Title

Clarifying the Transcriptional Profiles of Malignant Clones and Nonmalignant Cells of the Microenvironment through Multiscale and Multiomic Analysis of Individual Tumors

Author

Schupp, Patrick Georg

Publication Date

2023

Clarifying the Transcriptional Profiles of Malignant Clones and Nonmalignant Cells of the Microenvironment through Multiscale and Multiomic Analysis of Individual Tumors

by
Patrick Schupp

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Biomedical Sciences

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Joanna Phillips ______________________ Joanna Phillips
10E30A4140364D6...
Chair

DocuSigned by:
Michael Oldham ______________________ Michael Oldham
[illegible]430...

DocuSigned by:
______________________ Joseph Costello
917AF3FB58574BE...

____

____
Committee Members

Acknowledgements

I would like to thank my advisor and mentor Dr. Michael Oldham for his enduring support and guidance throughout my thesis. In addition, I would also like to acknowledge the support of Dr. Joanna Phillips and Dr. Joseph Costello. Their encouragement, helpful critiques, and contributions have shaped my work and me in many positive ways. In addition, I am deeply grateful to my lab members who always supported me, including Samuel Shelton, Daniel Brody, Rebecca Eliscu, Gugene Kang, and Elena Turkalj.

Finally, I would like to thank my wife, parents, and in-laws for their encouragement and support.

Chapters 1 and 2, in full, is a reprint of the material as it appears in “Clarifying the Transcriptional Profiles of Malignant Clones through Multiscale and Multiomic Analysis of Individual Tumors”, Patrick G. Schupp, Samuel J. Shelton, Daniel J. Brody, Rebecca Eliscu , Brett E. Johnson, Tali Mazor, Kevin W. Kelley, Matthew B. Potts, Michael W. McDermott, Eric J. Huang, Daniel A. Lim, Russell O. Pieper, Mitchel S. Berger, Joseph F. Costello, Joanna J. Phillips, Michael C. Oldham. Biorxiv 2023 (v1). The dissertation author was the primary investigator and author of this paper.

Chapter 3 represents a work currently in preparation titled “Metaanalysis of glioma samples reveals unique, cell-type specific dysregulated genes of the microenvironment”. The dissertation author was the primary investigator and author of this paper.

ABSTRACT

Clarifying the Transcriptional Profiles of Malignant Clones and Nonmalignant Cells of the Microenvironment through Multiscale and Multiomic Analysis of Individual Tumors

Patrick Schupp

Understanding the transcriptional consequences of oncogenic mutations is an important goal that may reveal new therapeutic targets for diverse cancers. Although single-cell methods hold promise for this task, it remains non-trivial to isolate and sequence DNA and RNA from the same cell at scale. Here we present a statistically motivated strategy that utilizes multiscale and multiomic analysis of individual human tumor specimens to deconstruct intra-tumoral heterogeneity by clarifying clonal populations of malignant cells and their transcriptional profiles. By combining deep, multiscale sampling of IDH-mutant astrocytomas with integrative, multiomic analysis, we reconstruct and validate the phylogenies, spatial distributions, and transcriptional profiles of distinct malignant clones. We identify a core set of genes that is consistently expressed by the truncal clone, including *AKR1C3*, whose expression is associated with poor outcomes in several types of cancer. Some derived clones exhibit significant enrichment with gene sets representing glioblastoma subtypes and nonmalignant cell types, including ependymal cells. Importantly, by genotyping nuclei for truncal mutations, we show that existing strategies for inferring malignancy from gene expression profiles of single cells may be inaccurate. Furthermore, we find that transcriptional phenotypes of malignancy persist despite loss of the mutant IDH1 protein following chr2q deletion in a subset of malignant cells. In summary, our study provides a generalizable strategy for precisely deconstructing intra-tumoral heterogeneity and clarifying the molecular profiles of malignant clones in any kind of solid tumor.

We extend this approach to a metaanalysis of the cell-type specific dysregulation in the glioma microenvironment. Using the same statistically motivated approach to take advantage of inherent patterns of cell-type specific coexpression, we characterize the differentially expressed cell-type specific transcriptome. We perform this process on thousands of samples and hundreds of datasets from both glioma and normal samples. Finally by taking the difference in the *in silico* derived differential expression metrics for each cell type in glioma and normal contexts, we identify ideal markers of each cell type specifically in glioma but not normal and validate and filter them using orthogonal datasets. In summary, our study provides a generalizable strategy for precise identification of cell-type specific dysregulated genes using abundant bulk transcriptome data for any disease state involving solid tissues.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The existence of different types of cells is readily apparent due to the morphological complexity of multi-cellular life on this planet, however a substantive definition of what truly defines a cell-type has eluded biologists. This difficulty is especially salient in the study of human tumors, where the identification and characterization of malignant cells provides the basis of medical intervention. Despite early progress in identifying molecular markers of malignancy, recent studies, especially using new single-cell RNA-seq approaches, have highlighted the underlying plasticity and heterogeneity of malignant cells. This inherent diversity stymies the development of effective treatments even as the promise of personalized therapies offers a highly labor intensive pathway to treatment.

1.1 The difficulty in targeting gliomas

Brain tumors (including the gliomas, which are the most common malignant, primary brain tumors) are a particularly difficult tumor to treat as they are inaccessible, heterogeneous, and known to phenocopy their surrounding nonmalignant cells[1]. Molecular characterization has been recently embraced by the WHO as the predominant classification of the various glioma types[2]. Mutations in key genes have also provided the basis for a number of pharmaceutical treatments[3–6]. However, some of these targets are not essential to the tumor, allowing escape from drug pressure.

Furthermore, the selected targets are not necessarily most highly expressed in the tumor and may also be found elsewhere in healthy tissue, leading to significant side-effects[7–9]. Fusion transcripts are the most attractive targets as novel epitopes ensure reduced off-target effects, but this approach leaves many potential targets undiscovered relative to targeting the most highly dysregulated genes[10].

Single-cell RNA-sequencing has begun to shed light on this area of differentially expressed genes of glioma cells. However, the inability to capture all malignant and nonmalignant cell-types means that this method is not yet optimal for identification of dysregulated, druggable targets[11,12]. Furthermore, the field of integrating multiple single-cell datasets is yet in its infancy and relatively few single-cell datasets have been published.

Bulk expression data inherently captures RNA from all cells with minimal bias and therefore does not suffer from any cell-type dropout[13,14]. Because this is a mature technology, multiple methods exist by which to aggregate datasets[15,16] and the number of datasets vastly exceeds that of single-cell data, meaning that more of the inherent heterogeneity of malignancies is captured in this data modality.

However, the deconvolution of malignant and nonmalignant cells remains the biggest hurdle. While efforts to deconvolute cell-types from bulk data have been published, they do not characterize the altered transcriptional space of malignant cells nor the transcriptionally dysregulated microenvironment of malignancies. Because these methods generally operate on strong priors of what genes drive expression for certain cell-types, they are unable to adequately characterize dysregulated genes in the glioma context[17–19].

1.2 A new paradigm for the derivation of therapeutic targets for glioma

We propose an unsupervised algorithm based on the premise that there exist genes that covary with the cellular abundance of a certain cell-type[20]. Through iterative rounds of clustering on the correlation space of gene expression, we are able to derive a set of modules composed of highly correlated gene expression patterns. Summarizing this module as the average expression of the constituent genes over all samples via its first principal component, we are able to derive a vector of relative abundance. Finally, we can assign biological meaning to the modules using gene set enrichment analysis (GSEA) or manual annotation. Mathematically, the genes which are members of a module with cell-type significance are high fidelity genes for that cell-type and their expression will reliably track the abundance of the cell-type. Fidelity is also a quantifiable metric, and we can thereby rate how well each gene's expression recapitulates the relative abundance of different cell-types.

In Chapter Two, which is a full reproduction of the manuscript "Deconstructing Intratumoral Heterogeneity through Multiomic and Multiscale Analysis of Serial Sections" available on BioRXiv, I demonstrate how it is possible to use Whole Exome Sequencing (WES) in tandem with amplicon sequencing and methylation arrays to infer phylogenies of malignant clones and track their relative abundance across serial sections of malignant tissue. These clonal abundance vectors can be matched to the abundance of coexpressed gene modules of corresponding expression arrays. We match the malignant clonal abundance to relative abundance of cellular expression programs specific to cell-types or states. Thereby, we can derive the unique transcriptional signature of the constituent malignant clones of a tumor at high resolution and across the entire tumor piece.

Due to covariance between malignant clones and tumor purity, GSEA reveals that certain clonal expression signatures are partially driven by tumor purity, not solely by the clone itself. Because tumor purity is driven largely by the exclusion of normal tissue from the tumor, we can use spatially matched normal samples to remove the tumor purity signal. We find that we greatly enrich the signal of tumor programs while removing enrichment driven by tumor purity.

I repeat the methods with another tumor piece, optimizing the approach and orthogonally validating it with multi-omic single-nucleus approaches. Using single-nucleus RNA- and amplicon-seq technologies I show that we can make incredibly accurate estimations of cellular abundance as well cellular transcriptional identity. It is important to note that while the single-nucleus data represents four sections interspersed throughout the tumor, the bulk data represents nearly 100 sections and therefore provides far more data on the relative abundances of the malignant clones and nonmalignant cell-types. This translates into greater accuracy in deriving the expression programs and fidelity of genes to their cell-type.

In Chapter Three, which is a summary of the manuscript “Metaanalysis of glioma samples reveals unique, cell-type specific dysregulated genes of the microenvironment” currently in preparation, we conduct the largest meta-analysis of adult glioma expression data to our knowledge. We use our algorithm to define the highest fidelity features for nonmalignant cell-types when they are part of the tumor microenvironment or part of healthy tissue. Subsequently, we can compare gene fidelity of these cell-types in these two contexts to determine which genes are dysregulated in nonmalignant cells of all gliomas.

We identify genes which are highly expressed in a nonmalignant cell-type only in the tumor microenvironment but not normal tissue, representing an ideal target for therapeutic intervention. We extend our analysis by subsetting the data to different tumor grades and providing an interactive web-application that makes use of publicly available data to provide the community a rich platform for target identification.

In these approaches, which represent a paradigm shift in the definition of cellular identity, we are able to track coherent cellular populations through space and identify markers which constitute their core cellular programming. This insight allows for the accelerated development of drug-targets and deepened understanding of the cellular dynamics in disease.

1.3 References

1. Tirosh, I. *et al.* Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**, 309–313 (2016).

2. Louis, D. N. *et al.* The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro Oncol.* **23**, 1231–1251 (2021).

3. Flaherty, K. T. *et al.* Inhibition of mutated, activated BRAF in metastatic melanoma. *N. Engl. J. Med.* **363**, 809–819 (2010).

4. Hahn, W. C. *et al.* An expanded universe of cancer targets. *Cell* **184**, 1142–1155 (2021).

5. Kim, D., Xue, J. Y. & Lito, P. Targeting KRAS(G12C): from inhibitory mechanism to modulation of antitumor effects in patients. *Cell* **183**, 850–859 (2020).

6. Mellinghoff, I. K. *et al.* Vorasidenib in IDH1- or IDH2-Mutant Low-Grade Glioma. *N. Engl. J. Med.* (2023) doi:10.1056/NEJMoa2304194.

7. Reardon, D. A. *et al.* A phase I/II trial of pazopanib in combination with lapatinib in adult patients with relapsed malignant glioma. *Clin. Cancer Res.* **19**, 900–908 (2013).

8. Wen, P. Y. *et al.* Phase I, open-label, multicentre study of buparlisib in combination with temozolomide or with concomitant radiation therapy and temozolomide in patients with newly diagnosed glioblastoma. *ESMO Open* **5**, (2020).

9. Nayak, L. *et al.* Phase I trial of aflibercept (VEGF trap) with radiation therapy and concomitant and adjuvant temozolomide in patients with high-grade gliomas. *J Neurooncol* **132**, 181–188 (2017).

10. Yang, K. *et al.* Glioma targeted therapy: insight into future of molecular approaches. *Mol. Cancer* **21**, 39 (2022).

11. Denisenko, E. *et al.* Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol.* **21**, 130 (2020).

12. Caglayan, E., Liu, Y. & Konopka, G. Neuronal ambient RNA contamination causes misinterpreted and masked cell types in brain single-nuclei datasets. *Neuron* **110**, 4043-4056.e5 (2022).

13. Pervez, M. T. *et al.* A Comprehensive Review of Performance of Next-Generation Sequencing Platforms. *Biomed Res. Int.* **2022**, 3457806 (2022).

14. Hong, M. *et al.* RNA sequencing: new technologies and applications in cancer research. *J. Hematol. Oncol.* **13**, 166 (2020).

15. Rau, A., Marot, G. & Jaffrézic, F. Differential meta-analysis of RNA-seq data from multiple studies. *BMC Bioinformatics* **15**, 91 (2014).

16. Toro-Domínguez, D. *et al.* A survey of gene expression meta-analysis: methods and applications. *Brief. Bioinformatics* **22**, 1694–1705 (2021).

17. Ajaib, S. A. *et al.* GBMdeconvoluteR accurately infers proportions of neoplastic and immune cell populations from bulk glioblastoma transcriptomics data. *BioRxiv* (2022) doi:10.1101/2022.11.19.517187.

18. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).

19. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Commun.* **11**, 5650 (2020).

20. Kelley, K. W., Nakao-Inoue, H., Molofsky, A. V. & Oldham, M. C. Variation among intact tissue samples reveals the core transcriptional features of human CNS cell classes. *Nat. Neurosci.* **21**, 1171–1184 (2018).

Chapter 2: Clarifying the Transcriptional Profiles of Malignant Clones through Multiscale and Multiomic Analysis of Individual Tumors

2.1 Introduction

Advances in high-throughput DNA sequencing have revealed the most frequently mutated genes underlying the most prevalent human cancers[1–5]. These findings have catalyzed development of novel therapies that precisely target oncoproteins produced by recurrent mutations[6–9]. However, some oncoproteins present difficult drug targets[10]. Furthermore, targeted monotherapies that are initially successful often induce acquired resistance, which can lead to recurrence of more virulent cancer[11,12]. To address these challenges, it is necessary to expand the therapeutic search space beyond oncoproteins to identify other molecules that distinguish malignant cells from their normal counterparts. Therefore, clarifying the transcriptional profiles of malignant cells is an important goal.

Many gene expression studies of human cancers have focused on defining molecular subtypes for patient stratification using bulk tumor specimens[13–17]. Efforts to identify gene expression signatures of malignant cells in these data are confounded by variable genetic background, tumor purity, and cellular composition, as well as the limited use of control samples. More recent efforts using single-cell methods hold greater promise for this task, but it remains non-trivial to isolate and sequence DNA and RNA from the same cell at scale. As such, malignancy is often inferred for single cells from the presence of copy-number variations (CNVs), which are themselves inferred from single-cell RNA-seq (scRNA-seq) data. However, scRNA-seq data are confounded by technical factors related to tissue dissociation, sampling bias, noise, contamination, and sparsity[18–22], which muddle the relationships between malignant cell genotypes and transcriptional phenotypes, particularly for cancers that lack consistent CNVs.

We have shown that variation in the cellular composition of intact tissue samples drives covariation of transcripts that are uniquely or predominantly expressed in specific kinds of cells[23,24]. We have also shown that the correlation between a gene's expression pattern and the abundance of a cell type is a proxy for the extent to which the same gene is differentially expressed by that cell type[23]. These findings suggest that transcriptional profiles of malignant cells can be identified by correlating genome-wide expression patterns with variant allele frequencies (VAFs), which represent the fraction of sequence reads for a given locus that carry an oncogenic mutation, over a large number of intact tumor samples. The same logic implies that transcriptional profiles of distinct malignant clones can be identified by correlating genome-wide expression patterns with clonal abundance, which can be determined through integrative analysis of VAFs[25,26]. In principle, such patterns should be highly robust since they derive from millions or even billions of cells and do not suffer from the technical and practical limitations imposed by quantifying gene expression in single cells.

Although such analyses are conceptually straightforward, they are difficult to apply to existing gene expression datasets from bulk human tumor specimens, for many reasons. First, many of these datasets lack information about mutations in the analyzed samples. Second, when mutations are analyzed, the most common approach is whole-exome sequencing (WES), which provides shallow coverage (100-200x) of coding regions that is adequate for binary calls of common mutations but inadequate for precisely estimating VAFs. As a result, many studies report mutation frequencies as dichotomous instead of continuous variables. Third, it is often unclear whether paired WES and gene expression data derive from exactly the same tumor sample or adjacent subsamples. Fourth, clonal heterogeneity among tumors from different individuals can obscure the transcriptional consequences of specific mutations. And fifth, many datasets are inadequately powered to identify robust correlations.

To address these challenges, we describe a novel approach for determining the transcriptional profiles of malignant clones through multiscale and multiomic analysis of individual tumor specimens. By amplifying a single tumor specimen into a large number of standardized biological replicates through serial sectioning, we exploit variation in the cellular composition of tumor sections to reveal molecular signatures of distinct malignant clones and nonmalignant cell types. Using a similar approach, we previously isolated transcriptional signatures of radial glia[27] and inhibitory neurons[28] by analyzing gene coexpression relationships over serial sections of human prenatal neocortical specimens. Here we deconstruct human IDH-mutant astrocytomas by precisely defining the evolutionary history and spatial distributions of malignant clones through integrative analysis of single-nucleotide variants (SNVs) and CNVs. By comparing these distributions to gene expression data derived from the same tumor sections, we reveal transcriptional profiles of distinct clones and validate them through comparisons with normal human brain and analyses of individual nuclei from interpolated tumor sections. Our findings suggest that a core set of genes is consistently expressed by the truncal clone of human astrocytomas, offering new therapeutic targets and a generalizable strategy for identifying robust molecular profiles of malignant clones in any kind of solid tumor.

2.2 Methods

2.2.1 Pseudobulk analysis of scRNA-seq data

Single-cell RNA-sequencing (scRNA-seq) data from Venteicher et al.[29] comprising 6243 cells from 10 IDH-mutant adult astrocytomas were downloaded from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/; accession ID = GSE89567). To generate a pseudobulk gene expression matrix from these data, 10% of all cells were randomly sampled and expression levels were summed for each gene from all sampled cells (this process was repeated 100x to generate a matrix with 100 pseudobulk samples). Using cell-class labels provided by the authors, the identities of all cells comprising each pseudobulk sample were tracked. Genome-wide differential expression analysis was performed by comparing all sampled malignant cells to all sampled nonmalignant cells using a two-sided t-test. In parallel, genome-wide gene coexpression analysis was performed as described[23]. Briefly, genome-wide biweight midcorrelations (bicor) were calculated using the WGCNA R package[30] and all genes were clustered using the flashClust[31] implementation of hierarchical clustering with complete linkage and 1 – bicor as a distance measure. The resulting dendrogram was cut at a static height of 0.277, corresponding to the top 1% of bicor values. All clusters consisting of at least 10 genes were identified and summarized by their module eigengene[32] (i.e., the first principal component obtained by singular value decomposition) using the moduleEigengenes function of the WGCNA R package[30]. Highly similar modules were merged if the Pearson correlation of their module eigengenes was > 0.85. This procedure was performed iteratively such that the pair of modules with the highest correlation > 0.85 was merged, followed by recalculation of all module eigengenes, followed by recalculation of all correlations, until no pairs of modules exceeded the threshold. The pseudobulk gene coexpression module most strongly associated with malignant cells was identified by maximizing the correlation between the module eigengene and the actual fraction of sampled malignant cells in each pseudobulk sample. Genome-wide Pearson correlations to this module eigengene (k_{ME} values)[32] were then calculated and compared to the results of single-cell differential expression analysis (t-values).

2.2.3 Sample acquisition

The tumor specimen from case one (WHO grade II primary astrocytoma, IDH-mutant) was obtained from a 40 y.o. female patient following surgical resection at the University of California, San Francisco (UCSF), along with the patient's blood (UCSF case ID: SF9495). The tumor specimen from case two (WHO grade II recurrent astrocytoma, IDH-mutant) was obtained from a 58 y.o. male patient following surgical resection at UCSF, along with the patient’s blood (UCSF case ID: SF10711). Four postmortem control human brain samples from two brain regions (anterior cingulate cortex [ACC] and entorhinal cortex [EC]) were also obtained from routine autopsies of two individuals (41 and 75 y.o. females) at UCSF. Control samples were examined by a neuropathologist (E.J.H.) and found to exhibit no evidence of brain disease. Tissue samples for nucleic acid isolation were immediately frozen on dry ice without fixation.

For tumor histology, a smaller subsample was formalin-fixed and paraffin-embedded (FFPE) using standard procedures. All tumor samples were obtained with donor consent in accordance with protocols approved on behalf of the UCSF Brain Tumor Center Tissue Core.

2.2.4 Serial sectioning

Tissue cryosectioning was performed on a Leica CM3050S cryostat at -20°C. Each sample was oversectioned to account for the possibility of low RNA quality or quantity from some cryosections; after excluding these (see below), most, but not all, analyzed sections were adjacent to one another. For the first case, 81 sections were cut and utilized as shown in **Fig. 2.2f**. For each of the four control samples, ~120 sections were cut and 94 were utilized for gene expression profiling. For the second case, 140 sections were cut and utilized as shown in **Fig. 2.7f**. In addition, the plane of sectioning for the second case was rotated 90 degrees at the halfway point to provide additional spatial variation (**Fig. 2.7f**). These sectioning strategies resulted in 73% power to detect weak correlations ($|r| > 0.3$, $P < .05$) for case one and 83% power for case two[33]. To control for differences in the cross-sectional area of each tissue sample, section thickness was varied as needed to ensure sufficient and comparable amounts of nucleic acids could be extracted from sections for multiomic analysis. Quality control and usage information for all sections can be found in **Table 2.1** (case one), **Table 2.12** (control samples), and **Table 2.15** (case two). Frozen sections were collected in RNase-free 1.7 ml tubes (Denville Scientific Inc, South Plainfield, NJ) and stored at -80°C.

2.2.5 Nucleic acid isolation and quality control

Tissue cryosections were thawed on ice and homogenized by pipette in QIAzol (Qiagen Inc., Valencia, CA). For control samples, RNA was extracted from each section with the miRNeasy mini kit (Qiagen Inc., Valencia, CA). For tumor samples, DNA and RNA were isolated simultaneously from each section with the AllPrep DNA / RNA / miRNA kit (Qiagen Inc., Valencia, CA). All nucleic acid isolation from tissue sections was performed using a QIAcube automated sample preparation system according to the manufacturer's instructions (Qiagen Inc., Valencia, CA). Sections were processed in random batches of 12 on the QIAcube to avoid confounding section number with potential technical sources of variation associated with nucleic acid isolation.

Frozen blood was thawed and resuspended in red blood cell lysis solution (Qiagen Inc., Valencia, CA). White blood cells were removed by centrifugation at 2000g for 5 mins and repeated until white blood cells were depleted. Remaining red blood cells were resuspended in extraction buffer (50 mM Tris [pH8.0], 1 mM EDTA [pH8.0], 0.5% SDS and 1 mg / ml Proteinase K [Roche, Nutley, NJ]) and incubated overnight at 55°C. The extracted DNA was RNAse treated (40 μg / ml) (Roche, Nutley, NJ) for 1 h at 37°C before being phenol chloroform extracted and ethanol precipitated. The resulting DNA was resuspended in TE buffer (10 mM 460 Tris, 1 mM EDTA [pH7.6]).

RNA and DNA were analyzed using a Nanodrop 1000 spectrophotometer (Thermo Scientific Inc., Waltham, MA) to quantify concentrations, OD 260 / 280 ratios, and OD 260 / 230 ratios. Further validation of RNA and DNA concentrations was performed using the Qubit RNA HS kit and Qubit dsDNA HS kit on the Qubit 2.0 Fluorometer (Life Technologies Inc., Carlsbad, CA). RNA integrity (RIN) was assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies Inc., Santa Clara, CA). Sections for which RIN ≥ 5 (case one median = 7.6, case two median = 8.3), OD 260 / 280 ratio ≥ 1.80 (case one median = 2.03, case two median = 1.94), and concentration by Nanodrop ≥ 9 ng / µl (case one median = 25.4 ng / µl, case two median = 9.25 ng / µl) were selected.

2.2.6 Whole exome sequencing (WES) and data preprocessing

WES was performed at the UCSF Institute for Human Genetics genomics core facility (San Francisco, CA). Exome libraries were prepared from 1 µg of genomic DNA from each analyzed section using the Nimblegen EZ Exome kit V3 (Roche, Nutley, NJ). Paired-end 100 bp sequencing was performed on a HiSeq2500 sequencer (Illumina Inc., San Diego, CA). The analysis of WES data was performed as previously described[34]. Briefly, paired-end sequences were aligned to the human genome (University of California, Santa Cruz build hg19) using the Burrows-Wheeler Aligner (BWA)[35]. Uniquely aligned reads were further processed to achieve deduplication, base quality recalibration, and multiple sequence-realignment with the Picard suite[36] and Broad Institute Genome Analysis ToolKit (GATK)[37]. After processing, a mean coverage of 131-151x and 104-122x was achieved for case one and case two, respectively.

2.2.7 Single-nucleotide variant (SNV) and small insertion / deletion (indel) calling workflow

SNVs were identified using MuTect[38] and indels were identified with Pindel[39] using default settings. SNVs were further filtered to only retain variants with frequency > 0.10

in at least one tumor section and < 6 variant reads in the patient's blood. Indels were filtered to only retain variants with > 5 variant reads in a given tumor section and < 13 total reads in the patient's blood. If multiple indels were detected at the same genomic location, only the indel with the most supporting reads was retained. Identified mutations were annotated for their mutational context using ANNOVAR[40] and were also cross-referenced with dbSNP[41] (Build ID: 132) and the 1000 Genomes[42] (Phase 1). SNV and indel events were converted to hg38 coordinates and assigned HGVS compliant names using Ensembl's Variant Effect Predictor[43].

2.2.8 Droplet Digital PCR (ddPCR)

Variant allele frequencies (VAFs) of the IDH1 R132H mutation were determined in 69 tumor sections from case one and the patient's blood using the PrimePCR IDH1 R132H mutant assay and the QX100 Droplet Digital PCR system (Bio-Rad Inc., Hercules, CA). An initial serial dilution of a positive control was performed to optimize the input concentration of genomic DNA from each section and to assess the reliability of the assay. Duplicate reactions were performed to quantify the reproducibility of the assay (**Fig. 2.3b**). Data were analyzed and 95% Poisson confidence intervals were calculated using QuantaSoft software (Bio-Rad Inc., Hercules, CA).

2.2.9 Amplicon sequencing (amp-seq) and data preprocessing

Groups of mutations with similar allele frequency distributions in WES data were identified by hierarchical clustering. Biweight mid-correlations (bicor) were used to estimate the proximities of somatic mutations and 1-bicor was used as a dissimilarity measure. A subset of representative mutations from distinct clusters was validated by Sanger sequencing and deep sequencing of PCR amplicons (amp-seq) derived from tumor sections and the patient's blood. Primers were designed using Primer-BLAST[44] to yield an amplicon of around 500 bp (case one) or 100 bp (case two) with the mutation located within the center of the amplicon (**Tables 2.3 and 2.17**). Amplicons were generated for 42 mutations in case one (n = 69 sections; **Table 2.3**) and 75 mutations in case two (n = 85 sections; **Table 2.17**). For case one, the mutation-containing region was amplified by PCR using the FastStart high-fidelity PCR system (Roche, Nutley, NJ) or the GC-Rich PCR system (Roche, Nutley, NJ) as instructed by the manufacturer using specific annealing temperatures (**Table 2.3**). The resulting amplicons were purified using the NucleoSpin gel and PCR cleanup kit following the manufacturer's instructions (Macherey-Nagel Inc., Bethlehem, PA) and submitted for Sanger sequencing with the same primers used to generate the amplicons. For case two, 50ng of gDNA was used as template per sample in each reaction and 35 cycles of PCR amplification were performed with KAPA HiFi HotStart Ready Mix (2x, KAPA Biosystems, Wilmington, MA). Multiplexed PCR reactions were purified using a 2X volume ratio of KAPA pure SPRI beads (KAPA Biosystems, Wilmington, MA). Purified PCR reactions were quantified using the Qubit dsDNA HS kit and Qubit 2.0 fluorometer. For both cases, the concentration of each amplicon was adjusted to 0.2 ng/µl. Barcoded libraries for each section were generated using the Nextera XT DNA Kit (Illumina Inc., San Diego, CA). After library preparation the barcoded libraries were pooled using bead-based normalization supplied with the Nextera XT kit. The pooled libraries were sequenced with paired-end 250 bp reads in a single flow cell on an Illumina MiSeq

(Illumina Inc., San Diego, CA) in case one and an Illumina HiSeq 4000 in case two. In case one, libraries were sequenced in two runs, whereas all amplicons were sequenced in the same run for case two. Sequence reads were demultiplexed and basecalled using "bcl2fastq" (Illumina Inc., San Diego, CA). FASTQ files were aligned to a custom genome (based on the amplicon sequences) using BWA-MEM[45]. The SAMtools suite[46] was used to create and index BAM files and create pileup files based on reads with a base quality score > 30. Read counts supporting the reference or variant within each amplicon were determined using the read counts function from VarScan 2[47] and these counts were used to calculate VAFs.

2.2.10 Downsampling analysis of amp-seq data

Amplicon reads originating from the reference or alternative alleles for *IDH1* or *TP53* were randomly downsampled to various coverage levels (n = 1000 random downsamples per coverage level) for each section to quantify the effect of reduced coverage on VAF estimates. VAFs were recalculated for each downsampled coverage level and compared to full coverage VAF estimates over all sections using Pearson's correlation or root-mean-square error (RMSE), as illustrated in **Fig. 2.3c-d** (case one) and **Fig. 2.8a-b** (case two).

2.2.11 Hierarchical clustering of variant allele frequencies (VAFs)

Groups of mutations with similar VAF patterns were identified by hierarchical clustering over all tumor sections. VAFs were clustered with Ward's D method and 1 – Pearson's correlation as a dissimilarity measure. The number of clusters was determined from the consensus of elbow[48] and silhouette plot[49] methods, using the cluster package in R[50].

2.2.12 DNA methylation data production and preprocessing

The sample order of genomic DNA from serial sections of case one was randomized to avoid confounding section number with potential sources of technical variation. DNA was concentrated with Genomic DNA Clean & Concentrator 10TM columns (Zymo Research, Irvine, CA) in batches of 12 samples, resulting in approximately two-fold concentration (median concentration after processing: 45ng / µl). The sample order was randomized again and concentrated DNA was shipped on dry ice to the University of California, Los Angeles (UCLA) Neurogenomics Core facility (Los Angeles, CA) for analysis using Illumina 450K microarrays (Illumina Inc., San Diego, CA).

Raw idat files were processed using the ChAMP R package[51]. Initial probe filtering was performed using the load.champ R function[52–54]. Probes with detection P-value > 0.01 (11,799 probes) or beadcount < 3 in at least 5% of samples were removed (n = 760), leaving 461,797 probes for analysis. The Illumina 450K microarrays contain two different assay types (Infinium I and Infinium II). Each assay has different sensitivity and dynamic range, which means that joint normalization leads to type II bias due to the lower sensitivity of the Infinium II assay[55]. We therefore performed beta-mixture quantile normalization (BMIQ) using the “champ.norm” function from ChAMP, which accounts for the different assay types[56].

Additional preprocessing of the methylation data was performed with the SampleNetwork R function[57], which identifies outlying samples, performs data normalization, and corrects for technical batch effects. The standardized sample network connectivity (Z.K) criterion was used to exclude one outlying sample (section #69, whose DNA concentration was substantially lower than other sections), leaving 68 sections. No batch effects associated with ArrayID or ArrayPosition were observed.

2.2.13 Gene expression data production and preprocessing

Total RNA from case one (n = 69 sections) was shipped on dry ice to the UCLA Neurogenomics Core facility (Los Angeles, CA) for analysis using Illumina HT-12 v4 human microarrays (Illumina Inc., San Diego, CA). The order of the sections was randomized prior to shipment to avoid confounding potential technical artifacts with potential biological gradients of gene expression. Two control samples from the same pool of total human brain RNA (Ambion FirstChoice human brain reference RNA Cat#AM6050, Life Technologies Inc., Carlsbad, CA) were included with each of the five datasets. For each of the five datasets (case one and four control samples), all microarray samples (n = 72 – 96 / dataset) were processed in the same batch for amplification, labeling, and hybridization. Amplification was performed using the Ambion TotalPrep RNA amplification kit (Life Technologies Inc., Carlsbad, CA). Raw bead-level data were minimally processed by the UCLA Neurogenomics Core facility (no normalization or background correction) using BeadStudio software (Illumina Inc., San Diego, CA).

For each dataset the minimally processed expression data were further preprocessed using the SampleNetwork R function[57]. Using the standardized sample network connectivity (Z.K) criterion[57], the following numbers of outliers were removed from each dataset: ACC1 (n = 2), ACC2 (n = 11), EC1 (n = 0), EC2 (n = 2), and case one (n = 1). Exclusion of outliers resulted in the following numbers of remaining sections in each dataset: ACC1 (n = 92), ACC2 (n = 83), EC1 (n = 94), EC2 (n = 92), and case one (n = 69). After removing outliers each dataset was quantile normalized[58] and technical batch effects were assessed[57]. Significant batch effects (P < .05 after Bonferroni correction for univariate ANOVA) were corrected using the ComBat R function[59] with no covariates as follows: ACC1 = ArrayID, ACC2 = ArrayID, EC1 = ArrayID and ArrayPosition, EC2 = QCBatch and ArrayID. No batch effects were observed for case one. Multiple technical batch effects were corrected sequentially. Analysis was restricted to 30,425 probes that were re-annotated[60] as having either "perfect" (n = 29,272) or "good" (up to two mismatches; n = 1,153) sequence alignment to their target transcripts. Probes were further collapsed to unique genes (n = 20,019) by retaining one probe per gene with the highest mean expression over all sections.

For case two, RNA-sequencing was used to profile gene expression for all sections (n = 96). Full-length RNA was made into libraries using the KAPA stranded mRNA library prep kit (Roche, Nutley, NJ) following the manufacturer’s instructions, with a mean insert size of 300 bp. One ng of library (composed of library and ERCC spike-in controls, Life Technologies Inc., Carlsbad, CA) was added as input, and all libraries were normalized according to the manufacturer's instructions. During this process samples were randomized in both section order and plane to avoid conflating biological and technical covariates. Sequencing was performed on eight lanes of a HiSeq4000 at the Center for Advanced Technology (CAT) at UCSF with single-end 50 bp sequencing using dual-index barcoding.

Reads were assessed with FastQC to ensure the quality of sequencing data by verifying high base quality scores, lack of GC bias, narrow distribution of sequencing lengths, and low levels of sequence duplication or adapter sequences[61]. Next, reads were subjected to adapter trimming using Cutadapt[62] with minimum length = 20 and a quality cutoff of 20. Reads were subsequently aligned using default settings with the Bowtie2 program[63] to the Genome Reference Consortium Human Build 37[64]. Finally, an expression matrix was generated using the FeatureCounts program with UCSC's library of genomic features[65] (n = 23,900 features). Genes with zero variance were removed (n = 30). Data were normalized with the RUVg package, regressing out 10 factors derived from principal component analysis of the ERCC spike-in control expression matrix[66]. The number of factors was determined empirically by evaluating relative log-expression (RLE) plots and gene-gene correlation distributions. Finally, the SampleNetwork R function[57] was used to identify and remove six outlier sections based on the standardized sample connectivity criterion (Z.K).

2.2.14 Copy number analysis by qPCR

The copy numbers for *TP53* and *ACCS* in case one were determined by SYBR Green-based qPCR. Primers were designed using Primer-BLAST[44] and positioned immediately adjacent to but not including the SNV (ACCS F: TCTCTATGGCAACATCCGGC, R: CAGCCATGCAGCAACAGAAG; RPPH1 F: CGGAGGGAAGCTCATCAGTG, R: CCGTTCTCTGGGAACTCACC, TERT F: CTCGGATCATGCTGAGGACC, R: TTGTGCAATTCTGTGCCAGC, TP53 F: CAGTCACAGCACATGACGGA, R: GGGCCAGACCTAAGAGCAAT). qPCR was performed on genomic DNA from all 69 tumor sections and the patient's blood using the LightCycler 480 SYBR Green I master mix and LightCycler 480 qPCR machine according to the manufacturer's recommendations (Roche, Nutley, NJ). Measurements were triplicated and data were analyzed using the standard curve method. Copy numbers were determined for *TP53* and *ACCS* and two control genes on different chromosomes: ribonuclease P RNA component H1 (*RPPH1*) and telomerase reverse transcriptase (*TERT*) (data not shown). Relative copy number was determined by dividing the mean copy number of *TP53* and *ACCS* by the mean copy number of each reference gene separately to get a ratio and multiplying the ratio by two to obtain the diploid chromosome number. The relative copy number normalized to one of the reference genes (*RPPH1*) is shown in **Fig. 2.3e**.

2.2.15 Copy number variation (CNV) calling (bulk data)

CNVs were quantified using multiple technologies and algorithms to generate reliable estimates. Although WES remains the gold-standard method for calling CNVs, DNA methylation and RNA-seq data provide cost-effective options that can be triangulated with sparse WES data to reduce false positives. Unless otherwise noted, default parameters were used. For case one we used the champ.CNA function, included with the ChAMP R package[67], to call CNVs from DNA methylation data. For both cases, we called CNVs from exome data using FACETS[68] with critical values of 25 (case one) and 450 (case two). Finally, we used CNVkit with circular binary segmentation to call CNVs from bulk RNA-seq data[69–71].

2.2.16 Generation of clonal trees with corresponding frequencies

CNVs were filtered to ensure that they were called in exome data and either DNA methylation data (case one) or RNA-seq data (case two) and covered more than 10% of a chromosomal arm. CNV coordinates were defined based on the intersection of ranges from both methods (**Tables 2.5** and **2.19**). Using the frequencies of CNV / SNV mutations and tumor purity estimated from the *TP53* locus as input to PyClone[25], we determined cluster membership for SNP and CNV events. We then used the PyClone output as the input to the CITUP algorithm[26] to generate the most likely clonal tree (i.e., the tree with the minimum objective value) and derive clonal frequencies. In cases where there was an approximate tie between objective values, the tree was manually chosen based on biologically plausible principles. To visualize results we used the data.tree[72] and DiagrammeR[73] packages in R.

2.2.17 Gene coexpression network analysis

Genome-wide biweight midcorrelations (bicor) were calculated using the WGCNA R package[30] for case one (n = 20,019 genes) and case two (n = 23,870 genes). All genes were clustered using the flashClust[31] implementation of hierarchical clustering with complete linkage and 1 – bicor as a distance measure. Each resulting dendrogram was cut at a static height (0.875 for case one and 0.562 for case two) corresponding to the top 30% and 20% of values of the correlation matrix for case one and case two, respectively. All clusters consisting of at least 15 members for case one or five members for case two were identified and summarized by their module eigengene[32] (i.e. the first principal component obtained by singular value decomposition) using the moduleEigengenes function of the WGCNA R package[30]. Highly similar modules were merged if the Pearson correlation of their module eigengenes was > 0.80. This procedure was performed iteratively such that the pair of modules with the highest correlation > 0.80 was merged, followed by recalculation of all module eigengenes, followed by recalculation of all correlations, until no pairs of modules exceeded the threshold (case one: **Table 2.7**; case two: **Table 2.22**).

2.2.18 Module enrichment analysis

The WGCNA measure of module membership, k_{ME}, was calculated for all genes with respect to each module. k_{ME} is defined as the Pearson correlation between the expression pattern of a gene and a module eigengene and therefore quantifies the extent to which a gene conforms to the characteristic expression pattern of a module[32] (case one: **Table 2.8**; case two: **Table 2.23**). For enrichment analyses, module definitions were expanded to include all genes with significant k_{ME} values, with significance adjusted for multiple comparisons by correcting for the false-discovery rate[74]. If a gene was significantly correlated with more than one module, it was assigned to the module for which it had the highest k_{ME} value. Enrichment analysis was performed for all modules using a one-sided Fisher's exact test as implemented by the fisher.test R function.

2.2.19 Lasso modeling of gene expression

The machine learning variable-selection method lasso (least absolute shrinkage and selection operator) and group lasso were performed using the R package Seagull[75–77].

Modeling was performed for each case with gene expression patterns as dependent variables and clonal frequency vectors as independent variables. For case one, clone 2 was excluded from modeling due to its low frequency and clone 6 was excluded since it was defined by a single CNV. Because clone 1 corresponds to the tumor purity vector, which represents the major vector of variation in this dataset, many genes experience inflated correlations to clone 1. To counteract this effect group lasso was performed. The truncal clone (clone 1) was placed in its own group and all remaining clones to be modeled were placed in a separate group. This procedure improved modeling performance for case one (**Fig. 2.5f-g**) but not case two (**Fig. 2.10f-g**), which may reflect the greater variance in tumor purity for case one. As such, modeling results for case two presented in the manuscript derive from the regular lasso model. For each gene, models were bootstrapped (n = 100) to address collinearity among clonal frequency vectors[78] (as shown in **Fig. 2.5h** and **Fig. 2.10h**). We also generated empirical null distributions for model performance by permuting each gene's expression profile prior to bootstrapping (n = 100).

When performing group-lasso modeling, only models with one surviving clonal frequency vector (not including the truncal clone) were considered. When performing lasso modeling, only models with one surviving clonal frequency vector were considered. To quantify model stability, we calculated the number of times out of 100 bootstraps that the most frequent surviving independent variable was the sole surviving variable. This stability metric was calculated for all gene models, including the permuted models. From the resulting distributions of stability values, a 5% FDR threshold was determined. For case one, the stability value of 73 represents the point beyond which

5% or fewer of the models were permuted models. Similarly, for case two the 5% FDR threshold for the stability metric was 45. Gene set enrichment analysis was performed via a one-sided Fisher’s exact test for all genes with significant model stability for the same clonal frequency vector (**Tables 2.6** and **2.20** for case one and case two, respectively), with genes separated by the sign of the coefficient for the independent variable (**Tables 2.10** and **2.24** for case one and case two, respectively).

2.2.20 Differential gene coexpression analysis

Using the WGCNA R package[30], pairwise biweight midcorrelations (bicor) were calculated among all 30,425 high-quality probes over all sections (n = 69–94) in each of five datasets (case one + four normal human brain samples), generating five identically proportioned correlation matrices (30,425 X 30,425). These correlations were then scaled to lie between [0,1] using the strategy of Mason et al.[79]. To identify gene coexpression relationships that were present in tumor but absent or weaker in normal human brain, each scaled bicor matrix produced from normal human brain was subtracted[80] from the scaled bicor matrix produced from case one, resulting in four “subtraction matrices", or SubMats. The consensus of the four SubMats was formed by taking the minimum value at each point in the four matrices using the parallel minimum (pmin) R function, and the resulting “Consensus SubMat" was used as input for gene coexpression analysis (**Fig. 2.6a**). By definition, gene coexpression modules identified with this strategy will consist of groups of genes with expression patterns that are highly correlated in the astrocytoma but not in any of the normal human brain samples (**Fig. 2.6a**).

Probes in the Consensus SubMat were clustered using the flashClust[31] implementation of a hierarchical clustering procedure with complete linkage and 1 – Consensus SubMat as a distance measure. The resulting dendrogram was cut at a static height of ~0.38, corresponding to the top 2% of values in the Consensus SubMat. All clusters consisting of at least 10 members were identified and summarized by their module eigengene[32] using the moduleEigengenes function of the WGCNA R package[30]. Highly similar modules were merged if the Pearson correlation of their module eigengenes was > 0.85. This procedure was performed iteratively such that the pair of modules with the highest correlation > 0.85 was merged, followed by recalculation of all module eigengenes, followed by recalculation of all correlations, until no pairs of modules exceeded the threshold. The WGCNA[30] measure of intramodular connectivity (k_{ME}) was calculated for all probes (n = 47,202) with respect to each module by correlating each probe's expression pattern across all 69 tumor sections with each module eigengene.

2.2.21 Single-nucleus DNA-sequencing and analysis

Three sections from case two (sections 29 and 113 / 115, which were combined) were analyzed by MissionBio, Inc. (MissionBio, San Francisco, CA) using their Tapestri microfluidics platform for single-nucleus DNA amplicon sequencing[81]. Using an in-house protocol, 4,433 (section 29) and 3,736 (sections 113 / 115) nuclei were extracted and recovered for analysis with the Mission Bio AML panel, which includes primers flanking one *IDH1* and two *TP53* loci. In addition, chr17 chromosomal copy number changes and *TP53* zygosity were inferred from a germline heterozygous intronic mutation upstream of TP53 G245V that happened to fall within the targeting panel (NC_000017.11:g.7674797T>A). Sequencing was performed on a MiSeq (Illumina Inc., San Diego, CA), yielding an average of 6,801 (section 29) or 6,433 (sections 113 / 115) reads per nucleus, with alignment rates of ~90%. Hierarchical clustering of nuclei for mutations of interest was performed separately for section 29 and sections 113 / 115 using complete linkage and Euclidean distance, with $k = 4$ chosen based on silhouette[49] and elbow plots[48]. Genotype calls for the clusters were manually annotated as described in **Fig. 2.8f** and **Table 2.21**.

2.2.22 Single-nucleus RNA-sequencing and analysis

2.2.22.1 Library prep and sequencing

Four sections (17, 53, 93, 117) from case two were used to generate single-nucleus RNA-seq (snRNA-seq) data. Our approach was adapted from TARGET-Seq[82], a protocol utilizing dual-indexing of sample barcodes and unique molecular identifiers (UMIs) of captured transcripts. Briefly, for each section, lysis was performed by dounce homogenization with staining of nuclei by Hoescht3342 and subsequent flow-sorting into three 96-well plates per section. Each plate was randomized and subsequently processed individually and in random order. We used the SmartScribe kit (Takara Bio USA, San Diego, CA) for RT-PCR, followed by PCR with the SeqAmp PCR kit (Takara Bio USA, San Diego, CA). Unlike TARGET-Seq, the RT reaction was performed using only polyA primers (**Table 2.26**). ERCC spike-in control RNA was added to the wells according to manufacturer’s instructions to facilitate identification and correction of batch effects. Wells for each plate were pooled in equivolume proportions and an

Agilent 2100 Bioanalyzer (Agilent Technologies Inc., Santa Clara, CA) was used to assess sample quality and cDNA concentrations were quantified using a Qubit 2.0 Fluorometer with the dsDNA-High Sensitivity kit (Life Technologies Inc., Carlsbad, CA), yielding mean cDNA concentration of 1ng/ul. Concentrations were normalized prior to tagmentation (Nextera Kit, Illumina Inc., San Diego, CA) and amplification of 3' ends, as in TARGET-Seq[82]. Sequencing was performed using the 150-cycle high-throughput kit on an Illumina NextSeq550 at SeqMatic (Fremont, CA) with dual-indexed sequencing and read parameters as in TARGET-seq.

2.2.22.2 Data preprocessing

snRNA-seq raw reads were demultiplexed and basecalled using “bcl2fastq” (Illumina Inc., San Diego, CA). Barcodes were filtered using the “umi_tools” package[83] whitelist function, with a Hamming distance of 2 and the density knee method to determine the number of true barcodes. 809 / 1152 nuclei (70.2%) passed this initial quality control step. Reads were assessed with FastQC to ensure the quality of sequencing data by verifying high base-quality scores, lack of GC bias, narrow distribution of sequencing lengths, and low levels of sequence duplication or adapter sequences[61]. Next, reads were subjected to adapter trimming using the Trimmomatic algorithm[84] with a minimum length of 30, a minimum quality of 4 with a 15 bp sliding window, and otherwise default settings. A mean of 445,082 reads / nucleus was achieved at this stage. Reads were subsequently aligned using ENCODE RNA-seq settings (except for outFilterScoreMinOverLread, which was set to 0) with the STAR program[85] to the Genome Reference Consortium Human Build 38[64]. Finally, an expression count matrix was generated using the FeatureCounts program[86] with Gencode's library of gene features (version 21)[87], subset using the "gene" attribute (n = 60,708 features). Deduplication of UMIs was performed using a custom R script, resulting in a mean number of 206,638 unique reads / nucleus, a 46% deduplication rate. Features with counts less than one in more than 90% of cells were removed (n = 57,021 final features). Data were normalized with the RUVg package, regressing out 10 factors derived from PCA of the ERCC spike-in control expression matrix[66]. Normalized counts were further processed using the Sanity package[19]**,** with 1000 bins and a minimum and maximum variance of 0.001 and 1000, respectively. Internuclear distance was determined using the Sanity_distance function with a signal to noise parameter of 1 and inclusion of error bars.

2.2.22.3 snRNA-seq clustering and differential expression analysis

snRNA-seq data were hierarchically clustered using the hclust function in R with Ward's method and the distance metric derived by Sanity[19]. This distance metric uses a Bayesian approach by giving less weight to gene expression estimates with large error bars when calculating cell distances. The optimal number of clusters (k = 12) was determined using elbow[48] and silhouette plots[49] with the cluster package in R[50]. Differential expression analysis (t.test) was performed between each cluster and all other clusters using Sanity-adjusted expression values for all genes. The resulting distributions of t-values were then compared for genes comprising the bulk coexpression modules most strongly associated with each malignant clone / nonmalignant cell type and all other genes (white and black distributions, respectively, in

Fig. 2.11a-j; significance was evaluated with a one-sided Wilcoxon rank-sum test). Module genes were defined as those that were significantly correlated with corresponding bulk coexpression module eigengenes as determined by the FDR threshold[74]. If a gene was significantly correlated with more than one module eigengene, it was assigned to the module for which it had the highest k_{ME} value.

2.2.22.4 CNV calling

The snRNA-seq count matrix was used as input to CopyKat[88]. Nuclei snRNA-seq clusters determined to be non-malignant by snAmp-seq were used as normal control cells. “KS.cut” was set to 2, “ngene.chr” was set to 20, and Ensembl gene names were used. InferCNV[89] was provided with a vector of nonmalignant cells (as previously determined) based on clustering and snAmp-seq in “subclusters” mode, with a cutoff parameter of 1, and denoising turned on, “ward.D” as clustering method, “qnorm” as subcluster partition method, and tumor subcluster p-value of 0.05. The Hidden Markov model was not used. The program CaSpER[90] was run with the raw snRNA-seq count matrix as input and default settings, again using snRNA-seq clusters of nuclei determined to be malignant by snAmp-seq as negative controls. For each of these algorithms, the outputs were clustered based on Euclidean distance using Ward’s D method. Clusters with no CNV signal were labeled nonmalignant while all other clusters were presumed to represent malignant cells.

Sensitivity and specificity were calculated using the snAmp-seq data as ground truth. True positives (TP) were defined as the intersection of malignant calls by the CNV calling algorithm and the snAmp-seq data. True negatives (TN) were defined as the intersection of nonmalignant calls by the CNV calling algorithm and the snAmp-seq data. False negatives (FN) and false positives (FP) were similarly defined. Nuclei with insufficient data were excluded from the analysis. Sensitivity was defined as: TP / (TP + FN), while specificity was defined as: TN / (TN + FP). Accuracy was defined as (TP + TN) / (TP + FP + TN + FN).

2.2.22.5 UMAP and trajectory analysis

UMAP was performed for all nuclei (n = 809) with a starting seed of 15, 30 neighbors, a spread of 3, a minimum distance of 2, and 1 – Pearson correlation as a distance metric using the "uwot" R package[91] after selecting the first 30 principal components of the Sanity-corrected expression matrix including all genes. UMAP was also performed separately for all cells associated with malignant clusters using the Sanity-corrected expression matrix. After selecting the first 15 principal components, the "uwot" package was used with a seed of 15, 20 neighbors, a spread of 3, a minimum distance of 2, and 1-Pearson correlation as the similarity metric. All other settings were left as defaults. Trajectory analysis was performed with the Slingshot R package[92] on the UMAP plot. The "simple" distance method was used and all other parameters were left as their default values.

2.2.22.6 Gene set enrichment analysis

Enrichment analysis (one-sided Fisher's exact test) was performed for each snRNA-seq cluster using genes that were differentially expressed in that cluster relative to all other clusters using a one-sided Wilcoxon rank-sum test. Resultant p-values were further FDR-corrected to q-values[74]. Gene sets used for enrichment analysis are listed in **Table 2.9.**

2.2.22.6 Amp-seq genotyping

Single-nucleus amplicon-seq (snAmp-seq) was adapted from the TARGET-seq protocol[82]. Primers flanking the following mutations (marking the truncal clone) were designed with Primer3[93]: IDH1 R132H, TP53 G245V, and RUFY1 K218N (**Table 2.26**). To overcome lack of heterogeneity in sequencing, random spacers were added to the beginning (5' end) with 0 - 5 nucleotides from the sequence CGTAC. Finally, a common sequence was added to the 5' end of the primer for a second round of PCR (**Table 2.26**). We selected wells that passed QC for snRNA-seq analysis and processed each plate separately and in random order. Amplification of the first round of PCR was performed with the KAPA 2G Ready Mix (Roche Inc., Nutley, NJ) with the same PCR program as for TARGET-Seq[82]. The program "Barcrawl"[94] was used to create custom dual-index barcodes for the amplification PCR. At this stage 10% of wells were checked using an Agilent 2100 Bioanalyzer (Agilent Inc., Santa Clara, California) to determine whether products of appropriate size were produced. All wells were quantified with a Qubit 2.0 Fluorometer using the dsDNA-High Sensitivity kit (Life Technologies Inc., Carlsbad, CA) and normalized prior to the next step. The second round of PCR used custom sequencing primers that were partially complementary to the previous sequences, with custom dual-index barcodes generated from BarCrawl[94] and Illumina

P5 / P7 sequences. Sequencing was performed using a 300 cycle Miseq v2 Nano kit on a MiSeq (Illumina Inc., San Diego, CA).

snAmp-seq data were demultiplexed and basecalled using “bcl2fastq” (Illumina Inc., San HDiego, CA). Reads were assessed with FastQC to ensure the quality of sequencing data by verifying high base quality scores, lack of GC bias, narrow distribution of sequencing lengths, and low levels of sequence duplication or adapter sequences[61]. Next, reads were subjected to adapter trimming using the Trimmomatic algorithm[84] with a minimum length of 30, a minimum quality of 4 with a 15 bp sliding window, and otherwise default settings[84]. Reads were subsequently aligned with the STAR program to a custom version of the genome containing only the amplicons of interest. Default parameters were altered such that no multiple alignments or splicing events were allowed. The median number of reads per nucleus for each amplicon was (IDH1 R132H: 177; TP53 G245V: 246; RUFY1 K218N: 209). Read counts supporting the reference or variant allele within each amplicon were determined using the read counts function from VarScan 2[47] and these counts were used to calculate variant frequencies. Nuclei were sorted into three categories: called nuclei (calls by VarScan 2 of two or more mutant or two or more wild-type [WT] calls of the three loci with either one or zero indeterminate calls), discrepant nuclei (two WT and one mutant call), and insufficient data nuclei (two or more loci in which VarScan 2 was unable to call a genotype). The breakdown for these categories is as follows: 75% called nuclei, 1% discrepant nuclei, and 24% insufficient data nuclei (**Table 2.27**).

2.2.23 Inter-case analysis

Combined Pearson correlations to tumor purity for the 15,288 genes shared between case one and case two were determined by calculating the weighted average of the z-scores produced by Fisher's transformation, dividing this value by the joint standard error, and applying the inverse Fisher transformation[23]. To define significant genes for enrichment analysis (**Fig. 2.16a-b**), a minimum absolute value for Pearson's correlation of > 0.3 or < -0.3 was required in both cases along with an FDR-corrected q-value of < 0.05. Enrichment analysis was performed as described above, with gene sets listed in **Table 2.9**. Significant positively correlated genes were subjected to protein-protein interaction (PPI) analysis using the STRING database[95]. We used the STRINGdb[95,96], network[97], intergraph[98], and ggnetwork[99] packages to visualize the results of STRING PPI analysis. The "physical" network flavor and minimum score of 900 was utilized to guarantee that all depicted interactions were actual PPIs with experimental evidence. Clusters with more than five members were chosen from the set of interaction clusters generated from all genes that had positive correlations and passed the correlation cutoffs listed above. Enrichment analysis of PPI clusters was performed as described above, with gene sets listed in **Table 2.9**.

2.2.24 Histology and immunostaining

Tumor tissue was fixed in 10% neutral-buffered formalin, processed, and embedded in paraffin. Tumor sections (5 μm) were prepared and stored at -20ºC prior to use. Hematoxylin and eosin staining was performed using standard methods. As part of clinical evaluation, the proliferative index and TP53 mutation status were estimated based on review of immunostained slides for KI67 or TP53, respectively. Briefly, in regions with increased signal the percent of tumor cells staining was estimated based on review of ten 200x fields.

Anti-AKR1C3 was selected based on statistical considerations pursuant to bioinformatic analyses and after preliminary validation of efficacy in human tissue via The Human Protein Atlas[100] (http://www.proteinatlas.org). Primary antibodies and conditions were IDH1 R132H (DIA-H09, Dianova, mouse clone H09, dilution 1:50); AKR1C3 (Catalog# AB84327, Abcam, rabbit polyclonal, dilution 1:600 for single immunohistochemistry and 1:1200 for dual immunofluorescence); and TP53 (1:25, Novocastra, catalog # P53-D07-L-CE-H). Heat antigen retrieval was performed in Tris-EDTA at pH8. Following antigen retrieval, sections for immunohistochemistry were treated with 3% methanol-hydrogen peroxide at 22°C for 16 min.

All immunostaining and multiplex immunostainings were performed using a Discovery XT autostainer or Benchmark XT (Ventana Medical Systems, Inc., USA). For signal detection, the Multimer HRP kit (Ventana Medical Systems, Inc., USA) followed by either DAB or fluorescent detection kits were used. Fluorophores with the least autofluorescence on FFPE tissue were selected to minimize false positives: Cyanine 5 (Cy5) (DISCOVERY CY5 Kit, Cat#760238, Roche Diagnostics Corporation, Indianapolis, USA) and rhodamine (DISCOVERY Rhodamine Kit, Cat#760233, Roche Diagnostics Corporation, Indianapolis, USA). Slides were then counterstained with DAPI (Sigma Aldrich, USA) at 5 µg/ml in PBS (Sigma Aldrich, USA) for 15 minutes, mounted with prolong Gold antifade mounting media reagent (Invitrogen, USA) and stored at -20ºC prior to imaging. Positive and negative controls were included for each marker.

Images of stained slides were acquired using either a light microscope (Olympus BX41 microscope using UC90 Cooled CCD 9 Megapixel camera) or Zeiss Cell Observer epifluorescence microscope equipped with an AxioCam 506M camera and an Excellitas X-Cite 120Q light source and processed with Photoshop CS6 (Adobe systems, San Jose, CA). Nonmalignant tissue analyzed in **Fig. 2.16** was obtained from a patient with epilepsy and corresponds to normal tissue adjacent to epileptic foci.

2.2.25 Data analysis and figure production

Unless otherwise stated, all analyses were performed in the R computing environment (https://www.r-project.org). Figures were produced with the aid of the R packages ggplot2[101], data.table[102], RColorBrewer[103], gridExtra[104], ComplexHeatmap[105], Circlize[106], and ggsignif[107].

2.22.26 Data and code availability

All data are publicly available for download under NCBI Bioproject ID PRJNA953039. Code for processing data and producing figures featured in this manuscript is available on GitHub: https://github.com/oldham-lab/Deconstructing-Intratumoral-Heterogeneity-through-Multiomic-and-Multiscale-Analysis…

2.3 Results

2.3.1 Rationale

Intuitively, genes whose expression patterns correlate most strongly with the abundance of a cell type should include optimal biomarkers. This intuition can also be proven mathematically and empirically. **Fig. 2.1a-c** illustrates a hypothetical example in which the goal is to identify optimal transcriptional markers of malignant cells in a human brain tumor. A conventional strategy would involve physically isolating individual cells, transcriptionally profiling them by single-cell RNA-seq (scRNA-seq), inferring the malignancy of individual cells from the scRNA-seq data based on the presence of driver mutations (CNVs and/or SNVs), and performing differential expression analysis for each gene between all malignant and nonmalignant cells (for example, using a t-test; **Fig. 2.1b**). **Fig. 2.1c** shows an alternative analytical path that leads to the same place: by correlating expression levels of the same hypothetical gene from **Fig. 2.1b** with a dichotomous variable denoting malignant cell abundance (1=malignant cells, 0=nonmalignant cells), the resulting statistical significance is identical to that obtained by differential expression analysis.

Although the t-test and correlation produce identical results when the independent variable is dichotomous, this is not the case when the independent variable is continuous. However, we have shown via pseudobulk analysis of scRNA-seq data from normal adult human brain that: i) the correlation between the expression pattern of a gene and the [continuous] abundance of a cell type accurately predicts differential expression of that gene in that cell type, and ii) cell-type-specific gene coexpression relationships accurately predict cellular abundance in pseudobulk samples[23]. To determine whether these findings extend to malignant cells, we repeated this analysis using scRNA-seq data from 10 adult human astrocytomas[29] (**Fig. 2.1d**). Genome-wide gene coexpression analysis of pseudobulk samples obtained by randomly aggregating scRNA-seq data revealed a malignant cell coexpression module whose eigengene[32] (i.e., first principal component, which summarizes the characteristic expression pattern of the module over all samples) closely tracked the actual abundance of sampled malignant cells (**Fig. 2.1e-f**). Furthermore, the genes that were most significantly up-regulated in malignant cells per differential expression analysis of the underlying scRNA-seq data (**Fig. 2.1d**) also had the highest correlations to malignant cell abundance in pseudobulk data (**Fig. 2.1g**). These results show that gene expression profiles of malignant cells can be revealed by correlating genome-wide expression patterns with malignant cell abundance in heterogeneous tumor samples.

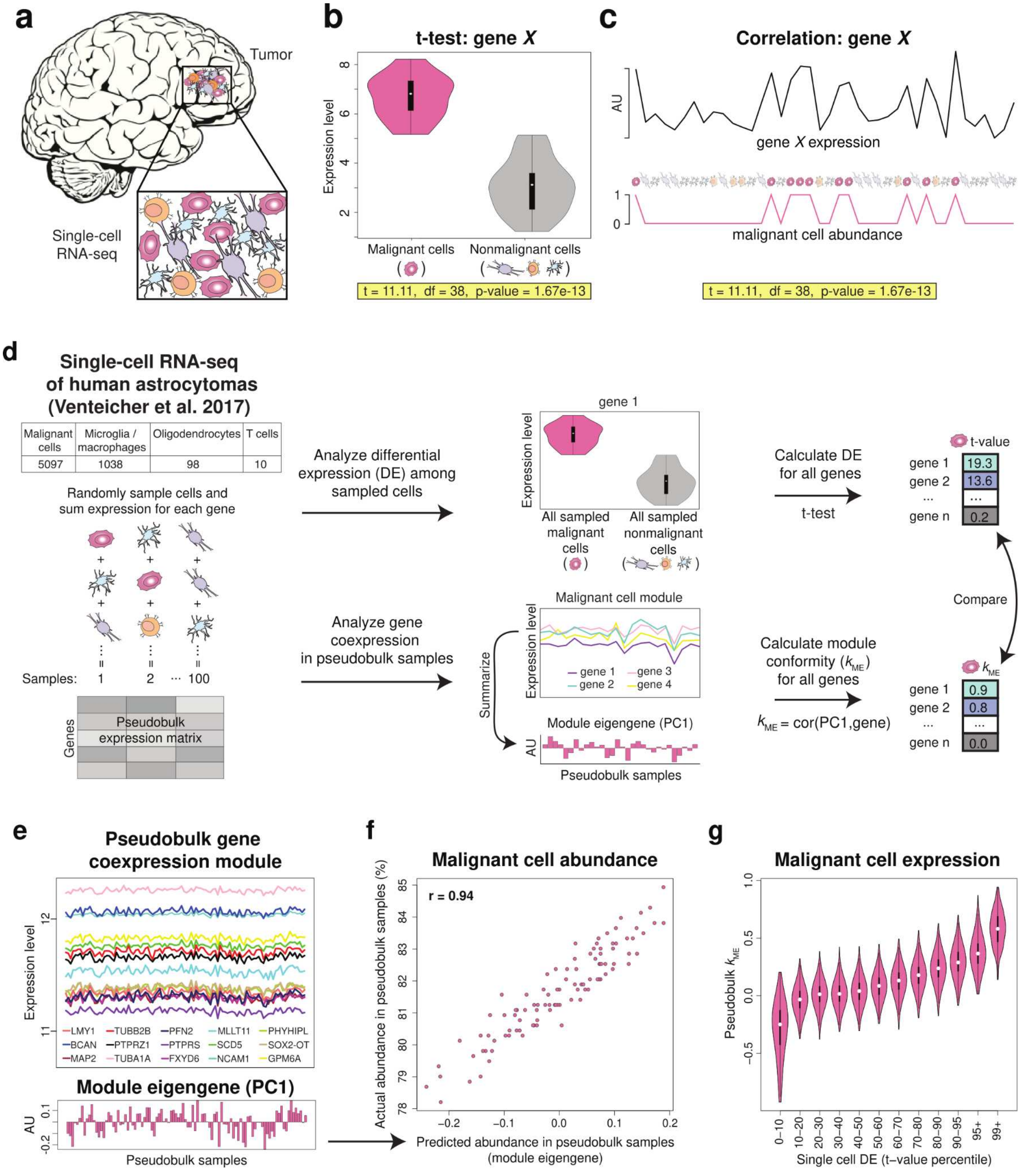


Figure 2.1 | Rationale: Differential expression in malignant vs. nonmalignant cells accurately predicts correlation to malignant cell abundance in pseudobulk samples.

a-d) Analysis schematic. An adult malignant glioma consisting of malignant cells (pink) interspersed with nonmalignant cells **(a)**. **b)** Single-cell RNA-seq (scRNA-seq) reveals a hypothetical gene (gene *X*) that is significantly up-regulated in malignant vs. nonmalignant cells. (Figure caption continued on the next page.)

(Figure caption continued from the previous page.) **c)** Correlating the same gene's expression pattern with a binary vector encoding malignant cell abundance (1 = malignant, 0 = nonmalignant) produces identical results. **d)** Left: scRNA-seq data from 10 adult human IDH-mutant astrocytomas[29] were randomly sampled and aggregated to create 100 pseudobulk samples. Right (top): Genome-wide differential expression (DE) was analyzed for all sampled cells. Right (bottom): Genome-wide gene coexpression was analyzed for all pseudobulk samples. Each pseudobulk module was summarized by its module eigengene (PC1), which was compared to malignant cell abundance, and the correlation between each gene and each module eigengene (module conformity, or k_{ME}) was calculated. **e)** A pseudobulk malignant cell module featuring the top 15 genes ranked by k_{ME}. By correlating the module eigengene to pseudobulk tumor purity **(f)**, we see that this module is driven by variation in malignant cell abundance among pseudobulk samples. **g)** The correlation between gene expression and malignant cell abundance (pseudobulk k_{ME}) predicts the extent of DE identified via scRNA-seq of malignant vs. nonmalignant cells.

2.3.2 Case 1: clonal composition

To put these ideas into practice, we obtained a resected specimen from a primary diffuse glioma that was removed from the left cerebral hemisphere of a 40 y.o. female who presented with language deficits (**Fig. 2.2a-c**). Molecular pathology revealed evidence for mutations in *IDH1* and *TP53* (**Fig. 2.2d-e**), no evidence for chromosome 1p/19q codeletion (data not shown), and KI67 labeling of 6% (data not shown), consistent with a CNS WHO grade 2 astrocytoma, IDH-mutant. We reasoned that sub-sampling (via serial sectioning) would introduce variation in cellular composition across sections, which would in turn drive covariation of molecular markers for distinct subpopulations of malignant and nonmalignant cells. We therefore cut 81 cryosections along the tumor specimen's longest axis (**Fig. 2.2f**), followed by automated DNA/RNA extraction from each section (**Table 2.1**). To identify mutations and characterize the clonal landscape, we performed WES on DNA isolated from sections 14, 39, 69, and the patient's blood. Mutations detected in blood or in genes with very low tumor expression levels were excluded. Of the remaining 33 mutations (**Table 2.2**), including an in-frame deletion in *ATRX*, which is often mutated in IDH-mutant astrocytomas[108], 18 were validated by Sanger sequencing, five were validated by deep sequencing of PCR amplicons spanning each mutation (amp-seq; **Table 2.3**), and ten (mostly indels) could not be validated (**Table 2.2**). Among the 23 validated mutations, 16 were detected by WES in all three tumor sections and seven were detected in only one section, suggesting clonal heterogeneity among malignant cells (**Fig. 2.3a**).

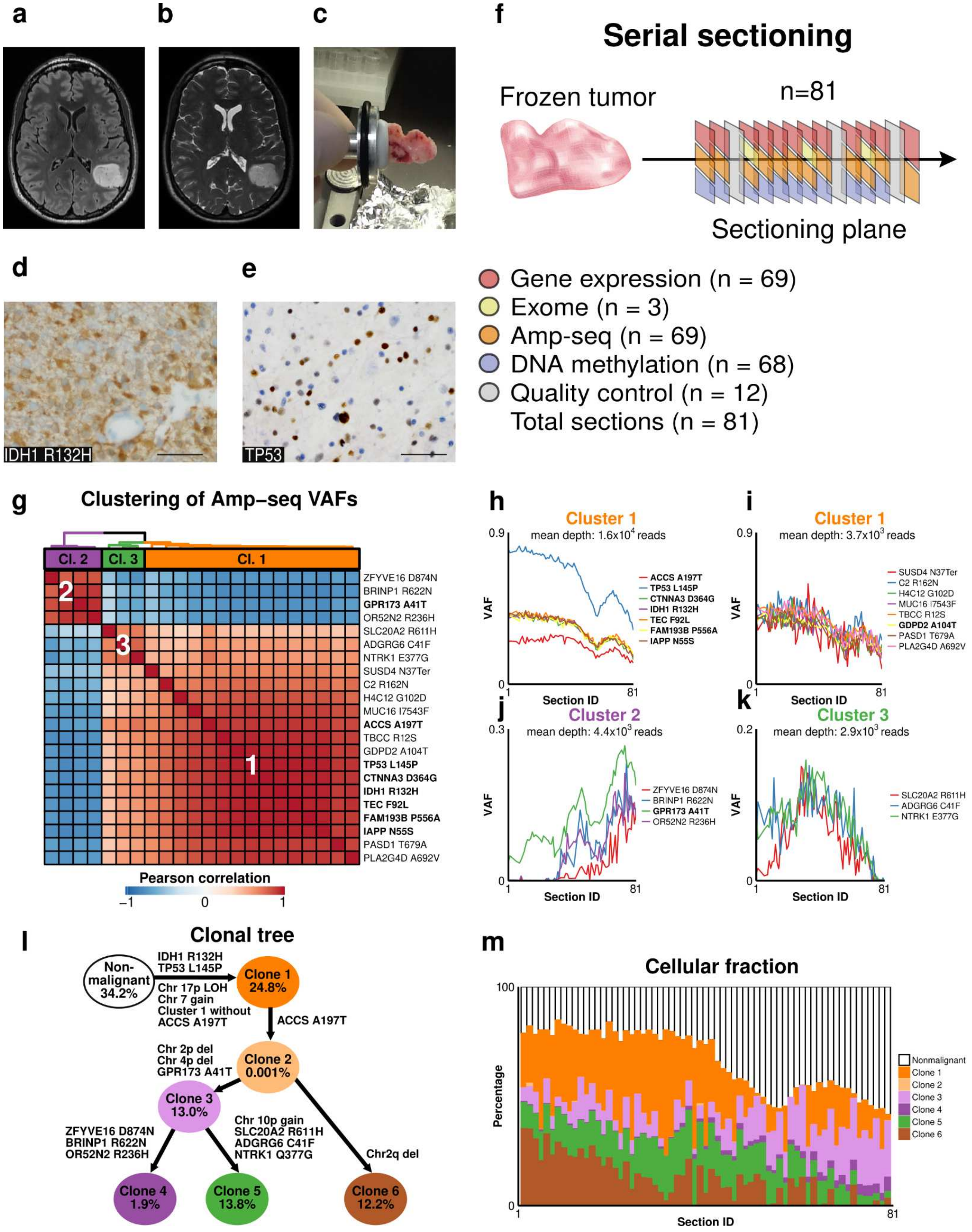


Figure 2.2 | Multiomic analysis of serial tumor sections reveals the clonal composition of a primary grade 2 IDH-mutant astrocytoma (case 1).
(Figure caption continued on the next page.)

(Figure caption continued on the next page.) Axial T2 **(a)** and axial FLAIR **(b)** images demonstrate a round, well-defined T2 and FLAIR hyperintense intraaxial left temporoparietal mass that is non-enhancing and consistent with a low-grade glial neoplasm. **c)** Image of the frozen tumor sample prior to cryosectioning and nucleic acid isolation. **d-e)** Immunostaining for IDH1 R132H (**d**) and TP53 (**e**). Images: 400x. Scale bars: 50 μm. **f)** Schematic of serial sectioning strategy and section usage plan. Amp-seq = deep sequencing of PCR amplicons spanning mutations identified by exome sequencing. **g)** Hierarchical clustering of mutations, using 1 – Pearson correlation of amp-seq variant allele frequencies (VAFs) over all tumor sections (n = 69) as a distance measure, reveals three clusters. Amp-seq was performed in two sequencing runs (denoted by bold and regular fonts). **h-k)** VAF patterns comprising cluster 1 **(h,i)**, cluster 2 **(j)**, and cluster 3 **(k)**. Cluster 1 was split to illustrate the effects of high **(h)** and low **(i)** coverage. **l)** Clone phylogeny (with arbitrary branch lengths) derived from integrated analysis of SNVs (from amp-seq data) and CNVs (from DNA methylation data). Percentages represent the average abundance of each cellular fraction over all analyzed sections (n = 68). **m)** Estimated cellular fractions for all clones and nonmalignant cells over all sections (n = 68).

To determine the relative abundance and spatial distributions of cells carrying mutations within the tumor specimen, we quantified VAFs for validated somatic mutations in each tumor section. We first used droplet digital PCR (ddPCR) to quantify VAFs for IDH1 R132H and observed that this method was highly reproducible (**Fig. 2.3b**). However, given the limited amount of DNA from each tumor section (**Table 2.1**), it was not feasible to quantify all VAFs in this fashion. We therefore tested whether amp-seq yielded VAFs for IDH1 R132H that were comparable to those obtained by ddPCR. We observed high concordance between these methods (**Fig. 2.3b**) and subsequently used amp-seq to quantify VAFs for all validated somatic mutations over all tumor sections, with theoretical VAF detection sensitivity of < 1%.

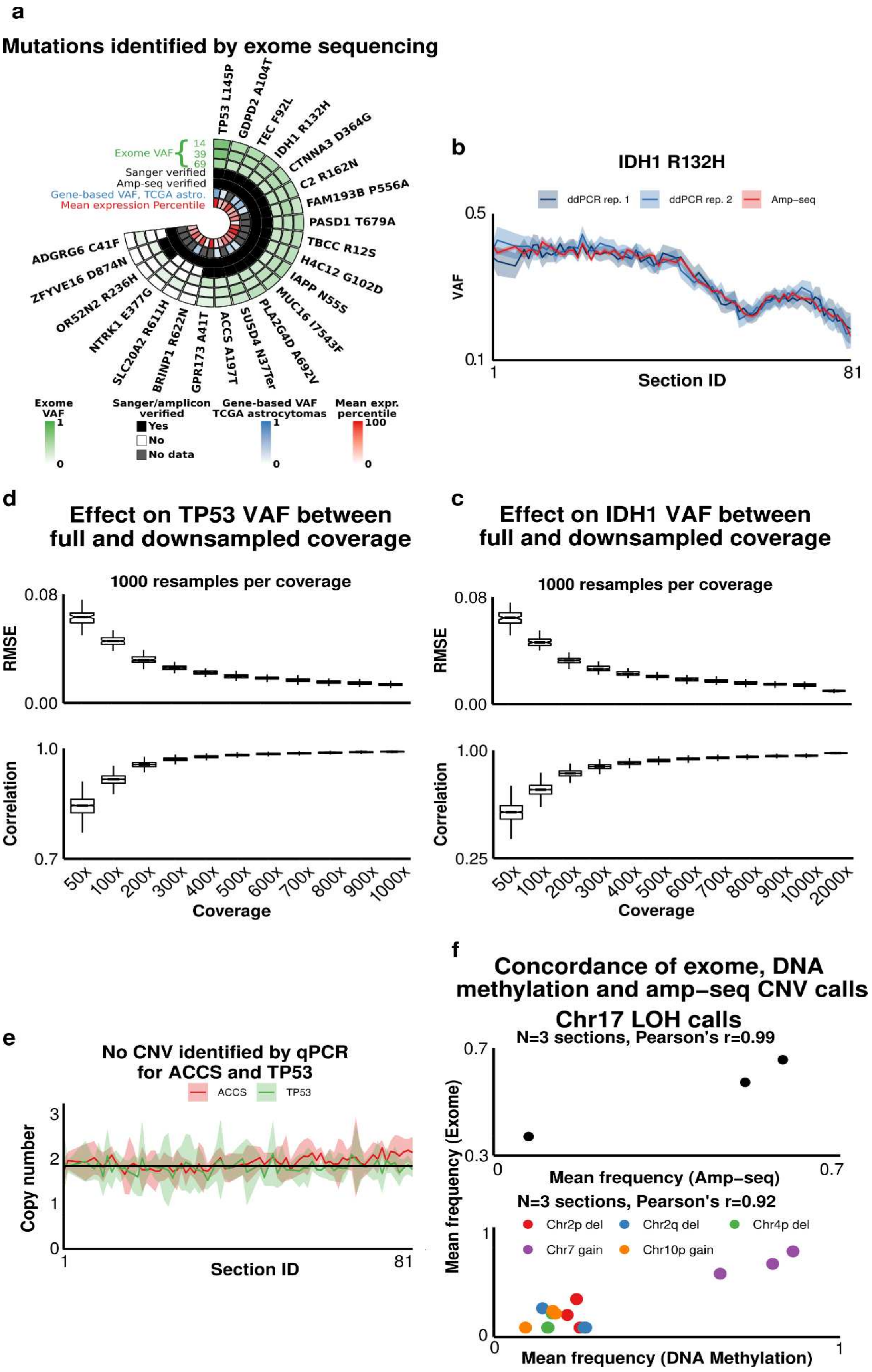

a
Mutations identified by exome sequencing
TP53 L145P
GDPD2 A104T
TEC F92L
IDH1 R132H
CTNNA3 D364G
C2 R162N
FAM193B P556A
PASD1 T679A
TBCC R12S
H4C12 G102D
IAPP N55S
MUC16 I7543F
PLA2G4D A692V
SUSD4 N377Ter
ACCS A197T
GPR173 A41T
BRINP1 R622N
SLC20A2 R611H
NTRK1 E377G
OR52N2 R236H
ZFYVE16 D874N
ADGRG6 C41F
Exome VAF
14
39
69
Sanger verified
Amp-seq verified
Gene-based VAF, TCGA astro.
Mean expression Percentile
Exome VAF
1
0
Sanger/amplicon verified
Yes
No
No data
Gene-based VAF TCGA astrocytomas
1
0
Mean expr. percentile
100
0
b
IDH1 R132H
ddPCR rep. 1
ddPCR rep. 2
Amp-seq
0.5
0.1
VAF
1
81
Section ID
d
Effect on TP53 VAF between full and downsampled coverage
1000 resamples per coverage
0.08
0.00
RMSE
1.0
0.7
Correlation
50x
100x
200x
300x
400x
500x
600x
700x
800x
900x
1000x
Coverage
c
Effect on IDH1 VAF between full and downsampled coverage
1000 resamples per coverage
0.08
0.00
RMSE
1.00
0.25
Correlation
50x
100x
200x
300x
400x
500x
600x
700x
800x
900x
1000x
2000x
Coverage
e
No CNV identified by qPCR for ACCS and TP53
ACCS
TP53
3
2
1
0
Copy number
1
81
Section ID
f
Concordance of exome, DNA methylation and amp-seq CNV calls
Chr17 LOH calls
N=3 sections, Pearson's r=0.99
0.7
0.3
0
0.7
Mean frequency (Amp-seq)
Mean frequency (Exome)
N=3 sections, Pearson's r=0.92
Chr2p del
Chr2q del
Chr4p del
Chr7 gain
Chr10p gain
1
0
0
1
Mean frequency (DNA Methylation)

Figure 2.3 | Mutation validation (case 1).
a) Nonsynonymous mutations were identified by exome sequencing of tumor sections 14, 39, 69, and the patient's blood. Green track: variant allele frequencies (VAF) for each mutation in each section. Black tracks: mutation validation by Sanger sequencing and amp-seq, which is more sensitive. Blue track: gene mutation frequencies in TCGA astrocytomas (n = 286). Red track: mean expression percentiles for each gene over all tumor sections. **b)** Amp-seq and droplet-digital PCR (ddPCR) yielded consistent estimates of IDH1 R132H variant frequencies (n = 69 tumor sections; rep. 1 and rep. 2 denote technical replicates using the same input DNA). Shaded areas represent two standard errors. **c-d)** Downsampling of amp-seq reads for IDH1 R132H **(c)** and TP53 L145P **(d)** was performed in each tumor section to achieve desired coverage levels (x-axis). For each downsampling (n = 1,000), the root mean square-error (RMSE; top) and Pearson's correlation (bottom) was calculated with respect to the true VAF (calculated using all reads) over all sections (n = 69). **e)** Relative copy number was determined by SYBR Green qPCR for *TP53* and *ACCS* loci using genomic DNA from 69 tumor sections and blood. The mean of triplicate measurements, normalized to RNaseP (*RPPH1*) copy number, is shown. Shaded areas represent two standard errors. **f)** Top: Concordant estimates of chr17p loss-of-heterozygosity (LOH) in the same tumor sections (n = 3) were obtained from exome data by analyzing changes in B-allele frequencies and from amp-seq data by analyzing TP53 L145P VAF, which is equivalent to chr17p LOH frequency since both events are truncal. Bottom: Concordant estimates of CNV frequencies in the same tumor sections (n = 3) were obtained using FACETS[68] and ChAMPS[67] to analyze exome and DNA methylation data, respectively.

Amp-seq was performed in two sequencing runs: an initial run consisting of 25 amplicons (mean coverage: $3.0\text{x}10^3$ reads/mutation/section) and a second run consisting of nine amplicons (mean coverage: $1.7\text{x}10^4$ reads/mutation/section). To analyze the stability of amp-seq-derived VAFs, we downsampled reads spanning IDH1 R132H or TP53 L145P and calculated the root-mean-square-error (RMSE) and Pearson correlation between VAFs from full and downsampled read depths. This analysis revealed monotonic improvement in VAF estimates as a function of read depth (**Fig. 2.3c-d**). Notably, VAFs derived from 100-200x coverage were far noisier than VAFs derived from full coverage, indicating that conventional WES data are inadequate for precisely estimating VAFs and malignant cell abundance.

We performed unsupervised hierarchical clustering of amp-seq data to identify mutations with similar VAF patterns within the tumor sample (**Fig. 2.2g** and **Table 2.4**). This analysis revealed three distinct clusters. Cluster 1 included 15 mutations with VAFs that decreased in the latter sections of the tumor sample, which were separated according to sequencing run to display the effects of read depth (**Fig. 2h,i**). Cluster 2 included four mutations with VAFs that increased in the latter sections of the tumor sample (**Fig. 2.2j**). Cluster 3 included three mutations

with VAFs that peaked in the middle sections of the tumor sample (**Fig. 2.2k**).

Table 2.1 | Mutations detected using exome sequencing along with their consequence and incidance in TCGA astrocytoma cases.

Gene	Chr	Consequence	Protein	Mean VAF	TCGA VAF
TP53	17	missense_variant	Leu145Pro	0.6629	0.6433
ATRX	X	inframe_deletion	Glu1464del	0.4180	0.5411
TEC	4	missense_variant	Phe92Leu	0.3824	NA
IDH1	2	missense_variant	Arg132His	0.3770	0.3378
IAPP	12	missense_variant	Asn55Ser	0.3254	NA
PASD1	X	missense_variant	Thr679Ala	0.3458	NA
GDPD2	X	missense_variant	Ala104Thr	0.4209	0.1509
CTNNA3	10	missense_variant	Asp364Gly	0.3653	0.5106
C2	6	missense_variant	Arg162Gln	0.3545	0.1991
H4C12	6	missense_variant	Gly102Asp	0.3343	NA
PLA2G4D	15	missense_variant	Ala692Val	0.3105	0.3438
TBCC	6	missense_variant	Arg12Ser	0.3344	NA
MUC16	19	missense_variant	Ile7543Phe	0.3211	0.2177
FAM193B	5	missense_variant	Pro556Ala	0.3469	NA
SUSD4	1	stop_gained	Gln37Ter	0.2961	NA
MUC21	6	missense_variant	Glu409Asp	0.1267	NA
ACCS	11	missense_variant	Ala197Thr	0.2202	0.2111
OR7C2	19	frameshift_variant	Phe104SerfsTer12	0.2720	NA
RECQL	12	frameshift_variant	Val41SerfsTer14	0.3214	NA
GPR173	X	missense_variant	Ala41Thr	0.1907	NA
PKD2	4	inframe_deletion	Glu102del	0.1223	NA
PHF8	X	3_prime_UTR_variant		0.0356	NA
IL7R	5	frameshift_variant	Arg267GlyfsTer28	0.0850	NA
WASL	7	inframe_deletion	Pro303del	0.0450	NA

Gene	Chr	Consequence	Protein	Mean VAF	TCGA VAF
CSAG1	X	frameshift_variant	Lys65ArgfsTer86	0.0207	NA
ADGRG6	6	missense_variant	Cys41Phe	0.0390	NA
BRINP1	9	missense_variant	Arg622Gln	0.0510	NA
CTNND2	5	inframe_deletion	Lys817del	0.0359	NA
NTRK1	1	missense_variant	Glu377Gly	0.0444	NA
OR52N2	11	missense_variant	Arg236His	0.0436	NA
PCLO	7	missense_variant	Ser506Pro	0.0357	0.1144
SLC20A2	8	missense_variant	Arg611His	0.0476	0.4178
ZFYVE16	5	missense_variant	Asp874Asn	0.0435	NA

Focusing on the sequencing run with higher coverage, we observed that five mutations in cluster 1 (including IDH1 R132H) had VAFs over all tumor sections that were statistically indistinguishable (**Fig. 2.2h**). Two other mutations (TP53 L145P and ACCS A197T) followed a similar pattern but at different scales. For example, VAFs for TP53 L145P were two-fold higher than VAFs for IDH1 R132H (**Fig. 2.2h**). We tested the hypothesis that CNVs might underlie these patterns by performing qPCR for these genes in each tumor section and the patient's blood. We observed approximately diploid copy numbers for both genes in all analyzed sections (**Fig. 2.3e**), indicating that observed VAFs for these mutations are unlikely to result from CNVs. Instead, VAFs for TP53 L145P appear to reflect copy-neutral loss of heterozygosity for chromosome 17p (chr17p LOH) that occurred early in the tumor's evolution (but after the L145P point mutation). Notably, the frequencies of chr17p LOH (derived from B-allele frequencies) were highly concordant between WES and amp-seq data (r=0.99, **Fig. 2.3f** [top]). In contrast, the lower VAFs for ACCS A197T suggest that this mutation appeared after the other mutations comprising cluster 1.

To determine the clonal composition and evolutionary history of the tumor specimen more precisely, we analyzed genome-wide CNVs and their relationships to

SNVs quantified by amp-seq. CNVs were called from WES (n=3 sections) and DNA methylation (n=68 sections) data using FACETS[68] and ChAMPS[67], respectively, yielding highly concordant frequencies for copy number changes (r=0.92, **Fig. 2.3f** [bottom] and **Table 2.5**). Through combined analysis of SNV and CNV frequencies over all tumor sections, we produced an integrated model of tumor evolution. Specifically, we used PyClone[25] to jointly analyze SNV and CNV frequencies, which identified seven distinct clusters and their overall prevalence. Subsequently, the evolutionary history of the tumor specimen was reconstructed using CITUP[26], which produced the most likely phylogenetic tree (**Fig. 2.2l**) and frequencies of six malignant clones over all sections (**Fig. 2.2m** and **Table 2.6**). These analyses confirmed the truncal nature of mutations in *IDH1* and *TP53*[108], while revealing wide variation in the purity of individual tumor sections (range: 38.3 - 84.8%; **Table 2.6**).

2.3.3 Case 1: gene expression

Because DNA and RNA were co-isolated from the same tumor sections (**Fig. 2.2f**), we explored the relationships between clonal abundance and bulk gene expression data. We first performed genome-wide gene coexpression analysis to identify groups of genes with similar expression patterns, which may reflect variation in the abundance of malignant clones and nonmalignant cell types. We identified 38 modules of coexpressed genes (arbitrarily labeled by colors), which were summarized by their eigengenes and hierarchically clustered (**Table 2.7, Fig. 2.4a-c**). As we have shown previously[23,24,27,28], many modules were significantly enriched with markers of distinct cell types (**Fig. 2.5a-d**). By comparing cumulative clonal abundance (**Fig. 2.2m**) to module eigengenes over all tumor sections, we identified five gene coexpression modules whose expression patterns closely tracked the abundance of clone 1 (turquoise: r = 0.97, **Fig. 2.4d**), clone 3 (blue: r = 0.84, **Fig. 2.4e**), clone 4 (black: r = 0.83, **Fig. 2.4f**), clone 5 (midnightblue: r = 0.71, **Fig. 2.4g**), and clone 6 (steelblue: r = 0.69, data not shown). We did not identify a module that was significantly correlated with clone 2, which represented only 0.001% of cells (**Fig. 2.2l**).

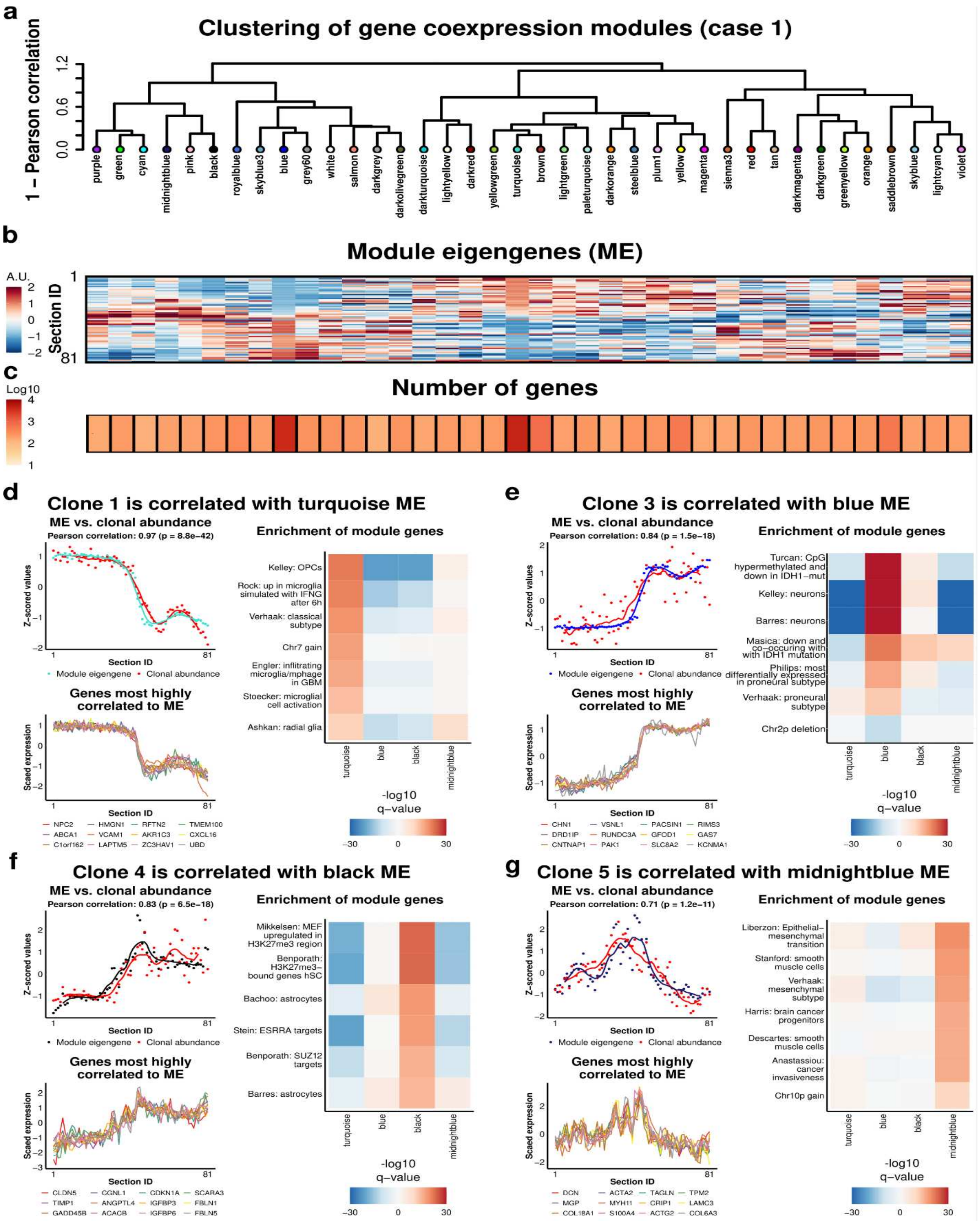

a
Clustering of gene coexpression modules (case 1)
1 – Pearson correlation
0.0
0.6
1.2
purple
green
cyan
midnightblue
pink
black
royalblue
skyblue3
blue
grey60
white
salmon
darkgrey
darkolivegreen
darkturquoise
lightyellow
darkred
yellowgreen
turquoise
brown
lightgreen
paleturquoise
darkorange
steelblue
plum1
yellow
magenta
sienna3
red
tan
darkmagenta
darkgreen
greenyellow
orange
saddlebrown
skyblue
lightcyan
violet
b
Module eigengenes (ME)
A.U.
2
1
0
−1
−2
Section ID
1
81
c
Number of genes
Log10
4
3
2
1
d
Clone 1 is correlated with turquoise ME
ME vs. clonal abundance
Pearson correlation: 0.97 (p = 8.8e−42)
Z-scored values
1
0
−1
−2
1
81
Section ID
• Module eigengene • Clonal abundance
Genes most highly correlated to ME
Scaed expression
1
0
−1
−2
1
81
Section ID
NPC2
HMGN1
RFTN2
TMEM100
ABCA1
VCAM1
AKR1C3
CXCL16
C1orf162
LAPTM5
ZC3HAV1
UBD
Enrichment of module genes
Kelley: OPCs
Rock: up in microglia simulated with IFNG after 6h
Verhaak: classical subtype
Chr7 gain
Engler: infiltrating microglia/mphage in GBM
Stoecker: microglial cell activation
Ashkan: radial glia
turquoise
blue
black
midnightblue
-log10 q−value
−30
0
30
e
Clone 3 is correlated with blue ME
ME vs. clonal abundance
Pearson correlation: 0.84 (p = 1.5e−18)
Z-scored values
2
1
0
−1
1
81
Section ID
• Module eigengene • Clonal abundance
Genes most highly correlated to ME
Scaed expression
1
0
−1
1
81
Section ID
CHN1
VSNL1
PACSIN1
RIMS3
DRD1IP
RUNDC3A
GFOD1
GAS7
CNTNAP1
PAK1
SLC8A2
KCNMA1
Enrichment of module genes
Turcan: CpG hypermethylated and down in IDH1−mut
Kelley: neurons
Barres: neurons
Masica: down and co−occuring with with IDH1 mutation
Philips: most differentially expressed in proneural subtype
Verhaak: proneural subtype
Chr2p deletion
turquoise
blue
black
midnightblue
-log10 q−value
−30
0
30
f
Clone 4 is correlated with black ME
ME vs. clonal abundance
Pearson correlation: 0.83 (p = 6.5e−18)
Z-scored values
2
1
0
−1
1
81
Section ID
• Module eigengene • Clonal abundance
Genes most highly correlated to ME
Scaed expression
2
1
0
−1
−2
−3
1
81
Section ID
CLDN5
CGNL1
CDKN1A
SCARA3
TIMP1
ANGPTL4
IGFBP3
FBLN1
GADD45B
ACACB
IGFBP6
FBLN5
Enrichment of module genes
Mikkelsen: MEF upregulated in H3K27me3 region
Benporath: H3K27me3−bound genes hSC
Bachoo: astrocytes
Stein: ESRRA targets
Benporath: SUZ12 targets
Barres: astrocytes
turquoise
blue
black
midnightblue
-log10 q−value
−30
0
30
g
Clone 5 is correlated with midnightblue ME
ME vs. clonal abundance
Pearson correlation: 0.71 (p = 1.2e−11)
Z-scored values
2
1
0
−1
−2
1
81
Section ID
• Module eigengene • Clonal abundance
Genes most highly correlated to ME
Scaed expression
2
0
−2
1
81
Section ID
DCN
ACTA2
TAGLN
TPM2
MGP
MYH11
CRIP1
LAMC3
COL18A1
S100A4
ACTG2
COL6A3
Enrichment of module genes
Liberzon: Epithelial−mesenchymal transition
Stanford: smooth muscle cells
Verhaak: mesenchymal subtype
Harris: brain cancer progenitors
Descartes: smooth muscle cells
Anastassiou: cancer invasiveness
Chr10p gain
turquoise
blue
black
midnightblue
-log10 q−value
−30
0
30

Figure 2.4 | Gene coexpression modules are highly correlated with clonal abundance (case 1).
a) Hierarchical clustering of gene coexpression modules over all tumor sections (n = 69). **b)** Module eigengenes (ME) illustrate the relative expression levels of genes in each module over all tumor sections. **c)** The number of genes used to form each ME. **d-g)** Top left: MEs with the strongest correlations to clonal abundance (defined cumulatively). Locally weighted smoothing (LOESS) lines are shown; correlation is based on data points. Bottom left: the 12 genes with the highest correlations to the ME (k_{ME}). Right: enrichment analysis of gene coexpression modules using published gene sets. FDR-corrected p-values (q-values) from one-sided Fisher's exact tests are shown. Positive values represent enrichments of genes that were significantly positively correlated to the ME, while negative values represent enrichments of genes that were significantly negatively correlated to the ME. Gene sets representing chromosomal gains or losses include all genes within affected regions (as described in **Fig. 2.2** and **Table 2.5**). See **Table 2.9** for descriptions and sources of featured gene sets.

To characterize these modules, we performed enrichment analysis with biologically relevant gene sets (**Fig. 2.4d-g**). We first asked whether genes within clonal CNV boundaries (**Fig. 2l** and **Table 2.5**) were significantly enriched (for gains) or depleted (for deletions) in the bulk coexpression modules most strongly associated with each clone (**Table 2.8**). Notably, all such gene sets were significantly enriched in the appropriate module and expected direction (e.g., chr7 gain for clone 1 [**Fig. 2.4d**]**,** chr2p deletion for clone 3 [**Fig. 2.4e**], and chr10p gain for clone 5 [**Fig. 2.4g**]). We next analyzed publicly available gene sets from diverse sources (**Table 2.9**). We found that the largest (turquoise) module, which closely tracked the abundance of clone 1 (i.e., tumor purity), was significantly enriched with markers of oligodendrocyte progenitor cells (OPCs) and radial glia, genes comprising the 'classical' subtype of glioblastoma proposed by Verhaak et al.[17] and numerous gene sets related to microglial infiltration and activation. The second largest (blue) module, which tracked clone 3, was significantly enriched with neuronal gene sets as well as genes that are down-regulated pursuant to *IDH1* mutations. The black module, which tracked clone 4, was enriched with astrocyte markers as well as genes that are differentially regulated during development and glioma. The midnightblue module, which tracked clone 5, was enriched with markers of smooth muscle cells, genes comprising the ‘mesenchymal’ subtype of glioblastoma[17,109], and gene sets related to epithelial-mesenchymal transition and invasiveness. The steelblue module, which tracked clone 6, was enriched with markers of non-resident immune cells (data not shown).

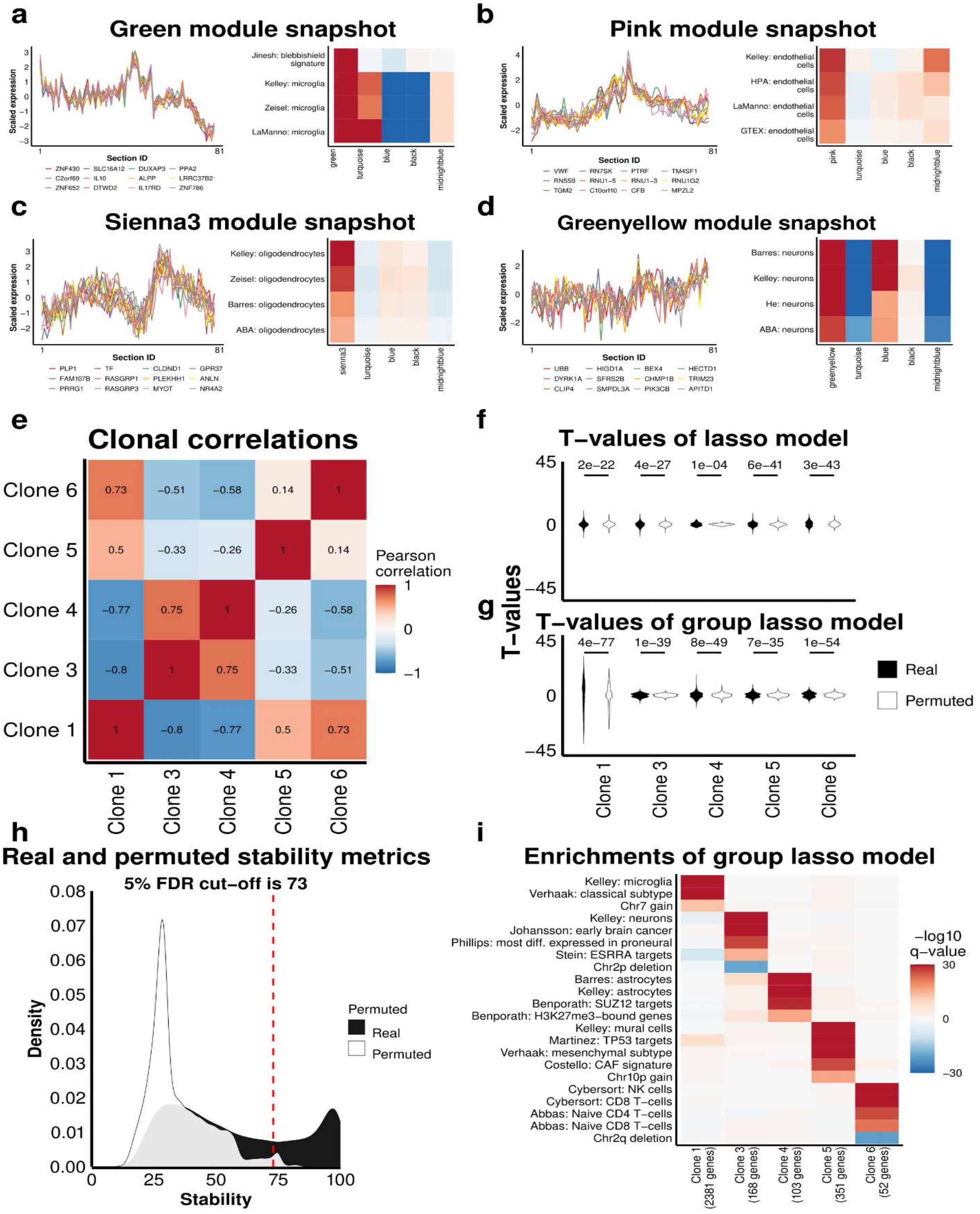


Figure 2.5 | Linear modeling of gene expression using clonal frequencies reveals concordant gene-set enrichments with coexpression modules (case 1).
(Figure caption continued on the next page.)

(Figure caption continued from the previous page.) **a-d)** Left: snapshots of additional gene coexpression modules enriched for markers of nonmalignant cell types (expression patterns for the top 12 genes ranked by k_{ME} are shown). Right: heatmaps of gene set enrichment results for each module. Modules included genes that were most specifically and significantly correlated (after FDR correction) to the module eigengene (ME), and enrichment was assessed with a one-sided Fisher's exact test (followed by FDR correction; see panel **i** for legend). **e)** Correlation heatmap for the cumulative frequency vectors of identified clones. **f-g)** Lasso regression[75] was used to model the expression of all genes (n = 20,018) as a function of clonal frequencies over all tumor sections (n = 69). Violin plots illustrate the distributions of t-values for all models where the indicated clone was the only explanatory variable that survived lasso selection. Permutations were performed by randomly scrambling clonal frequencies (n = 100) prior to lasso regression. Real and permuted clonal frequency vectors were bootstrapped (n = 100) to address collinearity. P-values denote the significance of the Anderson-Darling test, which evaluates whether two distributions are likely to be derived from the same distribution. **f)** Results of a standard lasso model. **g)** Results of a group lasso model where the truncal clone (equivalent to tumor purity) was placed in a separate group due to its strong effect on gene expression (**Fig. 2.4c**); note the general improvement in Anderson-Darling test P-values. **h)** Density plot showing the number of times (out of 100 bootstraps) that the same explanatory (clonal frequency vector) variable was retained by the group lasso regression model, or 'stability'. Only group lasso models where retained explanatory variables included the truncal clone and up to one other clone were considered. The vertical line demarcates the point to the right of which only 5% of values belong to the permuted distribution, i.e. a 5% FDR rate. **i)** Enrichment analysis (one-sided Fisher's exact test) of genes that were significantly (FDR < .05) and stably (FDR < .05) associated with each clone. Gene sets are described in **Table 2.9**. Heatmap depicts -log10 FDR-corrected p-values (q-values; shared legend for **a-d**) after comparing each gene set to all genes with stability > 73 for a given clone (one-sided Fisher's exact test). Positive values represent enrichments for genes with significant positive correlations to the ME (**a-d**) or significant positive modeling coefficients (**i**), while negative values represent enrichments for genes with significant negative correlations to the ME (**a-d**) or significant negative modeling coefficients (**i**).

To further characterize the transcriptional signatures associated with each clone, we used multiple linear regression to model genome-wide expression levels as a function of clonal abundance. To account for collinearity and the dominant effect of clone 1, we used a group lasso model with bootstrapped clonal abundance vectors (real or permuted) as predictors (**Fig. 2.5e-i**). We restricted our focus to genes that were significantly and stably modeled by a single clone (in addition to clone 1, per the group lasso model, **Table 2.10**). Enrichment analysis of these genes largely recapitulated enrichment analysis of gene coexpression modules associated with each clone, including the associations of different clones with different cell types (**Table 2.11** and **Fig. 2.5i**).

The associations of different clones with different cell types suggest two non-mutually exclusive possibilities. First, different clones may preferentially express different cell-type-specific transcriptional programs. Second, different clones may preferentially associate with different nonmalignant cell types in the tumor microenvironment, leading to correlated gene expression patterns. Although such possibilities are ideally studied at the level of individual cells, all sections from this case were consumed during bulk data production. However, we reasoned that bona fide transcriptional signatures of malignant clones should be absent from non-neoplastic human brains. To test this hypothesis, we profiled gene expression in 361 cryosections from four neurotypical adult human brain samples (**Table 2.12**) and performed genome-wide differential coexpression analysis by subtracting normal correlations from tumor correlations, such that tumor-specific gene coexpression relations would be retained (**Fig. 2.6a, Table 2.13, Table 2.14**). This analysis revealed tumor-specific gene coexpression modules that tracked the abundance of distinct clones and largely recapitulated the transcriptional signatures described in **Fig. 2.4** and **Fig. 2.5**, including preserved enrichment of clone-specific CNV gene sets (**Fig. 2.6b-e**). However, enrichment results for nonmalignant cell-type-specific gene sets became less significant, with the exception of OPCs and radial glia for clone 1, which became more significant (**Fig. 2.6b-e**). These results suggest that derived clones may occupy distinct microenvironments, while the truncal clone retains signatures of progenitor cells that may reflect the cell of origin.

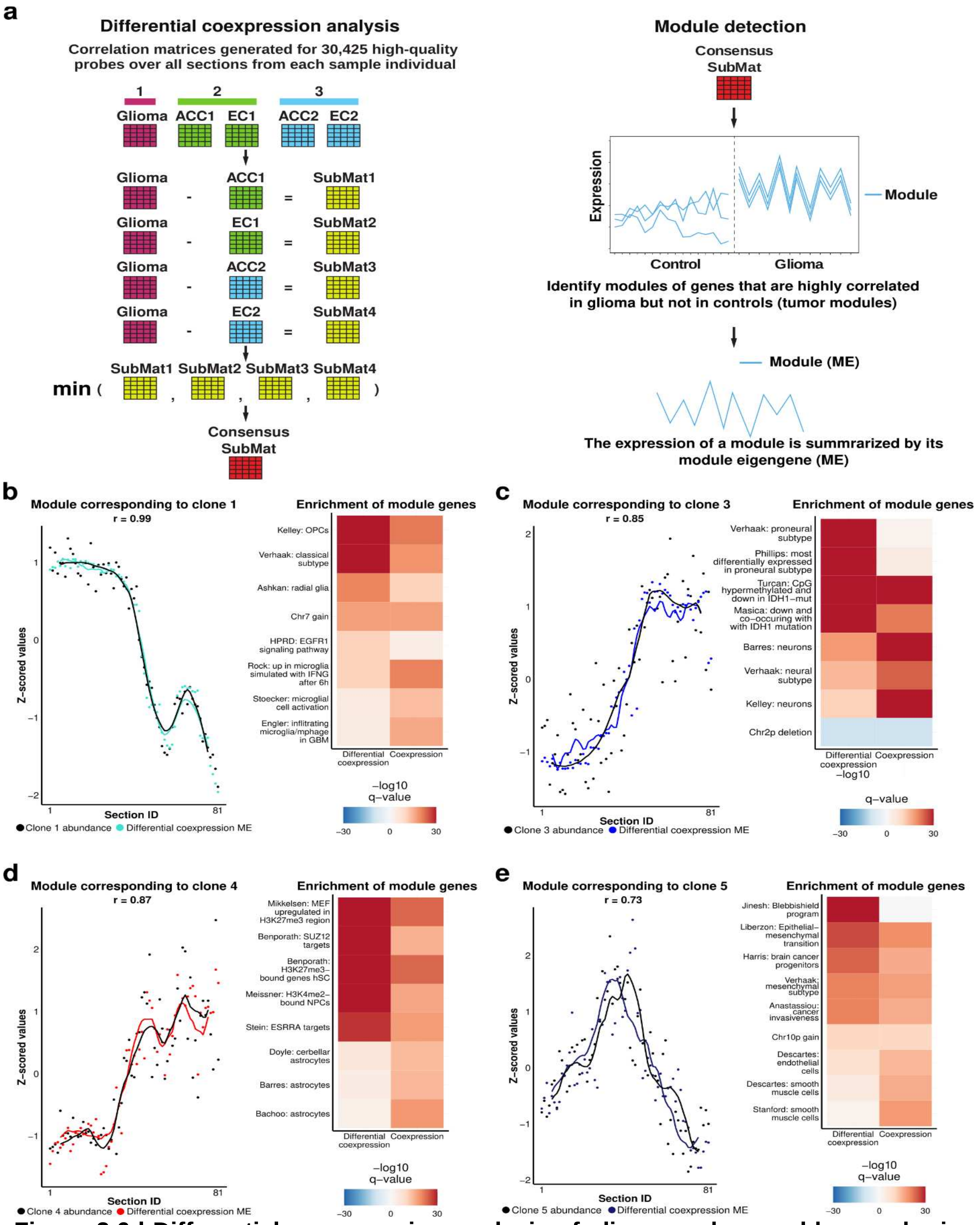


Figure 2.6 | Differential coexpression analysis of glioma and normal human brain preserves gene coexpression modules associated with malignant clones (case 1). (Figure caption continued on the next page.)

(Figure caption continued from the previous page.) **a)** Genome-wide gene coexpression relationships were calculated for each of the five tissue specimens (one astrocytoma and four normal brain controls) over all tissue sections, resulting in five correlation matrices with the same dimensions. Unbiased differential coexpression analysis was performed as illustrated. ACC = anterior cingulate cortex; EC = entorhinal cortex. **b-e)** Left: differentially coexpressed module eigengenes (ME) with the strongest correlations to clonal abundance (defined cumulatively). Locally weighted smoothing (LOESS) lines are shown; correlation is based on data points. Right: enrichment analysis of differentially coexpressed module genes using published gene sets. FDR-corrected p-values (q-values) from one-sided Fisher's exact tests are shown. Positive values represent enrichments of genes that were significantly positively correlated to the ME, while negative values represent enrichments of genes that were significantly negatively correlated to the ME. Gene sets representing chromosomal gains or losses include all genes within affected regions (as described in **Fig. 2.2** and **Table 2.5**). See **Table 2.9** for descriptions and sources of featured gene sets.

2.3.4 Case 2: clonal composition

To test our strategy on a more complex case, we obtained a resected specimen from a recurrent diffuse glioma that was removed from the right cerebral hemisphere of a 58 y.o. male (**Fig. 2.7a-c**) approximately 28 years after the primary resection. Molecular pathology revealed evidence for mutations in *IDH1* and *TP53* (**Fig. 2.7d-e**), no evidence for chromosome 1p/19q codeletion (data not shown), and KI67 labeling of 4% (data not shown), consistent with a recurrent CNS WHO grade 2 astrocytoma, IDH-mutant. Building on our observations from case 1, we applied the same strategy to case 2, with five modifications. First, we increased power by analyzing more sections (**Table 2.15**). Second, we rotated the sample 90° halfway through sectioning to capture intra-tumoral heterogeneity in orthogonal planes (**Fig. 2.7f**). Third, we inferred CNVs from RNA-seq data instead of DNA methylation data. Fourth, we increased the average sequencing depth for amp-seq data. And fifth, we analyzed single nuclei from interpolated sections to validate predictions from bulk sections (**Fig. 2.7f**).

To identify somatic mutations, we performed WES on DNA from two sections in each plane (22, 46, 85, 123; **Table 2.16**) and the patient's blood. 227 mutations were identified and 74 were selected for amp-seq by clustering WES VAFs to reveal candidate mutations most likely to mark distinct clones (**Table 2.17**). Of these, 58 mutations were verified by amp-seq (**Table 2.18**). As with case 1, downsampling reads spanning IDH1 R132H or TP53 G245V revealed monotonic improvements in VAF estimates as a function of read depth (**Fig. 2.8a-b**). We therefore restricted further analysis of amp-seq data to 27 mutations with high coverage over all tumor sections or strong VAF correlations to other mutations (**Fig. 2.8c**). Hierarchical clustering of these amp-seq data (**Table 2.18**) revealed five clusters of mutations with similar VAF patterns within the tumor sample (**Fig. 2.7g-l**), suggesting multiple malignant clones.

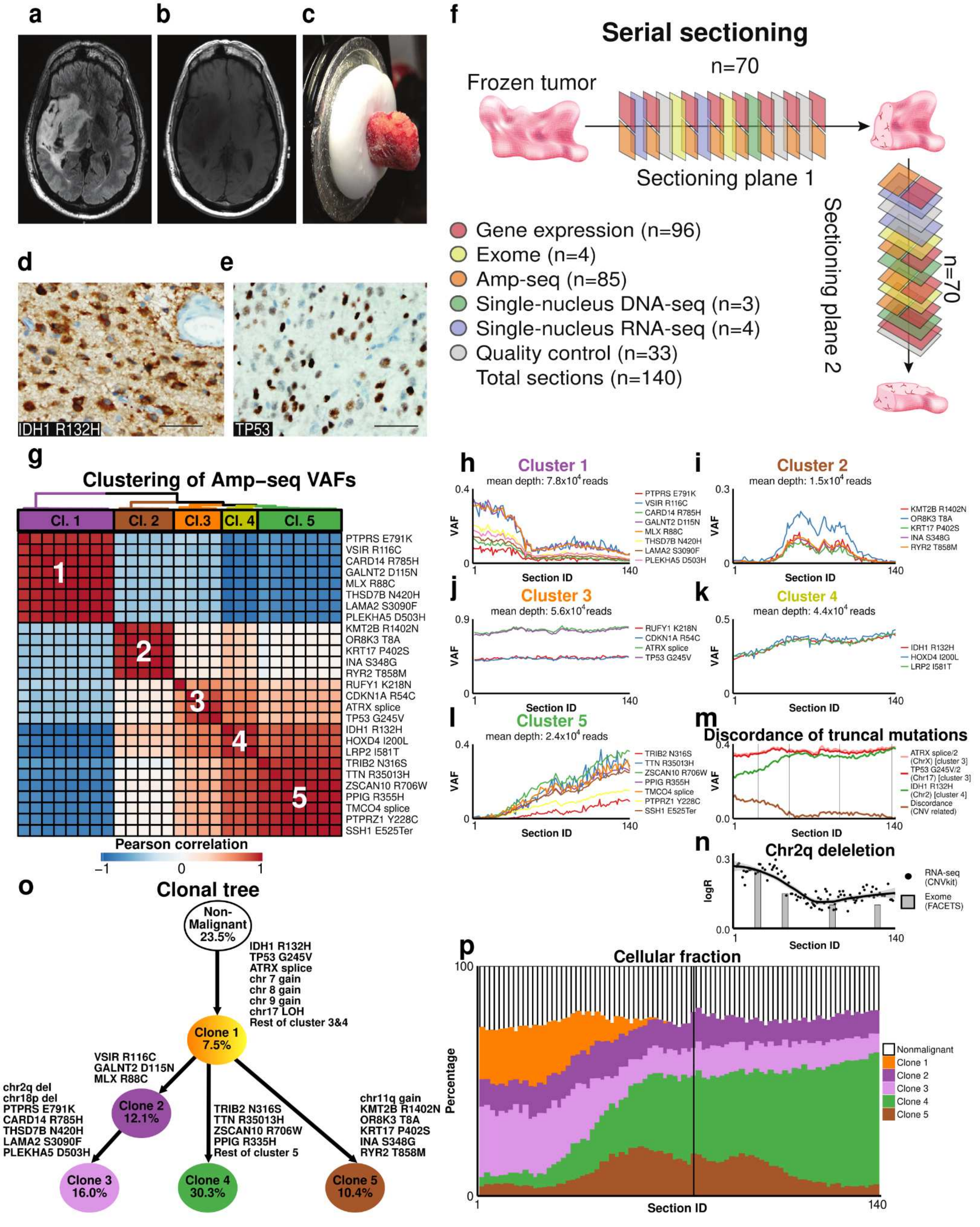


Figure 2.7 | Multiomic analysis of serial tumor sections reveals the clonal composition of a recurrent grade 2 IDH-mutant astrocytoma (case 2).
(Figure caption continued on the next page.)

(Figure caption continued from the previous page.) Axial T2 **(a)** and axial FLAIR **(b)** images demonstrate a non-enhancing, expansile, infiltrating glioma centered in the right insula and involving the basal ganglia, inferior frontal lobe, and temporal lobe. Cystic degeneration was present in the tumor. **c)** Image of the frozen tumor specimen prior to cryosectioning and nucleic acid isolation. **d)** The tumor was determined to harbor the IDH1 R132H mutation based on immunostaining with an antibody specific to the mutant protein. **e)** TP53 immunostaining demonstrated nuclear expression with an estimated staining index of 20%. All histological images were captured at 400x. Scale bars denote 50 µm. **f)** Schematic of serial sectioning strategy and section usage plan. **g)** Hierarchical clustering of mutations, using 1 – Pearson correlation of amp-seq VAFs over all tumor sections (n = 85) as a distance measure, reveals five clusters. **h-l)** VAF patterns comprising cluster 1 **(h)**, cluster 2 **(i)**, cluster 3 **(j)**, cluster 4 **(k)**, and cluster 5 **(l)**. **m)** Controlling for gene dosage reveals discordance of IDH1 R132H VAF with respect to truncal *ATRX* and *TP53* mutations, which is explained by a subclonal deletion of chromosome 2q (including *IDH1*) that occurred after the *IDH1* point mutation. **(n)** Heatmap of the chromosome 2q deletion event frequency (as determined by FACETS[68]), with LOESS fit line (black) and smoothed 95% confidence interval (gray envelope). **o)** Clone phylogeny (with arbitrary branch lengths) derived from integrated analysis of SNVs (from amp-seq data) and CNVs (from RNA-seq data). Percentages represent the average abundance of each cellular fraction over all analyzed sections (n = 85). **p)** Estimated cellular fractions for all clones and nonmalignant cells over all sections. Black vertical line denotes orthogonal sample rotation.

Because mutations in *IDH1*, *TP53*, and *ATRX* are considered diagnostic for astrocytoma[108], we expected these to be truncal and were therefore surprised that IDH1 R132H fell in a separate cluster from mutations in *TP53* and *ATRX* (**Fig. 2.7j-k**). To explore this discrepancy, we analyzed VAFs for all three mutations after controlling for gene dosage. This analysis revealed greater discordance between VAFs for *IDH1* and *TP53* / *ATRX* mutations in sectioning plane 1 vs. sectioning plane 2 (**Fig. 2.7m**). We also observed that all genes in mutation cluster 4 (including *IDH1*) are located on chr2q. These observations suggested that the discrepancy between *IDH1* and *TP53* / *ATRX* mutation VAFs might be explained by a subclonal deletion in chr2q pursuant to the IDH1 R132H mutation, as has been previously reported[110–112]. To test this hypothesis, we quantified CNVs from WES (n=4 sections) and RNA-seq (n=90 sections) data using FACETS[68] and CNVkit[69], respectively, which yielded highly concordant frequencies for copy number changes (r=0.97, **Fig. 2.8d** and **Table 2.19**), including a chr2q deletion event. As expected, frequencies of the chr2q deletion event were substantially higher in sectioning plane 1 vs. sectioning plane 2 (**Fig. 2.7n**) and almost perfectly correlated with the observed discordance between *IDH1* and *TP53* / *ATRX* mutation VAFs (r=0.98, **Fig. 2.8e**).

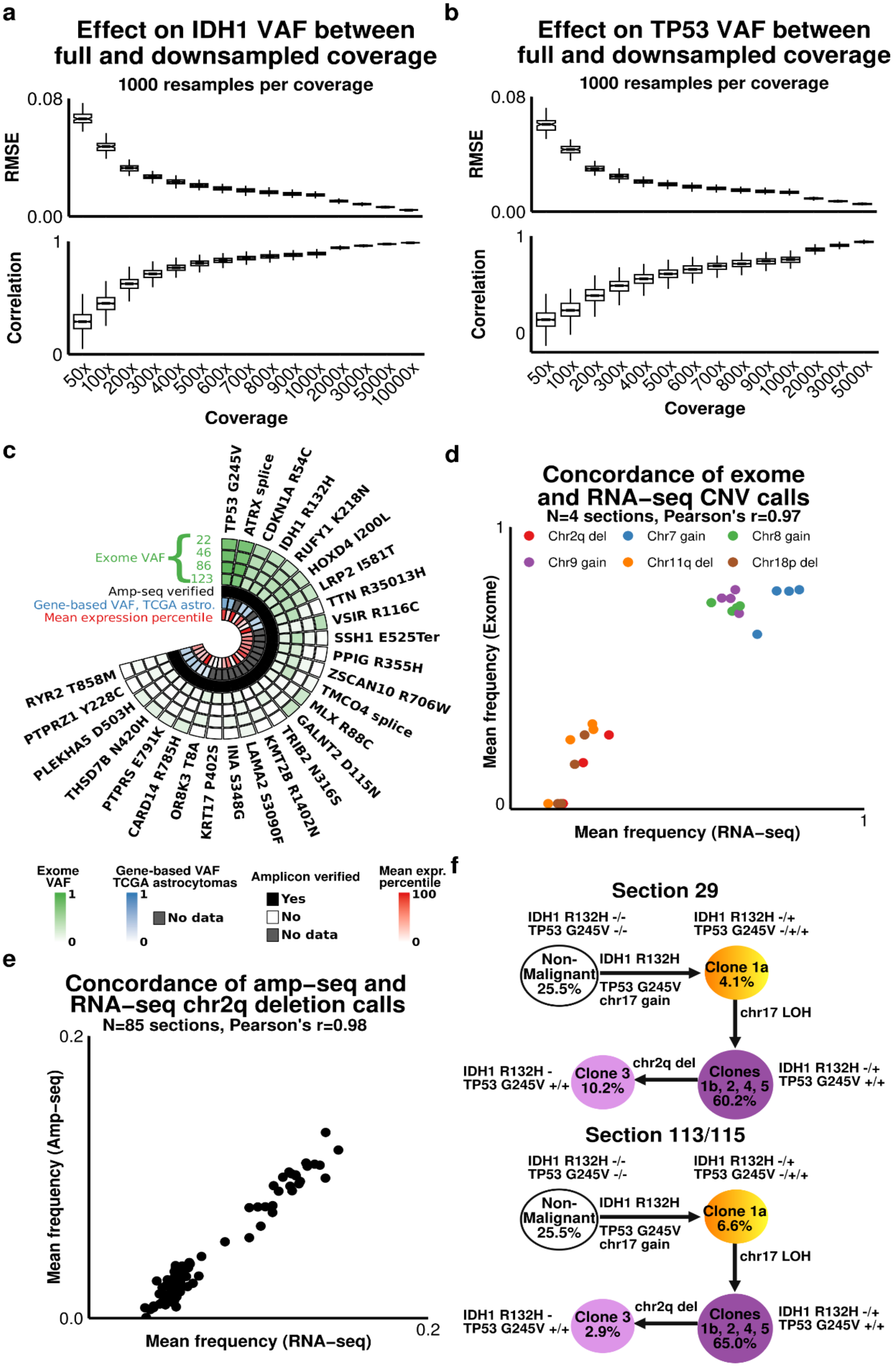

a
Effect on IDH1 VAF between full and downsampled coverage
1000 resamples per coverage
RMSE
0.08
0.00
Correlation
1
0
50x
100x
200x
300x
400x
500x
600x
700x
800x
900x
1000x
2000x
3000x
5000x
10000x
Coverage
b
Effect on TP53 VAF between full and downsampled coverage
1000 resamples per coverage
RMSE
0.08
0.00
Correlation
1
0
50x
100x
200x
300x
400x
500x
600x
700x
800x
900x
1000x
2000x
3000x
5000x
Coverage
c
Exome VAF
22
46
86
123
Amp-seq verified
Gene-based VAF, TCGA astro.
Mean expression percentile
TP53 G245V
ATRX splice
CDKN1A R54C
IDH1 R132H
RUFY1 K218N
HOXD4 I200L
LRP2 I581T
TTN R35013H
VSIR R116C
SSH1 E525Ter
PPIG R355H
ZSCAN10 R706W
TMCO4 splice
MLX R88C
GALNT2 D115N
TRIB2 N316S
KMT2B R1402N
LAMA2 S3090F
INA S348G
KRT17 P402S
OR8K3 T8A
CARD14 R785H
PTPRS E791K
THSD7B N420H
PLEKHA5 D503H
PTPRZ1 Y228C
RYR2 T858M
Exome VAF
1
0
Gene-based VAF TCGA astrocytomas
1
0
No data
Amplicon verified
Yes
No
No data
Mean expr. percentile
100
0
d
Concordance of exome and RNA–seq CNV calls
N=4 sections, Pearson's r=0.97
Chr2q del
Chr7 gain
Chr8 gain
Chr9 gain
Chr11q del
Chr18p del
Mean frequency (Exome)
1
0
1
Mean frequency (RNA–seq)
e
Concordance of amp–seq and RNA–seq chr2q deletion calls
N=85 sections, Pearson's r=0.98
0.2
0.0
0.2
Mean frequency (Amp–seq)
Mean frequency (RNA–seq)
f
Section 29
IDH1 R132H -/-
TP53 G245V -/-
IDH1 R132H -/+
TP53 G245V -/+/+
Non-Malignant 25.5%
IDH1 R132H
TP53 G245V
chr17 gain
Clone 1a 4.1%
chr17 LOH
IDH1 R132H -
TP53 G245V +/+
Clone 3 10.2%
chr2q del
Clones 1b, 2, 4, 5 60.2%
IDH1 R132H -/+
TP53 G245V +/+
Section 113/115
IDH1 R132H -/-
TP53 G245V -/-
IDH1 R132H -/+
TP53 G245V -/+/+
Non-Malignant 25.5%
IDH1 R132H
TP53 G245V
chr17 gain
Clone 1a 6.6%
chr17 LOH
IDH1 R132H -
TP53 G245V +/+
Clone 3 2.9%
chr2q del
Clones 1b, 2, 4, 5 65.0%
IDH1 R132H -/+
TP53 G245V +/+

Figure 2.8 | Mutation validation (case 2).
a-b) Downsampling of amp-seq reads for IDH1 R132H (**a**) and TP53 G245V (**b**) was performed in each tumor section to achieve desired coverage levels (x-axis). For each downsampling (n = 1,000), the root mean square-error (RMSE; top) and Pearson's correlation (bottom) was calculated with respect to the true VAF (calculated using all reads) over all sections (n = 85). **c)** Nonsynonymous mutations were identified by exome sequencing of tumor sections 22, 46, 85, 123, and the patient's blood. Green track: variant allele frequencies (VAF) for each mutation in each section. Black track: mutations validation by amp-seq. Blue track: gene mutation frequencies in TCGA astrocytomas (n = 286). Red track: genome-wide mean expression percentiles over all sections (n = 90). **d)** Concordant estimates of CNV frequencies in the same tumor sections (n = 4) were obtained using FACETS[68] and CNVkit[69] to analyze exome and RNA-seq data, respectively. **e)** Concordant estimates of chromosome 2q deletion frequencies in the same tumor sections (n = 85) were obtained using amp-seq (**Fig. 2.7m**) and RNA-seq, which was analyzed by CNVkit. **f)** Clone phylogeny (with arbitrary branch lengths) derived from single-nucleus amp-seq (snAmp-seq) of mutations affecting the *IDH1* and *TP53* loci for section 29 (n = 4,433 nuclei) and sections 113/115 (n = 3,736 nuclei). Clone names are derived from **Fig. 2.7o**, and the percentages of nuclei assigned to each clone are shown.

Through combined analysis of SNV and CNV frequencies over all tumor sections, we generated an integrated model of tumor evolution using the same approach described for case 1, including the most likely phylogenetic tree (**Fig. 2.7o**) and frequencies of five malignant clones over all sections (**Fig. 2.7p** and **Table 2.20**). Compared to case 1, there was substantially less variation in the purity of individual tumor sections (range: 71.4 - 81.6%; **Table 2.20**). We confirmed the truncal nature of mutations in *IDH1*, *TP53*, and *ATRX*, along with gains of chr7, chr8, and chr9. To more closely examine the sequence of early mutational events, we performed single-nucleus DNA sequencing using MissionBio's Tapestri microfluidics platform[81]. We took advantage of an existing panel of cancer genes, which included primers flanking one *IDH1* and two *TP53* loci. We were also able to infer chr17 and chr2q copy-number changes using mutations that fell within the targeting panel. We analyzed 4,433 nuclei from plane 1 (section 29) and 3,736 nuclei from plane 2 (sections 113 and 115).

Clustering nuclei from each plane revealed clonal frequencies that broadly matched those obtained by bulk analysis (**Fig. 2.8f, Table 2.21**). Interestingly, we observed a subpopulation of clone 1 (clone 1a: 4.1 - 6.6%) with IDH1 R132H -/+ and TP53 G245V -/+/+ genotypes (**Fig. 2.8f**). These genotypes suggest that *TP53* LOH occurred mechanistically in this case through duplication of the mutant allele prior to loss of the wild-type allele, and may also explain the slightly lower VAFs for TP53 G245V compared to the mutation in *ATRX* (**Fig. 2.7j**).

Table 2.2 | Top 64 mutations detected using exome sequencing along with their consequence and indicance in TCGA astrocytoma cases.

Gene	Chr	Consequence	Protein	Mean VAF	TCGA VAF
TP53	chr17	missense_variant	Gly245Val	0.8005	0.643
ATRX	chrX	splice_acceptor_variant		0.7717	0.541
MAGEA12	chrX	missense_variant	Ala28Val	0.7387	NA
CASP1	chr11	missense_variant	Arg161His	0.4261	NA
PAX7	chr1	missense_variant	His335Tyr	0.4082	NA
ADGRE1	chr19	stop_gained	Ser101Ter	0.3944	NA
MYH3	chr17	missense_variant	Glu410Val	0.3837	0.358
CACNA1I	chr22	missense_variant	Ser897Asn	0.3824	0.309
EPB41L3	chr18	missense_variant	Thr759Ser	0.3816	0.551
CDKN1A	chr6	missense_variant	Arg54Cys	0.3803	NA
F2RL1	chr5	missense_variant	Thr301Ile	0.3785	NA
IDH1	chr2	missense_variant	Arg132His	0.3681	0.338
WDR90	chr16	missense_variant	Arg596Cys	0.3632	NA
C14orf37	chr14	missense_variant	Asn428Ser	0.3594	0.624
DLGAP4	chr20	missense_variant	Gln403His	0.3524	NA
RUFY1	chr5	missense_variant	Lys218Asn	0.3420	0.342
ZNF175	chr19	missense_variant	Glu388Gln	0.3380	NA
HOXD4	chr2	missense_variant	Ile200Leu	0.3348	NA
PALB2	chr16	missense_variant	Thr993Met	0.3340	NA
LRP2	chr2	missense_variant	Ile581Thr	0.3248	0.329
OR11L1	chr1	stop_gained	Arg54Ter	0.3230	NA
SLC39A6	chr18	missense_variant	Arg351Gln	0.2966	NA

Gene	Chr	Consequence	Protein	Mean VAF	TCGA VAF
OR52H1	chr11	missense_variant	Ala269Thr	0.2714	NA
LPAR1	chr9	missense_variant	Met64Ile	0.2712	NA
TSPAN13	chr7	missense_variant	Ala32Val	0.2430	NA
TTN	chr2	missense_variant	Arg35013His	0.2173	0.286
HRCT1	chr9	missense_variant	Leu100His	0.1972	NA
FAM90A1	chr12	missense_variant	Arg123Lys	0.1912	0.312
TNS3	chr7	missense_variant	Ile305Thr	0.1830	0.223
VSIR	chr10	missense_variant	Arg116Cys	0.1660	NA
SSH1	chr12	stop_gained	Glu525Ter	0.1588	NA
PPIG	chr2	missense_variant	Arg355His	0.1484	NA
ZSCAN10	chr16	missense_variant	Arg706Trp	0.1399	NA
MUC17	chr7	missense_variant	Ile3060Ser	0.1323	0.137
CHRM3	chr1	missense_variant	Gly358Ala	0.1248	0.213
TMCO4	chr1	splice_region_variant		0.1243	NA
KRTAP5-1	chr11	missense_variant	Ser105Cys	0.1228	NA
MLX	chr17	missense_variant	Arg88Cys	0.1160	NA
CYP2D6	chr22	missense_variant	Arg329Leu	0.1115	NA
HS3ST4	chr16	missense_variant	Thr416Pro	0.1074	0.196
GALNT2	chr1	missense_variant	Asp115Asn	0.1030	NA
FUT9	chr6	missense_variant	Trp318Arg	0.1022	NA
SLC39A7	chr6	missense_variant	Gly457Val	0.0976	NA
TRIB2	chr2	missense_variant	Asn316Ser	0.0947	NA
KMT2B	chr19	missense_variant	Arg1402Gln	0.0882	NA
ALDH1A3	chr15	missense_variant	Val335Met	0.0861	NA
LAMA2	chr6	missense_variant	Ser3090Phe	0.0800	NA
PLIN4	chr19	missense_variant	Val917Met	0.0800	NA
GP1BA	chr17	missense_variant	Ser441Pro	0.0787	NA
INA	chr10	missense_variant	Ser348Gly	0.0780	NA
PPBP	chr4	missense_variant	Ala46Val	0.0777	NA
KRT17	chr17	missense_variant	Pro402Ser	0.0734	NA
OR8K3	chr11	missense_variant	Thr8Ala	0.0732	NA
MUC17	chr7	missense_variant	Val3083Ile	0.0694	0.137
SSPO	chr7	non_coding		0.0675	0.362
CAPN3	chr15	missense_variant	Arg489Trp	0.0675	0.208
KRT86	chr12	missense_variant	Val249Ile	0.0650	NA

Gene	Chr	Consequence	Protein	Mean VAF	TCGA VAF
CARD14	chr17	missense_variant	Arg785His	0.0644	0.195
PTPRS	chr19	missense_variant	Glu791Lys	0.0604	0.171
THSD7B	chr2	missense_variant	Gln420His	0.0599	0.356
PLEKHA5	chr12	missense_variant	Asp503His	0.0592	0.394
FAM149B1	chr10	missense_variant	Ala504Val	0.0570	NA
PTPRZ1	chr7	missense_variant	Tyr228Cys	0.0529	0.327
RPGR	chrX	splice_acceptor_variant		0.0511	NA

2.3.5 Case 2: gene expression

We explored relationships between clonal abundance and bulk gene expression data using the same strategies described for case one. Genome-wide gene coexpression analysis identified 68 modules of coexpressed genes, which were summarized by their eigengenes and hierarchically clustered (**Fig. 2.9a-c**). As expected[23,24,27,28], many modules were significantly enriched with markers of distinct cell types (**Fig. 2.10a-d**). By comparing clonal abundance (**Fig. 2.7p, Table 2.20**) to module eigengenes over all tumor sections, we identified five gene coexpression modules whose expression patterns closely tracked the abundance of clone 1 (red: r = 0.65, **Fig. 2.9d**), clone 2 (violet: r = 0.82, **Fig. 2.9e**), clone 3 (black: r = 0.8, **Fig. 2.9f**), clone 4 (ivory: r = 0.86, **Fig. 2.9g**), and clone 5 (lightcyan: r = 0.82, data not shown).

Enrichment analysis using gene sets defined by clonal CNV boundaries (**Fig. 2.7o** and **Table 2.19**) confirmed expected over-representation (for gains) or under-representation (for deletions) in the bulk coexpression modules most strongly associated with each clone (**Fig. 2.9d-g, Table 2.22, Table 2.23)**. Further analysis using publicly available gene sets from diverse sources (**Table 2.9**) revealed that the red module, which tracked the abundance of clone 1 (i.e., tumor purity), was significantly enriched with markers of radial glia and microglia, as well as genes comprising the mesenchymal subtype of glioblastoma. The violet module, which closely tracked the abundance of clone 2, was significantly enriched with genes from reported astrocytoma expression programs, as well as TNFalpha signaling and extracellular matrix components. The black module, which closely tracked the abundance of clone 3, was significantly enriched with markers of neurons and genes involved in chromatin remodeling. The ivory module, which closely tracked the abundance of clone 4, was enriched with markers of ependymal cells and myeloid cells. The lightcyan module, which closely tracked the abundance of clone 5, was significantly enriched with genes involved in EGFR and NF-kB signaling, as well as genes comprising the proneural subtype of glioblastoma (data not shown).

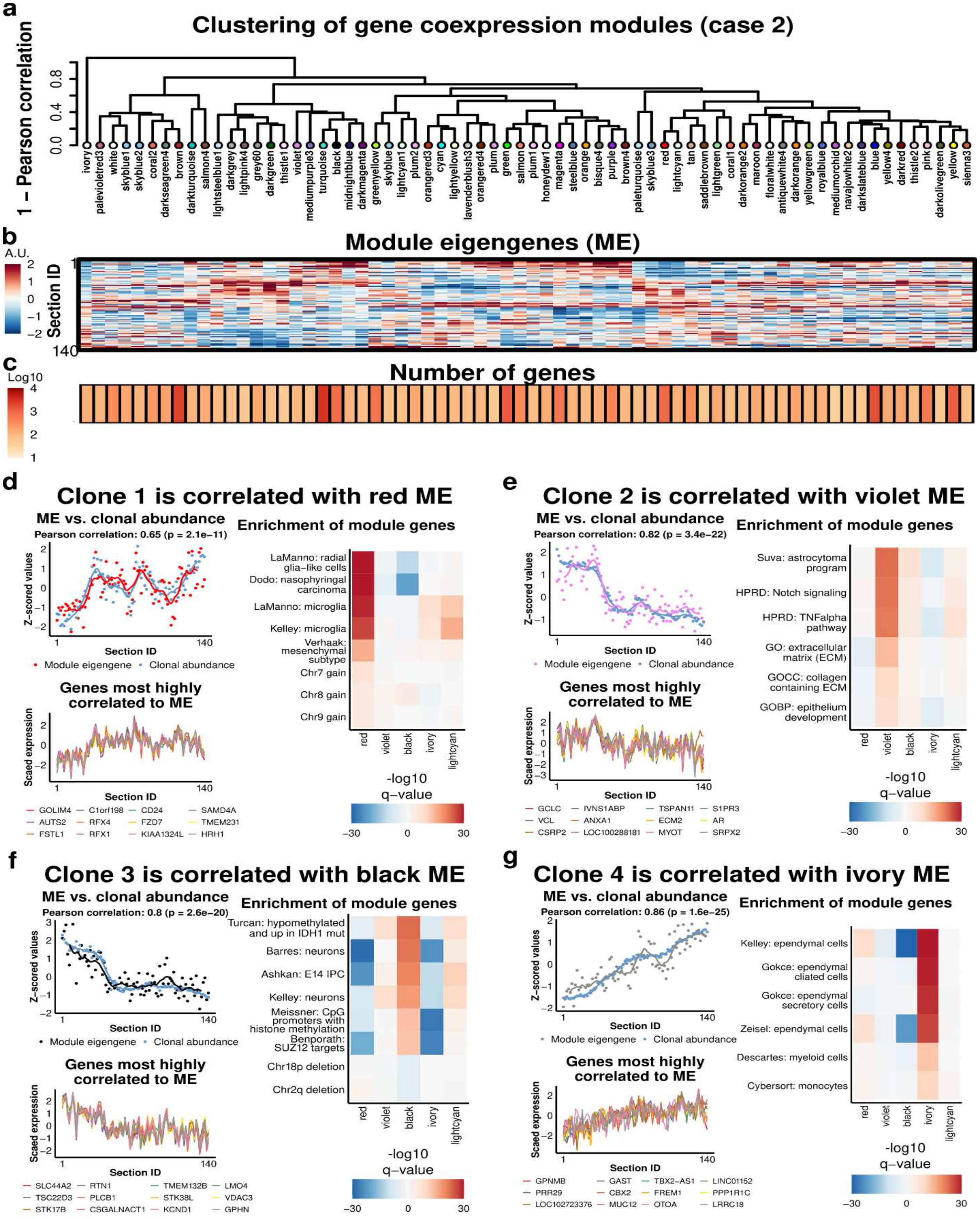
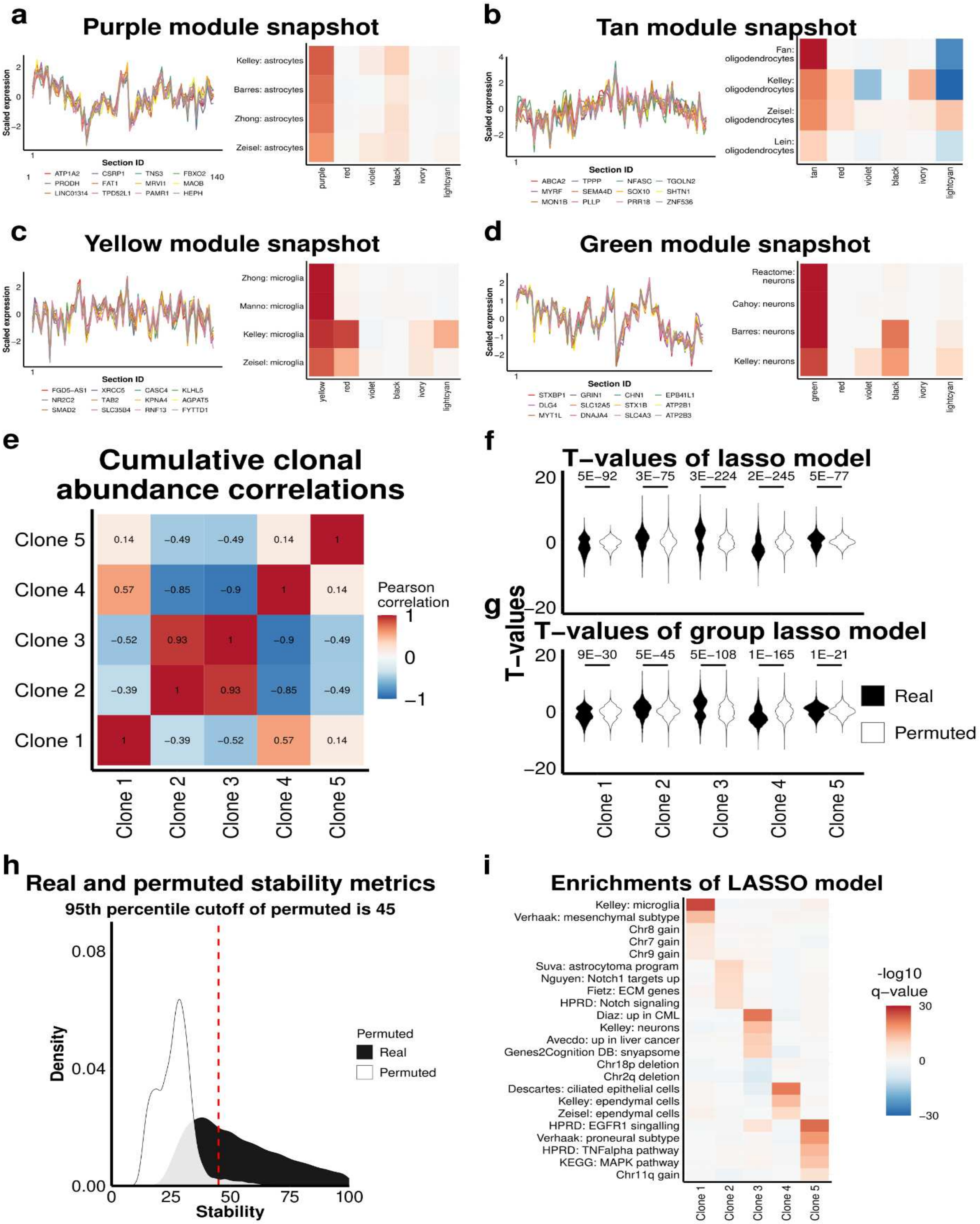


Figure 2.9 | Gene coexpression modules are highly correlated with clonal abundance (case 2).

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.) **a)** Hierarchical clustering of gene coexpression modules over all tumor sections (n = 90). **b)** Module eigengenes (ME) illustrate the relative expression levels of genes in each module over all tumor sections. **c)** The number of genes that formed each ME. **d-g)** Top left: MEs with the strongest correlations to clonal abundance (defined cumulatively). Locally weighted smoothing (LOESS) lines are shown; correlation is based on data points. Bottom left: the 12 genes with the highest correlations to the ME (k_{ME}). Right: enrichment analysis of gene coexpression modules using published gene sets. FDR-corrected p-values (q-values) from one-sided Fisher's exact tests are shown. Positive values represent enrichments of genes that were significantly positively correlated to the ME, while negative values represent enrichments of genes that were significantly negatively correlated to the ME. Gene sets representing chromosomal gains or losses include all genes within affected regions (as described in **Fig. 2.7** and **Table 2.19**). See **Table 2.9** for descriptions and sources of featured gene sets.

To further characterize the transcriptional signatures associated with each clone, we used multiple linear regression to model genome-wide expression levels as a function of clonal abundance. To account for collinearity, we used a regular lasso model with bootstrapped clonal abundance vectors (real or permuted) as predictors (**Fig. 2.10e-i**). We restricted our focus to genes that were significantly and stably modeled by a single clone (**Table 2.24)**. Enrichment analysis of these genes largely recapitulated enrichment analysis of gene coexpression modules associated with each clone, including CNVs and the associations of different clones with different cell types (**Fig. 2.10i, Table 2.25**).



a Purple module snapshot

b Tan module snapshot

c Yellow module snapshot

d Green module snapshot

e Cumulative clonal abundance correlations

f T−values of lasso model

g T−values of group lasso model

h Real and permuted stability metrics

95th percentile cutoff of permuted is 45

i Enrichments of LASSO model

Figure 2.10 | Linear modeling of gene expression using clonal frequencies reveals concordant gene-set enrichments with coexpression modules (case 2).
a-d) Left: snapshots of additional gene coexpression modules enriched for markers of nonmalignant cell types (expression patterns for the top 12 genes ranked by k_{ME} are shown). Right: heatmaps of gene set enrichment results for each module. Modules included genes that were most specifically and significantly correlated (after FDR correction) to the module eigengene (ME), and enrichment was assessed with a one-sided Fisher's exact test (followed by FDR correction; see panel **i** for legend). **e)** Correlation heatmap for the cumulative frequency vectors of identified clones. **f-g)** Lasso regression[75] was used to model the expression of all genes (n = 20,246) as a function of clonal frequencies over all tumor sections (n = 85). Violin plots illustrate the distributions of t-values for all models where the indicated clone was the only explanatory variable that survived lasso selection. Permutations were performed by randomly scrambling clonal frequencies (n = 100) prior to lasso regression. Real and permuted clonal frequency vectors were bootstrapped (n = 100) to address collinearity. P-values denote the significance of the Anderson-Darling test, which evaluates whether two distributions are likely to be derived from the same distribution. **f)** Results of a standard lasso model. **g)** Results of a group lasso model where the truncal clone (equivalent to tumor purity) was placed in a separate group. Unlike case 1, the group lasso model did not outperform the standard lasso model. **h)** Density plot showing the number of times (out of 100 bootstraps) that the same explanatory (clonal frequency vector) was retained by the standard lasso regression model, or 'stability'. The vertical line demarcates the point to the right of which only 5% of values belong to the permuted distribution, i.e. a 5% FDR rate. **i)** Heatmap of FDR-corrected p-values (q-values; shared legend for panels **a-d**) after comparing each gene set to all genes with stability > 45 for a given clone (one-sided Fisher's exact test). Positive values represent enrichments of genes with significant positive correlations to the ME (**a-d**) or significant positive modeling coefficients (**i**), while negative values represent enrichments of genes with significant negative correlations to the ME (**a-d**) or significant negative modeling coefficients (**i**).

To validate gene expression signatures of malignant clones and nonmalignant cell types identified from bulk tumor sections, we performed single-nucleus RNA-seq (snRNA-seq) on tumor sections 17, 53, 93, and 117 (**Fig. 2.7f, Table 2.26**). Using a protocol adapted from TARGET-Seq82,113, we profiled gene expression in 288 flow-sorted nuclei per section. Following data preprocessing and quality control, 809 nuclei (70.2%) with an average of >200K unique reads/nucleus were retained for further analysis. Uniform manifold approximation and projection (UMAP) analysis revealed that nuclei did not segregate by section ID (**Fig. 2.14a, Table 27**).

To determine whether nuclei segregated by cancerous state, we analyzed the malignancy of each nucleus. Unlike some tumors, astrocytomas are not defined by truncal CNVs, which can drive gene expression changes that are used to infer malignancy in snRNA-seq data[88–90,108]. We therefore genotyped all nuclei through single-nucleus amplicon sequencing (snAmp-seq) of cDNA spanning mutations in the truncal clone (**Fig. 2.7o**). This analysis provided sufficient information to call malignancy for 75% of nuclei. Projecting malignancy status onto the UMAP plot revealed clear segregation of malignant and nonmalignant nuclei (**Fig. 2.14b**).

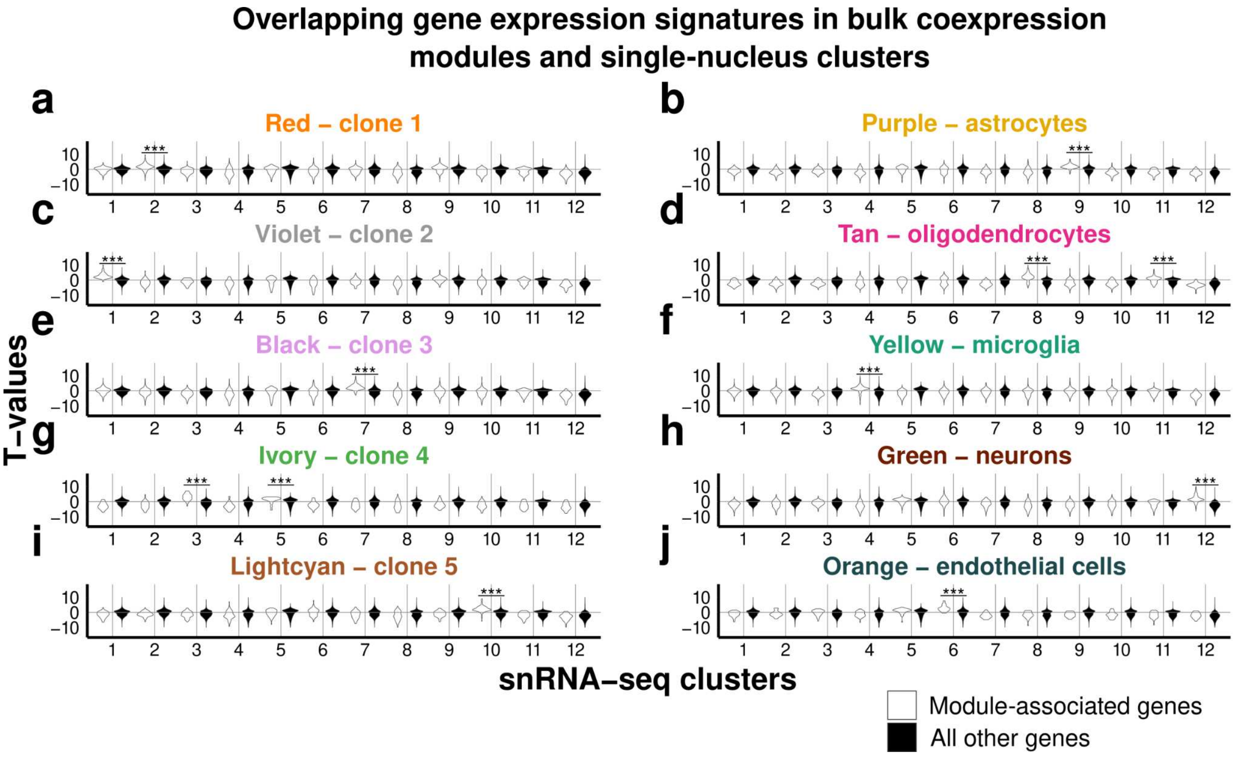


Figure 2.11 | Bulk coexpression module genes map definitively onto single-nucleus clusters.
a-j) Modules of coexpressed genes from bulk tumor sections (n = 90) that were most strongly associated with specific clones (**Fig. 2.9**) or nonmalignant cell classes (**Fig. 2.10**) were evaluated for differential expression in each snRNA-seq cluster vs. all other clusters (white distributions: t-test results for all module genes). (Figure caption continued on the next page.)

(Figure caption continued from the previous page.) Genes that were not associated with each module were evaluated in the same fashion (black distributions), and a one-sided Wilcoxon rank-sum test was used to determine whether module genes were significantly upregulated in a given snRNA-seq cluster relative to all other genes (*** = P < 1e-10).

To further classify nuclei as specific malignant clones or nonmalignant cell types, we took a two-step approach. First, we hierarchically clustered all nuclei using a Bayesian distance metric calculated by Sanity[19] that downweights genes with large error bars, revealing 12 clusters. Second, we asked whether genes in the bulk coexpression modules most strongly associated with each malignant clone or nonmalignant cell type were upregulated in distinct snRNA-seq clusters compared to all other genes (**Fig. 2.11a-j**). This analysis revealed specific and significant upregulation of genes from the red (**Fig. 2.9d**), violet (**Fig. 2.9e**), black (**Fig. 2.9f**), and lightcyan (data not shown) modules in snRNA-seq clusters 2, 1, 7, and 10 (**Fig. 2.12a**), suggesting that these clusters correspond to malignant clones 1, 2, 3, and 5, respectively. Genes in the ivory module (**Fig. 2.9g)** were significantly upregulated in snRNA-seq clusters 3 and 5, suggesting that both of these clusters represent clone 4 (**Fig. 2.12a)**. Similarly, we observed specific and significant upregulation of genes from the purple **(Fig. 2.10a)**, yellow (**Fig. 2.10c**), green (**Fig. 2.10d**), and orange (data not shown) modules in snRNA-seq clusters 9, 4, 12, and 6 (**Fig. 2.12a),** suggesting that these clusters correspond to nonmalignant astrocytes, microglia, neurons, and endothelial cells, respectively. Genes in the tan module (**Fig. 2.10b**) were significantly upregulated in snRNA-seq clusters 8 and 11, suggesting that both of these clusters represent nonmalignant oligodendrocytes (**Fig. 2.12a**).

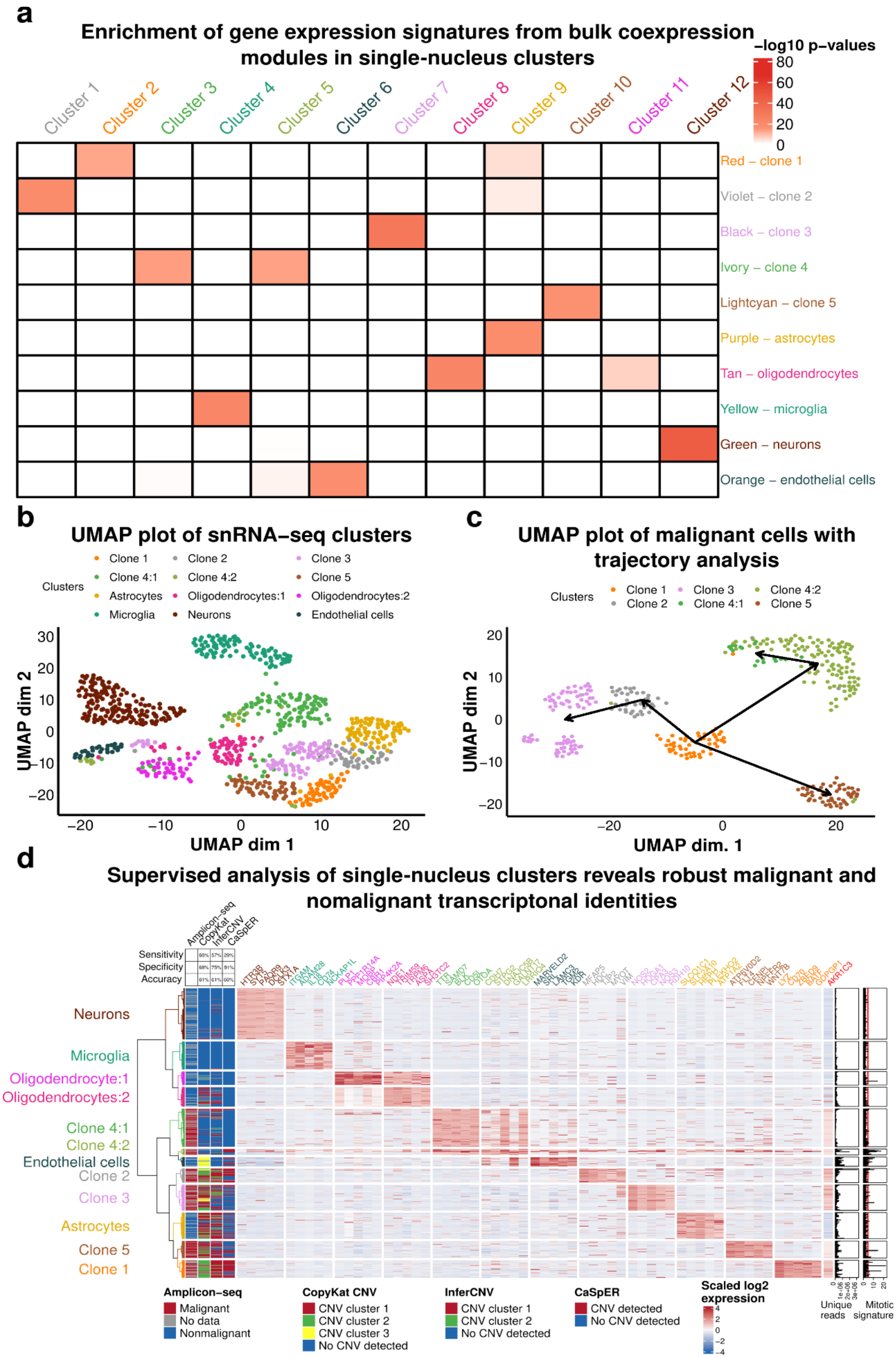

a
Enrichment of gene expression signatures from bulk coexpression modules in single-nucleus clusters
−log10 p−values
80
60
40
20
0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
Cluster 10
Cluster 11
Cluster 12
Red – clone 1
Violet – clone 2
Black – clone 3
Ivory – clone 4
Lightcyan – clone 5
Purple – astrocytes
Tan – oligodendrocytes
Yellow – microglia
Green – neurons
Orange – endothelial cells
b
UMAP plot of snRNA−seq clusters
Clusters
Clone 1
Clone 2
Clone 3
Clone 4:1
Clone 4:2
Clone 5
Astrocytes
Oligodendrocytes:1
Oligodendrocytes:2
Microglia
Neurons
Endothelial cells
UMAP dim 2
30
20
10
0
−10
−20
−20
−10
0
10
20
UMAP dim 1
c
UMAP plot of malignant cells with trajectory analysis
Clusters
Clone 1
Clone 2
Clone 3
Clone 4:1
Clone 4:2
Clone 5
UMAP dim 2
20
10
0
−10
−20
−20
0
20
UMAP dim. 1
d
Supervised analysis of single-nucleus clusters reveals robust malignant and nomalignant transcriptonal identities
Amplicon−seq
CopyKat
InferCNV
CaSpER
Sensitivity
66%
57%
29%
Specificity
58%
75%
91%
Accuracy
61%
61%
60%
HTR3B
SYT5
PAQR9
DCLK3
STX1A
ITGAM
ADAM28
IL18
CD74
NCKAP1L
PLP1
PPP1R14A
MOBP
CBR1
PIP4K2A
NDE1
TRIM59
TRPM6
ASPA
SH3TC2
TTR
SAMD7
BLK
CD5L
OTOA
CDH7
STPG2
DNAJC5B
GALNT17
LRRIQ4
MARVELD2
SRL
LAMC3
TGM2
KDR
MFAP5
HDC
TJP2
MYOT
VIM
NOS2
HOGA1
CPA3
NOS3
RDH10
SLCO1C1
SLC2A10
HEPH
PLEKHO2
ATP1A2
ATP6V0D2
FLT4
CENPL
NPFFR2
WNT7B
LYZ
CD70
ZBED9
BATF
GDPGP1
AKR1C3
Neurons
Microglia
Oligodendrocyte:1
Oligodendrocytes:2
Clone 4:1
Clone 4:2
Endothelial cells
Clone 2
Clone 3
Astrocytes
Clone 5
Clone 1
Amplicon−seq
Malignant
No data
Nonmalignant
CopyKat CNV
CNV cluster 1
CNV cluster 2
CNV cluster 3
No CNV detected
InferCNV
CNV cluster 1
CNV cluster 2
No CNV detected
CaSpER
CNV detected
No CNV detected
Scaled log2 expression
4
2
0
−2
−4
0
1e−06
2e−06
3e−06
0
10
20
Unique reads
Mitotic signature

Figure 2.12 | Single-nucleus analysis supports and refines inferences from bulk data.
a) Heatmap of P-values (one-sided Wilcoxon rank-sum test) comparing differential expression t-values for genes comprising each bulk coexpression module (colors, x-axis) to all other genes in each SN cluster versus all other clusters. **b)** UMAP plot of all nuclei (n = 809) with characterizations of clusters from (**a**) superimposed. **c)** UMAP plot of malignant nuclei (n = 360), with results of Slingshot trajectory analysis[92] superimposed. **d)** Heatmap of scaled $\log_2$ expression vectors for the five most upregulated genes in each snRNA-seq cluster vs. all other clusters (one-sided Wilcoxon rank-sum test). Far left: malignancy vector determined by snAmp-seq of cDNA spanning mutations in the truncal clone. Left: malignancy vectors inferred from CNV analysis of snRNA-seq data using the CopyKat[88], InferCNV[89], or CaSpER[90] algorithms (blue = nonmalignant; all other colors = malignant). Right: bar plots depict the total number of unique reads (UMIs) for each nucleus and the average number of UMIs for genes comprising the Gene Ontology category ‘mitotic chromosome condensation’ (GO: 0030261). Red vertical line: max expression of mitotic genes in neurons, which presumably represents background noise.

We performed several additional analyses to verify these findings. First, we projected snRNA-seq cluster assignments onto the UMAP plot (**Fig. 2.12b)** and observed that cluster assignments were consistent with the malignancy map produced by genotyping nuclei via snAmp-seq (**Fig. 2.14b**). Second, we performed UMAP analysis for malignant cells only, followed by trajectory analysis with Slingshot[92] (**Fig. 2.12c**). This analysis revealed patterns of clonal evolution that recapitulated the phylogenetic tree inferred from integrative analysis of bulk tumor sections (**Fig. 2.7o**). Third, we compared estimates of cellular abundance obtained from bulk and single-nucleus data for adjacent tissue sections. This analysis revealed highly consistent estimates for the relative abundance of malignant clones (r ≥ 0.94; **Fig. 2.14c**) and nonmalignant cell types (r ≥ 0.90; **Fig. 2.14d**).

Supervised clustering with differentially expressed genes revealed clear separation of snRNA-seq clusters (**Fig. 2.12d**). Overall, malignant clones were more transcriptionally active than nonmalignant cell types, with the exceptions of clone 4:1

and endothelial cells (**Fig. 2.12d, right**). Enrichment analysis of genes that were significantly up-regulated in snRNA-seq clusters confirmed the identities of nonmalignant cell types (**Fig. 2.14, Table 2.28)**. For malignant clones, enrichment analysis of snRNA-seq clusters supported and refined inferences from bulk data (**Fig. 2.9d-g, Fig. 2.10i, Fig. 2.14, Table 2.9**). For clone 1, consistent enrichments for markers of radial glia and genes comprising the mesenchymal subtype of glioblastoma were observed in bulk and snRNA-seq data. In contrast, markers of microglia were less significantly enriched in clone 1 nuclei from snRNA-seq data versus bulk data, and markers of oligodendrocyte progenitor cells (OPCs) were more significantly enriched. For clone 2, markers of astrocytes were more significantly enriched in snRNA-seq data versus bulk data. Clone 3 was consistently enriched with genes involved in chromatin remodeling, but neuronal markers were less significantly enriched in snRNA-seq data. Clone 4 showed strong enrichment for markers of ependymal cells in all analyses, while clone 5 was significantly enriched with genes comprising the proneural subtype of glioblastoma in all analyses. Interestingly, genes involved in mitosis were most highly expressed by clone 1, clone 4:2, and endothelial cells (**Fig. 2.12d, right)**.

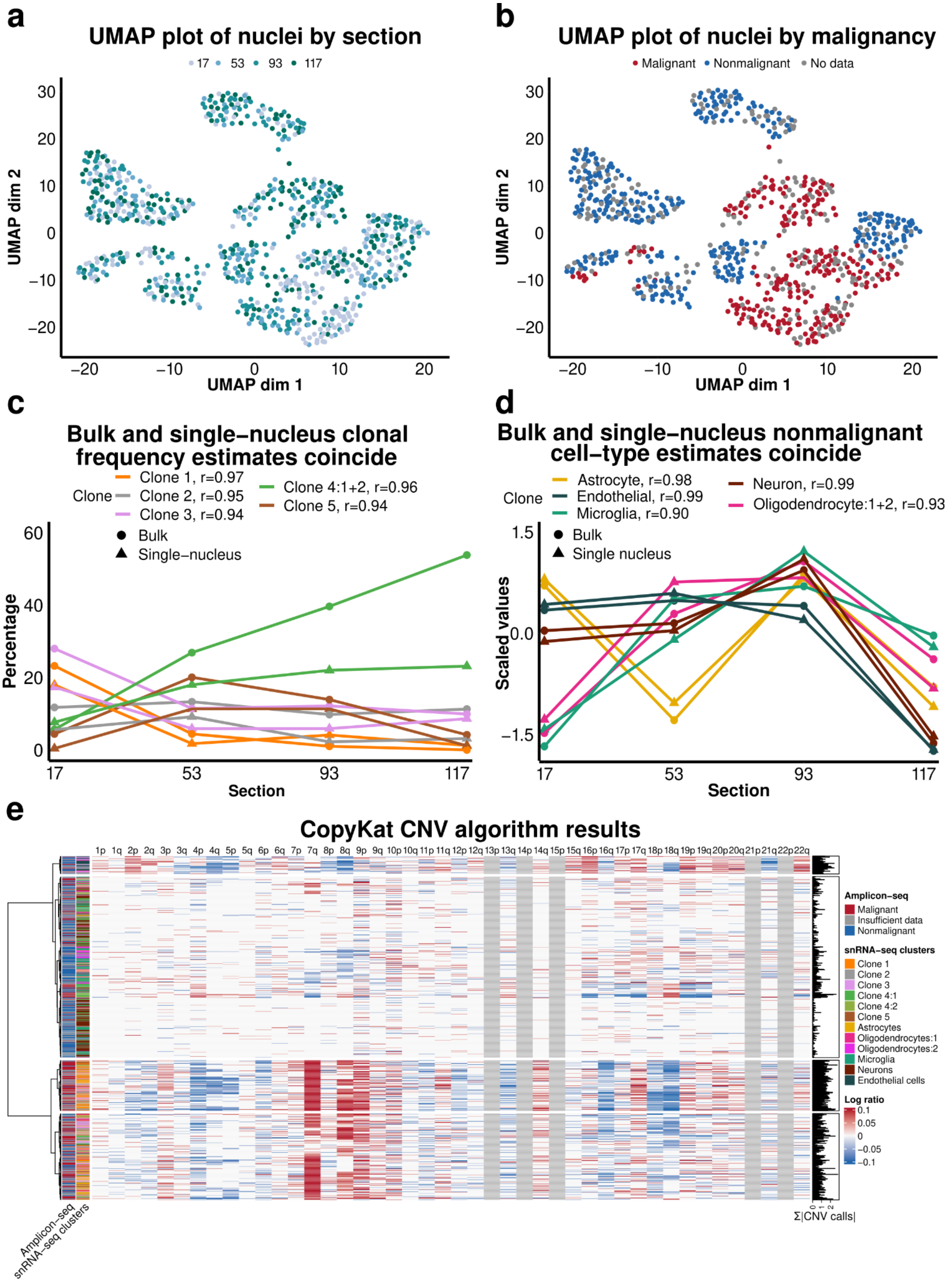

a
UMAP plot of nuclei by section
17
53
93
117
UMAP dim 2
UMAP dim 1
30
20
10
0
−10
−20
−20
−10
0
10
20
b
UMAP plot of nuclei by malignancy
Malignant
Nonmalignant
No data
UMAP dim 2
UMAP dim 1
c
Bulk and single−nucleus clonal frequency estimates coincide
Clone
Clone 1, r=0.97
Clone 2, r=0.95
Clone 3, r=0.94
Clone 4:1+2, r=0.96
Clone 5, r=0.94
Bulk
Single−nucleus
Percentage
60
40
20
0
17
53
93
117
Section
d
Bulk and single−nucleus nonmalignant cell−type estimates coincide
Clone
Astrocyte, r=0.98
Endothelial, r=0.99
Microglia, r=0.90
Neuron, r=0.99
Oligodendrocyte:1+2, r=0.93
Bulk
Single nucleus
Scaled values
1.5
0.0
−1.5
17
53
93
117
Section
e
CopyKat CNV algorithm results
1p 1q 2p 2q 3p 3q 4p 4q 5p 5q 6p 6q 7p 7q 8p 8q 9p 9q 10p 10q 11p 11q 12p 12q 13p 13q 14p 14q 15p 15q 16p 16q 17p 17q 18p 18q 19p 19q 20p 20q 21p 21q 22p 22q
Amplicon−seq
Malignant
Insufficient data
Nonmalignant
snRNA−seq clusters
Clone 1
Clone 2
Clone 3
Clone 4:1
Clone 4:2
Clone 5
Astrocytes
Oligodendrocytes:1
Oligodendrocytes:2
Microglia
Neurons
Endothelial cells
Log ratio
0.1
0.05
0
−0.05
−0.1
0
1
2
Σ|CNV calls|
Amplicon−seq
snRNA−seq clusters

Figure 2.13 | Single-nucleus analysis supports and refines inferences from bulk data.

a) UMAP plot of snRNA-seq data (n = 809 nuclei) with the tumor section IDs that served as the source for each nucleus superimposed. **b)** UMAP plot of snRNA-seq data with malignancy superimposed. Malignancy was determined by genotyping all nuclei via single-nucleus amplicon sequencing (snAmp-seq) of cDNA spanning mutations in the truncal clone. **c)** Frequencies of malignant clones in snRNA-seq data (n = 360 nuclei from four tumor sections) and bulk data (n = 16 tumor sections), with correlations in legend. **d)** Relative abundance of nonmalignant cell types in snRNA-seq data (n = 449 nuclei from four tumor sections) and bulk data (n = 16 tumor sections), with correlations in legend. Estimates were scaled and centered for comparability. Bulk estimates for (**c-d**) are derived from clonal abundance and module eigengene values featured in **Fig. 2.7p** and **Fig. 2.10a-d**, respectively, averaged across the four sections flanking each section analyzed by snRNA-seq (snRNA-seq section 17: bulk sections 14, 16, 18, 19; snRNA-seq section 53: bulk sections 50, 51, 54, 55; snRNA-seq section 93: bulk sections 91, 92, 94, 95; snRNA-seq section 117: bulk sections 114, 116, 118, 119). **e)** Log-ratio output of the CopyKat CNV algorithm[88]. Left: snAmp-seq malignancy assignments and snRNA-seq cluster assignments. Right: sum of the absolute value of CopyKat CNV calls (chromosomal arms in gray could not be called due to inadequate gene coverage).

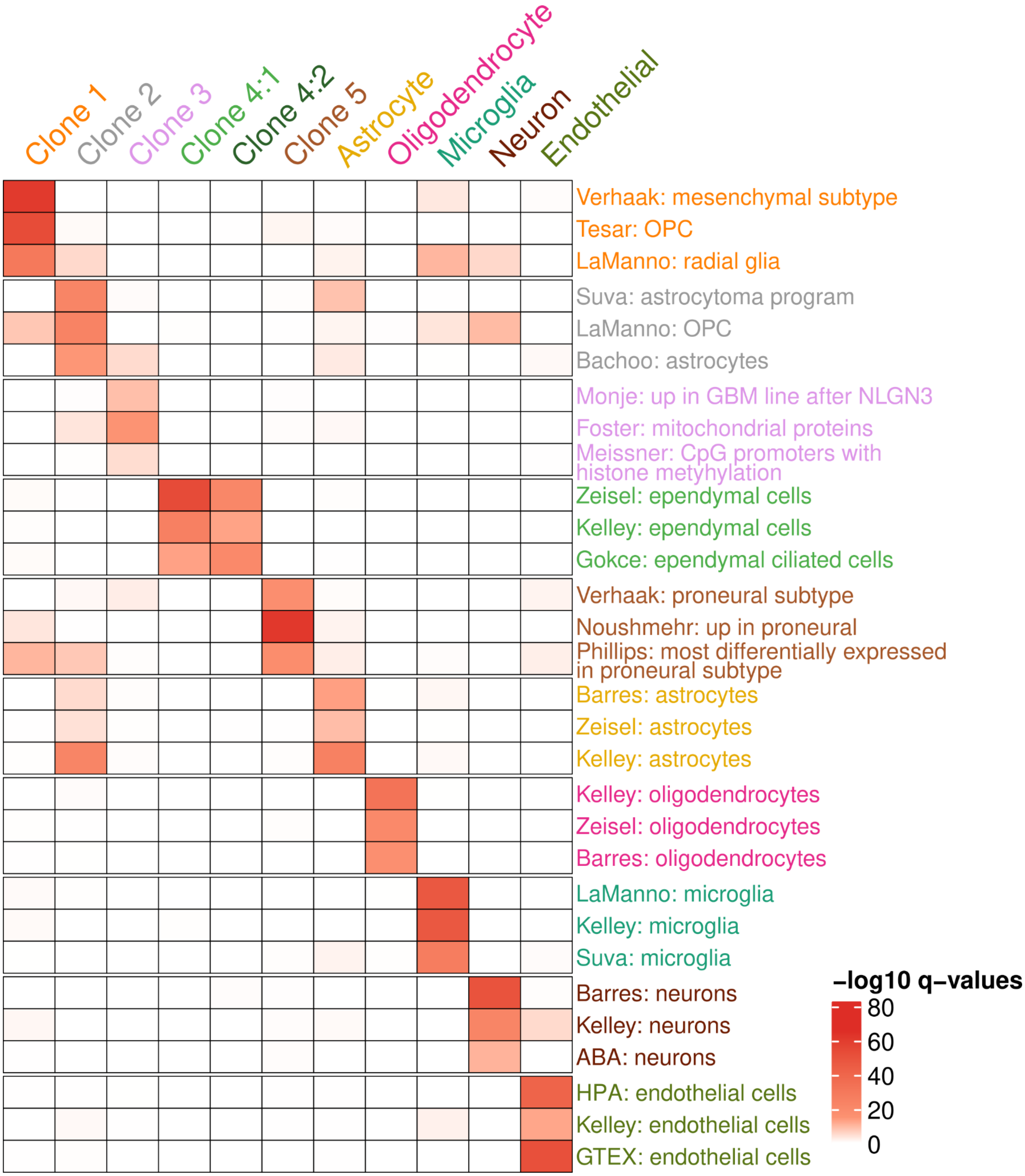


Figure 2.14 | Gene set enrichment analysis supports the functional distinctness of snRNA-seq clusters.
Clustered heatmap of FDR-corrected p-values (q-values) from one-sided Fisher's exact tests comparing featured gene sets with genes that were significantly upregulated (FDR < .05) in each snRNA-seq cluster vs. all other clusters by the one-sided Wilcoxon rank-sum test.

Because clones in this case were characterized by disparate CNVs (**Fig. 2.7o**), we asked how malignancy calls compared between algorithms that infer CNVs from snRNA-seq data and malignant genotypes derived from snAmp-seq data. We used CopyKat[88], InferCNV[89], and CaSpER[90] to call CNVs from snRNA-seq data. These analyses revealed substantial variation in malignancy calls for different algorithms (**Fig. 2.12d**) as well as differences from bulk CNV calls (e.g., no gains in chr7p, chr8p, and chr9q; **Fig. 2.14e**). Taking the snAmp-seq genotyping as ground truth, CopyKat and InferCNV were more sensitive but less specific than CaSpER, leading to discrepant calls. For example, nonmalignant astrocytes and oligodendrocytes:2 were mostly called malignant by CopyKat and InferCNV, while clone 4:2 was mostly called nonmalignant by these two algorithms. CaSpER's classification of nuclei from these populations was mostly correct, but it failed to recognize most malignant nuclei for clones 3 and 5. In addition, clone 4:1 was mostly classified as nonmalignant by all three algorithms. Overall, no method for inferring malignancy from CNVs achieved accuracy > 61% (**Fig. 2.12d)**.

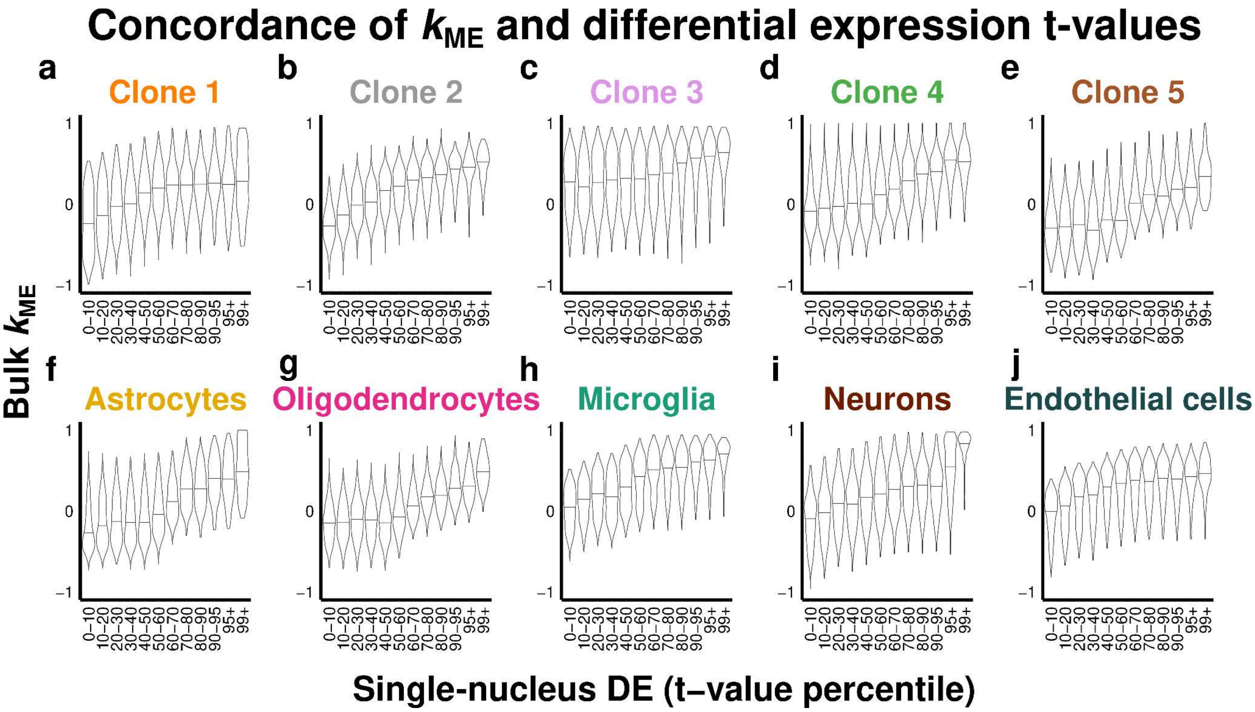


Figure 2.15 | Concordance of k_{ME} and differential expression t-values from bulk and single-nucleus experiments.
a-j) Violin plots reveal the relationship between differential expression t-values for each snRNA-seq cluster (calculated by t-test for all genes between each cluster and all other clusters) and the k_{ME} values of the bulk coexpression module most strongly associated with each clone or nonmalignant cell type.

Multiscale gene expression profiling of this case allowed us to compare the consistency of transcriptional signatures associated with distinct malignant clones and nonmalignant cell types in bulk and snRNA-seq data. Based on our previous findings in normal human brain[23] and **Fig. 1**, we expected differential gene expression in snRNA-seq data to predict gene expression correlations to cellular abundance in bulk RNA-seq data, and vice versa. Genome-wide analysis confirmed this relationship for all malignant clones and nonmalignant cell types (**Fig. 2.16**), further demonstrating that gene expression profiles of distinct cellular populations can be revealed by correlating genome-wide expression patterns with cellular abundance in heterogeneous samples.

2.3.6 Integrative analysis

We next sought to compare transcriptional profiles of malignant cells between case one and case two through integrative analysis. However, despite the fact that both tumors were diagnosed as grade 2 IDH-mutant astrocytomas, only one SNV was shared between the cases. Furthermore, the shared SNV (IDH1 R132H) was absent in ~21% of malignant cells in case 2 following loss of chr2q (**Fig. 2.7o**). We therefore asked whether the truncal clones (i.e., clone 1), which presumably included all of the mutations required to initiate these tumors along with passenger mutations, had consistent transcriptional profiles in case 1 and case 2. For each case, we analyzed genome-wide correlations to the cumulative abundance of clone 1 (equivalent to tumor purity). Comparing these results between cases, we observed a highly significant relationship (**Fig. 2.16a, Table 2.29**). Enrichment analysis of genes whose expression patterns were most positively correlated with clone 1 in both cases implicated gene sets comprising the 'classical' subtype of glioblastoma proposed by Verhaak et al.[17], markers of radial glia, infiltrating monocytes, and extracellular matrix components (**Fig. 2.16a-b**; red). In contrast, genes whose expression patterns were most negatively correlated with clone 1 in both cases largely implicated gene sets related to neurons and neuronal function (**Fig. 2.16a-b**; blue).

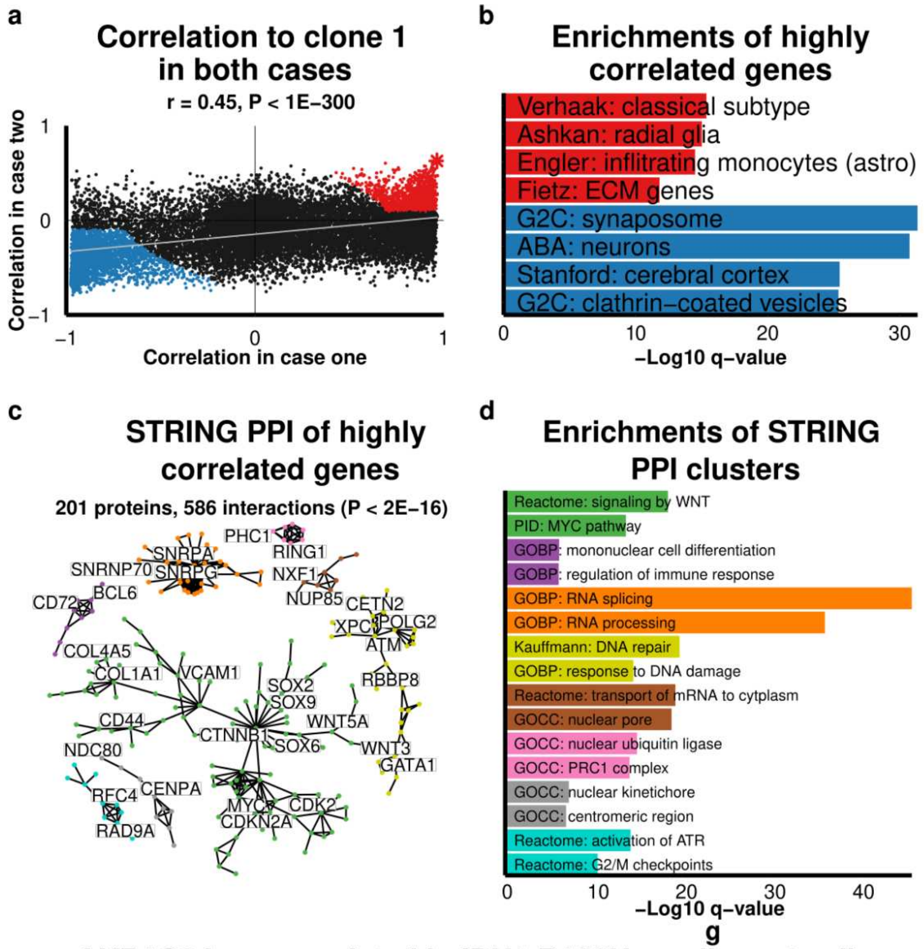
We further characterized genes whose expression patterns were most positively correlated with the truncal clone in both cases (**Fig. 2.16a**; red) by cross-referencing them with human protein-protein interaction (PPI) data from the STRING database[95,96]. This analysis revealed eight distinct clusters of interacting proteins (**Fig. 2.16c**). The largest of these (green) included several SOX transcription factors and was significantly enriched with genes involved in WNT and MYC signaling (**Fig. 2.16d**). The second largest cluster (yellow) was significantly enriched with genes involved in DNA repair, and the third largest cluster (orange) was significantly enriched with genes involved in RNA splicing (**Fig. 2.16d**). The remaining clusters were significantly enriched with genes involved in mRNA transport (brown), DNA replication (turquoise), specific cellular compartments and protein complexes (pink, gray), and immune

response (purple) (**Fig. 2.16d**).

Table 2.3 | Most highly correlated genes to the tumor purity in each case, along with joint correlations and associated p-values.

Gene	Correlation to purity in case one	Correlation to purity in case two	Joint correlation after Fisher's method	FDR-correction of P-value using q-value
AKR1C3	0.9642	0.6278	0.8787	0.0000
NMB	0.9590	0.6275	0.8705	0.0000
VCAM1	0.9674	0.4723	0.8568	0.0000
C1orf94	0.9528	0.5371	0.8428	0.0000
GLIPR2	0.9491	0.5468	0.8390	0.0000
S100A10	0.9399	0.6013	0.8385	0.0000
NUPR1	0.9422	0.5830	0.8373	0.0000
CTSH	0.9553	0.4697	0.8334	0.0000
CD24	0.9551	0.4615	0.8315	0.0000
DFNA5	0.9430	0.5492	0.8309	0.0000
SOX3	0.9465	0.5213	0.8299	0.0000
DYNLT1	0.9608	0.3865	0.8281	0.0000
C1orf194	0.9451	0.5080	0.8248	0.0000
FAM181B	0.9617	0.3495	0.8231	0.0000
TMEM163	0.9355	0.5557	0.8222	0.0000
RND2	0.9657	0.2814	0.8200	0.0000
TMEM218	0.9520	0.4210	0.8180	0.0000
TRAF3IP2	0.9651	0.2717	0.8167	0.0000
ANP32B	0.9549	0.3780	0.8149	0.0000
MARVELD3	0.9507	0.4165	0.8148	0.0000
C1orf198	0.9414	0.4820	0.8137	0.0000
SULF1	0.9320	0.5373	0.8134	0.0000

Gene	Correlation to purity in case one	Correlation to purity in case two	Joint correlation after Fisher’s method	FDR-correction of P-value using q-value
PLTP	0.9554	0.3586	0.8119	0.0000
NES	0.9507	0.3969	0.8108	0.0000
SNCAIP	0.9466	0.4222	0.8090	0.0000
ABI3BP	0.9274	0.5189	0.8031	0.0000
CDCA7	0.9275	0.5166	0.8027	0.0000
GINS3	0.9508	0.3543	0.8024	0.0000
MIF4GD	0.9400	0.4370	0.8017	0.0000



AKR1C3 is up-regulated in IDH1 R132H+ malignant cells

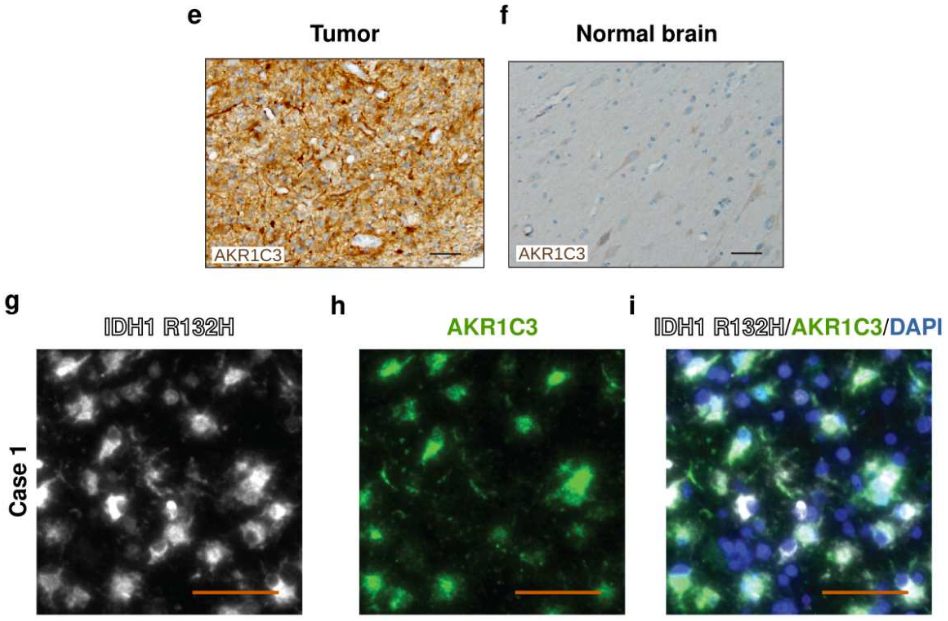


Figure 2.16 | Integrating correlations to malignant cell abundance reveals core transcriptional features of astrocytomas.
a) Gene expression correlations (n = 15,288 genes) to malignant cell abundance in case 1 and case 2. Red and blue denote significantly correlated genes that were used for enrichment analysis **(b)**, and the star denotes *AKR1C3*. **b)** -Log_{10} FDR-corrected p-values (q-values) from one-sided Fisher's exact tests analyzing gene set enrichment in red and blue genes from **(a)**. **c)** Validated protein-protein interactions (PPI) from STRINGdb[96] for red genes from **(a)**. The 201 proteins shown formed networks of five or more proteins, with the number of interactions equal to the number of edges. **d)** -Log_{10} FDR-corrected p-values (q-values) from one-sided Fisher's exact tests analyzing gene set enrichment for each STRINGdb interaction cluster in **(c)**. **e-f)** AKR1C3 immunostaining in FFPE tissue adjacent to the sectioned region of case 1 **(e)** and non-neoplastic human brain **(f)**. Image: 200x; scale bar: 50 µm. **g-i)** Immunofluorescent co-staining of IDH1 R132H (white), AKR1C3 (green), and nuclei (blue [DAPI]) in case one demonstrating expression of AKR1C3 in malignant cells carrying the truncal IDH1 R132H mutation. Scale bar denotes 50µm.

To provide further validation for these findings, we performed immunostaining for AKR1C3. Out of 15,288 genes, AKR1C3 bulk expression correlations to tumor purity ranked fifth in case one and first in case two (**Fig. 2.16a [asterisk], Table 2.29**). AKR1C3 was also significantly upregulated in malignant vs. nonmalignant nuclei per snRNA-seq (**Fig. 2.12m, right**). Immunostaining confirmed substantial upregulation of AKR1C3 in tumor vs. normal human brain at the protein level (Fig. 2.16e, f). To provide cellular resolution, we co-stained for AKR1C3 and IDH1 R132H using an antibody that recognizes the mutated IDH1 protein. As expected, this analysis revealed broad overlap between cells expressing AKR1C3 and cells expressing IDH1 R132H (**Fig. 2.16g-i)**.

2.4 Discussion

Understanding how oncogenic mutations alter gene expression to induce and maintain malignancy is an important goal that may reveal new therapeutic targets for diverse cancers. However, it has been difficult to isolate transcriptional profiles of malignant cells through differential expression analysis of bulk tumor and normal human tissue samples due to variation in the purity, microenvironment, and clonal architecture of tumor specimens. Single-cell methods hold promise for this task but suffer from limited scalability, potentially inaccurate annotation of malignant cells, and technical factors related to tissue dissociation, sampling bias, noise, contamination, and sparsity[18–22]. In this study, we have described an alternative approach for clarifying the transcriptional profiles of malignant clones through multiscale and multiomic analysis of individual tumor specimens.

The premise of our approach is straightforward: variation in the abundance of malignant clones in bulk tumor sections should drive covariation of transcripts that optimally distinguish those malignant clones. The same premise underlies our efforts to determine the core transcriptional identities of nonmalignant CNS cell types through integrative gene coexpression analysis of bulk human brain samples[23]. However, unlike normal brain samples, tumor samples typically include distinct malignant clones defined by partially overlapping sets of mutations. Because most of these mutations are not shared between clones from different individuals[1–5], they may differentially impact gene expression in malignant cells. We therefore sought to apply our strategy to individual tumor specimens and evaluate associations between malignant cell genotypes and gene expression.

By amplifying each tumor specimen into a large number of standardized biological replicates through serial sectioning, we obtained representative subsamples of each tumor with variable cellular composition. Because section size and number can be tailored to experimental needs, this strategy provides flexibility for a variety of concurrent assays while preserving spatial information. We performed WES to identify mutations in a small number of distant sections, followed by deep sequencing of PCR amplicons spanning mutation sites to quantify SNV frequencies with high confidence in a large number of sections. Although clusters of SNVs with highly correlated VAFs suggested distinct clones, we found that integrative analysis of SNV and CNV frequencies (inferred from bulk DNA methylation data [case 1] or bulk RNA-seq data [case 2]) was required to accurately reconstruct clonal phylogenies. Using this approach, we identified the six most prevalent clonal populations of malignant cells in case 1 and five in case 2 and quantified their abundance in all tumor sections.

By comparing clonal abundance to genome-wide expression patterns over all tumor sections, we identified transcriptional profiles of distinct malignant clones in each case. Clone expression profiles were validated through comparisons with normal human brain (case 1) and snRNA-seq using nuclei isolated from interpolated tumor sections (case 2). Enrichment analysis of these profiles revealed several interesting findings. First, gene sets defined by clonal CNV boundaries were significantly enriched (for gains) or depleted (for deletions) in the expected clone expression profiles, providing independent validation of clonal identities. Second, gene sets representing transcriptional subtypes of glioblastoma[17] were significantly associated with distinct clones in each case, suggesting stereotyped patterns of malignant cell differentiation that may reflect different microenvironments[114]. Third, in both cases, markers of neural stem cells (radial glia) were most significantly enriched in the truncal clone. And fourth, markers of ependymal cells were significantly and specifically enriched in clone 4 from case 2. To our knowledge, malignant ependymal cells have not previously been described in human astrocytomas. Because ependymal cells differentiate from neural stem cells during normal brain development[115], the presence of malignant ependymal cells is consistent with a neural stem cell as the cell of origin for case two.

Although both cases were diagnosed as IDH-mutant grade 2 astrocytomas, they shared only one SNV (IDH1 R132H), which was truncal in both cases but lost from 21% of malignant cells (clone 3) in case 2 due to chr2q deletion. The extent of clonal heterogeneity, even for the same type of tumor, begs the question of how gene expression correlations to malignant cell abundance should be aggregated across cases. Here, we reasoned that aggregating gene expression correlations to the truncal clone (equivalent to tumor purity) would identify the most specific and consistent transcriptional features of all malignant cells in both astrocytomas. An alternative strategy is to aggregate correlations to VAFs for specific mutations that are shared among many cases. Both strategies are cumulative in nature and based upon analysis of bulk tumor samples that collectively may represent millions or even billions of cells.

We observed a highly significant genome-wide correlation between gene expression profiles of the truncal clone in both cases, which suggests that a core set of genes is consistently expressed by the founding population of malignant cells in human astrocytomas. This result is particularly striking given the biological and technical differences between case 1 (primary astrocytoma, microarray gene expression data) and case 2 (recurrent astrocytoma, RNA-seq gene expression data). Cross-referencing these genes with human PPI data[96] revealed distinct groups of interacting proteins that were significantly enriched with cancer-related pathways and processes, including WNT and MYC signaling, RNA splicing, and DNA repair. Furthermore, many of the genes whose expression patterns correlated most strongly with malignant cell abundance in both cases (**Table 2.29**) have been implicated in other types of cancer. For example, *AKR1C3*, which encodes a prostaglandin synthase involved in androgen production[116], is significantly upregulated and associated with poor outcomes in hepatocellular carcinoma[117], prostate cancer[118], and pediatric T-cell acute lymphoblastic leukemia[119]. These findings point to the exciting possibility that malignant cells from diverse cancers caused by distinct mutations may nevertheless share transcriptional dependencies that can be exploited therapeutically.

It is also important to note that transcriptional phenotypes of malignancy, including upregulation of *AKR1C3*, persisted in clone 3 from case 2 despite loss of the driver mutation IDH1 R132H following chr2q deletion. IDH1 R132H perturbs genome-wide expression patterns by increasing production of the oncometabolite D-2-hydroxyglutarate[120], which competes with endogenous a-ketoglutarate to alter the activities of enzymes that are required to maintain normal DNA methylation[121]. Our findings suggest that altered DNA methylation patterns can persist and perpetuate malignant phenotypes despite loss of the mutated protein that caused them. This example is illustrative because it highlights the limitations of conventional gene panels for cancer diagnostics, which provide binary calls for the presence or absence of common oncogenic mutations. In this case, such panels would indicate the presence of IDH1 R132H and recommend treatment that targets this mutation[9]. However, with knowledge of this tumor’s clonal phylogeny, we can see that such treatment will be entirely ineffective for one-fifth of malignant cells, since the mutated IDH1 protein is no longer there.

There are several important methodological implications and limitations of our approach. First, each tumor specimen analyzed in this study represents a small fraction of overall tumor volume; future efforts will analyze multiple, geographically distinct tumor subsamples to evaluate the consistency of clonal architecture. Second, our approach requires a large number of sections to detect meaningful correlations (for example: 25 sections provide ~85% power to detect moderate correlations [$|r| > 0.5$, $P < .05$])[33]. Third, DNA and RNA must be co-isolated from each section (i.e., from the same population of cells). Fourth, deep sequencing is required to establish high-confidence VAFs for SNVs, which are in turn required to estimate clonal frequencies. Fifth, limited variability in clonal frequencies may impact the ability to detect corresponding molecular signatures. Sixth, some types of mutations are not yet captured by our approach (e.g., noncoding SNVs, rearrangements, chromothripsis, etc.). And seventh, collinearity in the abundance of malignant and / or nonmalignant cell types may produce spurious correlations (which can be mitigated by differential coexpression analysis with normal tissue, as done for case one, or sectioning in multiple planes, as done for case two). For this reason, we recommend validating transcriptional profiles of malignant clones using one or more independent techniques. We found that multiscale integration of bulk sections and single nuclei allowed us to leverage the complementary strengths of each sampling strategy. Specifically, bulk sections facilitate multiomic integration while yielding robust molecular signatures driven by millions of cells, while single nuclei enable precise validation of predictions made from bulk data. However, the success of this approach depends on accurate classification of malignant nuclei. As we have shown, existing algorithms for identifying malignant nuclei based on inferred CNVs from gene expression data may be inaccurate. Therefore, our approach will benefit from scalable methods for profiling gene expression and malignant cell genotypes in parallel.

In summary, we have described a novel approach for clarifying the transcriptional profiles of malignant clones through multiscale and multiomic analysis of individual tumor specimens. Importantly, our approach is generalizable to other molecular phenotypes and any kind of solid tumor. Ongoing efforts seek to incorporate additional cases, molecular species, and data modalities, while increasing efficiency through automation. By shining a bright light on the most robust molecular properties of malignant clones, we hope that these efforts will expand the therapeutic search space for human cancers.

2.5 References

1. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).

2. Chalmers, Z. R. *et al.* Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med.* **9**, 34 (2017).

3. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).

4. Martincorena, I. *et al.* Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029-1041.e21 (2017).

5. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).

6. Flaherty, K. T. *et al.* Inhibition of mutated, activated BRAF in metastatic melanoma. *N. Engl. J. Med.* **363**, 809–819 (2010).

7. Hahn, W. C. *et al.* An expanded universe of cancer targets. *Cell* **184**, 1142–1155 (2021).

8. Kim, D., Xue, J. Y. & Lito, P. Targeting KRAS(G12C): from inhibitory mechanism to modulation of antitumor effects in patients. *Cell* **183**, 850–859 (2020).

9. Mellinghoff, I. K. *et al.* Vorasidenib in IDH1- or IDH2-Mutant Low-Grade Glioma. *N. Engl. J. Med.* (2023) doi:10.1056/NEJMoa2304194.

10. Dang, C. V., Reddy, E. P., Shokat, K. M. & Soucek, L. Drugging the “undruggable” cancer targets. *Nat. Rev. Cancer* **17**, 502–508 (2017).

11. Russo, M. *et al.* Adaptive mutability of colorectal cancers in response to targeted therapies. *Science* **366**, 1473–1480 (2019).

12. Boumahdi, S. & de Sauvage, F. J. The great escape: tumour cell plasticity in resistance to targeted therapy. *Nat. Rev. Drug Discov.* **19**, 39–56 (2020).

13. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21**, 1350–1356 (2015).

14. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315–322 (2014).

15. Ferreira, P. G. *et al.* Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia. *Genome Res.* **24**, 212–226 (2014).

16. Shen, H. *et al.* Integrated molecular characterization of testicular germ cell tumors. *Cell Rep.* **23**, 3392–3406 (2018).

17. 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).

18. Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* **9**, giaa151. (2020).

19. Breda, J., Zavolan, M. & van Nimwegen, E. Bayesian inference of gene expression states from single-cell RNA-seq data. *Nat. Biotechnol.* **39**, 1008–1016 (2021).

20. Zhang, M. J., Ntranos, V. & Tse, D. Determining sequencing depth in a single-cell RNA-seq experiment. *Nat. Commun.* **11**, 774 (2020).

21. Denisenko, E. *et al.* Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol.* **21**, 130

(2020).

22. Caglayan, E., Liu, Y. & Konopka, G. Neuronal ambient RNA contamination causes misinterpreted and masked cell types in brain single-nuclei datasets. *Neuron* **110**, 4043-4056.e5 (2022).

23. Kelley, K. W., Nakao-Inoue, H., Molofsky, A. V. & Oldham, M. C. Variation among intact tissue samples reveals the core transcriptional features of human CNS cell classes. *Nat. Neurosci.* **21**, 1171–1184 (2018).

24. Oldham, M. C. *et al.* Functional organization of the transcriptome in human brain. *Nat. Neurosci.* **11**, 1271–1282 (2008).

25. Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* **11**, 396–398 (2014).

26. Malikic, S., McPherson, A. W., Donmez, N. & Sahinalp, C. S. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics* **31**, 1349–1356 (2015).

27. Lui, J. H. *et al.* Radial glia require PDGFD-PDGFRβ signalling in human but not mouse neocortex. *Nature* **515**, 264–268 (2014).

28. Raju, C. S. *et al.* Secretagogin is expressed by developing neocortical gabaergic neurons in humans but not mice and increases neurite arbor size and complexity. *Cereb. Cortex* **28**, 1946–1958 (2018).

29. Venteicher, A. S. *et al.* Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science* **355**, (2017).

30. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

31. Langfelder, P. & Horvath, S. Fast R Functions for Robust Correlations and

Hierarchical Clustering. *J. Stat. Softw.* **46**, (2012).

32. Horvath, S. & Dong, J. Geometric interpretation of gene coexpression network analysis. *PLoS Comput. Biol.* **4**, e1000117 (2008).

33. Faul, F., Erdfelder, E., Buchner, A. & Lang, A.-G. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behav. Res. Methods* **41**, 1149–1160 (2009).

34. Johnson, B. E. *et al.* Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science* **343**, 189–193 (2014).

35. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

36. Broad Institute, G. Repository. *Picard Toolkit.* (Broad Institute, 2019).

37. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **11**, 11.10.1-11.10.33 (2013).

38. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).

39. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).

40. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).

41. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).

42. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

43. McLaren, W. *et al.* The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).

44. Ye, J. *et al.* Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13**, 134 (2012).

45. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* (2013) doi:10.48550/arxiv.1303.3997.

46. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

47. Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).

48. Thorndike, R. L. Who belongs in the family? *Psychometrika* **18**, 267–276 (1953).

49. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53–65 (1987).

50. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. & Hornik, K. cluster: Cluster Analysis Basics and Extensions. Available online at: https://CRAN.R-project.org/package=cluster. (2022).

51. Morris, T. J. *et al.* Champ: 450k chip analysis methylation pipeline. *Bioinformatics* **30**, 428–430 (2014).

52. 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).

53. Fortin, J.-P., Triche, T. J. & Hansen, K. D. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics* **33**, 558–560 (2017).

54. Zhou, W., Laird, P. W. & Shen, H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res.* **45**, e22 (2017).

55. Dedeurwaerder, S. *et al.* Evaluation of the Infinium Methylation 450K technology. *Epigenomics* **3**, 771–784 (2011).

56. Teschendorff, A. E. *et al.* A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29**, 189–196 (2013).

57. Oldham, M. C., Langfelder, P. & Horvath, S. Network methods for describing sample relationships in genomic datasets: application to Huntington's disease. *BMC Syst. Biol.* **6**, 63 (2012).

58. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).

59. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

60. Barbosa-Morais, N. L. *et al.* A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. *Nucleic Acids Res.* **38**, e17 (2010).

61. Andrews, S. *et al.* FastQC: a quality control tool for high throughput sequence

data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc. (2010).

62. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet j.* **17**, 10 (2011).

63. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

64. International Human Genome Sequencing Consortium *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

65. Rosenbloom, K. R. *et al.* ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic Acids Res.* **40**, D912-7 (2012).

66. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32**, 896–902 (2014).

67. Feber, A. *et al.* Using high-density DNA methylation arrays to profile copy number alterations. *Genome Biol.* **15**, R30 (2014).

68. Shen, R. & Seshan, V. E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.* **44**, e131 (2016).

69. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput. Biol.* **12**, e1004873 (2016).

70. Olshen, A. B. *et al.* Parent-specific copy number in paired tumor-normal studies using circular binary segmentation. *Bioinformatics* **27**, 2038–2046 (2011).

71. Venkatraman, E. S. & Olshen, A. B. A faster circular binary segmentation

algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657–663 (2007).

72. Glur, C. data.tree: General Purpose Hierarchical Data Structure. Available online at: https://CRAN.R-project.org/package=data.tree. (2020).

73. Iannone, R. DiagrammeR: Graph/Network Visualization. Available online at: https://CRAN.R-project.org/package=DiagrammeR. (2022).

74. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* **100**, 9440–9445 (2003).

75. Klosa, J., Simon, N., Westermark, P. O., Liebscher, V. & Wittenburg, D. Seagull: lasso, group lasso and sparse-group lasso regularization for linear regression models via proximal gradient descent. *BMC Bioinformatics* **21**, 407 (2020).

76. Tibshirani, R. Regression Shrinkage and Selection via the Lasso . *J R Stat Soc Series B Stat Methodol* **58**, 267–288 (1995).

77. Yuan, M. & Lin, Y. Model selection and estimation in regression with grouped variables. *J. Royal Statistical Soc. B* **68**, 49–67 (2006).

78. Laurin, C., Boomsma, D. & Lubke, G. The use of vector bootstrapping to improve variable selection precision in Lasso models. *Stat. Appl. Genet. Mol. Biol.* **15**, 305–320 (2016).

79. Mason, M. J., Fan, G., Plath, K., Zhou, Q. & Horvath, S. Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. *BMC Genomics* **10**, 327 (2009).

80. Tesson, B. M., Breitling, R. & Jansen, R. C. DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinformatics* **11**, 497 (2010).

81. Eastburn, D. J., Sciambi, A. & Abate, A. R. Identification and genetic analysis of cancer cells with PCR-activated cell sorting. *Nucleic Acids Res.* **42**, e128 (2014).

82. Rodriguez-Meira, A., O'Sullivan, J., Rahman, H. & Mead, A. J. TARGET-Seq: A Protocol for High-Sensitivity Single-Cell Mutational Analysis and Parallel RNA Sequencing. *STAR Protocols* **1**, 100125 (2020).

83. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).

84. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

85. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

86. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

87. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).

88. Gao, R. *et al.* Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nat. Biotechnol.* **39**, 599–608 (2021).

89. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).

90. Serin Harmanci, A., Harmanci, A. O. & Zhou, X. CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data. *Nat.*

Commun. **11**, 89 (2020).

91. Melville, J. uwot: The Uniform Manifold Approximation and Projection (UMAP) Method for Dimensionality Reduction. (2021).

92. Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).

93. Untergasser, A. *et al.* Primer3--new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115 (2012).

94. Frank, D. N. BARCRAWL and BARTAB: software tools for the design and implementation of barcoded primers for highly multiplexed DNA sequencing. *BMC Bioinformatics* **10**, 362 (2009).

95. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

96. Szklarczyk, D. *et al.* The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605–D612 (2021).

97. Butts, C. T. network : A Package for Managing Relational Data inR. *J. Stat. Softw.* **24**, (2008).

98. Bojanowski, M. intergraph: Coercion Routines for Network Data Objects. Available online at: http://mbojan.github.io/intergraph. (2015).

99. Briatte, F. ggnetwork: Geometries to Plot Networks with “ggplot2”. Available online at: https://CRAN.R-project.org/package=ggnetwork. (2021).

100. Karlsson, M. *et al.* A single-cell type transcriptomics map of human tissues. *Sci.*

Adv. **7**, (2021).

101. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis (Use R!)*. 276 (Springer, 2016).

102. Dowle, M. & Srinivasan, A. data.table: Extension of `data.frame`. Available online at: https://CRAN.R-project.org/package=data.table. (2021).

103. Neuwirth, E. RColorBrewer: ColorBrewer Palettes. Available online at: https://CRAN.R-project.org/package=RColorBrewer. (2022).

104. Auguie, B. gridExtra: Miscellaneous Functions for “Grid” Graphics. Available online at: https://CRAN.R-project.org/package=gridExtra. (2017).

105. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).

106. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize Implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).

107. Ahlmann-Eltze, C. & Patil, I. ggsignif: R Package for Displaying Significance Brackets for “ggplot2.” (2021) doi:10.31234/osf.io/7awm6.

108. Louis, D. N. *et al.* The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro Oncol.* **23**, 1231–1251 (2021).

109. Phillips, H. S. *et al.* Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* **9**, 157–173 (2006).

110. Mazor, T. *et al.* Clonal expansion and epigenetic reprogramming following deletion or amplification of mutant IDH1. *Proc Natl Acad Sci USA* **114**, 10743–10748 (2017).

111. Favero, F. *et al.* Glioblastoma adaptation traced through decline of an IDH1 clonal driver and macro-evolution of a double-minute chromosome. *Ann. Oncol.* **26**, 880–887 (2015).

112. Pusch, S. *et al.* IDH1 mutation patterns off the beaten track. *Neuropathol. Appl. Neurobiol.* **37**, 428–430 (2011).

113. Rodriguez-Meira, A. *et al.* Unravelling Intratumoral Heterogeneity through High-Sensitivity Single-Cell Mutational Analysis and Parallel RNA Sequencing. *Mol. Cell* **73**, 1292-1305.e8 (2019).

114. Garcia-Diaz, C. *et al.* Glioblastoma cell fate is differentially regulated by the microenvironments of the tumor bulk and infiltrative margin. *Cell Rep.* **42**, 112472 (2023).

115. Spassky, N. *et al.* Adult ependymal cells are postmitotic and are derived from radial glial cells during embryogenesis. *J. Neurosci.* **25**, 10–18 (2005).

116. Penning, T. M. AKR1C3 (type 5 17β-hydroxysteroid dehydrogenase/prostaglandin F synthase): Roles in malignancy and endocrine disorders. *Mol. Cell. Endocrinol.* **489**, 82–91 (2019).

117. Zhou, Q. *et al.* A Positive Feedback Loop of AKR1C3-Mediated Activation of NF-κB and STAT3 Facilitates Proliferation and Metastasis in Hepatocellular Carcinoma. *Cancer Res.* **81**, 1361–1374 (2021).

118. Liu, C. *et al.* Intracrine androgens and AKR1C3 activation confer resistance to enzalutamide in prostate cancer. *Cancer Res.* **75**, 1413–1422 (2015).

119. Bortolozzi, R. *et al.* AKR1C enzymes sustain therapy resistance in paediatric T-ALL. *Br. J. Cancer* **118**, 985–994 (2018).

120. Dang, L. *et al.* Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. *Nature* **462**, 739–744 (2009).

121. Xu, W. *et al.* Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of α-ketoglutarate-dependent dioxygenases. *Cancer Cell* **19**, 17–30 (2011).

Chapter 3: Metaanalysis of glioma samples reveals unique, cell-type specific dysregulated genes of the microenvironment.

Having established the significance of correlative deconvolution methods for malignant samples, I wish to expand this paradigm to the deconvolution of the nonmalignant microenvironment.

Chapter 3.1: Introduction

Having established the utility of using coexpression analysis to track the abundance and transcriptional output of malignant cells, we now wish to interrogate the relative abundance and dysregulation of expression in nonmalignant cells. The fundamental framework for this method has been established in a previous publication of the lab[1], though we will be expanding this methodology to allow direct comparison between normal and malignant brain tissue.

Nomalignant cells have been shown to be key in the tumor growth and survival, adopting both tumor suppressive[2] and tumor inhibitory[3] phenotypes. Identifying these factors of dysregulation in the microenvironment represents an exciting, new realm for drug discovery as malignancies are more likely to share microenvironment phenotypes than their mutational background[4,5], as well as having less opportunity to evolve resistance to microenvironmental changes[6,7].

The central thesis of this approach is to characterize the difference between the correlation matrices of normal and malignant brains to identify modules of coexpressed genes which are specific to the malignant context. Instead of using serial-sections I have used the wealth of publicly-available transcriptomic datasets to compute dataset-specific networks of coexpressed genes. Using gene set enrichment analyses I have chosen modules maximally and uniquely enriched in cell-types of interest. Finally, I use statistical techniques to average the correlations of genes to the module eigengene. This correlation statistic has been shown to be highly correlated to the degree of cell-type specific expression and it allows for the quantification of the degree to which each gene is expressed in specific cell types across thousands of samples[1]. By comparing the averaged correlations between malignant and normal samples, a total sample number exceeding 13,000, it is possible to identify genes which are highly dysregulated in specific cell-types in glioma, but not in a normal context. These genes represent ideal targets for therapeutic intervention targeting the glioma microenvironment.

In order to allow for greater community access to our data, as well as to allow easy orthogonal validation of our findings using publicly available single-cell and single-nucleus transcriptomic data, we have also developed an R-shiny application. Overlaying our data with functionally relevant database annotations, such as cellular localization of the gene product[8] and essentiality of genes from knock-out screens in cancer cell lines[9], will allow this application to be useful tool in identifying potential targets for drug development against dysregulated genes in the glioma microenvironment.

Chapter 3.2: Methods

3.2.1 Sample Sourcing

Our glioma cohort is composed of 47 datasets (5450 samples) and includes grades two through four of high and low grade glioma (including IDH mutant and wild-type gliomas). This broad sampling ensures that the conclusions drawn from the data are minimally biased towards any histological subtype and generalizable across glioma. Our normal cohort is composed of 62 datasets (7269 samples) from the human brain and includes the frontal, temporal and parietal regions, reflecting the spatial bias of gliomas. The exact breakdown of samples across platform, source, type and grade is shown in **Table 3.1**.

Table 3.1 | Distribution of normal and glioma samples by platform and source. Additionally, distribution of glioma samples by type and grade.
In some cases grade was indicated as low grade and indicated in the table as grade 2/3. Furthermore, there were samples simply annotated as low-grade glioma (of undetermined type) which are listed as "Low-grade glioma" for type and 2/3 for grade.

	Glioma	**Normal**
Platform		
Affy Hg U133	998	422
Affy Hg U133	1879	202
Affy HuEx	242	1329
Agilent	448	3633
Illumina	1883	1683
Source		
TCGA	1956	0
EMTAB	404	0
CGGA	301	0
GSE	2829	2579
GTEX	0	877
ABI	0	3633
Other		180
Total	**5450**	**7269**

Type	**Grade**	**Count**
GBM	4	2271
Astrocytoma	4	455
Astrocytoma	3	369
Astrocytoma	2	281
Astrocytoma	2/3	85
Oligodendroglioma	3	505
Oligodendroglioma	2	217
Oligodendroglioma	2/3	40
Low-grade glioma	2/3	1227

3.2.2 Sample preprocessing

Sample preprocessing was performed as described in Kelley et al.[1]. Briefly, batch information and other metadata were derived from supplementary files on GEO, when available, as well as from ".CEL" files when applicable. Care was taken to process each unique sample only once, and when a sample was included in multiple datasets, it was only analyzed in the original dataset.

Each dataset was processed using the SampleNetwork function[10]. Sample outliers were removed if their connectivity was more than 4 standard deviations below the mean connectivity. Connectivity is a measure of the degree of relatedness between samples and is roughly equivalent to correlation.

Data was quantile normalized and batch correction was performed if a significant association between batch and the first principal component was found via ANOVA. Batch correction was performed using the ComBat R function[11]. Depictions of the dataset before and after preprocessing can been seen in **Figure 3.1**.

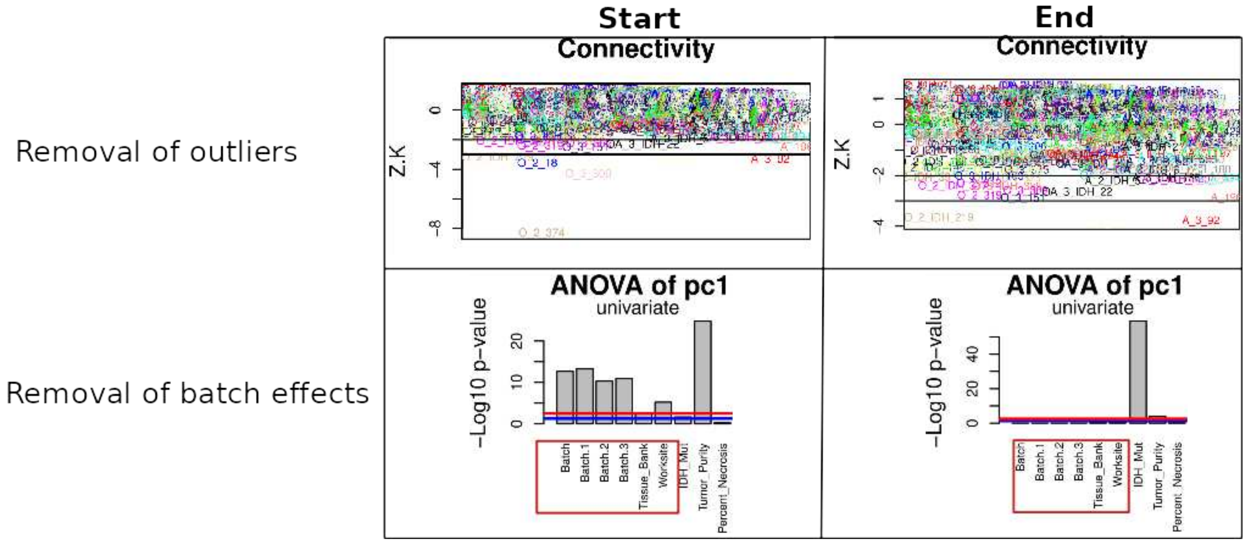


Figure 3.1 | TCGA-LGG cohort data before and after dataset preprocessing. Sample connectivity and significance values of ANOVA of PC1 to batch effects before and after sample processing. Samples with connectivity values (Z.k) more than four standard deviations from the mean are removed and batch effects corrected using the ComBat package. The lack of highly disconnected samples (as quantified by Z.k) and elimination of significant association of PC1 with batch effect shows successful sample processing.

3.2.3 Construction of Coexpressed modules for each dataset

Correlations within datasets were calculated as biweight midcorrelation for all features[12], as this is a correlation metric with reduced sensitivity to outliers while maintaining power to detect correlations[12]. Clustering on the correlation space was performed using flashClust[12] using 1-correlation as a distance metric.

The hierarchically clustered dendrogram was cut a series of heights defined by hyperparameters for the top 0.01%, 0.1%, 1%, 2%, 3%, 4%, or 5% of pairwise correlations for the entire data set and a minimum module size of 8, 10, 12, 15, or 20 members. We used the first principal component (via singular-value decomposition) of each resulting module to determine the degree of correlation between modules. Modules with a first principal component (which I refer to as the module eigengene, ME) correlated above 0.85 were merged to focus on identifying unique patterns of coexpression.

The correlation of each gene to the MEs was calculated as a quantitative measure of belongingness of the genes to the modules, a metric defined as intramodular connectivity (k_{ME})[13,14].

3.2.4 Assignment of Genes to Modules

Module membership was assigned to all unique genes with positive kME values that were significant after applying a False-discovery rate correction[15] for multiple comparisons. If a probe/gene was significantly correlated with more than one module, it was assigned to the module for which it had the highest kME value. Probe/gene identifiers from all data sets were mapped to a common identifier (HomoloGene ID data build 68).

3.2.5 Enrichment Analysis

To identify modules enriched in specific cell-types I used the top 150 genes with the highest fidelity for distinct brain cell-types, derived from the lab's publication[1]. Enrichment analysis in each dataset was conducted using a one-sided Fisher's exact test as implemented by the fisher.test R function[16]. The module with the most significant enrichment for each cell-class gene set was identified. Only modules that were significantly enriched after applying a False-discovery rate correction[15] were used in subsequent analyses.

3.2.6 Mathematical Derivation of Fidelity

Because correlation coefficients, like k_{ME} are correlation coefficients, they cannot be averaged over independent data sets of different sample sizes. Instead Fisher's method can be used [16]. The first step is to transform the correlation coefficient kME values into Z-scores:

$$z_{gdc} = \frac{1}{2} ln \left(\frac{1 + k_{ME.gdc}}{1 - k_{ME.gdc}} \right)$$

where g indexes the gene, d indexes the data set, and c indexes the cell class. An average of the resulting z-scores (weighted by sample size) was then determined with the following equation:

$$\bar{z}_{gc} = \frac{\sum_{d=1}^{D} z_{gdc}\left(n_d - 3\right)}{\sum_{d=1}^{D}\left(n_d - 3\right)}$$

where n denotes the number of samples in data set d. The sampling s.d. of $\bar{z}_{gc}$ is:

$$SD\left(\bar{z}_{gc}\right) = \sqrt{\frac{1}{\sum_{d=1}^{D}\left(n_d - 3\right)}}$$

Dividing the ‘average’ z-scores by the sampling s.d. yielded the genome-wide statistics or gene expression fidelity:

$$fidelity = Z_{gc} = \frac{z_{gc}}{SD\left(\bar{z}_{gc}\right)}$$

For interpretability, we also converted z_{gc} into an ‘average’ correlation coefficient by performing the reverse Fisher transformation:

$$\bar{r}_{gc} = \frac{e^{2\bar{z}_{gc}} - 1}{e^{2\bar{z}_{gc}} - 1}$$

which is reported as ‘Mean.r’ along with expression fidelity for all genes with respect to all cell classes for humans. It is important to note that gene expression fidelity, as defined here, is robust to the choice of gene set used for enrichment analysis.

3.2.7 Expression

The mean rank expression percentile was calculated by first converting the published quantification metric of the dataset (TPM, CPM, FPKM, etc.) into ranks, and then taking an average for each gene across all samples.

3.2.8 Outside Data Sources

We used multiple orthogonal data-sources to validate our analysis regarding dysregulation in nonmalignant cells and provide context for putative therapeutic targets.

3.2.8.1 COMPARTMENTS

Cellular localization for mouse and human genes were extracted from the COMPARTMENTS resource, utilizing only level 5 confidence annotations[8].

3.2.8.2 DepMap

We used the DepMap "Gene Dependency" metric for cell lines of the lineage subtype "Glioma" to calculate the mean dependency metric for each gene[9].

3.2.8.3 GTEX Human Tissue Map

We used the Genotype-Tissue Expression (GTEx) database to quantify expression of genes across tissues[17]. Expression was averaged across replicates and two standard error values were calculated. Tissues with fewer than 10 replicates were excluded.

3.2.8.4 IVY-GAP GBM Atlas

We used the Ivy-Gap Glioblastoma Atlas to measure gene expression across tumor regions[18]. Replicates were averaged and two standard errors were calculated.

Single-cell datasets

We used the single-cells from the control dataset, along with the author's annotation of cell identity to quantify average expression of genes at the single-cell level along with two standard error measurements of variance[19]. We used the purified single-cells GBM dataset to quantify average expression per cell-type as well as two standard errors of variation[20].

Chapter 3.3: Results

Our pipeline's (**Fig. 2.4.2**) goal is finding genes which are excellent markers for cell-types in the tumor microenvironment, but are non expressed in the normal brain. The initial steps of using SampleNetwork for batch correction and outlier removal and FindModules for the creation of coexpression networks of broad resolution is detailed in the methods. The enrichment analysis portion is extremely flexible and can be adapted to target genesets for any relevant biological context or cell-type of interest. Furthermore, datasets with no significant geneset enrichment are not included in the analysis, ensuring that priors regarding gene expression for certain cell-types do not introduce bias into this analysis.

Next, the Fisher transformation allows averaging of correlation from differently-sized datasets. This is a very efficient transformation and recomputing fidelity after the addition of additional data is trivial. Finally, computing differential fidelity gives a direct measurement of which genes are highly expressed in a cell-type in the glioma but not normal brain context.

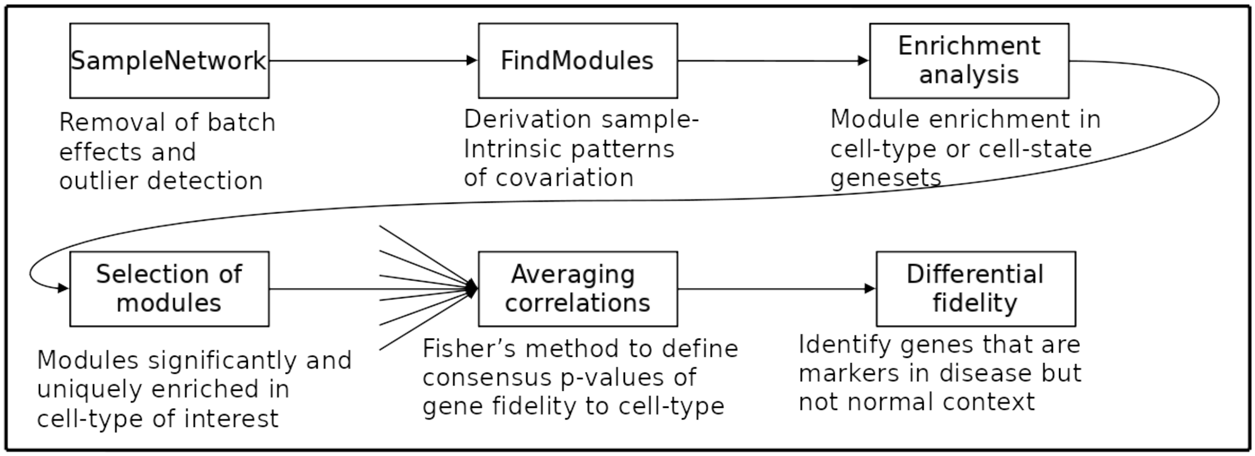

SampleNetwork
Removal of batch effects and outlier detection
FindModules
Derivation sample-Intrinsic patterns of covariation
Enrichment analysis
Module enrichment in cell-type or cell-state genesets
Selection of modules
Modules significantly and uniquely enriched in cell-type of interest
Averaging correlations
Fisher's method to define consensus p-values of gene fidelity to cell-type
Differential fidelity
Identify genes that are markers in disease but not normal context

Figure 3.2 | High-level overview of the pipeline to compute differential fidelity between glioma and normal brain samples.
Starting with the SampleNetwork function to remove batch effects and outliers, the resulting expression matrix is fed into FindModules, which uses a range of hyperparameters to identify networks of coexpressed genes at numerous resolutions. In order to assign biological meaning to the modules, gene set enrichment is used to determine whether the modules are significantly enriched in either cell-state or cell-type gene sets. Finally, modules with enrichments for cell-types of interest are chosen for each dataset. Constituent genes' k_{ME} values to the chosen module for each dataset are summarized via Fisher's method. This process is performed for both normal and glioma datasets. The difference between the resultant fidelity between normal and glioma is termed differential fidelity.

Our consideration of differential fidelity is best coupled with differential expression between glioma and normal contexts. We have plotted differential fidelity (glioma minus normal) for microglia **(a)**, oligodendrocytes **(b),** neurons **(c)**, and endothelial cells **(d)** against differential expression (as quantified by the differences in mean rank expression percentile), while including coloring for the mean expression percentile in normal samples (**Fig. 3.3**). The 13,702 genes shown represent the consensus of 5297 glioma and 7221 normal samples. Genes represented in the top right quadrant are of greatest interest as they represent genes whose fidelity and expression values have increased for that cell-type in glioma versus normal. The most likely explanation for this pattern is increased, dysregulated expression of the gene likely due to being in the proximity of the tumor as part of its microenvironment. The converse of this situation is located in the bottom left quadrant. These genes have experienced decreased fidelity to the cell-type as well as decreased bulk expression. This can be caused by the converse dysregulation event where a gene that is a high fidelity marker for the cell-type in the normal context is no longer a high fidelity marker. This is likely due to a loss in expression in the cell-type and/or increased expression in some other cell-type, either another dysregulated nonmalignant cell or malignant cells (which often have broad dysregulation of their transcriptomic profiles), coupled with a decrease of expression of the gene in the cell-type of interest.

Another confounding factor is the change in relative abundance of the cell-types. For example, both increased expression in the relevant cell-type as well as increased relative abundance of the cell-type can drive increased bulk expression of a gene and contribute to its presence in the upper right quadrant. However, if a gene gains fidelity in a cell-type but the cell-type experiences a decrease in relative abundance, the gain in fidelity and increased expression in the relevant cell-type might not be enough to compensate for the loss of in abundance of the cell type and place the gene in the lower right quadrant.

We see two major patterns in the scatterplots featured in **Fig. 3.3**, positively correlated and negatively correlated. Positively correlated plots include those of microglia (**Fig. 3.3a**) and endothelial (**Fig. 3.3d**), and are suggestive of increased relative abundance of these cell-types. Both of these cell-types are known to be enriched in the both tumor core and periphery: microglia due to immune recruitment[21] and endothelial cells due to increased vascularization[22] by the tumor. On the other hand, negatively correlated scatterplots are suggestive that these cell-types are experiencing a decrease in relative abundance, like those of oligodendrocytes (**Fig. 3.3b**) and neurons (**Fig. 3.3c**). Both of these cell-types have been shown to be largely excluded from the tumor mass[23].

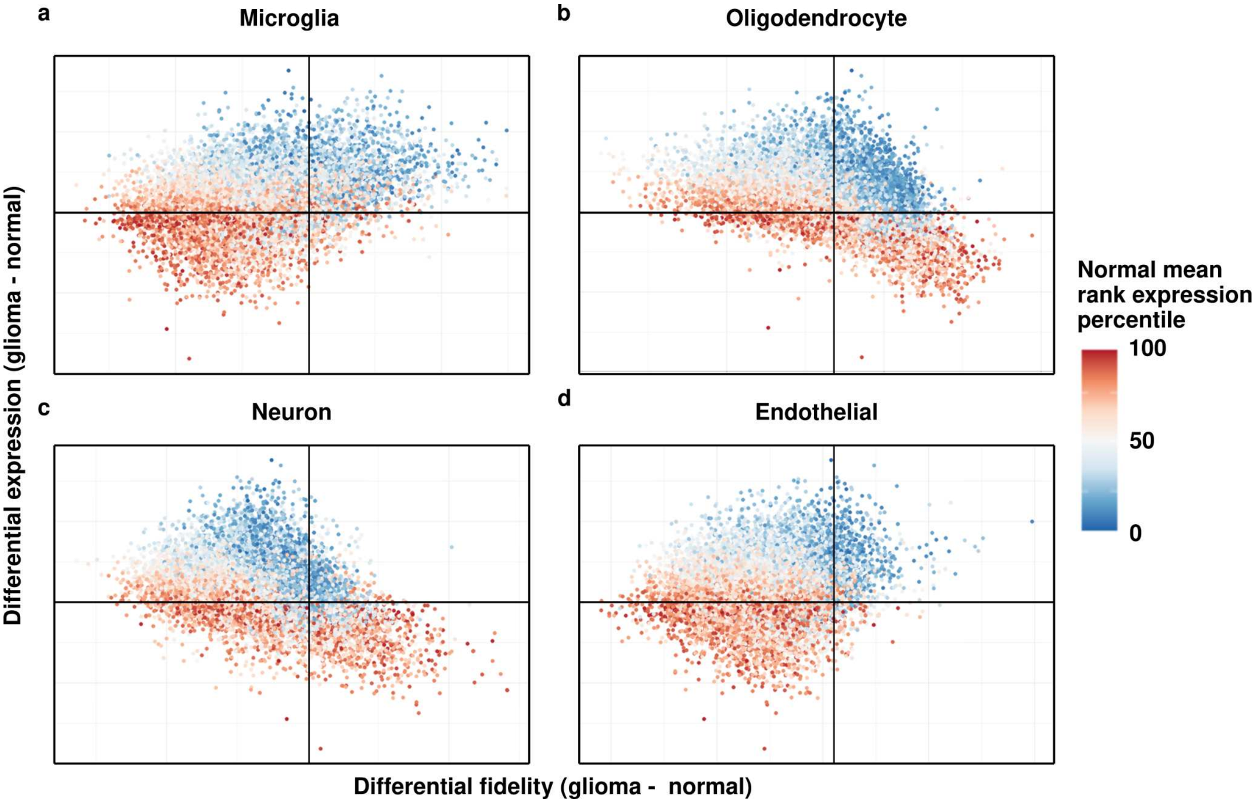


Figure 3.3 | Scatterplots of differential expression against differential fidelity. Scatterplots of differential expression against differential fidelity for all genes corresponding to microglia (a), oligodendrocytes (b), neurons (c), and endothelial cells (d). Color of the points corresponds to the normal mean rank expression percentile.

Ultimately, the area of the plot of the most interest is the top right, representing genes with both high differential fidelity and high differential expression. Of the plots featured in **Fig 3.3**, we were particularly interested in endothelial cells as they would be directly accessible to the blood.

To zoom in on this region of interest, we set arbitrary thresholds in **Fig. 3.4a** of an absolute value of fidelity in the normal data of less than 30 and in the glioma data of more than 30. In addition we can use resources like the COMPARTMENTS database[8] to overlay information on whether the genes have an extracellular component, which makes them accessible to targeted therapeutics.

An outlier with extreme differential fidelity and expression is present in (**Fig 3.4a**), which also does have an extracellular component. We can further investigate the identified target, *ENPEP*, by subsetting mean rank expression percentile (**Fig. 3.4b**) and k_{ME} (**Fig. 3.4c**) by grade. Both low (LGG) and high (HGG) grade gliomas have significantly elevated expression of *ENPEP* and increased fidelity of *ENPEP* to endothelial cells in glioma versus normal datasets. Furthermore there is an nonsignificant increase in expression from LGG to HGG and a significant increase in fidelity to endothelial cells from LGG to HGG.

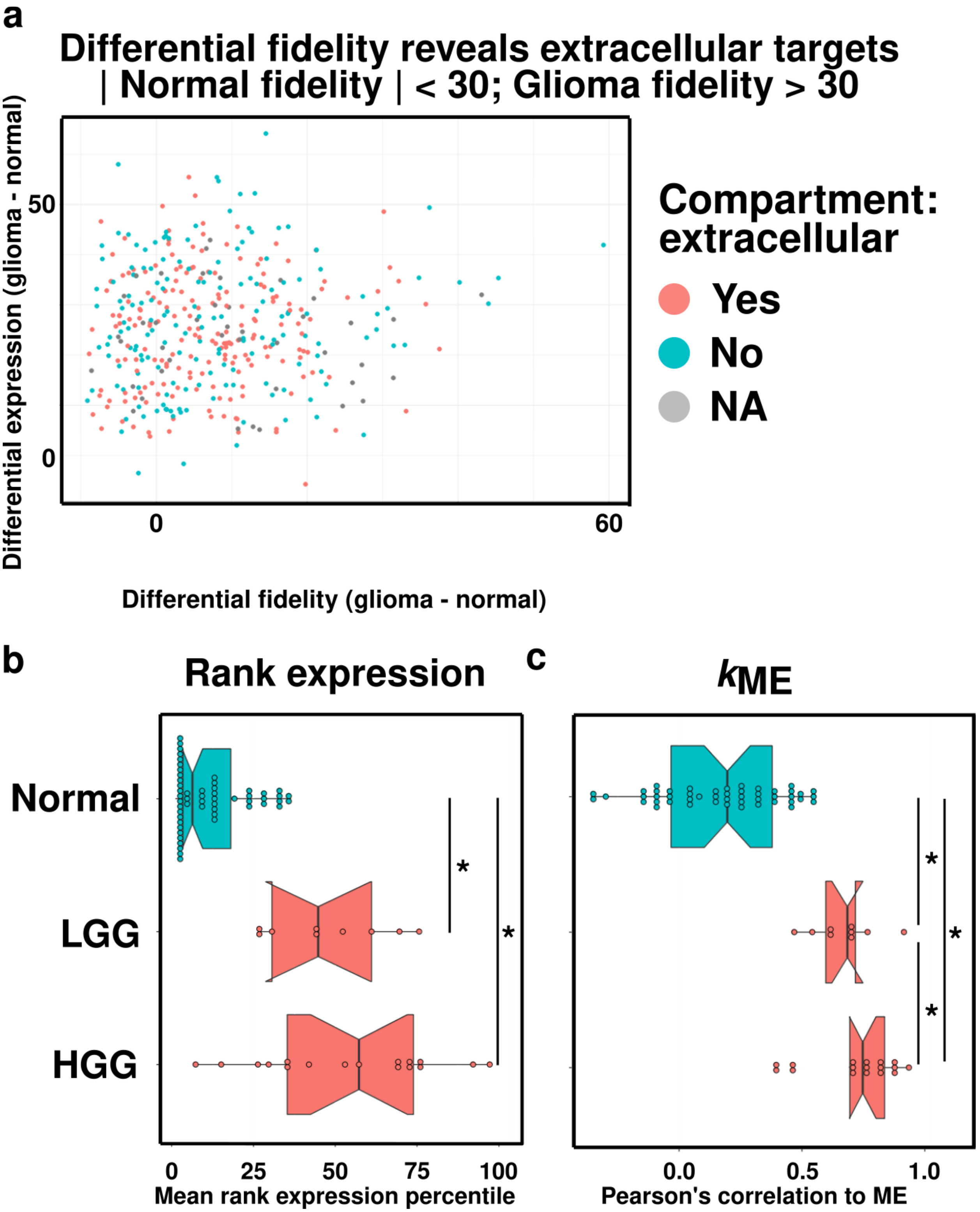


Figure 3.4 | Target identification using subsetting of the data in endothelial cells. Filtering genes with normal absolute fidelity to endothelial cells less than 30 and glioma fidelity greater than 30, while overlaying whether the genes have extracellular components allows for easy target identification (a).(Figure caption continued on the next page.)

(Figure caption continued from the previous page.) The identified target, *ENPEP*, has an extracellular component and has a high differential fidelity for endothelial cells as well as high differential expression in the datasets. *ENPEP* can be further studied by breaking down the data to normal, low-grade glioma (LGG), and high-grade glioma (HGG) types for measuring mean rank expression percentile (b) and k_{ME} (c). * indicates significance of (p<0.05) by the two-sided Wilcoxon rank-sum test.

Finally, we use orthogonal data sources to validate our predictions regarding *ENPEP* expression in endothelial cells of the glioma microenvironment. In **Fig. 3.5a**, we use the GTEX Human Tissue Atlas[17] to determine whether *ENPEP* has significant expression in any other tissues. While there are low levels of expression of *ENPEP* in the cerebellum, expression in the frontal cortex is very low, around 1 TPM. Other tissues with elevated levels of expression are the musco-skeleltal system and liver, which have significant mass or favorable tolerance profiles for therapeutics, respectively[24]. Using single cell data of normal cortex from Velmeshev et al.[19] (**Fig. 3.5b**) we find very negligible expression of *ENPEP* in endothelial cells (0.2 CPM) and even lower expression in all other cell types. This matches data from the GTEX Human Tissue Atlas which found low expression of *ENPEP* in the frontal cortex.

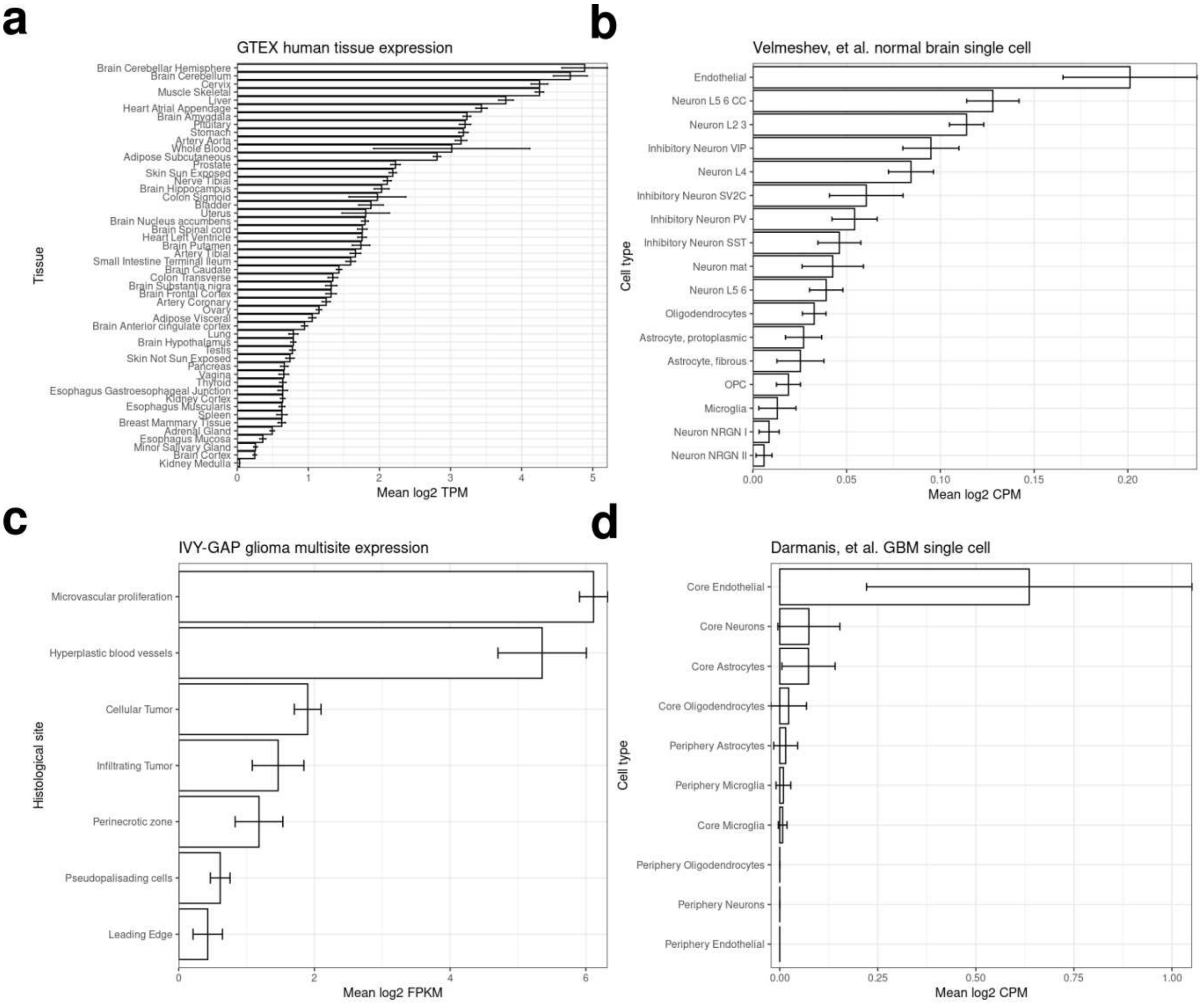


Figure 3.5 | Orthogonal data recapitulates *in silico* predictions of *ENPEP* expression.
(a) Data from the GTEX human tissue atlas details the expression of *ENPEP* in various human tissues. **(b)** Similarly, using single-cell data from Velmeshev et al., we can see the cell-type distribution of *ENPEP* expression in the human brain. **(c)** The IVY-GAP Atlas shows distribution of *ENPEP* expression in various histologic areas of GBM tumors. **(d)** Single-cell data from Darmanis et al. reveals distribution of *ENPEP* expression in various cell-types of the tumor microenvironment in the tumor core and periphery. Error bars in all figures represent two standard errors.

To investigate expression in the Glioblastoma (GBM) setting, we use the IVY-GAP Atlas[18] (**Fig 3.5c**) which performed histologically-guided multi-site sampling to selectively transcriptomically profile biologically important regions. We find that expression is highest in areas of microvascular proliferation and hyperplastic blood vessels and low elsewhere, in line with our prediction of *ENPEP* expression in endothelial cells. Single cell transcriptomic profiling of GBMs by Darmanis et al.[20] (**Fig. 3.5d**) has adapted the multi-site sampling methodology to the single cell workflow. Interestingly, we find highest expression of *ENPEP* in tumor core endothelial cells, with low expression everywhere else. Surprisingly, peripheral endothelial cells do not show any expression of *ENPEP*, suggesting that the dysregulation of endothelial cells by the tumor that results in *ENPEP* expression occurs only in the tumor core and not periphery.

Chapter 3.4: Discussion

Through the use of massive amounts of data and leveraging of our correlational algorithms we were able profile cells of the tumor microenvironment in terms of differential expression in glioma versus normal settings. This effort reflects a key advance in how we can further leverage the vast amounts of existing data bulk transcriptomic data to identify dysregulated genes for therapeutic intervention in any definable cell type.

We chose endothelial cells as a cell-type to investigate more deeply as its direct contact with the bloodstream makes it an attractive starting point to find a target. It was immediately apparent that *ENPEP* had both high differential fidelity and expression, suggesting that *ENPEP* was expressed uniquely in endothelial cells and was poorly expressed in the normal brain. Orthogonal validation in bulk and single-cell datasets bore out predictions of expression of *ENPEP* uniquely in endothelial cells close to the tumor in vascular regions.

ENPEP, known as glutamyl aminopeptidase, catalyzes the cleavage of glutamatic and aspartic residues from the N-terminus of proteins[25]. It is part of the renin-angiotensin system and is thus a critical part of regulating blood pressure[26]. It has also been implicated as being part of neovascularization pathways[22,27,28].

Compellingly, literature does exist in highlighting the use of anti-ENPEP antibodies to target tumor vasculature[29]. However, this work has been performed in mouse models and the specificity of *ENPEP* expression in brain malignancies has thus far not been satisfyingly clarified. It is our hope that this work provides the necessary support for further development of *ENPEP* as a therapeutic target for the following reasons: 1) it is uniquely expressed in endothelial cells close to gliomas and not in normal brain, 2) it is highly expressed in endothelial cells in both high and low-grade gliomas, 3) because endothelial cells are adjacent to the blood, this target can be bound without requiring further penetration into the tissue.

While this approach has thus far been limited to only a single target, it is clear that this process could be for any cell-type yielding a wealth of potential therapeutic targets for gliomas or any other disease with a detectable transcriptomic phenotype.

3.5: References

1. Kelley, K. W., Nakao-Inoue, H., Molofsky, A. V. & Oldham, M. C. Variation among intact tissue samples reveals the core transcriptional features of human CNS cell classes. *Nat. Neurosci.* **21**, 1171–1184 (2018).

2. van der Woude, L. L., Gorris, M. A. J., Halilovic, A., Figdor, C. G. & de Vries, I. J. M. Migrating into the Tumor: a Roadmap for T Cells. *Trends Cancer* **3**, 797–808 (2017).

3. Christofides, A. *et al.* The complex role of tumor-infiltrating macrophages. *Nat. Immunol.* **23**, 1148–1156 (2022).

4. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).

5. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).

6. Vitale, I., Shema, E., Loi, S. & Galluzzi, L. Intratumoral heterogeneity in cancer progression and response to immunotherapy. *Nat. Med.* **27**, 212–224 (2021).

7. Pribluda, A., de la Cruz, C. C. & Jackson, E. L. Intratumoral heterogeneity: from diversity comes resistance. *Clin. Cancer Res.* **21**, 2916–2923 (2015).

8. Binder, J. X. *et al.* COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database (Oxford)* **2014**, bau012 (2014).

9. Gillani, R. *et al.* Gene fusions create partner and collateral dependencies essential to cancer cell survival. *Cancer Res.* **81**, 3971–3984 (2021).

10. Oldham, M. C., Langfelder, P. & Horvath, S. Network methods for describing sample relationships in genomic datasets: application to Huntington's disease. *BMC Syst. Biol.* **6**, 63 (2012).

11. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

12. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

13. Oldham, M. C. *et al.* Functional organization of the transcriptome in human brain. *Nat. Neurosci.* **11**, 1271–1282 (2008).

14. Horvath, S. & Dong, J. Geometric interpretation of gene coexpression network analysis. *PLoS Comput. Biol.* **4**, e1000117 (2008).

15. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* **100**, 9440–9445 (2003).

16. Fisher, R. A. *Statistical Methods for Research Workers*. (1970).

17. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

18. Cantanhede, I. G. & de Oliveira, J. R. M. PDGF Family Expression in Glioblastoma Multiforme: Data Compilation from Ivy Glioblastoma Atlas Project Database. *Sci. Rep.* **7**, 15271 (2017).

19. Velmeshev, D. *et al.* Single-cell genomics identifies cell type-specific molecular changes in autism. *Science* **364**, 685–689 (2019).

20. Darmanis, S. *et al.* Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma. *Cell Rep.* **21**, 1399–1410 (2017).

21. Hambardzumyan, D., Gutmann, D. H. & Kettenmann, H. The role of microglia and macrophages in glioma maintenance and progression. *Nat. Neurosci.* **19**, 20–27 (2016).

22. Hardee, M. E. & Zagzag, D. Mechanisms of glioma-associated neovascularization. *Am. J. Pathol.* **181**, 1126–1141 (2012).

23. Hoogstrate, Y. *et al.* Transcriptome analysis reveals tumor microenvironment changes in glioblastoma. *Cancer Cell* **41**, 678-692.e7 (2023).

24. Navarro, V. J. & Senior, J. R. Drug-related hepatotoxicity. *N. Engl. J. Med.* **354**, 731–739 (2006).

25. Wu, Q., Lahti, J. M., Air, G. M., Burrows, P. D. & Cooper, M. D. Molecular cloning of the murine BP-1/6C3 antigen: a member of the zinc-dependent metallopeptidase family. *Proc Natl Acad Sci USA* **87**, 993–997 (1990).

26. Nehme, A., Zouein, F. A., Zayeri, Z. D. & Zibara, K. An update on the tissue renin angiotensin system and its role in physiology and pathology. *J. Cardiovasc. Dev. Dis.* **6**, (2019).

27. Alliot, F., Rutin, J., Leenen, P. J. & Pessac, B. Pericytes and periendothelial cells of brain parenchyma vessels co-express aminopeptidase N, aminopeptidase A, and nestin. *J. Neurosci. Res.* **58**, 367–378 (1999).

28. Dieterich, L. C. *et al.* Transcriptional profiling of human glioblastoma vessels indicates a key role of VEGF-A and TGFβ2 in vascular abnormalization. *J. Pathol.* **228**, 378–390 (2012).

29. Marchiò, S. *et al.* Aminopeptidase A is a functional target in angiogenic blood vessels. *Cancer Cell* **5**, 151–162 (2004).

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
Patrick Schupp
B5FFE9EA775A48F...
Author Signature

8/30/2023
Date