explicit permissions from local authorities. Animals were housed under standard conditions in pathogen-free environments, with temperature- and humidity-controlled housing and access to food and water in a 12-h light–dark cycle. For xenograft experiments, the Institutional Animal Care and Use Committee established guidelines based on indications of morbidity, with mice killed if they displayed signs of neurological morbidity or lost 15% or more of their body weight.

Orthotopic xenografting of patient-derived low- and high-neural glioblastoma cells

NSG mice (NOD-SCID-IL2Rγ-chain-deficient, The Jackson Laboratory) were used for experiments conducted at Stanford University, with equal distribution of male and female mice. Primary patient-derived low- ('UCSF-UKE-1') or high-neural ('UCSF-UKE-2') glioblastoma neurospheres were prepared in sterile Hanks balanced salt solution (HBSS) and stereotactically implanted into the premotor cortex (M2) of mice at postnatal day (P) 28–30. Mice survival analyses were performed on NMRI-Foxn1nu immunodeficient mice (Janvier-Labs) at the University Medical Center Hamburg-Eppendorf. Neurospheres from cultured primary patient-derived low- ('GS-8', 'GS-10', 'GS-73' and 'GS-80') or high-neural ('GS-57', 'GS-74', 'GS-75' and 'GS-101') glioblastoma were injected into the striatum. External validation of mice survival data was conducted using publicly available datasets from Vaubel et al.[28] and Golebiewska et al.[29].

Perfusion and immunofluorescence staining

Eight weeks post-xenografting, low and high-neural glioblastoma-bearing mice were anesthetized with intraperitoneal avertin and transcardially perfused with PBS followed by fixation in 4% paraformaldehyde (PFA) overnight at 4 °C. After cryoprotection in 30% sucrose for 48 h, brains were embedded in Tissue-Tek O.C.T. and sectioned coronally at 40 µm using a sliding microtome. For immunofluorescence, sections were blocked in a solution of 3% normal donkey serum and 0.3% Triton X-100 in TBS, followed by incubation with primary antibodies overnight at 4 °C. Antibodies used included mouse anti-human nuclei clone 235-1, rabbit anti-Ki67, rat anti-MBP, mouse anti-nestin, guinea pig anti-synapsin-1/2, chicken anti-neurofilament or anti-PSD95. After rinsing, sections were incubated with appropriate secondary antibodies and mounted with ProLong Gold Mounting medium.

Confocal imaging and quantification of cell proliferation and infiltration

Cell quantification within xenografts was conducted by a blinded investigator using a Zeiss LSM980 scanning confocal microscope. A 1-in-6 series of coronal brain sections were selected, with four consecutive slices analyzed at approximately 1.1–0.86 mm anterior to bregma. HNA-positive tumor cells were quantified in each field to determine the proliferation index, calculated as the percentage of HNA-positive cells co-labeled with Ki67. Infiltration into the corpus callosum was assessed in the same sections, with HNA-positive tumor cells co-labeled with Ki67 and divided by the total number of DAPI-marked nuclei.

Confocal puncta quantification

Images were captured using a ×63 oil-immersion objective on a Zeiss LSM980 confocal microscope. Colocalization analysis of synaptic puncta images from both low and high-neural glioblastoma xenograft samples was performed by a blinded investigator. A custom ImageJ processing script, developed at the Stanford Shriram Cell Science Imaging Facility, was utilized for this purpose. The script defined each pre- and postsynaptic puncta and assessed colocalization within a defined proximity of 1.5 µM. To subtract local background, the ImageJ rolling ball background subtraction method was applied. Peaks were identified using the imglib2 DogDetection plugin, which employs the difference of Gaussians to enhance the signal of interest. The plugin then assigned ROIs to each channel based on predefined parameters. Neuron and glioma ROIs were quantified, and the script extracted the number of glioma ROIs within 1.5 µm of the neuron ROIs. This script was implemented in Fiji/ImageJ using the ImgLib2 and ImageJ Ops libraries.

Sample preparation and image acquisition for electron microscopy

Twelve weeks post-xenografting of low- ($n = 3$, 'UCSF-UKE-1') and high-neural glioblastoma cells ($n = 3$, 'UCSF-UKE-2'), mice were killed via transcardial perfusion with Karnovsky's fixative: 2% glutaraldehyde and 4% PFA in 0.1 M sodium cacodylate (pH 7.4). Transmission electron microscopy (TEM) analysis was conducted on tumor masses within the CA1 region of the hippocampus. Samples were post-fixed in 1% osmium tetroxide, washed and en bloc-stained overnight. Dehydration was performed using graded ethanol and acetonitrile. Samples were then infiltrated with EMbed-812 resin, followed by embedding in TAAB capsules and oven curing. Sections of 40–60 nm were cut on a Leica Ultracut S and mounted on 100-mesh Ni grids. For immunohistochemistry, grids underwent microetching with periodic acid and osmium elution with sodium metaperiodate. Grids were blocked, incubated with primary goat anti-RFP antibody overnight, rinsed and incubated with secondary antibodies. Grids were contrast stained with uranyl acetate and lead citrate. Imaging was conducted using a JEOL JEM-1400 TEM at 120 kV, with image capture facilitated by a Gatan Orius digital camera.

Cell culture

Fresh glioblastoma samples were obtained from patients operated in the Department of Neurosurgery, University Medical Center Hamburg-Eppendorf. Samples were immediately placed in HBSS (Invitrogen), transferred to the laboratory and processed within 20 min. The tissue was cut into <1-mm^3 fragments, washed with HBSS and digested with 1 mg ml^{-1} collagenase/dispase (Roche) for 30 min at 37 °C. Digested fragments were filtered using a 70-µm cell mesh (Sigma-Aldrich) and the cells were seeded into T25 flasks at 2,500–5,000 cells per cm^2. The culture medium consisted of neurobasal medium (Invitrogen) with B27 supplement (20 µl ml^{-1}, Invitrogen), Glutamax (10 µl ml^{-1}, Invitrogen), fibroblast growth factor-2 (20 ng ml^{-1}, Peprotech), epidermal growth factor (20 ng ml^{-1}, Peprotech) and heparin (32 IE ml^{-1},

Ratiopharm). Growth factors and heparin were renewed twice weekly. Spheres were split by mechanical dissociation when they reached a size of 200–500 µm. In this study, analyzed cell cultures with clinical data are represented in Extended Data Fig. 4. Long-term cultivation cell cultures were used from a publicly available dataset ($n = 7$, GSE181314) and one in-house cell line ($n = 1$).

Neuron-glioma co-culture experiments

Neurons were isolated from CD1 (The Jackson Laboratory) mice at P0 using the Neural Tissue Dissociation kit - Postnatal Neurons (Miltenyi) and followed by the Neuron Isolation kit, Mouse (Miltenyi). After isolation, 150,000 neurons were plated onto glass coverslips (Electron Microscopy Services) after pre-treatment with poly-L-lysine (Sigma) and mouse laminin (Thermo Fisher)[4]. Neurons are cultured in BrainPhys neuronal medium (StemCell Technologies) containing B27 (Invitrogen), BDNF (10 ng ml^{-1}, Shenandoah), GDNF (5 ng ml^{-1}, Shenandoah), TRO19622 (5 µM; Tocris) and β-mercaptoethanol (Gibco). Half of the medium was replenished on days in vitro (DIV) 1 and 3. On DIV 5, half of the medium was replaced in the morning. In the afternoon, the medium was again replaced with half serum-free medium containing 75,000 cells from patient-derived low- ('UCSF-UKE-1') or high-neural ('UCSF-UKE-2') cell cultures. Cells were cultured with neurons for 72 h and then fixed with 4% PFA for 20 min at room temperature and stained for puncta quantification as described above.

EdU proliferation assay

For EdU proliferation assays, coverslips were prepared as described above. Again, at DIV 5, low-neural ('UCSF-UKE-1') or high-neural ('UCSF-UKE-2') glioblastoma cells were added to the neuron cultures. Forty-eight hours after addition of glioblastoma cells, slides were treated with 10 µM EdU. Cells were fixed after an additional 24 h using 4% PFA and stained using the Click-iT EdU kit and protocol (Invitrogen). Proliferation index was then determined by quantifying the percentage of EdU-labeled glioblastoma cells (identified by EdU$^+$/DAPI$^+$) over total number of glioblastoma cells using confocal microscopy.

3D migration assay

3D migration experiments were performed as previously introduced[70] with some modifications. In brief, 96-well flat-bottomed plates (Falcon) were coated with 2.5 µg per 50 µl laminin per well (Thermo Fisher) in sterile water. After coating, a total of 200 µl of culture medium per well was added to each well. A total of 100 µl of medium was taken from 96-well round-bottom ULA plates containing ~200-µm diameter neurospheres of low- ('UCSF-UKE-1') and high-neural ('UCSF-UKE-2') glioblastoma lines and the remaining medium, including neurospheres was transferred into the pre-coated plates. Images were then acquired using an EVOS M5000 microscope (Thermo Fisher Scientific) at time 0, 24, 48 and 72 h after encapsulation. Image analysis was performed using ImageJ by measuring the diameter of the invasive area. The extent of cell migration on the laminin was measured for six replicate wells normalized to the diameter of each spheroid at time zero and the data are presented as a mean ratio for three biological replicates.

Bioinformatic and statistical analysis

DNA methylation profiling and processing. DNA was extracted from tumors, extracellular vesicles and bulk plasma, and analyzed for genome-wide DNA methylation patterns using the Illumina EPIC (850k) array. The processing of DNA methylation data was performed with custom approaches[71]. Methylation profiling results from the first surgery were submitted to the molecular neuropathology methylation classifier v.12.5 hosted by the German Cancer Research Center[18]. Patients were included if the calibrated score for the specific methylation class was >0.84 at the time of diagnosis[71]. For IDH-wild-type glioblastoma, patients (scores between 0.7 and 0.84) with a combined gain of chromosome 7 and loss of chromosome 10 or amplification of *EGFR* were included in accordance with cIMPACT-NOW criteria[72]. A class member score of ≥0.5 for one of the glioblastoma subclasses was required. Evaluation of the *MGMT* promoter methylation status was made from the classifier output v.12.5 using the *MGMT*-STP27 method[73].

All IDAT files were processed using the preprocess Illumina (minfi, v.1.40.0)[74]. Probes with detection P values <0.01 were kept for further analysis. Probes with <3 beads in at least 5% of samples, all non-CpG probes, SNP-related probes and probes located on X and Y chromosomes were discarded.

Dichotomization of tumors into low- and high-neural subgroups. We used the cell-type-specific methylation signature available from Moss et al.[17] consisting of 25 cell-type components. We used the original implementation of Moss et al. to perform cell-type deconvolution using non-negative least square linear regression.

We deciphered the neural signature in GBM using a combined dataset ($n = 1{,}058$) from Capper et al.[18] ($n = 624$) and our institutional cohorts from Hamburg, Berlin and Frankfurt (all Germany) ($n = 434$). The combined dataset was dichotomized into low- ($n = 529$) and high-neural ($n = 529$) tumors using the median neural proportion of 0.41. This cutoff value was used to classify GBM into low- and high-neural tumors for all analyses. External validation was performed using the publicly available dataset from the TCGA-GBM database ($n = 178$)[19].

Reproducibility of differential methylation sites between low- and high-neural groups. We performed differential methylation analysis of 363 samples of the internal cohort using dmpFinder function from minfi R package[74] (v.1.40.0). In total, we identified 1,289 CpG sites differentiating low- and high-neural groups. To estimate the predictive power of these sites, we trained a logistic regression model using scikit-learn package (v.1.2.2) on the clinical cohort using the differentially methylated sites as input features. The model was subsequently applied to the other cohorts.

Cell state composition analysis. To infer cell-type and cell state abundance, we conducted a bulk DNA methylation assay using EPIC arrays and applied the reference-free deconvolution method by Silverbush et al.[75]. This method, trained on the DKFZ glioblastoma cohort and tested on TCGA-GBM data, successfully infers cell types (immune, glia and neuron) and malignant cell states (stem-like and differentiated). We followed the protocol of Silverbush et al.[75], using the EpiDISH package[76], utilizing the provided encoding and RPC method with 2,000 maximum iterations.

DNA tumor purity. Tumor purity was predicted in silico from DNA methylation data using the RF_purify Package in R[77]. This package uses the 'absolute' method, which measures the frequency of somatic mutations within the tumor sample and relates this to the entire DNA quantity[78].

Integrative analysis of methylation and gene expression. WGCNA was performed using the hdWGCNA[22] R package. Methylation-derived neural subgroup labels were considered as a trait. The optimal soft power was determined to be 16. For dimension reduction and visualization of the coexpression network, we employed the UMAP via the ModuleUMAPPlot function. Gene Ontology analysis was subsequently performed on the top 100 module-associated genes using the compareCluster function. Visualization of module-associated pathway activations was accomplished using the clusterProfiler package.

To contextualize the identified modules at a single-cell level, we utilized GBMap[23] and the reference dataset of human motor cortex (Allen Institute). Both datasets were integrated by alignment of the latent space representation. Based on the zero-inflated nature of

single-cell data, we estimated the module enrichment by the frequency of each gene (g) being detected and the expression values as follows:

$$m_{\exp} = \frac{\sum_{i=1}^{n} x_n \cdot n\#i = 1(x_n = 0)}{2n}$$

$m_{\exp}$ refers to the module expression score per cell which is estimated by the mean of x the log normalized and scaled expression values of n genes from the WGCNA modules. The mean is normalized by the frequency of nonzero-determined genes.

SRT data analysis. Computational analysis of spatially resolved transcriptomics (SRT) data was performed by the SPATA2 R package (v.2.01). An SPATA object was prepared for the SRT data.

Single-cell deconvolution. Single-cell deconvolution was performed using Cell2location[79] with the GBMap single-cell data[23] as a reference. The SPATA object was converted into the AnnData format and mitochondrial genes were sequestered into the obsm['MT'] matrix of the object before training the model for 500 epochs on the GPU. After training, we invoked export_posterior on the model to extract the posterior distribution of cell-type abundances, drawing 1,000 samples to robustly estimate these abundances across spatial locations. The cell-type abundances were exported back to the SPATA object by the addFeature function of SPATA2.

RNA deconvolution. We utilized the GBMapExtended single-cell RNA-seq (scRNA-seq) dataset and the human neocortex dataset from the Allen Institute to perform cell-type deconvolution. Data preparation involved loading and transforming the scRNA-seq data into a SingleCellExperiment object with Seurat and SingleCellExperiment libraries in R, annotated with relevant cell and gene identifiers. We leveraged the digitalDLSorteR package to train a deconvolution model, initiating with the setting of a random seed for reproducibility, followed by loading scRNA-seq profiles into the digitalDLSorteR framework. Key parameters, including cell and gene identifiers and cell-type annotations, were specified. The digitalDLSorteR's zinbwave parameters were estimated to simulate single-cell profiles, incorporating previous knowledge of cell-type distributions to refine the simulation. A bulk cell matrix was generated based on probabilistic design from simulated cell profiles, and a digitalDLSorter model was trained on this matrix with standard scaling. Post-training, the model was applied to deconvolve a dataset comprising RNA-seq and methylation data, processed to extract counts and metadata. The deconvolution results were then visualized using ggplot2, with sample types and percentage compositions graphed, showcasing the cellular heterogeneity across different samples.

Construction of spatial graphs from Visium SRTs. The SRT object was preprocessed with SPATA2, including log transformation of the count matrix and alignment of the imaging dataset (H&E Image). Nucleus positions were annotated using an automated ilastik pretrained segmentation algorithm. For samples with low image quality, we adapted CytoSpace[80] in our workflow. Spot coordinates were extracted via the getCoordsDf function and a pairwise distance matrix was computed based on the 'x' and 'y' coordinates of cells. The zero values in the distance matrix were replaced with a constant value of 1,000 to avoid computational issues. This ensured that subsequent thresholding steps would not falsely consider a cell as its own neighbor. A distance threshold (one unit greater than the smallest nonzero distance) was employed to construct an adjacency matrix, where cells within the threshold distance were designated '1' for adjacency and cells beyond the threshold were assigned '0' for no adjacency. Unique cell barcodes were used to label the rows and columns of the adjacency matrix, obtained from getCoordsDf. The adjacency matrix was then transformed into an undirected graph using the graph_from_adjacency_matrix function from the igraph package. We obtained the gene expression matrix with 5,000 most variable genes from our object and transposed it to align with the graph's vertices. Using the graphical representation, we characterized the local topology around a specific location, termed a 'query spot,' by identifying its n-hop neighborhood. Specifically, the three-hop neighborhood of a query spot was defined as the set of all spots reachable within three edges from the query spot in the graph.

GNN architecture. We used a deep neural network combining a graph isomorphism network (GIN) backbone with multiple multilayer perceptron (MLP) prediction heads. We used the Pytorch Geometric library and defined each spot as a node and edges were defined as the direct neighbors of each individual spot within a three-hop neighborhood. Node features were log-scaled and normalized expression values from the 5,000 most variably expressed genes. Non-expressed genes within a subgraph were masked. Edge features were defined based on each node's direct neighbors, with each node having a maximum of six neighbors. Subgraphs with fewer than 15 nodes were excluded. Self-loop edges were added to input graphs before forward pass.

We employed a three-layer GIN, and in the kth graph convolutional layer to process batches (size of 32) of SRT data, messages were computed using MLPs,

$$m_{uv} = \mathrm{MLP}(h_u)$$

where $u, v \in N(v)$ and then aggregated for each node v over neighborhood $N(v)$,

$$a_v = \sum_{u \in N(v)} m_{uv}$$

The updated embedding of node v was updated on the basis of all incoming messages to v,

$$h'_v = \mathrm{MLP}(a_v)$$

The GIN layers are represented as follows: x_v defines the expression vector of node v and N(v) is the set of its neighbors. The GIN convolution operation updates the feature vector of node v by aggregating features from N(v) and combining them with x_v own features. The updated feature vector x'_v is computed with ReLU (rectified linear unit) as follows:

$$x'_v = \mathrm{ReLU}\left(\left((1+\epsilon) \times x_v + \sum_{u \in N(v)} \mathrm{ReL}(x_u)\right)\right)$$

we define ϵ as a learnable parameter that allows the model to weigh the importance of a node's own features versus the features of its neighbors. This operation is stacked multiple times ($k = 2$) in the kth GIN to allow for deeper aggregation of neighborhood information. After each GIN convolutional layer, batch normalization and LeakyReLU activation with a negative slope of 0.2 are applied, followed by a dropout layer with a dropout rate of 0.5 for regularization. The latent space representation of the graph is obtained by passing the output of the second GIN convolutional layer through a linear transformation (self.merge) with weights initialized using the Xavier uniform method. The resulting features are merged into a latent space and then global mean pooling is applied to create graph-level representations.

For the prediction tasks, separate MLP modules are employed. Each MLP consists of a linear layer, a ReLU activation, batch normalization, dropout and a final linear layer that outputs the predictions. The MLPs are structured as follows:

$$h(x) = W_2 \times D \times B \times \phi(W_1 \times x + b_1) + b_2$$

Where x is the latent space vector to the MLP, W_1 and W_2 are the weight matrices for the first and second linear transformations, respectively, b_1 and b_2 are the bias vectors for the first and second linear transformations, respectively, ϕ denotes the ReLU activation function, applied element-wise, where $\phi_z = ma(0, z)$, B represents the batch normalization operation applied to the activated output and D represents the dropout operation, which randomly zeroes some of the elements of its input with a certain probability to prevent overfitting.

For neural score prediction tasks, we minimized the squared L1 norm loss between predictions and score (torch.nn.L1loss).

Data split and evaluation metrics. We evaluated the GNN and comparative methods on both our proprietary Visium dataset and additional public domain datasets. We split the data into training and evaluation subsets using a stratified procedure. For the training dataset, we selected 20,000 subgraphs from spatial transcriptomics samples across 20 patients, incorporating clinical attributes such as tumor type and epigenetic neural score. For the evaluation dataset, we reserved samples from the remaining four patients, covering a range of neural scores. Additionally, we included a validation set of 24,000 subgraphs from all 24 patients, ensuring independence from the training set.

This approach ensured robust evaluation across diverse clinical and molecular features, with the neural score used as the prediction task, evaluated by R^2 against the neural score from EPIC methylation profiling.

Evaluation of the subgraph cell composition. We commenced by retrieving the spatial coordinates of each nucleus using the getNucleusPosition function from the SPATAwrappers package. The spatial coordinates representing the nuclei positions were obtained as $P = \{p_i | i = 1, \ldots, N\}$ where p_i is the coordinate pair for the ith nucleus and N is the total number of nuclei. Spatial grid coordinates corresponding to the transcriptomics data points were retrieved, denoted as $G = \{g_j | j = 1, \ldots, M\}$, with each g_j representing the coordinate pair for the jth grid point. For each grid point g_j, a vector of deconvolution scores $D_j = \{d_{jk} | k = 1, \ldots, T\}$ was extracted, where d_{jk} represents the score for the kth cell type at grid point j and T is the number of cell types. The scores were normalized to a range of [0, 1], and the number of cells of each type at each grid point was estimated as:

$$C_{jk} = \text{round}\left(\frac{d'_{jk} \times N_j}{\sum_{k=1}^{T} d'_{jk}}\right)$$

where d'_{jk} is the normalized score and N_j is the number of cells at grid point j. Cell types were assigned to each grid point g_j to create a mapping M_j, correlating grid points with their respective cell types. The cell-type mapping was integrated with nucleus position data to produce a comprehensive spatial map of cell-type distribution: $S = \{(p_i, M_j) | p_i \in P, M_j \in M\}$. This methodology facilitates the visualization and analysis of the cellular composition within the tissue section, providing insights into the complex spatial organization of the cellular environment.

Proteomic data processing. Proteomic samples ($n = 28$) were measured with liquid chromatography–tandem mass spectrometry (LC–MS/MS) systems and processed with Proteome Discoverer v.3.0. and searched against a reviewed FASTA database (UniProtKB[81]: Swiss-Prot, *Homo sapiens*, February 2022, 20,300 entries). The protein abundances were normalized at the peptide level. Perseus v.2.0.3 was used to obtain $\log_2$ transformed intensities. The imputation was performed using the random forest imputation algorithm (hyperparameters, 1,000 trees and ten repetitions) in RStudio v.4.3.

WGCNA for proteomics. We used hdWGCNA[82] to identify gene coexpression modules, employing a soft power of 9 and minimum module size of 10. After correcting for technical batch effects, significant modules ($P < 0.05$) were selected based on their correlation with traits. Overrepresentation analysis of gene sets within these modules was performed using clusterProfiler[67]. Cell-type enrichment within modules was identified using gene sets from PanglaoDB through the Python package enrichr[68]. Module scores on single cells were calculated using Scanpy's score_genes function with the core GBM single-cell atlas (GBMap)[23].

Electron microscopy data analysis. Sections from xenografted hippocampi of mice were imaged using TEM imaging. The xenografts were originally generated for a study by Krishna et al.[12] and mouse tissue was re-analyzed after epigenetic profiling and assignment to low- or high-neural glioblastoma groups. Here, 42 sections of high-neural glioblastoma across three mice and 45 sections of low-neural glioblastoma across three mice were analyzed. Electron microscopy images were taken at ×6,000 with a field of view of 15.75 µm^2. Glioma cells were counted and analyzed after identification of immunogold particle labeling with three or more particles. Furthermore, to determine synaptic structures all three of the following criteria had to be clearly met as previously described[4]: (1) presence of synaptic vesicle clusters; (2) visually apparent synaptic cleft; and (3) identification of postsynaptic density in the glioma cell. To quantify the percentage of glioma cells forming synaptic structures, the number of glioma-to-neuron synapses identified was divided by the total number of glioma cells analyzed.

Statistical analysis. Gaussian distribution was confirmed using the Shapiro–Wilk test. Parametric data were analyzed with an unpaired two-tailed Student's t-tests or one-way ANOVA with Tukey's post hoc tests. Survival curves were generated using the Kaplan–Meier method, with statistical significance determined by two-tailed log-rank analyses. Multivariate analysis for overall survival and PFS included computing hazard ratios and 95% confidence intervals using Cox proportional hazards regression models. Variables with $P < 0.05$ in univariate analysis were included. Significance was set at $P < 0.05$. GraphPad Prism v.10 was used for statistical analyses and data illustrations and R Studio was used for alluvial plots.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

IDAT files of the clinical cohort (363 patients with GBM) are available at the Gene Expression Omnibus under accession code GSE240704. The methylation data provided by Capper et al.[18] as illustrated in Extended Data Fig. 1 are accessible under accession code GSE109381. The TCGA-GBM cohort analyzed for external validation and as shown in Fig. 1d is accessible at https://portal.gdc.cancer.gov/projects/TCGA-GBM. Data files used in the spatial transcriptomic analyses are accessible at *Zenodo* at https://doi.org/10.5281/zenodo.10863736 (ref. 83). The single-cell RNA-seq dataset GBMap is available from the original publication and can be accessed through cellXgene (https://cellxgene.cziscience.com/collections/999f2a15-3d7e-440b-96ae-2c806799c08c) and the human motor cortex single-cell RNA-seq dataset is available from the Allen Brain Institute at https://portal.brain-map.org/atlases-and-data/rnaseq/human-m1-10x. Source data are provided with this paper.

Code availability

The code used to perform DNA methylation and proteomics analysis is available at https://github.com/imsb-uke/epigenetic-neural-glioblastoma. Codes used for performing transcriptomic analyses in Figs. 2 and 3 and Extended Data Figs. 3 and 4f are available at

https://github.com/heilandd/GBNeural. Additionally, the code for the non-reference-based multi-dimensional single-cell deconvolution from DNA methylation data as presented in Fig. 6f and Supplementary Fig. 4i can be found at https://github.com/danasilv/Deconvolution_of_GBM_bulk_DNA_methylation_profiles.

References

55. Capper, D. et al. Practical implementation of DNA methylation and copy-number-based CNS tumor diagnostics: the Heidelberg experience. *Acta Neuropathol.* **136**, 181–210 (2018).
56. Brat, D. J. et al. cIMPACT-NOW update 3: recommended diagnostic criteria for 'Diffuse astrocytic glioma, IDH-wildtype, with molecular features of glioblastoma, WHO grade IV'. *Acta Neuropathol.* **136**, 805–810 (2018).
57. Bady, P., Delorenzi, M. & Hegi, M. E. Sensitivity analysis of the MGMT-STP27 model and impact of genetic and epigenetic context to predict the MGMT methylation status in gliomas and other tumors. *J. Mol. Diagn.* **18**, 350–361 (2016).
58. Aryee, M. J. et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
59. Van Paemel, R. et al. Genome-wide study of the effect of blood collection tubes on the cell-free DNA methylome. *Epigenetics* **16**, 797–807 (2021).
60. Neuberger, E. W. I. et al. Physical activity specifically evokes release of cell-free DNA from granulocytes thereby affecting liquid biopsy. *Clin. Epigenetics* **14**, 29 (2022).
61. Zheng, S. C. & Teschendorff, A. E. EpiDISH - epigenetic dissection of intra-sample-heterogeneity. *Bioconductor* https://www.bioconductor.org/packages/devel/bioc/vignettes/EpiDISH/inst/doc/EpiDISH.html (2023).
62. Johann, P. D., Jäger, N., Pfister, S. M. & Sill, M. RF_Purify: a novel tool for comprehensive analysis of tumor-purity in methylation array data based on random forest regression. *BMC Bioinform.* **20**, 428 (2019).
63. Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
64. Maire, C. L. et al. Genome-wide methylation profiling of glioblastoma cell-derived extracellular vesicle DNA allows tumor classification. *Neuro-Oncol.* **23**, 1087–1099 (2021).
65. Hughes, C. S. et al. Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nat. Protoc.* **14**, 68–85 (2019).
66. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform* **9**, 559 (2008).
67. Wu, T. et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation* **2**, 100141 (2021).
68. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).
69. Garcia, M. et al. Sarek: a portable workflow for whole-genome sequencing analysis of germline and somatic variants. *F1000Research* **9**, 63 (2020).
70. Vinci, M., Box, C., Zimmerman, M. & Eccles, S. A. Tumor spheroid-based migration assays for evaluation of therapeutic agents. *Methods Mol. Biol.* **986**, 253–266 (2013).
71. Hanssen, F. et al. Scalable and efficient DNA sequencing analysis on different compute infrastructures aiding variant discovery. Preprint at *bioRxiv* https://doi.org/10.1101/2023.07.19.549462 (2023).
72. Garcia M. U. et al. nf-core/sarek: Sarek 3.4.0 - Pårtetjåkko. *Zenodo* https://doi.org/10.5281/zenodo.3476425 (2023).
73. Louis, D. N. et al. The 2016 World Health Organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol.* **131**, 803–820 (2016).
74. Wen, P. Y. et al. Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. *J. Clin. Oncol.* **28**, 1963–1972 (2010).
75. Scheffer, I. E. et al. ILAE classification of the epilepsies: position paper of the ILAE Commission for Classification and Terminology. *Epilepsia* **58**, 512–521 (2017).
76. Friston, K. J. *Statistical Parametric Mapping: The Analysis of Functional Brain Images* (Academic Press, 2011).
77. Jütten, K. et al. Dissociation of structural and functional connectomic coherence in glioma patients. *Sci. Rep.* **11**, 16790 (2021).
78. Jütten, K. et al. Asymmetric TUMOR-RELATED alterations of NETWORK-SPECIFIC intrinsic functional connectivity in glioma patients. *Hum. Brain Mapp.* **41**, 4549–4561 (2020).
79. Yushkevich, P. A. et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *NeuroImage* **31**, 1116–1128 (2006).
80. Vahid, M. R. et al. High-resolution alignment of single-cell and spatial transcriptomes with CytoSPACE. *Nat. Biotechnol.* **41**, 1543–1548 (2023).
81. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).
82. Morabito, S. et al. Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. *Nat. Genet.* **53**, 1143–1155 (2021).
83. Heiland, D. H. Visium spatially resolved transcriptomics of glioblastoma samples. *Zenodo* https://doi.org/10.5281/zenodo.10863736 (2024).

Acknowledgements

We thank L. Stegat for contributing data to the DMG cohort. The authors acknowledge S. Wutke for graphical assistance. Initial drafts of Figs. 1, 4 and 5 were made with BioRender.com. We thank all the patients who gave informed consent and without whom this research would not have been possible. This study was supported by numerous grants and research funds. F.L.R. received funding from the German research foundation (DFG RI2616/3-1), the Erich and Gertrud Roggenbuck foundation and from Illumina. U.S. was supported by the Fördergemeinschaft Kinderkrebszentrum Hamburg. Experiments conducted for investigating functional connectivity were supported by National Institutes of Health grants K08NS110919 and P50CA097257; Robert Wood Johnson Foundation grant 74259; the UCSF LoGlio Collective and Resonance Philanthropies; and U19 CA264339, Tom Paquin Brain Cancer Research Fund to S.H.J. and the Sullivan Brain Cancer Fund. R.K. and F.H. are funded by the EU eRare project Maxomod (01GM1917B). S.H. and T.B.H. received funding from SFB 1192 B8 and STOP-FSGS (01GM2202A). S.B. was supported by SFB 1192 C3. M.M. was supported by grants from the National Institute of Neurological Disorders and Stroke (R01NS092597), National Institutes of Health Director's Pioneer Award (DP1NS111132), National Cancer Institute (P50CA165962, R01CA258384 and U19CA264504). B.W. was supported by the DFG, SFB 824 and subproject B12. F.S. and P.G. received funding from SFB 1389 UNITE. D.H.H. received funding from the TRANSCAN (BMBF 01KT2328), German Research Foundation (Heisenberg Program DFG HE 8145/6-1, funding, HE 8145/5-1), the DKTK Partner Site Freiburg (DKTK-PI) and Joint Funding Program (HematoTrac). T.W. received support from Promedica Foundation, Baasch-Medicus Foundation, Helmut Horten Foundation and the Swiss Cancer League (KFS-5763-02-2023). M.W. was supported by the Swiss National Science Foundation (310030_185155/1).

Author contributions

R.D. and F.L.R. conceptualized, designed, conducted and interpreted all experiments and analyses. R.K. and S.H. designed the model and the computational framework, including deconvolution,

copy-number variation and proteomics analyses. F.H. assisted with data handling and conducted mutation analyses. R.K., F.H., S.H. and D.H.H. edited and gave input to the manuscript. S.H., D.H.H. and F.L.R. supervised the project. T.B.H. and S.B. provided critical review and commentary. M.M. and A.S.R. performed immunoassay quantification of BDNF serum levels. F.L.R., C.M., A.S. and K.L. contributed to cell culture and extracellular vesicle experiments. A.K.W. and U.S. contributed to DMG cohorts. H.B. calculated DNA tumor purity. J.N. conducted MS proteomic profiling. K.J. and D.D. contributed to functional connectivity measured by resting-state MRI. R.D., T.S., L.D., Y.Z., M.W., F.L.R., K.W., P.N.H., D.C., J.O. and P.V. contributed GBM cohorts of each institution. B.W. and J.G. performed stereotactic biopsies for spatial sample collection of human patients with GBM. M.M. contributed scRNA-seq data of DMG and provided equipment for in vivo analyses. R.D. and M.B.K. conducted in vivo experiments for proliferation and puncta synapse quantification and C.M. performed xenografting for survival analysis. R.D. performed co-culture experiments and migration assays. L.N. performed electron microscopy images, which were evaluated by R.D. D.S., V.H. and M.L.S. performed cell-state composition analysis. S.K. and S.H.J. contributed to functional connectivity measured by MEG. J.A.M., D.N.Z. and D.H.H. performed spatial transcriptomics. M.W., B.S., A.B. and T.W. conducted drug sensitivity analysis. P.C.G. and F.S. performed NGS. R.D. and F.L.R. wrote the manuscript. All authors contributed to manuscript editing and approved the final manuscript version.

Funding
Open access funding provided by Universitätsklinikum Hamburg-Eppendorf (UKE).

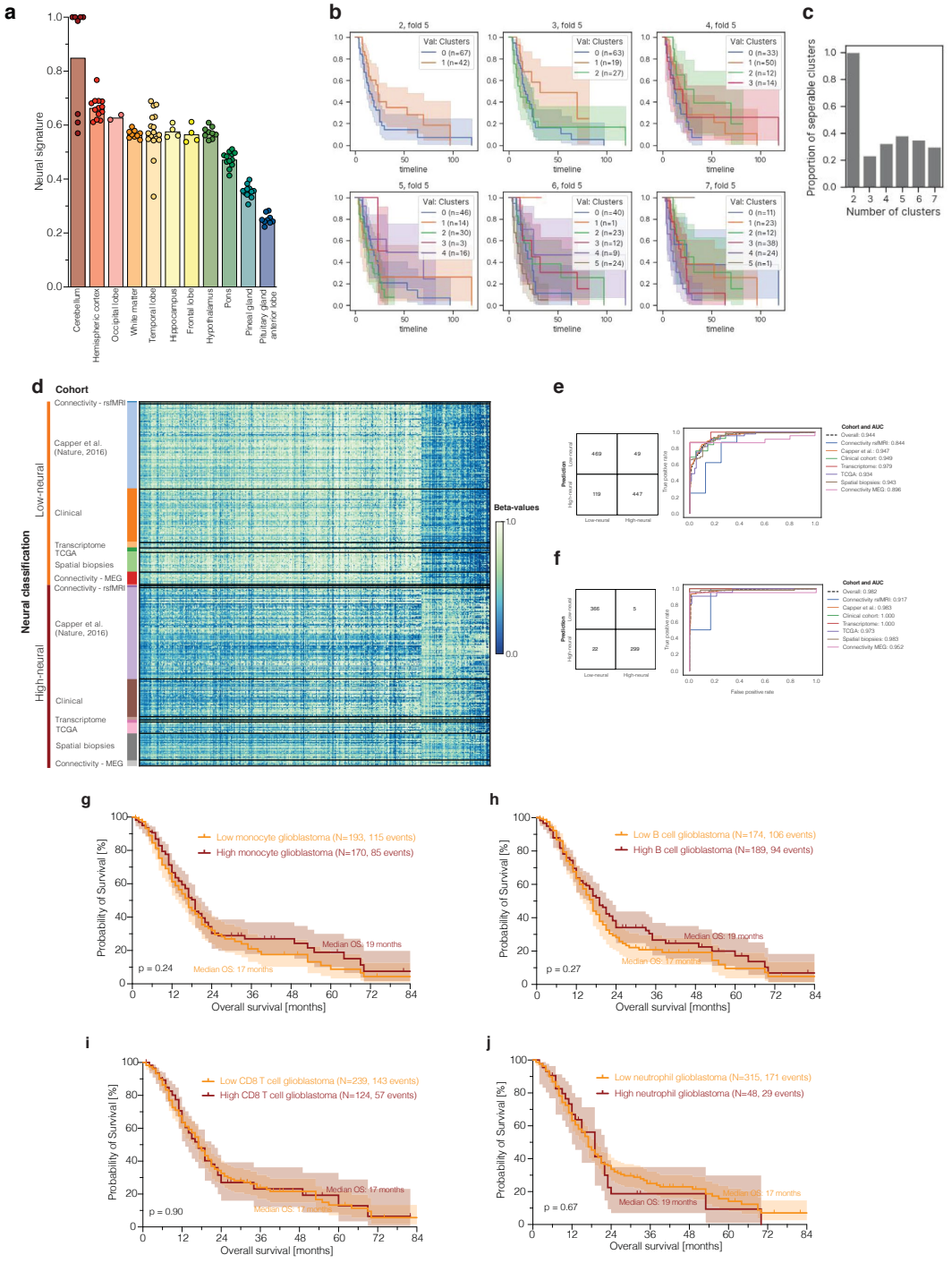
Competing interests
M.L.S. is an equity holder, scientific co-founder and advisory board member of Immunitas Therapeutics. M.M. holds equity in MapLight Therapeutics. The other authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41591-024-02969-w.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-024-02969-w.

Correspondence and requests for materials should be addressed to Franz L. Ricklefs.

Peer review information *Nature Medicine* thanks Maya Graham and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Ulrike Harjes, in collaboration with the *Nature Medicine* team.

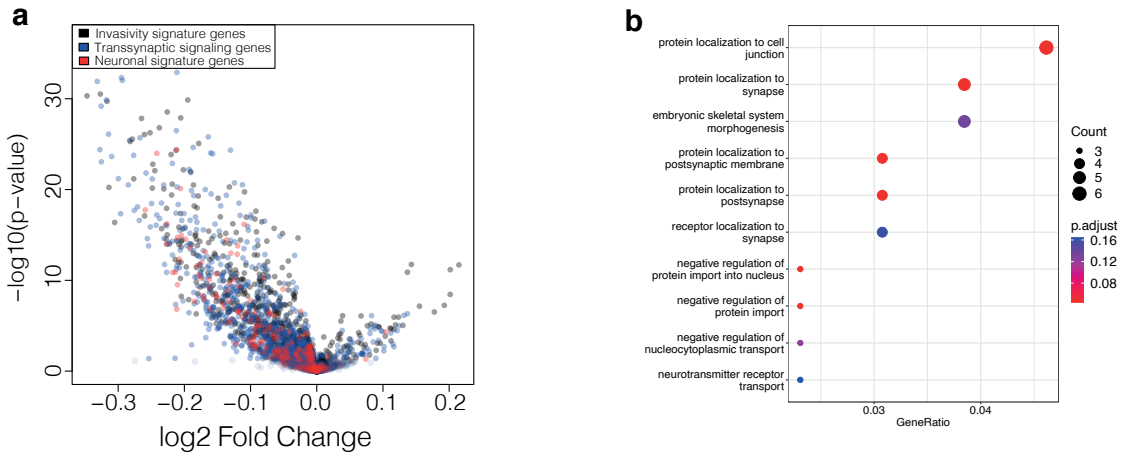


Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s41591-024-02969-w

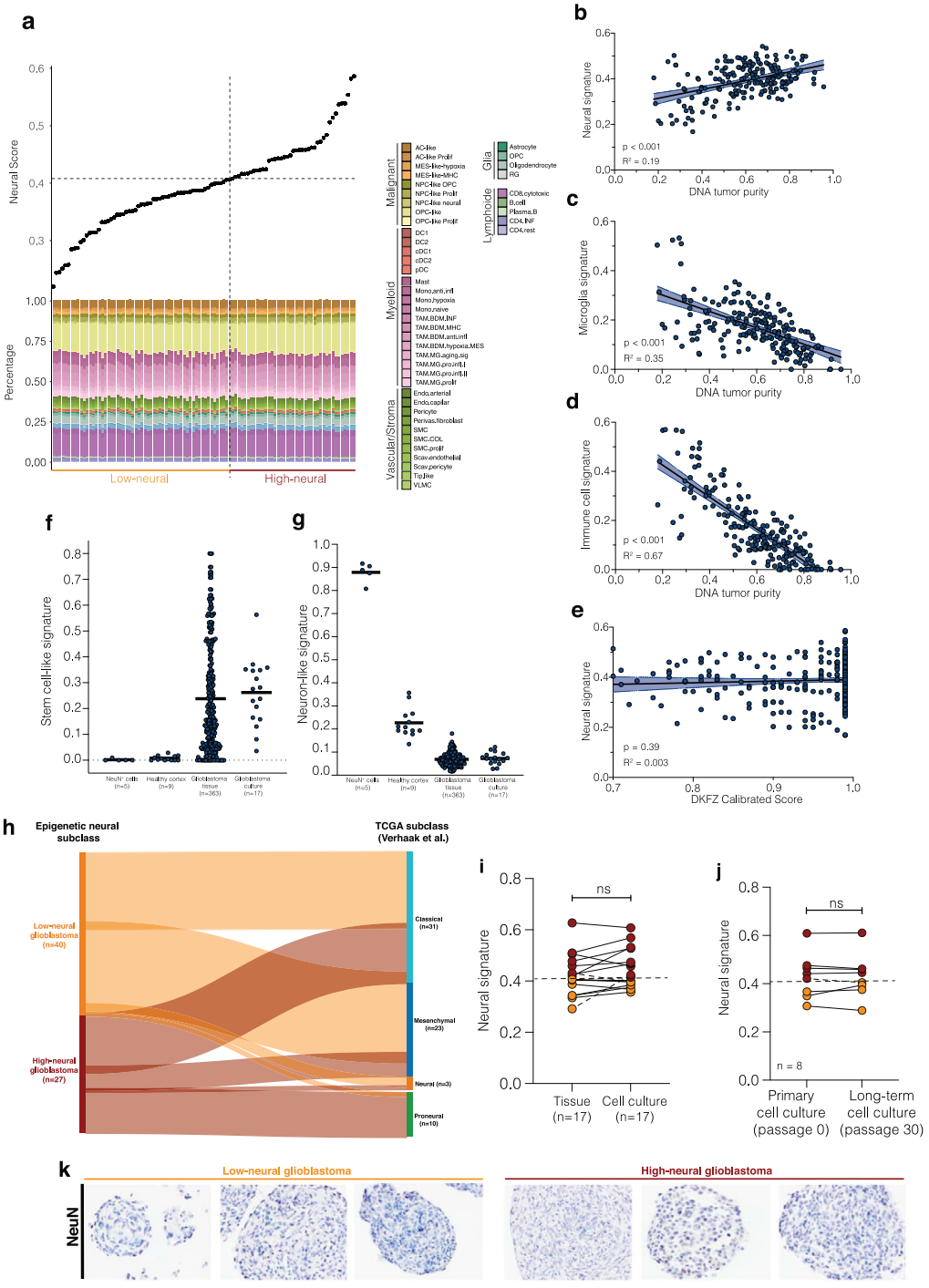
Extended Data Fig. 1 | Implementation of the epigenetic neural signature and validation of low- and high-neural subclassification of glioblastoma samples. a). Epigenetic neural signature in healthy brain tissues obtained from the Capper dataset[40]. **b, c)**. Analysis of different number of neural clusters that can predict differential survival outcome in the clinical cohort (n=363) by using 10-fold cross-validation with Kmeans. The figure displays Kaplan–Meier curves of the clusters in the validation set of the 5th fold. The survival curves demonstrate that the best results are obtained with two clusters (low- versus high-neural). Log rank test was used for the survival difference between the clusters. *Error bands representing the 95% confidence interval.* **d)**. Validation of the cut off for the neural signature across multiple cohorts used in the manuscript. Beta-values for CpGs differentially methylated between the low-neural and high-neural groups. The selection was made using the clinical cohort (n=363). **e)**. Using the clinical cohort as the training set, a logistic regression model was trained. The logistic regression model trained on the clinical cohort on the identified signature classifies across cohorts with overall AUC of 0.944 and > 0.84 in all cohorts. **f)**. Same as in e.) but a threshold on the prediction score was set (0.9) to keep only high confidence predictions. The AUC of the classifier is > 0.91 in the external cohorts when using only high probability predictions. **g, j)**. Survival analysis of patients with glioblastoma applying brain tumor-related cell signatures of the Moss signature. *Log-rank test, g) P = 0.2415, h.) P = 0.2703, i) P = 0.9010, j) P = 0.6646. Error bands representing the 95% confidence interval. OS: overall survival.*

https://doi.org/10.1038/s41591-024-02969-w



Extended Data Fig. 2 | Differentially methylated CpG sites of high- and low-neural glioblastomas. a). Volcano plot showing differentially methylated CpG sites of genes of the invasivity signature, neuronal signature, and trans-synaptic signaling signature in high-neural glioblastoma. **b)**. Gene set enrichment analysis of differentially methylated CpG sites in high-neural glioblastoma compared to low-neural glioblastoma samples.

https://doi.org/10.1038/s41591-024-02969-w

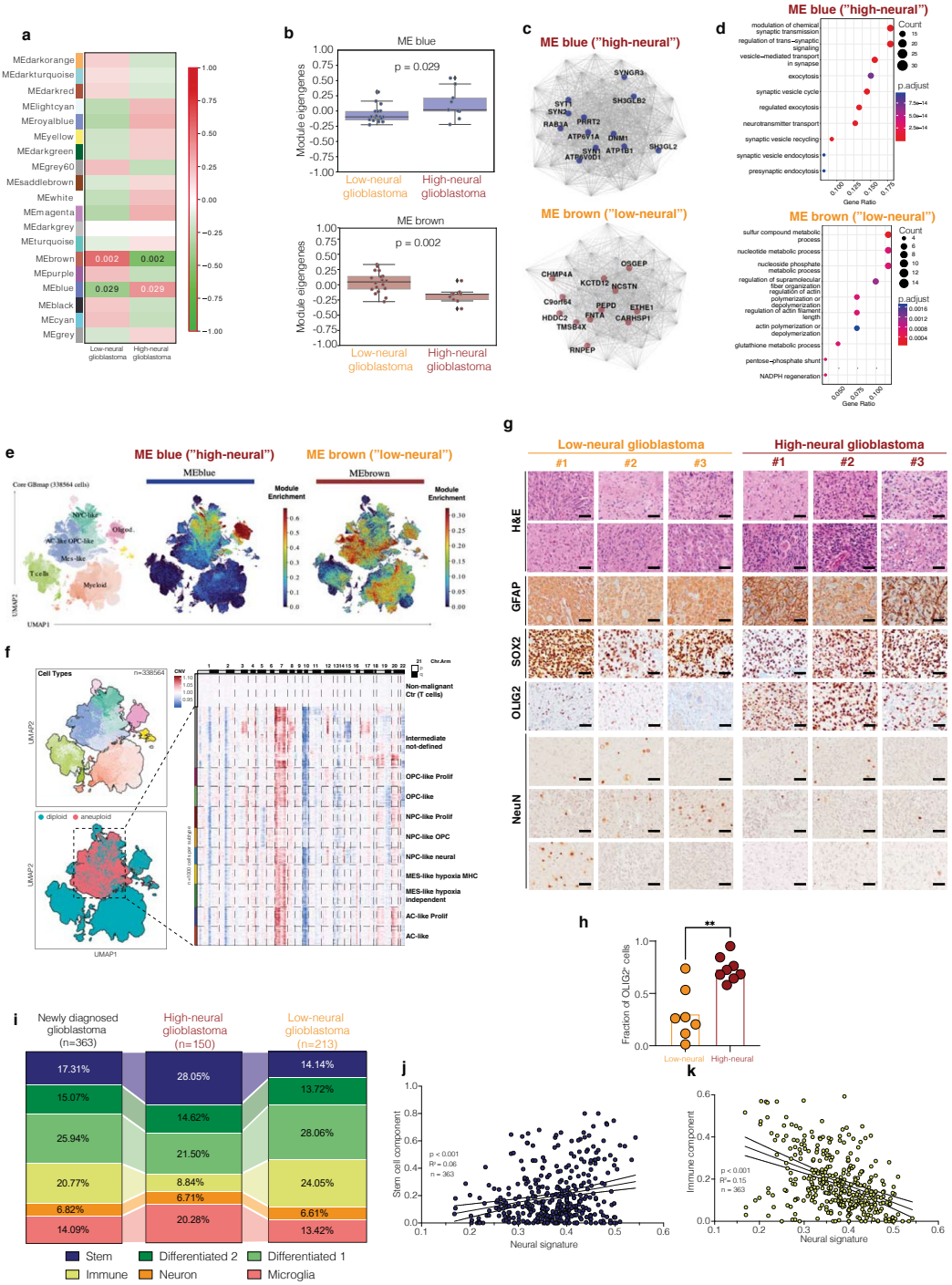


Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s41591-024-02969-w

Extended Data Fig. 3 | Quality measurements and reliability of the epigenetic neural signature in glioblastoma samples. a). Integrated analysis of the individual patients' neural scores and the corresponding cell proportions obtained from RNA sequencing deconvolution. **b**). Correlation between the neural signature and DNA tumor purity. *Simple linear regression P = 0.000000000063765, error bands representing the 95% confidence interval*. **c**). Correlation between the microglia signature and DNA tumor purity. *Simple linear regression P = 0.00000000041872, error bands representing the 95% confidence interval*. **d**). Correlation between the immune cell signature and DNA tumor purity. *Simple linear regression P = 0.000000000019814, error bands representing the 95% confidence interval*. **e**). Correlation between the DKFZ calibrated score for the diagnosis 'IDH-wild-type glioblastoma' and the neural signature. *Simple linear regression P = 0.2803, error bands representing the 95% confidence interval*. **f, g**). Single-cell deconvolution of DNA methylation profiles compare **f**). stem cell-like and **g**). neuron-like signatures in $NeuN^+$ cells, healthy cortex, glioblastoma tissue samples, and glioblastoma cell cultures. **h**). Overlap between the epigenetic neural classification and TCGA subtypes after integrated RNA sequencing analysis. **i**). Comparison of neural signature between patient's tumor tissue and cell culture in 17 glioblastomas. *Two-sided t-test P = 0.2593*. **j**). Stability of the epigenetic neural signature during long-term cell culturing. Data were obtained from a publicly available dataset (n = 6, GSE181314) and in-house (n = 1). *Two-sided t-test P = 0.8471*. **k**). Demonstration of $NeuN^+$ staining in glioblastoma neurospheres. *n=15 biological replicates*.

123
https://doi.org/10.1038/s41591-024-02969-w

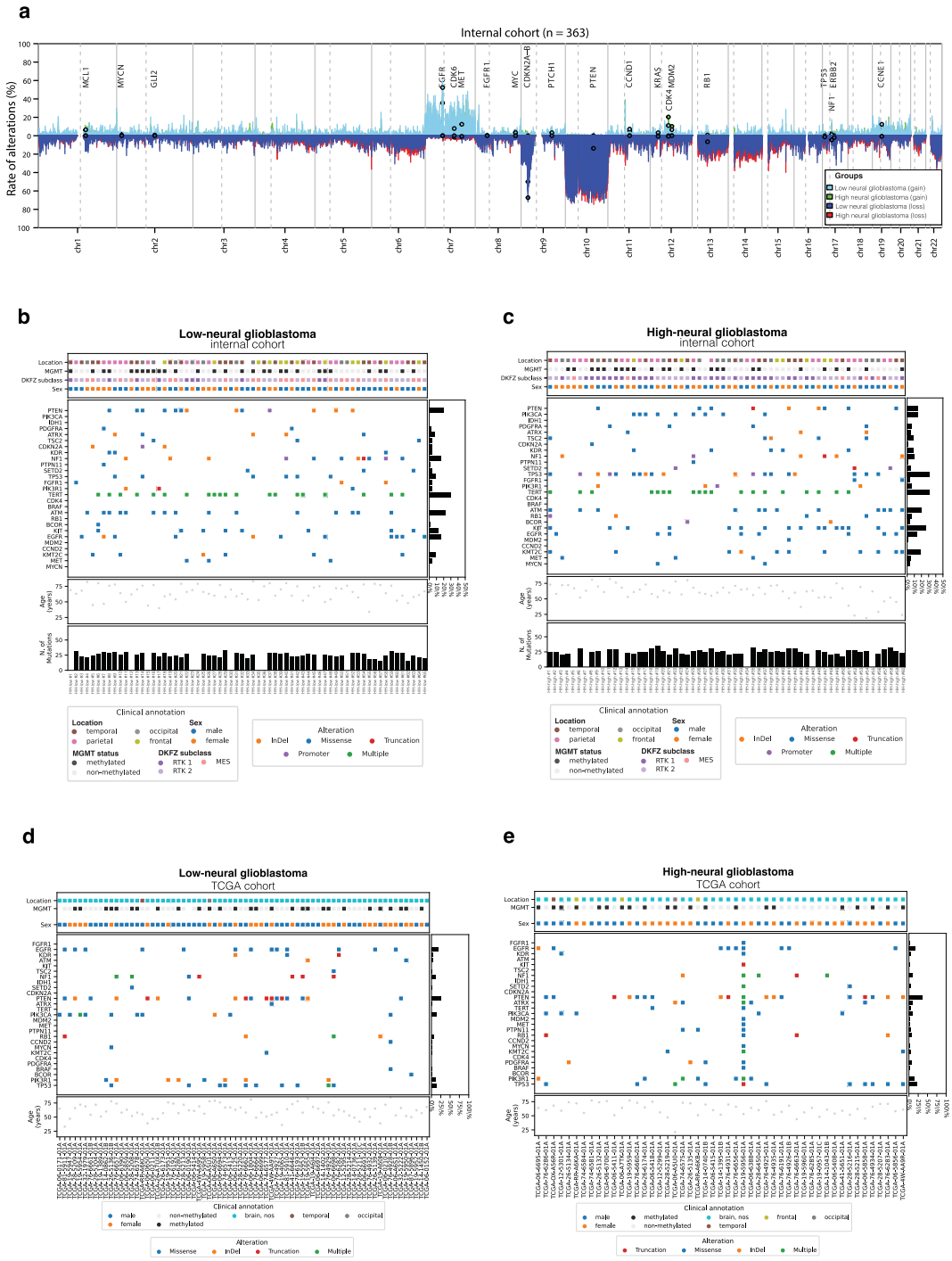


Extended Data Fig. 4 | See next page for caption.

https://doi.org/10.1038/s41591-024-02969-w

Extended Data Fig. 4 | High-neural glioblastoma is linked with synapse formation and trans-synaptic signaling from proteomic profiling. **a – e) Proteomic profiling of low- (n=19) and high-neural (n=9) glioblastoma. a)**. WGCNA analysis showed differentially abundant proteome modules between both neural subgroups. **b)**. High-neural glioblastomas are clustered to module 'blue' (top figure), while low-neural glioblastomas have a higher abundance in module 'brown' (bottom figure). Data are mean ± s.e.m. Two-sided t-test P = 0.0.029 (top figure) and P = 0.002 (bottom figure). **c, d)**. Network analysis revealed **e)**. most expressed proteins and **f)**. associated gene ontology terms for each neural subgroup (high-neural: top, low-neural: bottom). **e)**. Integrating transcriptomic single-cell data showed an OPC-/NPC-like character in high-neural tumors ('ME blue'). **f)**. Transcriptomic single-cell copy number variation plot analysis of glioblastomas with a high-neural signature. **g)**. Immunohistostaining of representative low- and high-neural glioblastoma samples. n=10 biological replicates. **h)**. Analysis of $OLIG2^+$ cells between low- and high-neural glioblastoma samples. **P < 0.01, two-tailed Student's t-test. **i)**. Comparison of abundance of cell states analyzed by reference-free deconvolution between newly diagnosed, high-neural, and low-neural glioblastomas. **j)**. Stem cell-like state significantly correlated with an increase of the neural signature in glioblastoma samples. Simple linear regression, P = 0.000003024480. Error bands representing the 95% confidence interval. **k)**. An anticorrelation was seen between the abundance of the immune compartment and the neural signature. Simple linear regression, P = 0.000000000005. Error bands representing the 95% confidence interval.

https://doi.org/10.1038/s41591-024-02969-w

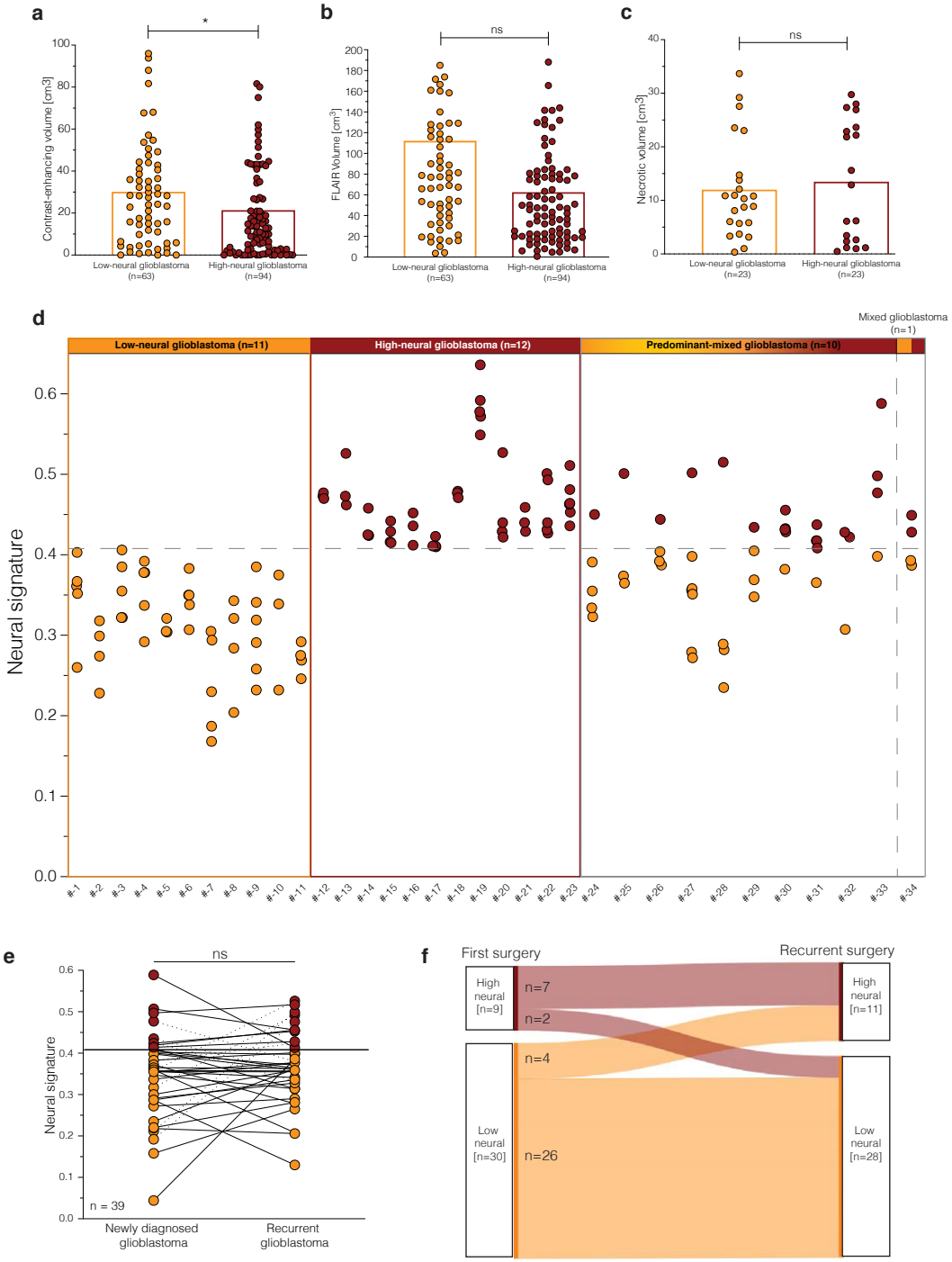


Extended Data Fig. 5 | See next page for caption.

https://doi.org/10.1038/s41591-024-02969-w

Extended Data Fig. 5 | Copy number variations and next-generation sequencing of gene mutations between low- and high-neural glioblastoma samples. a). Copy number variation plots for all samples stratified into low- and high-neural glioblastoma. **b, c)**. Oncoprint illustrating clinical characteristics and gene mutational status of **b)**. low-neural and **c)**. high-neural glioblastoma samples of our internal cohort. Of note, rarely detectable IDH mutations did not include the pathogenic R132H mutation. **d, e)**. Oncoprint illustrating clinical characteristics and gene mutational status of **d)**. low-neural and **e)**. high-neural glioblastoma samples of the TCGA dataset.

https://doi.org/10.1038/s41591-024-02969-w

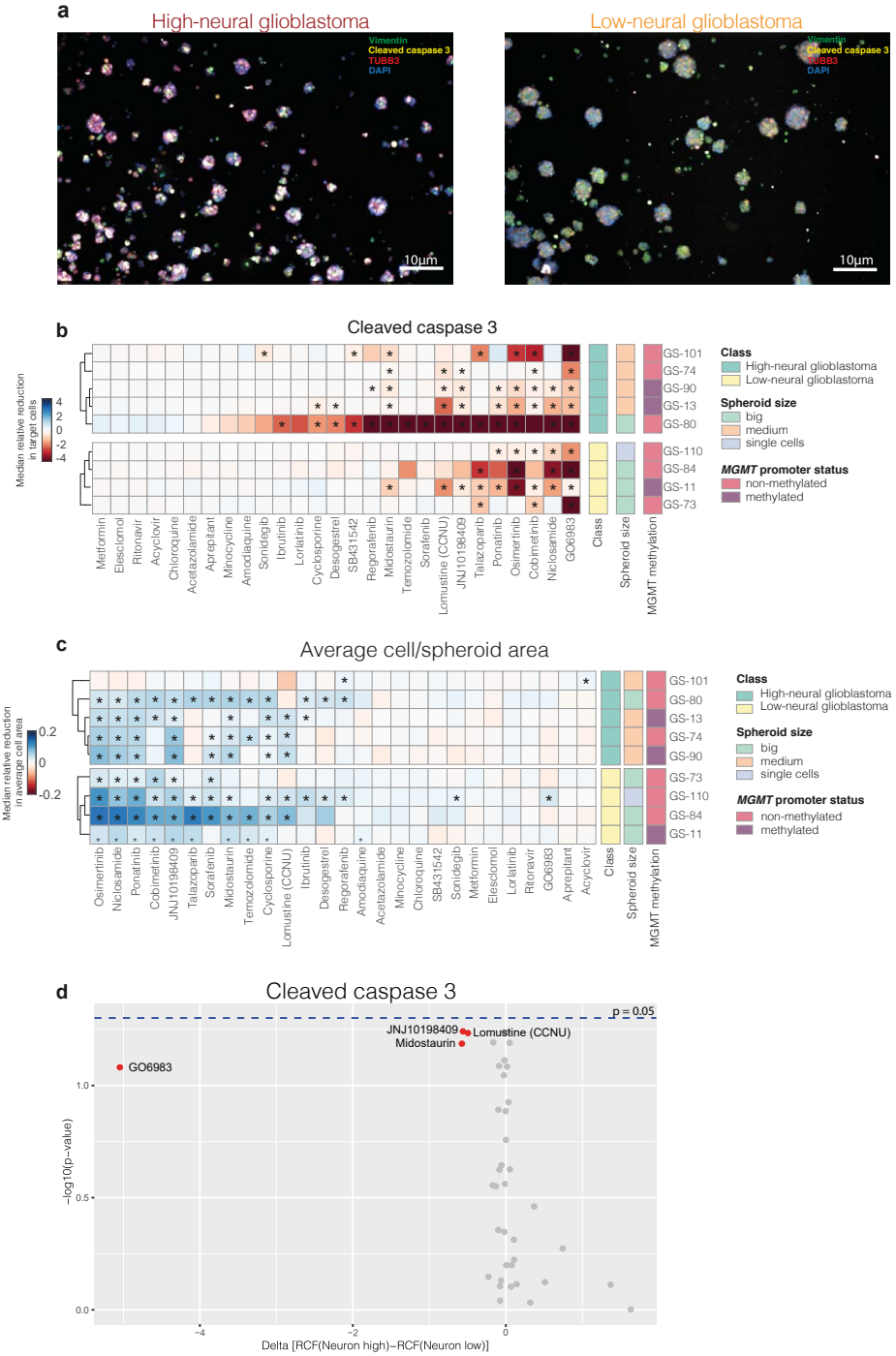


Extended Data Fig. 6 | See next page for caption.

https://doi.org/10.1038/s41591-024-02969-w

Extended Data Fig. 6 | Radiographic parameters and spatiotemporal tumor sampling. a – c). Association of neural glioblastoma group with volume of **a).** contrast enhancement, **b).** FLAIR, and **c).** tumor necrosis measured by preoperative magnetic resonance imaging. A) $P = 0.0374$, b) $P = 0.1767$, and c) $P = 0.6373$, two-tailed Student's t-test. **d).** Analysis of intertumoral difference of neural signature within 34 newly diagnosed glioblastomas with spatial collection of 3 to 7 samples per tumor. 23 (67.6 %) of these tumors had a pure low- or high-neural signature in all individual biopsies with additional 10 (29.4 %) tumors being predominantly low or high. **e).** Neural signature in 39 patients with matched tumor tissue obtained from surgery at first diagnosis and recurrence. ns: $P > 0.05$, two-tailed Student's t-test. **f).** Sankey plot illustrating a potential switch of the neural subgroup between first diagnosis and recurrence.

129
https://doi.org/10.1038/s41591-024-02969-w

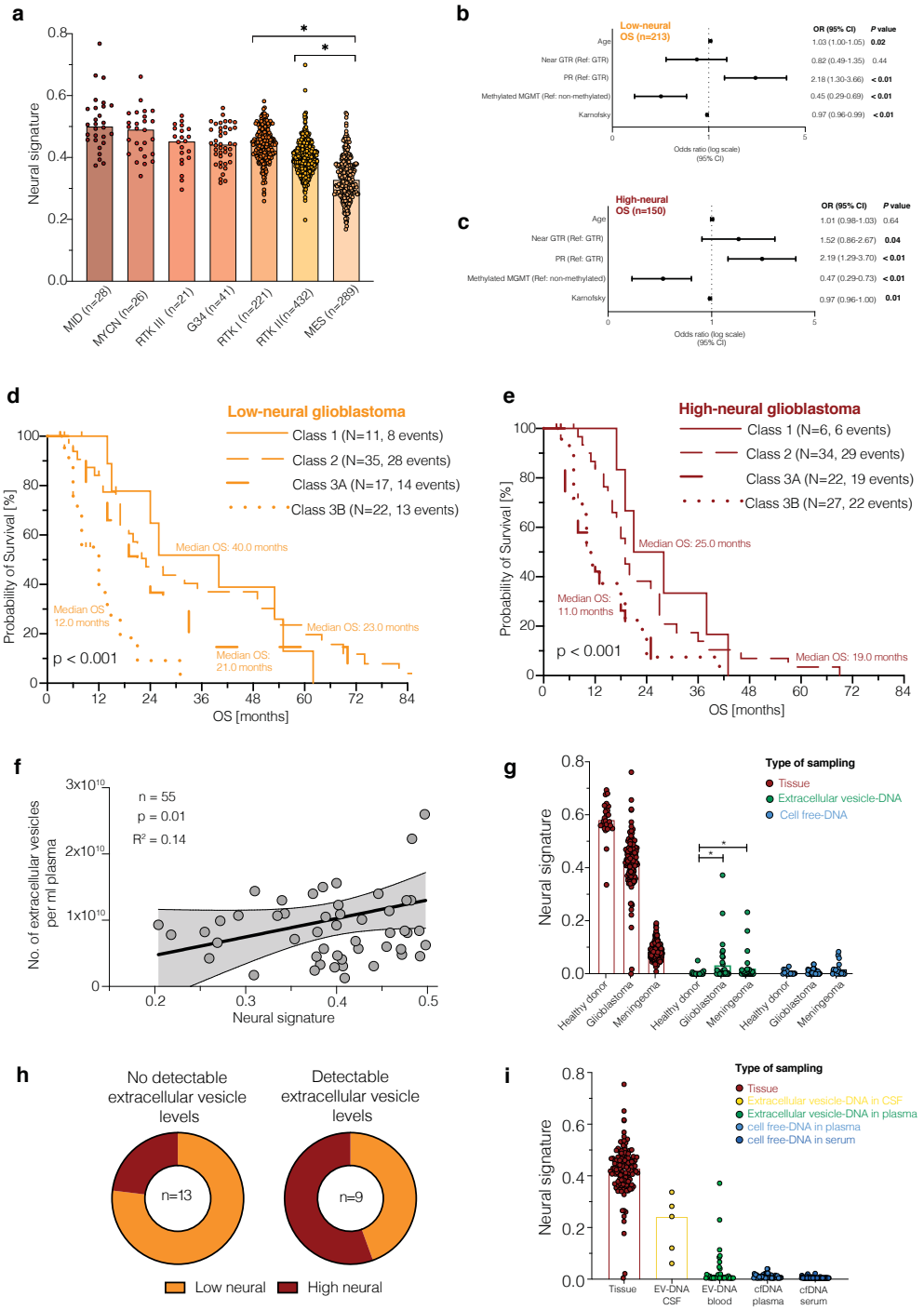


Extended Data Fig. 7 | See next page for caption.

https://doi.org/10.1038/s41591-024-02969-w

Extended Data Fig. 7 | Drug sensitivity analysis of low- and high-neural glioblastoma cells. a). Representative microscopic images for high- (left image) and low-neural (right image) glioblastoma cells. Green: Vimentin, yellow: cleaved caspase 3, TUBB3: red, DAPI: blue. Scale bars: 10µm. n=9 biological replicates. **b)**. Drug sensitivity of low- and high-neural glioblastoma cells measured by cleaved caspase 3. *$P < 0.05$, Mann–Whitney test. **c)**. Drug sensitivity of low- and high-neural glioblastoma cells measured by average cell area. *$P < 0.05$, Mann–Whitney test. **d)**. Statistical difference of sensitivity to various drugs between low- and high-neural glioblastoma cells. Mann–Whitney test.

131
https://doi.org/10.1038/s41591-024-02969-w

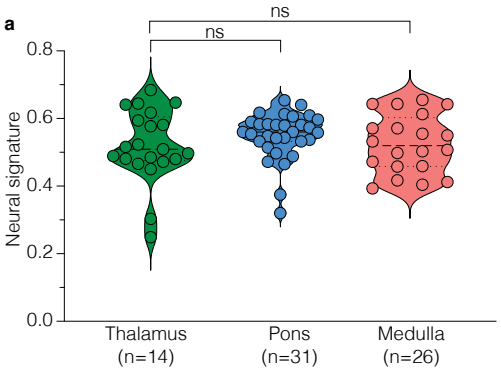


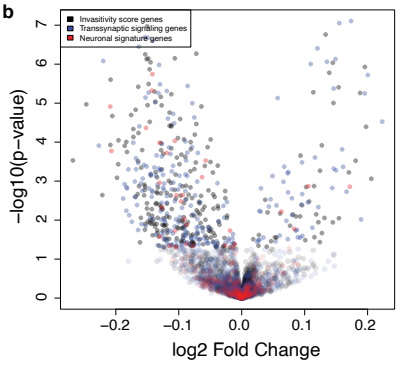
Extended Data Fig. 8 | See next page for caption.

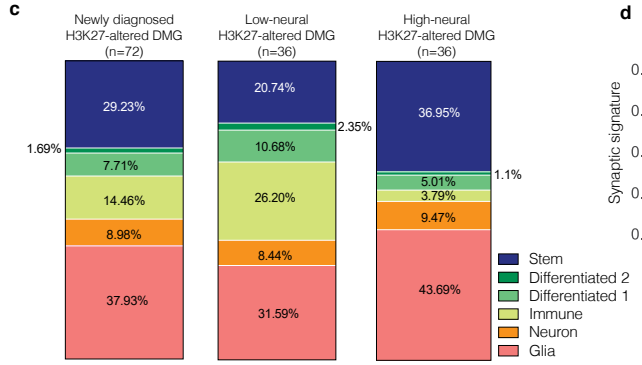
https://doi.org/10.1038/s41591-024-02969-w

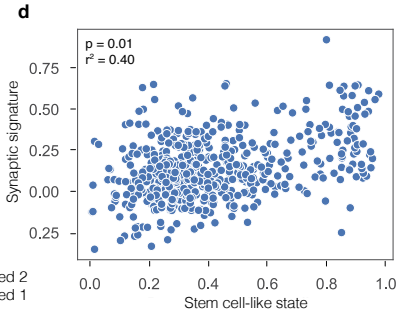
Extended Data Fig. 8 | Clinical prognostic and circulating biomarkers of epigenetic neural glioblastomas. a). Neural signature in DNA methylation subclasses of newly diagnosed IDH-wild-type glioblastoma. *P < 0.05, two-tailed Student's t-test. **b)**. Forest plot illustrating the multivariate analysis of low-neural patients with glioblastoma. Means are shown by closed circles and whiskers representing the 95% confidence interval. Cox proportional hazards regression model. **c)**. Forest plot illustrating the multivariate analysis of high-neural patients with glioblastoma. Means are shown by closed circles and whiskers representing the 95% confidence interval. Cox proportional hazards regression model. **d – e)**. Survival outcome categorized after RANO categories for extent of resection in patients with glioblastoma treated by radiochemotherapy with a low- and high-neural signature. Class 1: 0 cm^3 CE + ≤5 cm^3 nCE tumor, Class 2: ≤1 cm^3 CE, Class 3A: ≤5 cm^3 CE, Class 3B: ≥5 cm^3. Log-rank test, d) $P = 0.0002$, and e) P = 0.0011. **f.)** Correlation of neural signature and number of extracellular vesicles in patient serum at time of diagnosis. Simple linear regression P = 0.01. Error bands representing the 95% confidence interval. **g.)** Comparison of neural signature in healthy individuals, patients with glioblastoma, and meningeoma patients between matched tumor tissue, extracellular vesicle-associated DNA in serum, and cell-free DNA in serum. *$P < 0.05$, two-tailed Student's t-test. **h.)** Comparison of patients with no detectable (left panel) and detectable (right panel) extracellular vesicle levels in serum stratified to their epigenetic neural glioblastoma type. **i.)** Illustration of the neural signature in different types of sampling in patients with glioblastoma.

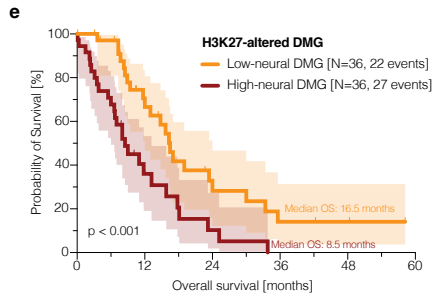
133
https://doi.org/10.1038/s41591-024-02969-w

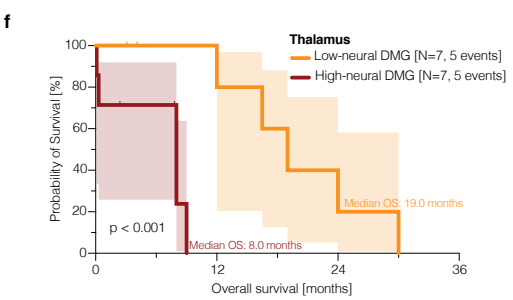


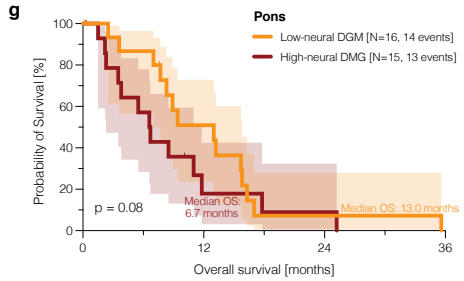


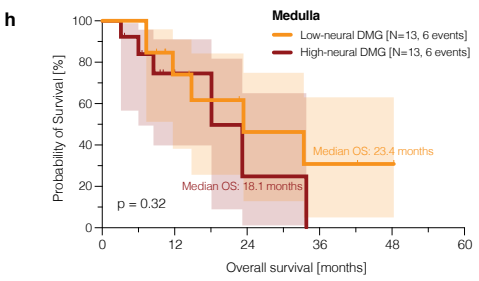












Extended Data Fig. 9 | See next page for caption.

https://doi.org/10.1038/s41591-024-02969-w

Extended Data Fig. 9 | Relevance of neural classification in pediatric and adolescent patients diagnosed with *H3K27*-altered diffuse midline glioma (DMG). a). Association of tumor location with neural signature. Two-tailed Student's t-test. **b)**. Volcano plot showing differentially methylated CpG sites of genes of the invasivity signature, neuronal signature, and trans-synaptic signaling signature. **c)**. Cell state composition analysis in low- and high-neural DMG. **d)**. Synaptic gene expression (*PTPRS*, *ARHGEF2*, *GRIK2*, *DNM3*, *LRRTM2*, *GRIK5*, *NLGN4X*, *NRCAM*, *MAP2*, *INA*, *TMPRSS9*)[6] is significantly correlated with the stem cell-like state of DMG cells calculated by an overlap of single-cell DNA methylation and single-cell RNA sequencing (599 cells from 3 study participants) measurements. Simple linear regression. **e** – **h)**. Kaplan–Meier survival analysis of 72 DMG patients under 18 years of age with a low- and high-neural DMG. Error bands representing the 95% confidence interval. Log-rank test, e) P = 0.0017, f) P = 0.0022, g) P = 0.0882, and h) P = 0.3236.

Extended Data Table 1 | Clinical Characteristics

Characteristic	N	Low-neural glioblastoma (n=213)	High-neural glioblastoma (n=150)	*P* value
Age, mean (SD), years	61.4 (10.0)	60.9 (10.2)	61.8 (9.7)	0.41
Sex, n (%)				
Female	139 (38.3)	83 (39.0)	56 (37.3)	0.83
Male	224 (61.7)	130 (61.0)	94 (62.7)	
Location, n (%)				
Frontal	106 (29.2)	68 (31.9)	38 (25.3)	0.19
Parietal	145 (39.9)	80 (37.6)	65 (43.3)	0.31
Temporal	141 (38.8)	71 (33.3)	70 (46.7)	**0.02**
Occipital	55 (15.2)	30 (14.1)	25 (16.7)	0.55
Hemisphere, n (%)				
Left	166 (45.7)	99 (46.5)	67 (44.7)	0.55
Right	174 (47.9)	103 (48.4)	71 (47.3)	
Both	23 (6.3)	11 (5.2)	12 (8.0)	
Karnofsky prior surgery, mean (SD), %	84.6 (12.4)	83.0 (12.9)	86.7 (11.4)	**< 0.01**
Extent of resection, n (%)				
Gross total	142 (39.1)	92 (43.2)	50 (33.3)	0.08
Near gross total	99 (27.3)	59 (27.7)	40 (26.7)	
Partial	122 (33.6)	62 (29.1)	60 (40.0)	
***MGMT* promoter methylation status, n (%)**				
Non-methylated	174 (47.9)	107 (50.2)	67 (44.7)	0.38
Methylated	189 (52.1)	106 (49.8)	83 (55.3)	
Karnofsky prior adjuvant treatment, mean (SD), %	85.4 (12.7)	84.4 (13.0)	86.7 (12.1)	0.09

Clinical characteristics of patients with glioblastoma who were treated with combined radio chemotherapy after surgical resection. SD: standard deviation, MGMT: O6-methylguanine-DNA-methyltransferase.

5

DISCUSSION

The individual findings of each of the three works are discussed in detail as part of each publication presented in chapters 2-4. In this chapter, we situate the findings of these works within the context of computational biology, immunology, and neuroscience with a focus on open questions and potential works.

5.1. CELL TYPE DECONVOLUTION, CHALLENGES AND POTENTIAL DIRECTIONS

As detailed in Chapter 2, there have been several works in cell type deconvolution - both the estimation of cell type proportions and gene expression profiles. Cell type deconvolution benefits from the bulk studies with larger cohorts in comparison to the single-cell studies. Further, bulk and single-cell studies in combination provide a better understanding of tissue heterogeneity and architecture (George et al., 2024; Zeng et al., 2024). In our work with DISSECT, we used the simple formulation of the cell deconvolution task as described in Chapter 2 and extended it to enable semi-supervised learning with gene expression profiles from both the bulk and single-cell reference. The method achieves a consistent performance across a range of experiments. Some challenges and open questions remain as discussed below:

5.1.1. BETTER AND CONSISTENT CELL TYPE DEFINITIONS

In DISSECT, we relied on cell type definitions based on molecular profiles, where cell types are defined by some marker genes that show preferential enrichment in a given cell type (Zeng, 2022; Domínguez Conde et al., 2022). Cell type definitions based on molecular profiles may differ across single-cell studies and may convolute the insights gained from cell type deconvolution. In our experiment with the estimation of granular level cell types, we harmonize the cell type annotations from two sources based on cell type definition similarities (e.g. mapping different memory CD4 T cell subsets to form a single cluster memory CD4 T cells). This limits the evaluation of algorithms on specific subsets. Efforts into homogeneous cell ontologies and cell type marker genes may help alleviate this problem (Bernstein et al., 2021; Börner et al.). The human cell atlas is a prominent example of this direction (Osumi-Sutherland et al., 2021). Further, the validation of novel cell types identified in previously published studies consisting of large-scale atlases may help in producing good training data for reference, which is a prominent challenge in cell type deconvolution as argued in Garmire et al. (2024). In a similar direction, accessibility to good quality reference data for cell type deconvolution can be aided by integrating single-cell data portals, e.g. CellXGene (Megill et al., 2021) with the deconvolution software for ease of choosing appropriate reference by tissue, disease, and organism.

From a technical perspective, this challenge directly impacts the learning framework of DISSECT's neural architecture since the consistency regularization framework assumes that marker genes remain invariant across conditions. Further, in a semi-supervised setting where some real bulk data with cell type fractions is available, inconsistent cell type definitions across references can violate this assumption, potentially increasing the effective dimension of the learning problem beyond our theoretical $\mathcal{O}(\sqrt{nc})$ bound. Future work could explore adaptive architectures that learn robust representations across different annotation schemes, perhaps by incorporating hierarchical cell ontologies directly into the network structure.

5.1.2. INCORPORATION OF ADDITIONAL INFORMATION FOR THE MIXTURES

In DISSECT, we incorporate gene expression information from the bulk data, by generating mixtures of bulk samples and samples simulated from a scRNA-seq reference. This allows for a seamless training procedure that is capable of learning from real gene expression as well. DISSECT, thus, does not rely only on the simulated data that borrows from the limitations of single-cell data such as the presence of dropouts (Lähnemann et al., 2020). This also allows better deconvolution from a mismatched reference such as the case when reference data is from a different condition. Previously, in MuSiC2

(Fan et al., 2022), authors proposed to train on multi-condition references to solve this problem. However, it is unlikely that for a desired condition, the single-cell reference is always available.

While in bulk RNAseq, observations are limited to the gene expression, spatial transcriptomics data such as Visium from 10x genomics, provides additional information such as the location of each sequenced mixture as well as the hematoxylin and eosin (H&E) image staining of the underlying tissue. However, the algorithms found best in our comparisons as well as in a benchmarking study from 2022 do not take either of these additional information into account. In contrast, algorithms like SONAR (Liu et al., 2023) and CARD (Ma and Zhou, 2022) that take into account the mixture locations do not outperform these algorithms and DISSECT. This seems counterintuitive as there is information embedded in the additional information about tissue as given by mixture-locations and pathology images. We believe that this observation is a multimodal learning problem rather than a lack of information (Gao et al., 2020). Further works into incorporating location and image information in frameworks like DISSECT have the potential to improve over current spatial transcriptomics deconvolution state-of-the-art and could also provide a more holistic landscape of cell type heterogeneity in tissues.

From an algorithmic perspective, extending DISSECT's framework to incorporate spatial and imaging data presents interesting theoretical challenges. The current consistency regularization framework relies heavily on the linear mixing assumption, which may not hold directly for spatial relationships. One potential direction is to extend our theoretical framework to include manifold regularization that preserves spatial structure. An alternative would also be to enforce the consistency term locally in selected neighborhoods, such as achieved by adding a spatial consistency term:

$$L_{spatial} = ||f(B_i^{mix}) - \sum_{j \in \mathcal{N}(i)} w_{ij} f(B_j)||^2,$$

where $\mathcal{N}(i)$ represents spatial neighbors and w_{ij} are spatial weights. The challenge lies in proving that such additional constraints maintain spatial coherence and in demonstrating that a model can be trained to achieve such objectives without compromising on an overall prediction quality.

5.1.3. ESTIMATION OF RARE POPULATIONS

In our experiments with DISSECT, we found that DISSECT identifies rare cell type populations (<1% true abundance) with better accuracy compared to other tested algorithms. However, Dissect has a high relative error rate when compared to the ground truth. The potential causes for this observation likely stem from the low enrichment of markers for these cell types in the bulk expression profiles. There is also a lack of enough cells to capture within-cell type heterogeneity, marker gene consensus, and cell type annotation artifacts in the single-cell reference (Cheng et al., 2023; Garmire et al., 2024). Further work into careful cell type annotation while preparing references as well as utilizing large single cell datasets with enough heterogeneity may provide more trustworthy results from DISSECT even on these rare cell populations.

The challenge of rare cell types presents an interesting theoretical tension with our current framework. While DISSECT's consistency regularization helps reduce sample complexity, rare populations may require more samples than our $O(\sqrt{nc})$ bound suggests for reliable estimation. This connects to fundamental questions in learning theory about the sample complexity required for tail estimation. Future work could explore adaptive weighting schemes in the loss function:

$$L_{weighted} = \sum_{k=1}^{c} w_k L_k,$$

where w_k could be inversely proportional to cell type frequency, potentially with theoretical guarantees for rare population estimation.

5.1.4. NOVEL BENCHMARKS FOR GENE EXPRESSION ESTIMATION

In evaluating deconvolution algorithms for the task of cell type-specific gene expression estimation, we relied only on simulations. Real bulk data, as we have argued in the DISSECT manuscript is a different modality than the simulations from a single-cell reference. An appropriate experimental setup would be a paired generation of purified cell populations and bulk data from the same or adjacent tissue samples. This would allow direct benchmarking by comparing the predicted expression profile with the measured expression profiles of the purified cell populations. In the case of spatial transcriptomics data, this is possible now to some extent with paired Xenium and Visium breast cancer data (Janesick et al., 2023). However, the number of samples and tissue diversity are still limited.

From a methodological perspective, the lack of ground truth for gene expression estimation raises questions about the validity of our theoretical guarantees in practice. While our consistency regularization framework provides formal guarantees for preservation of distances between cell type signatures, these guarantees assume the training distribution matches the test distribution. Future work could explore domain adaptation techniques that explicitly account for the shift between simulated and real data, perhaps by incorporating ideas from optimal transport theory to minimize the distribution shifts (Courty et al., 2016, 2017).

5.2. Ustekinumab as a treatment for ANCA-GN, and cell type abundance and heterogeneity in glomerulonephritis

As detailed in Chapter 3, despite significant therapeutic advancements, patients with ANCA-GN continue to face substantial risks of kidney failure and mortality. ANCA-GN is a severe autoimmune disease characterized by the production of autoantibodies targeting neutrophil proteins, leading to inflammation and damage in small blood vessels, particularly in the kidneys (Jennette and Falk, 2013). The primary causes of death in ANCA-GN patients - infections, cardiovascular disease, and malignancies - are often linked to the use of non-specific immunosuppressive treatments. This underscores the critical need for a balanced therapeutic approach that effectively controls the disease while minimizing potentially life-threatening side effects.

Current standard therapies for ANCA-GN typically involve a combination of high-dose corticosteroids and cyclophosphamide or rituximab for induction of remission, followed by maintenance therapy with azathioprine or rituximab (Yates et al., 2016). While these treatments have improved outcomes for many patients, they are associated with significant toxicities and do not address the underlying pathogenic mechanisms specific to ANCA-GN. This highlights the urgent need for more targeted therapies that can effectively control disease activity while reducing the risk of treatment-related complications.

5.2.1. Multi-Omics Approach for ANCA-GN Treatment

To address these challenges and gain deeper insights into the pathogenesis of ANCA-GN, we employed advanced multi-omics techniques and machine learning algorithms to study ANCA-GN patients. Our comprehensive approach integrated scRNA-seq and

spatial transcriptomics from Visum (from 10x Genomics) to provide a multidimensional view of the disease tissue environment.

Our findings suggested a significant accumulation of pro-inflammatory T cells, particularly T helper 1 (Th1) and T helper 17 (Th17) cells, in affected kidney areas. This observation is consistent with previous studies suggesting a crucial role for these T cell subsets in the pathogenesis of ANCA-GN (Gan et al., 2010; Krebs et al., 2016). The presence of these cells in inflamed glomerular and tubulointerstitial tissue compartments indicates their potential involvement in both the initiation and progression of kidney damage in ANCA-GN.

In addition to Th1 and Th17 cells, we observed an increased abundance of other memory T cell populations, including natural killer T (NKT) cells and T follicular helper (Tfh) cells. These findings suggest a complex interplay of various T cell subsets in the pathogenesis of ANCA-GN. Furthermore, we noted an increased number of macrophages in affected kidney areas, consistent with the known role of these cells in mediating tissue damage and fibrosis in glomerulonephritis (Guiteras et al., 2016).

Through integrative analysis of single-cell and spatial transcriptomics data, we identified ustekinumab as a potential treatment option for ANCA-GN. Ustekinumab is a monoclonal antibody that targets the p40 subunit shared by interleukin-12 (IL-12) and interleukin-23 (IL-23), cytokines crucial for the differentiation and maintenance of Th1 and Th17 cells, respectively (Teng et al., 2015). By inhibiting these pathways, ustekinumab could potentially modulate the pro-inflammatory T cell response observed in ANCA-GN patients, offering a more targeted therapeutic approach.

5.2.2. Cell type architecture and glomerulonephritis

While our study provides valuable insights into the cellular and molecular landscape of ANCA-GN, several limitations and areas for future research should be noted. First, the exact cell type proximities and niches within glomerular and tubulointerstitial compartments remain unclear due to the limited resolution of the Visium spatial transcriptomics data. To address this limitation, employing newly developed approaches such as Xenium or CosMx may provide a better understanding of cell type composition and networks at a higher resolution (Abedini et al., 2024).

A more detailed spatial analysis is particularly necessary for understanding glomerular crescent staging in ANCA-GN. Identifying glomeruli in different stages of crescent formation would help in constructing crescent pseudotime trajectories similar to single-cell data and has the potential to interrogate genes and pathways involved in glomerular

crescent progression (Isnard et al., 2024). Due to the supra-cellular nature of Visium data, we were unable to investigate the involvement of individual glomerular cell types such as parietal epithelial cells, mesangial cells, and podocytes in the crescent formation. Future studies using higher-resolution spatial transcriptomics techniques could provide valuable insights into the roles of these specific cell types in ANCA-GN pathogenesis (Sultana et al., 2024).

Our observation of increased fibronectin expression, elevated inflammatory signatures, and associated interferon pathways presents an interesting but non-specific finding. With single-cell information, future research has the potential to go deeper into identifying the key glomerular and immune cell type players involved in triggering crescent formation. This would also allow for better targeting of these cell types and pathways, potentially enabling the rescue of glomeruli in the early stages of crescent formation.

5.2.3. Comparative analysis of glomerulonephritis categories

ANCA-GN represents a particular pathogenesis of glomerulonephritis, but it is important to consider how it compares to other categories of GN. For instance, infection-related GN, where the cause of active GN is an underlying infection, leads to adaptive immune responses against the pathogen antigen (Anders et al., 2023). It remains unclear how glomeruli differ molecularly across various GN categories and which pathways are shared among them.

To address this knowledge gap, a comprehensive high-dimensional multimodal atlas covering multiple GN categories is warranted. Such an atlas could allow for better treatment proposals for different types of GN, similar to our ANCA-GN study. Furthermore, it is important to note that GN categories based solely on the pathology of the lesions may not offer sufficient information about molecular differences in these diseases (Lerner et al., 2021; Smith et al., 2022). Thus, it is necessary to compare these diseases beyond their pathology and identify dysregulated pathways and cell types.

5.3. Glioblastoma (GBM) heterogeneity

In our study with multi-omics analysis of glioblastoma, presented in chapter 4), we observed two distinct groups of glioblastoma based on DNA methylation across many cohorts. These groups, low- and high-neural, differ in their DNA methylation and transcriptomic profiles. The differentially methylated CpGs on gene promoter sites were enriched for synaptic pathways and integration with scRNA-seq highlighted an neural progenitor cells (NPC) and oligodendrocyte progenitor cells (OPC)-like gene expression

profile. This study highlights the heterogeneity of the GBM tumor microenvironment. The neural signature and its predictive significance are not limited to glioblastoma; similar patterns can be found in H3 K27-altered diffuse midline gliomas. Such broad applicability underlines the plausibility of mechanistically important effects of neural integration across several aggressive brain neoplasms.

5.3.1. Role of immune cells and oligodendrocyte-like cells

High-neural GBMs significantly express a phenotype of malignant stem cells that defines them as closely congruent with NPC or OPC. A surprising observation is the very low presence of immune cells in these tumors, which may indicate the existence of possible mechanisms for immune evasion. In contrast, we found low-neural GBMs to be immune-enriched based on our analysis of transcriptomics and cell state composition. This dichotomy suggests the existence of two opposing glioblastoma subtypes, each potentially requiring distinct therapeutic approaches. These differences in immune-enrichment are crucial for evaluating treatment options in these tumors using immunotherapy.

The high-neural GBMs exhibit upregulation and reduced methylation of genes associated with invasiveness and the formation and signaling of neuron-to-glioma synapses. This specific molecular profile indicates the mechanisms underlying the assimilation of these tumors into neural circuits (Venkataramani et al., 2019, 2022). The observation of transcriptomic oligodendrocyte modules in high-neural GBMs may reflect the influence of surrounding healthy oligodendrocyte progenitor cells on neuronal activity-driven mechanisms affecting glioma cells or the interplay between immune cells and oligodendrocytes (Hide et al., 2019; Moore et al., 2015).

5.3.2. Origin of cells implicated in high-neural GBM

Although in our work we observed consistent detection of two neural subgroups of GBM across multiple cohorts, we could not comprehensively characterize them. First, the term "neural" here is broad and may refer to GBM cells with profiles similar to various neural lineage cell types, including NPCs, OPCs, and astrocyte precursor cells (APCs). While high-neural GBMs show increased NPC or OPC phenotype, it is also possible that the total neural abundance may reflect a combination of these cell type profiles.

Given that with cell type deconvolution at the mRNA level, we could not find differences in the proportion of these cell types between the two neural groups, this may indicate either upstream changes or subtle differences that are not detectable at the mRNA level, particularly in bulk RNA-seq. This relates to limitations detailed in section 5.1.3, particularly regarding the estimation of rare cell populations. Further, studying DNA methylation with

greater granularity would require generation and analysis of single-cell DNA methylation data or the generation of DNA methylation reference profiles from diverse neural cell lineages and construction of more expansive cell type signatures to be used in cell type deconvolution.

5.3.3. Clinical implications of the neural subgroups of glioblastoma

In patients with high-neural GBMs, increased concentrations of brain-derived neurotrophic factor (BDNF) are noted. This observation indicates the potential utility of BDNF as a biomarker for the identification and surveillance of high-neural gliomas. BDNF is a neurotrophin whose expression is regulated by neuronal activity and has recently been described to exert functions that promote the growth of gliomas (Radin and Patel, 2017). Inhibiting BDNF-TrkB signaling pathway is considered an attractive therapeutic target for IDH-wildtype GBM (Taylor et al., 2023). Our research makes some key additions regarding BDNF in the context of high-neural glioblastoma: First, we observed elevated serum BDNF levels in adult patients with high-neural tumors, suggesting its potential use as a biomarker for this tumor subtype. Increased BDNF levels could originate from neurons in a glioma-induced hyperexcitable state or from the glioblastoma cells themselves, as a subpopulation of these cells express and secrete BDNF as clear in the integrated scRNA-seq data from glioblastoma and healthy brain, presented in our study. Consistent with preclinical models, high serum levels of BDNF corresponded to a greater seizure frequency within the glioblastoma population. This effect is consistent with the role of BDNF in regulating α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptor trafficking and upregulation of AMPA genes within highly epileptogenic glioblastoma subclasses (Nakata and Nakamura, 2007). Neuronal activity resulting from glioma-to-neuron interactions during growth or the onset of seizures appears to be a major stimulus for BDNF release. This is evidenced by studies that have shown rising serum concentrations of BDNF after the artificial induction of activity with ECT (van Buel et al., 2015; Ryan and McLoughlin, 2018). Further, inhibiting the BDNF-TrkB signaling pathway is considered an attractive therapeutic target for GBM.

We observed a significant benefit of maximal surgical resection of the GBM tumor in high-neural samples. In contrast, for the low-neural group, the benefit was already observed with partial resection, and a complete resection did not lead to a clear advantage, suggesting that for low-neural samples, complete surgical resection is not necessary. Taken together, serum BDNF levels can be used as a biomarker to identify high-neural subgroups that may require complete tumor resection.

Given the stability and different outcomes of the two neural subclasses of GBM, it can be hypothesized that these groups may also respond differently to various pharmacological interventions. In the future, exploration of the impact of different drugs on the two neural groups could be particularly useful, especially considering, for example, that trials with programmed cell death protein-1 (PD-1) inhibitors in GBM treatment have shown heterogeneous responses (Reardon et al., 2020; Lim et al., 2018).

5.4. CONCLUSION

In conclusion, this dissertation presents a series of findings that collectively advance our understanding of cellular heterogeneity in complex diseases through computational and multi-omics approaches. From the development of an improved cell type deconvolution method to the application of cell type deconvolution in estimating cellular compositions in ANCA-GN and GBM, our work demonstrates the utility of integrating diverse data modalities to understand disease mechanisms. The methodological advances described in this work—particularly our semi-supervised learning framework for deconvolution—provide a novel approach that can be applied across various disease contexts to reveal cellular heterogeneity that would otherwise remain hidden in bulk RNA-seq analysis.

Our studies led to the identification of distinct cellular subtypes in both ANCA-GN and GBM with significant prognostic and therapeutic implications. Building on these findings, several important research directions could be important for future work: (1) developing cell type deconvolution methods that incorporate modality-specific knowledge, particularly for spatial transcriptomics data; (2) exploring cell type heterogeneity across GN subtypes using high-resolution single-cell spatial technologies; and (3) investigating the origin of neural cells highly abundant in high-neural GBM and evaluating differential responses to pharmacological interventions between low- and high-neural GBM subtypes. Addressing these will further advance our understanding of complex diseases and potentially lead to better therapeutic approaches.

BIBLIOGRAPHY

James D Watson and Francis HC Crick. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.

Roger D. Kornberg. The molecular basis of eukaryotic transcription. *Proceedings of the National Academy of Sciences*, 104(32):12955–12961, 2007.

Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

Hongkui Zeng. What is a cell type and how to define it? *Cell*, 185(15):2739–2755, 2022.

Jonas Simon Fleck, J. Gray Camp, and Barbara Treutlein. What is a cell type? *Science*, 381 (6659):733–734, 2023.

Ian A. Hatton, Eric D. Galbraith, Nono S. C. Merleau, Teemu P. Miettinen, Benjamin McDonald Smith, and Jeffery A. Shander. The human cell count and size distribution. *Proceedings of the National Academy of Sciences*, 120(39):e2303077120, 2023.

Karthik A. Jagadeesh, Kushal K. Dey, Daniel T. Montoro, Rahul Mohan, Steven Gazal, Jesse M. Engreitz, Ramnik J. Xavier, Alkes L. Price, and Aviv Regev. Identifying disease-critical cell types and cellular processes by integrating single-cell rna-sequencing and human genetics. *Nature genetics*, 54(10):1479–1492, 2022.

Wenjun Ju, Casey S. Greene, Felix Eichinger, Viji Nair, Jeffrey B. Hodgin, Markus Bitzer, Young suk Lee, et al. Defining cell-type specificity at the transcriptional level in human disease. *Genome research*, 23(11):1862–1873, 2013.

Kimberly R. Kukurba and Stephen B. Montgomery. Rna sequencing and analysis. *Cold Spring Harbor Protocols*, 2015(11), 2015. pdb-top084970.

Mohan S. Rao, Terry R. Van Vleet, Rita Ciurlionis, Wayne R. Buck, Scott W. Mittelstadt, Eric A. G. Blomme, and Michael J. Liguori. Comparison of rna-seq and microarray gene expression platforms for the toxicogenomic evaluation of liver from short-term rat toxicity studies. *Frontiers in genetics*, 9:636, 2019.

Markus Wolfien, Robert David, and Anne-Marie Galow. Single-cell rna sequencing procedures and data analysis. *Bioinformatics*, 2021.

Efthymia Papalexi and Rahul Satija. Single-cell rna sequencing to explore immune cell heterogeneity. *Nature Reviews Immunology*, 18(1):35–45, 2018.

Xiliang Wang, Yao He, Qiming Zhang, Xianwen Ren, and Zemin Zhang. Direct comparative analyses of 10x genomics chromium and smart-seq2. *Genomics, Proteomics and Bioinformatics*, 19(2):253–266, 2021.

Justine Hsu, Julien Jarroux, Anoushka Joglekar, Juan P. Romero, Corey Nemec, Daniel Reyes, Ariel Royall, et al. Comparing 10x genomics single-cell 3'and 5'assay in short-and long-read sequencing.

Lambda Moses and Lior Pachter. Museum of spatial transcriptomics. *Nature methods*, 19 (5):534–546, 2022.

Ran Zhou, Gaoxia Yang, Yan Zhang, and Yuan Wang. Spatial transcriptomics in development and disease. *Molecular Biomedicine*, 4(1):32, 2023.

Cameron G. Williams, Hyun Jae Lee, Takahiro Asatsuma, Roser Vento-Tormo, and Ashraful Haque. An introduction to spatial transcriptomics for biomedical research. *Genome Medicine*, 14(1):68, 2022.

Haoyang Li, Juexiao Zhou, Zhongxiao Li, Siyuan Chen, Xingyu Liao, Bin Zhang, Ruochi Zhang, Yu Wang, Shiwei Sun, and Xin Gao. A comprehensive benchmarking with practical guidelines for cellular deconvolution of spatial transcriptomics. *Nature Communications*, 14(1):1548, 2023.

Amanda Janesick, Robert Shelansky, Andrew D. Gottscho, Florian Wagner, Stephen R. Williams, Morgane Rouault, Ghezal Beliakoff, et al. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature Communications*, 14(1):8353, 2023.

Kerry Lee Tucker. Methylated cytosine and the brain: a new base for neuroscience. *Neuron*, 30(3):649–652, 2001.

Peter A. Jones. Functions of dna methylation: islands, start sites, gene bodies and beyond. *Nature reviews genetics*, 13(7):484–492, 2012.

Marina Bibikova, Bret Barnes, Chan Tsan, Vincent Ho, Brandy Klotzle, Jennie M. Le, David Delano, et al. High density dna methylation array with single cpg site resolution. *Genomics*, 98(4):288–295, 2011.

E. Magda Price, Allison M. Cotton, Lucia L. Lam, Pau Farré, Eldon Emberly, Carolyn J. Brown, Wendy P. Robinson, and Michael S. Kobor. Additional annotation enhances potential for biologically-relevant analysis of the illumina infinium humanmethylation450 beadchip array. *Epigenetics chromatin*, 6:1–15, 2013.

Vladimir N. Babenko, Irina V. Chadaeva, and Yuriy L. Orlov. Genomic landscape of cpg rich elements in human. *BMC evolutionary biology*, 17:1–11, 2017.

Nora Fernandez-Jimenez, Catherine Allard, Luigi Bouchard, Patrice Perron, Mariona Bustamante, Jose Ramon Bilbao, and Marie-France Hivert. Comparison of illumina 450k and epic arrays in placental dna methylation. *Epigenetics*, 14(12):1177–1182, 2019.

Sebastian Moran, Carles Arribas, and Manel Esteller. Validation of a dna methylation microarray for 850,000 cpg sites of the human genome enriched in enhancer sequences. *Epigenomics*, 8(3):389–399, 2016.

Andrew E. Teschendorff and Caroline L. Relton. Statistical and integrative system-level analysis of dna methylation data. *Nature Reviews Genetics*, 19(3):129–147, 2018.

Cesar P. Canales and Katherina Walz. The mouse, a model organism for biomedical research. In *Cellular and animal models in human genomics research*, pages 119–140. Academic Press, 2019.

Sébastien A. Smallwood, Heather J. Lee, Christof Angermueller, Felix Krueger, Heba Saadeh, Julian Peat, Simon R. Andrews, Oliver Stegle, Wolf Reik, and Gavin Kelsey. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature methods*, 11(8):817–820, 2014.

Hongshan Guo, Ping Zhu, Xinglong Wu, Xianlong Li, Lu Wen, and Fuchou Tang. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome research*, 23(12):2126–2135, 2013.

Chongyuan Luo, Christopher L. Keown, Laurie Kurihara, Jingtian Zhou, Yupeng He, Junhao Li, Rosa Castanon, et al. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science*, 357(6351):600–604, 2017.

Malte D. Luecken and Fabian J. Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019.

Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. Single-cell rna sequencing technologies and bioinformatics pipelines. *Experimental molecular medicine*, 50(8):1–14, 2018.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint, 2018.

Vladimir Yu Kiselev, Tallulah S. Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics*, 20(5):273–282, 2019.

Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15:1–21, 2014.

Dvir Aran, Agnieszka P. Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature immunology*, 20(2):163–172, 2019.

Hannah A. Pliner, Jay Shendure, and Cole Trapnell. Supervised classification enables rapid annotation of cell atlases. *Nature methods*, 16(10):983–986, 2019.

Laleh Haghverdi, Aaron T. L. Lun, Michael D. Morgan, and John C. Marioni. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, 36(5):421–427, 2018.

Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12): 1289–1296, 2019.

Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *cell*, 177(7):1888–1902, 2019.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Diederik P. Kingma. Adam: A method for stochastic optimization. arXiv preprint, 2014.

Kevin Menden, Mohamed Marouf, Sergio Oller, Anupriya Dalmia, Daniel Sumner Magruder, Karin Kloiber, Peter Heutink, and Stefan Bonn. Deep learning–based cell composition analysis from tissue expression profiles. *Science advances*, 6(30), 2020. eaba2619.

Yoshiaki Yasumizu, Masaki Hagiwara, Yuto Umezu, Hiroaki Fuji, Keiko Iwaisako, Masataka Asagiri, Shinji Uemoto, et al. Neural-net-based cell deconvolution from dna methylation reveals tumor microenvironment associated with cancer prognosis. *NAR cancer*, 6(2), 2024. zcae022.

Andrew Ng. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19, 2011.

Duc Tran, Hung Nguyen, Bang Tran, Carlo La Vecchia, Hung N. Luu, and Tin Nguyen. Fast and precise single-cell data analysis using a hierarchical autoencoder. *Nature communications*, 12(1):1029, 2021.

Yanshuo Chen, Yixuan Wang, Yuelong Chen, Yuqi Cheng, Yumeng Wei, Yunxiang Li, Jiuming Wang, Yingying Wei, Ting-Fung Chan, and Yu Li. Deep autoencoder for interpretable tissue-adaptive deconvolution and cell-type-specific gene analysis. *Nature Communications*, 13(1):6735, 2022.

Xiaoshu Zhu, Jian Li, Yongchang Lin, Liquan Zhao, Jianxin Wang, and Xiaoqing Peng. Dimensionality reduction of single-cell rna sequencing data by combining entropy and denoising autoencoder. *Journal of Computational Biology*, 29(10):1074–1084, 2022.

Aaron M. Newman, Chloé B. Steen, Chih Long Liu, Andrew J. Gentles, Aadel A. Chaudhuri, Florian Scherer, Michael S. Khodadoust, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature biotechnology*, 37(7):773–782, 2019.

Christian W. Hesse and Christopher J. James. On semi-blind source separation using spatial constraints with applications in eeg analysis. *IEEE Transactions on Biomedical Engineering*, 53(12):2525–2534, 2006.

Aaron M. Newman, Chih Long Liu, Michael R. Green, Andrew J. Gentles, Weiguo Feng, Yue Xu, Chuong D. Hoang, Maximilian Diehn, and Ash A. Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*, 12(5):453–457, 2015.

Xuran Wang, Jihwan Park, Katalin Susztak, Nancy R. Zhang, and Mingyao Li. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature communications*, 10(1):380, 2019.

Kai Kang, Qian Meng, Igor Shats, David M. Umbach, Melissa Li, Yuanyuan Li, Xiaoling Li, and Leping Li. Cdseq: A novel complete deconvolution method for dissecting heterogeneous samples using gene expression data. *PLoS computational biology*, 15 (12):e1007510, 2019.

Meichen Dong, Aatish Thennavan, Eugene Urrutia, Yun Li, Charles M. Perou, Fei Zou, and Yuchao Jiang. Scdc: bulk gene expression deconvolution by multiple single-cell rna sequencing references. *Briefings in bioinformatics*, 22(1):416–427, 2021.

Shai S. Shen-Orr, Robert Tibshirani, Purvesh Khatri, Dale L. Bodian, Frank Staedtler, Nicholas M. Perry, Trevor Hastie, Minnie M. Sarwal, Mark M. Davis, and Atul J. Butte. Cell type–specific gene expression differences in complex tissues. *Nature methods*, 7(4): 287–289, 2010.

Maria Chikina, Elena Zaslavsky, and Stuart C. Sealfon. Cellcode: a robust latent variable approach to differential expression analysis for heterogeneous cell populations. *Bioinformatics*, 31(10):1584–1591, 2015.

Francisco Avila Cobos, José Alquicira-Hernandez, Joseph E Powell, Pieter Mestdagh, and Katleen De Preter. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nature communications*, 11(1):5650, 2020.

Gregor Sturm, Francesca Finotello, Florent Petitprez, Jitao David Zhang, Jan Baumbach, Wolf H Fridman, Markus List, and Tatsiana Aneichyk. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics*, 35(14):i436–i445, 2019.

Konstantin Zaitsev, Monika Bambouskova, Amanda Swain, and Maxim N Artyomov. Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. *Nature communications*, 10(1):2209, 2019.

Haijing Jin and Zhandong Liu. A benchmark for rna-seq deconvolution analysis under dynamic testing environments. *Genome biology*, 22:1–23, 2021.

Brandon Jew, Marcus Alvarez, Elior Rahmani, Zong Miao, Arthur Ko, Kristina M Garske, Jae Hoon Sul, Kirsi H Pietiläinen, Päivi Pajukanta, and Eran Halperin. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nature communications*, 11(1):1971, 2020.

Ellis Patrick, Mariko Taga, Ayla Ergun, Bernard Ng, William Casazza, Maria Cimpean, Christina Yung, Julie A Schneider, David A Bennett, Chris Gaiteri, et al. Deconvolving the contributions of cell-type heterogeneity on cortical gene expression. *PLOS Computational Biology*, 16(8):e1008120, 2020.

Francesca Finotello and Zlatko Trajanoski. Quantifying tumor-infiltrating immune cells from transcriptomics data. *Cancer Immunology, Immunotherapy*, 67(7):1031–1040, 2018.

Robin Khatri, Pierre Machart, and Stefan Bonn. Dissect: deep semi-supervised consistency regularization for accurate cell type fraction and gene expression estimation. *Genome Biology*, 25(1):112, 2024.

Daniel Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2000.

Mike E. Davies and Christopher J. James. Source separation using single channel ica. *Signal Processing*, 87(8):1819–1832, 2007.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A. Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

Vladimir N. Vapnik and A. Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity: festschrift for alexey chervonenkis*, pages 11–30. Springer International Publishing, Cham, 2015.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4): 929–965, 1989.

William B. Johnson. Extensions of lipshitz mapping into hilbert space. In *Conference modern analysis and probability, 1984*, pages 189–206, 1984.

J. C. Jennette and R. J. Falk. Pathogenesis of anca-associated vasculitis: observations, theories and speculations. *Presse Med.*, 42(4):493–498, 2013. Pt 2.

A. Richard Kitching, Hans-Joachim Anders, Neil Basu, Elisabeth Brouwer, Jennifer Gordon, David R. Jayne, Joyce Kullman, et al. Anca-associated vasculitis. *Nature reviews Disease primers*, 6(1):71, 2020.

David J. Radford, N. Thin Luu, Peter Hewins, Gerard B. Nash, and Caroline O. S. Savage. Antineutrophil cytoplasmic antibodies stabilize adhesion and promote migration of flowing neutrophils on endothelial cells. *Arthritis Rheumatism: Official Journal of the American College of Rheumatology*, 44(12):2851–2861, 2001.

Kai Kessenbrock, Markus Krumbholz, Ulf Schönermarck, Walter Back, Wolfgang L. Gross, Zena Werb, Hermann-Josef Gröne, Volker Brinkmann, and Dieter E. Jenne. Netting neutrophils in autoimmune small-vessel vasculitis. *Nature medicine*, 15(6):623–625, 2009.

R. Kettritz. How anti-neutrophil cytoplasmic autoantibodies activate neutrophils. *Clinical Experimental Immunology*, 169(3):220–228, 2012.

Divi Cornec, Emilie Cornec-Le Gall, Fernando C. Fervenza, and Ulrich Specks. Anca-associated vasculitis—clinical utility of using anca specificity to classify patients. *Nature Reviews Rheumatology*, 12(10):570–579, 2016.

Daniel Söderberg, Tino Kurz, Atbin Motamedi, Thomas Hellmark, Per Eriksson, and Mårten Segelmark. Increased levels of neutrophil extracellular trap remnants in the circulation of patients with small vessel vasculitis, but an inverse correlation to anti-neutrophil cytoplasmic antibodies during remission. *Rheumatology*, 54(11):2085–2094, 2015.

Benjamin Wilde, Pieter Van Paassen, Jan Damoiseaux, Petra Heerings-Rewinkel, Henk Van Rie, Oliver Witzke, and Jan Willem Cohen Tervaert. Dendritic cells in renal biopsies of patients with anca-associated vasculitis. *Nephrology Dialysis Transplantation*, 24(7): 2151–2156, 2009.

Wayel H. Abdulahad, Coen A. Stegeman, Ymke M. van der Geld, Berber Doornbos van der Meer, Pieter C. Limburg, and Cees G. M. Kallenberg. Functional defect of circulating regulatory cd4+ t cells in patients with wegener's granulomatosis in remission. *Arthritis Rheumatism*, 56(6):2080–2091, 2007.

Estela Nogueira, Sally Hamour, Devika Sawant, Scott Henderson, Nicholas Mansfield, Konstantia-Maria Chavele, Charles D. Pusey, and Alan D. Salama. Serum il-17 and il-23 levels and autoantigen-specific th17 cells are elevated in patients with anca-associated vasculitis. *Nephrology Dialysis Transplantation*, 25(7):2209–2217, 2010.

Nikola Lepse, Wayel H. Abdulahad, Cees G. M. Kallenberg, and Peter Heeringa. Immune regulatory mechanisms in anca-associated vasculitides. *Autoimmunity reviews*, 11(2): 77–83, 2011.

Anqun Chen, Kyung Lee, Tianjun Guan, John Cijiang He, and Detlef Schlondorff. Role of cd8+ t cells in crescentic glomerulonephritis. *Nephrology Dialysis Transplantation*, 35 (4):564–572, 2020.

Charlotte Boud'hors, Jérémie Riou, Nicolas Fage, Clément Samoreau, Alice Desouche, Philippe Gatault, Frank Bridoux, et al. Adding 6-month parameters for the prediction of kidney prognosis in anca-associated glomerulonephritis. *Clinical Kidney Journal*, 16 (12):2530–2541, 2023.

Roger Stupp, Warren P. Mason, Martin J. Van Den Bent, Michael Weller, Barbara Fisher, Martin J. B. Taphoorn, Karl Belanger, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *New England journal of medicine*, 352(10):987–996, 2005.

Roel G. W. Verhaak, Katherine A. Hoadley, Elizabeth Purdom, Victoria Wang, Yuan Qi, Matthew D. Wilkerson, C. Ryan Miller, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer cell*, 17(1):98–110, 2010.

Qianghu Wang, Baoli Hu, Xin Hu, Hoon Kim, Massimo Squatrito, Lisa Scarpace, Ana C. DeCarvalho, et al. Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment. *Cancer cell*, 32(1): 42–56, 2017.

David Capper, David T. W. Jones, Martin Sill, Volker Hovestadt, Daniel Schrimpf, Dominik Sturm, Christian Koelsche, et al. Dna methylation-based classification of central nervous system tumours. *Nature*, 555(7697):469–474, 2018.

Humsa S. Venkatesh, Wade Morishita, Anna C. Geraghty, Dana Silverbush, Shawn M. Gillespie, Marlene Arzt, Lydia T. Tam, et al. Electrical and synaptic integration of glioma into neural circuits. *Nature*, 573(7775):539–545, 2019.

Fanghui Lu, Ying Chen, Chuntao Zhao, Haibo Wang, Danyang He, Lingli Xu, Jincheng Wang, et al. Olig2-dependent reciprocal shift in pdgf and egf receptor signaling regulates tumor phenotype and mitotic growth in malignant glioma. *Cancer cell*, 29(5):669–683, 2016.

Toshiyuki Kawashima, Masakazu Yashiro, Hiroaki Kasashima, Yuzo Terakawa, Takehiro Uda, Kosuke Nakajo, Ryoko Umaba, Yuta Tanoue, Samantha Tamrakar, and Kenji Ohata. Oligodendrocytes up-regulate the invasive activity of glioblastoma cells via the angiopoietin-2 signaling pathway. *Anticancer research*, 39(2):577–584, 2019.

Emily P. Harrington, Dwight E. Bergles, and Peter A. Calabresi. Immune cell modulation of oligodendrocyte lineage cells. *Neuroscience letters*, 715:134601, 2020.

Dieter Henrik Heiland, Vidhya M. Ravi, Simon P. Behringer, Jan Hendrik Frenking, Julian Wurm, Kevin Joseph, Nicklas W. C. Garrelfs, et al. Tumor-associated reactive astrocytes aid the evolution of immunosuppressive environment in glioblastoma. *Nature communications*, 10(1):2541, 2019.

Anoop P. Patel, Itay Tirosh, John J. Trombetta, Alex K. Shalek, Shawn M. Gillespie, Hiroaki Wakimoto, Daniel P. Cahill, et al. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, 2014.

Michael Weller, Nicholas Butowski, David D. Tran, Lawrence D. Recht, Michael Lim, Hal Hirte, Lynn Ashby, et al. Rindopepimut with temozolomide for patients with newly diagnosed, egfrviii-expressing glioblastoma (act iv): a randomised, double-blind, international phase 3 trial. *The lancet oncology*, 18(10):1373–1385, 2017.

Michael Lim, Yuanxuan Xia, Chetan Bettegowda, and Michael Weller. Current state of immunotherapy for glioblastoma. *Nature reviews Clinical oncology*, 15(7):422–442, 2018.

Nancy George, Silvie Fexova, Alfonso Munoz Fuentes, Pedro Madrigal, Yalan Bi, Haider Iqbal, Upendra Kumbham, et al. Expression atlas update: insights from sequencing data at both bulk and single cell level. *Nucleic Acids Research*, 52(D1):D107–D114, 2024.

Zehua Zeng, Yuqing Ma, Lei Hu, Bowen Tan, Peng Liu, Yixuan Wang, Cencan Xing, Yuanyan Xiong, and Hongwu Du. Omicverse: a framework for bridging and deepening insights across bulk and single-cell sequencing. *Nature Communications*, 15(1): 5983, 2024.

C. Domínguez Conde, C. Xu, L. B. Jarvis, D. B. Rainbow, S. B. Wells, T. Gomes, S. K. Howlett, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594), 2022. eabl5197.

Matthew N. Bernstein, Zhongjie Ma, Michael Gleicher, and Colin N. Dewey. Cello: Comprehensive and hierarchical cell type classification of human cells with the cell ontology. *Iscience*, 24(1), 2021.

Katy Börner, Philip D. Blood, Jonathan C. Silverstein, Matthew Ruffalo, Rahul Satija, Sarah A. Teichmann, Gloria Pryhuber, et al. Human biomolecular atlas program (hubmap): 3d human reference atlas construction and usage.

David Osumi-Sutherland, Chuan Xu, Maria Keays, Adam P. Levine, Peter V. Kharchenko, Aviv Regev, Ed Lein, and Sarah A. Teichmann. Cell type ontologies of the human cell atlas. *Nature cell biology*, 23(11):1129–1135, 2021.

Lana X. Garmire, Yijun Li, Qianhui Huang, Chuan Xu, Sarah A. Teichmann, Naftali Kaminski, Matteo Pellegrini, Quan Nguyen, and Andrew E. Teschendorff. Challenges and perspectives in computational deconvolution of genomics data. *Nature Methods*, 21(3): 391–400, 2024.

Colin Megill, Bruce Martin, Charlotte Weaver, Sidney Bell, Lia Prins, Seve Badajoz, Brian McCandless, Angela Oliveira Pisco, Marcus Kinsella, Fiona Griffin, et al. Cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices. *BioRxiv*, pages 2021–04, 2021.

David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21: 1–35, 2020.

Jiaxin Fan, Yafei Lyu, Qihuang Zhang, Xuran Wang, Mingyao Li, and Rui Xiao. Music2: cell-type deconvolution for multi-condition bulk rna-seq data. *Briefings in Bioinformatics*, 23(6), 2022. bbac430.

Zhiyuan Liu, Dafei Wu, Weiwei Zhai, and Liang Ma. Sonar enables cell type deconvolution with spatially weighted poisson-gamma model for spatial transcriptomics. *Nature Communications*, 14(1):4727, 2023.

Ying Ma and Xiang Zhou. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nature biotechnology*, 40(9):1349–1359, 2022.

Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5):829–864, 2020.

Yuqi Cheng, Xingyu Fan, Jianing Zhang, and Yu Li. A scalable sparse neural network framework for rare cell type annotation of single-cell transcriptome data. *Communications Biology*, 6(1):545, 2023.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Advances in neural information processing systems*, 30, 2017.

M. Yates, Richard A. Watts, I. M. Bajema, M. C. Cid, B. Crestani, T. Hauser, B. Hellmich, et al. Eular/era-edta recommendations for the management of anca-associated vasculitis. *Annals of the rheumatic diseases*, 75(9):1583–1594, 2016.

Poh-Yi Gan, Oliver M. Steinmetz, Diana S. Y. Tan, Kim M. O'Sullivan, Joshua D. Ooi, Yoichiro Iwakura, A. Richard Kitching, and Stephen R. Holdsworth. Th17 cells promote autoimmune anti-myeloperoxidase glomerulonephritis. *Journal of the American Society of Nephrology*, 21(6):925–931, 2010.

Christian F. Krebs, Hans-Joachim Paust, Sonja Krohn, Tobias Koyro, Silke R. Brix, Jan-Hendrik Riedel, Patricia Bartsch, et al. Autoimmune renal disease is exacerbated by s1p-receptor-1-dependent intestinal th17 cell migration to the kidney. *Immunity*, 45(5): 1078–1092, 2016.

Roser Guiteras, Maria Flaquer, and Josep M. Cruzado. Macrophage in chronic kidney disease. *NDT Plus*, 9(6):765–771, 2016.

Michele W. L. Teng, Edward P. Bowman, Joshua J. McElwee, Mark J. Smyth, Jean-Laurent Casanova, Andrea M. Cooper, and Daniel J. Cua. Il-12 and il-23 cytokines: from discovery to targeted therapies for immune-mediated inflammatory diseases. *Nature medicine*, 21(7):719–729, 2015.

Amin Abedini, Jonathan Levinsohn, Konstantin A. Klötzer, Bernhard Dumoulin, Ziyuan Ma, Julia Frederick, Poonam Dhillon, et al. *Single-cell multi-omic and spatial profiling of human kidneys implicates the fibrotic microenvironment in kidney disease progression*. 1-13, Nature Genetics, 2024.

Pierre Isnard, Dian Li, Qiao Xuanyuan, Haojia Wu, and Benjamin D. Humphreys. Histopathological-based analysis of human kidney spatial transcriptomics data: toward precision pathology. *The American Journal of Pathology*, 2024.

Zeba Sultana, Robin Khatri, Behnam Yousefi, Nikhat Shaikh, Saskia L Jauch-Speer, Darius P Schaub, Jonas Engesser, Malte Hellmig, Arthur L Hube, Varshi Sivayoganathan, et al. Spatio-temporal interaction of immune and renal cells determines glomerular crescent formation in autoimmune kidney disease. *bioRxiv*, pages 2024–12, 2024.

Hans-Joachim Anders, A. Richard Kitching, Nelson Leung, and Paola Romagnani. Glomerulonephritis: immunopathogenesis and immunotherapy. *Nature Reviews Immunology*, 23(7):453–471, 2023.

Gabriel B. Lerner, Samarth Virmani, Joel M. Henderson, Jean M. Francis, and Laurence H. Beck Jr. A conceptual framework linking immunology, pathology, and clinical features in primary membranous nephropathy. *Kidney International*, 100(2):289–300, 2021.

Kelly D. Smith, David K. Prince, Kammi J. Henriksen, Roberto F. Nicosia, Charles E. Alpers, and Shreeram Akilesh. Digital spatial profiling of collapsing glomerulopathy. *Kidney international*, 101(5):1017–1026, 2022.

Varun Venkataramani, Dimitar Ivanov Tanev, Christopher Strahle, Alexander Studier-Fischer, Laura Fankhauser, Tobias Kessler, Christoph Körber, et al. Glutamatergic synaptic input to glioma cells drives brain tumour progression. *Nature*, 573(7775): 532–538, 2019.

Varun Venkataramani, Yvonne Yang, Marc Cicero Schubert, Ekin Reyhan, Svenja Kristin Tetzlaff, Niklas Wißmann, Michael Botz, et al. Glioblastoma hijacks neuronal mechanisms for brain invasion. *Cell*, 185(16):2899–2917, 2022.

Takuichiro Hide, Ichiyo Shibahara, and Toshihiro Kumabe. Novel concept of the border niche: Glioblastoma cells use oligodendrocytes progenitor cells (gaos) and microglia to acquire stem cell-like features. *Brain tumor pathology*, 36:63–73, 2019.

Craig S. Moore, Qiao-Ling Cui, Nebras M. Warsi, Bryce A. Durafourt, Nika Zorko, David R. Owen, Jack P. Antel, and Amit Bar-Or. Direct and indirect effects of immune and central nervous system–resident cells on human oligodendrocyte progenitor cell differentiation. *The Journal of Immunology*, 194(2):761–772, 2015.

Daniel P. Radin and Parth Patel. Bdnf: an oncogene or tumor suppressor? *Anticancer research*, 37(8):3983–3990, 2017.

Kathryn R. Taylor, Tara Barron, Alexa Hui, Avishay Spitzer, Belgin Yalçin, Alexis E. Ivec, Anna C. Geraghty, et al. Glioma synapses recruit mechanisms of adaptive plasticity. *Nature*, 623(7986):366–374, 2023.

Hiroko Nakata and Shun Nakamura. Brain-derived neurotrophic factor regulates ampa receptor trafficking to post-synaptic densities via ip3r and trpc calcium signaling. *FEBS letters*, 581(10):2047–2054, 2007.

Erin M. van Buel, Kostas Patas, Marcia Peters, Fokko J. Bosker, Ulrich L. M. Eisel, and Hans C. Klein. Immune and neurotrophin stimulation by electroconvulsive therapy: is some inflammation needed after all? *Translational psychiatry*, 5(7):e609–e609, 2015.

Karen M. Ryan and Declan M. McLoughlin. Vascular endothelial growth factor plasma levels in depression and following electroconvulsive therapy. *European Archives of Psychiatry and Clinical Neuroscience*, 268:839–848, 2018.

David A Reardon, Alba A Brandes, Antonio Omuro, Paul Mulholland, Michael Lim, Antje Wick, Joachim Baehring, Manmeet S Ahluwalia, Patrick Roth, Oliver Bähr, et al. Effect of nivolumab vs bevacizumab in patients with recurrent glioblastoma: the checkmate 143 phase 3 randomized clinical trial. *JAMA oncology*, 6(7):1003–1010, 2020.

EIDESSTATTLICHE VERSICHERUNG/DECLARATION

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Sofern im Zuge der Erstellung der vorliegenden Dissertationsschrift generative Künstliche Intelligenz (gKI) basierte elektronische Hilfsmittel verwendet wurden, versichere ich, dass meine eigene Leistung im Vordergrund stand und dass eine vollständige Dokumentation aller verwendeten Hilfsmittel gemäß der Guten wissenschaftlichen Praxis vorliegt. Ich trage die Verantwortung für eventuell durch die gKI generierte fehlerhafte oder verzerrte Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate.

Hamburg, 31.01.2025 Unterschrift/Signature

APPENDIX A

Details of contributions to individual publications.

Publication 1: **Khatri R**, Machart P, Bonn S. DISSECT: deep semi-supervised consistency regularization for accurate cell type fraction and gene expression estimation. **Genome Biol.** 2024 Apr 30;25(1):112. doi: 10.1186/s13059-024-03251-5. PMID: 38689377; PMCID: PMC11061925.

This publication as presented in Chapter 2 has been published in Genome Biology. I am the first author of the paper. I conceptualized the idea together with Stefan Bonn and Pierre Machart. I wrote the code and implemented the algorithm. I analyzed the results and prepared all the figures. The manuscript was written together with Stefan Bonn.

Publication 2: Engesser J*, **Khatri R***, Schaub DP*, Zhao Y, Paust HJ, Sultana Z, Asada N, Riedel JH, Sivayoganathan V, Peters A, Kaffke A, Jauch-Speer SL, Goldbeck-Strieder T, Puelles VG, Wenzel UO, Steinmetz OM, Hoxha E, Turner JE, Mittrücker HW, Wiech T, Huber TB, Bonn S[#], Krebs CF[#], Panzer Ulf[#]. Immune profiling-based targeting of pathogenic T cells with ustekinumab in ANCA-associated glomerulonephritis. **Nat Commun.** 2024 Sep 19;15(1):8220. doi: 10.1038/s41467-024-52525-w. PMID: 39300109; PMCID: PMC11413367.

This publication included in chapter 3 has been published in Nature Communications. I am a co-first author of the paper together with clinician Dr. Jonas Engesser and bioinformatician Darius P. Schaub. I supervised the analysis and performed the individual analysis of spatial transcriptomics data (Figure 2, Supplementary Figure 1 and Supplementary Tables 1-7 and 9) and its integrative analysis with scRNA-seq data (Figure 3e-f). I made contributions to the main findings of the paper, manuscript writing and figure designs during the first submission and subsequent revisions.

Publication 3: Drexler R*, **Khatri R***, Sauvigny T, Mohme M, Maire CL, Ryba A, Zghaibeh Y, Dührsen L, Salviano-Silva A, Lamszus K, Westphal M, Gempt J, Wefers AK, Neumann JE, Bode H, Hausmann F, Huber TB, Bonn S, Jütten K, Delev D, Weber KJ, Harter PN, Onken J, Vajkoczy P, Capper D, Wiestler B, Weller M, Snijder B, Buck A, Weiss T, Göller PC, Sahm F, Menstel JA, Zimmer DN, Keough MB, Ni L, Monje M, Silverbush D, Hovestadt

V, Suvà ML, Krishna S, Hervey-Jumper SL, Schüller U, Heiland DH[#], Hänzelmann S[#], Ricklefs FL[#]. A prognostic neural epigenetic signature in high-grade glioma. **Nat Med.** 2024 Jun;30(6):1622-1635. doi: 10.1038/s41591-024-02969-w. Epub 2024 May 17. PMID: 38760585; PMCID: PMC11186787.

This publication included in chapter 4 has been published in Nature Medicine. I am a first author of the paper together with clinician scientist Dr. Richard Drexler. I implemented and ran the computational framework for the processing and analysis of multi-omics datasets (DNAm, RNA-seq, proteomics, and clinical data). I did the benchmark of DNAm deconvolution algorithms, signature preparations and bioinformatics analysis in the paper. I made contributions to the main finding of the paper, manuscript writing, and figure designs in the first submission and subsequent revisions.

Hamburg, 31.01.2025 Signature