Gene regulatory network landscape of Group 3/4 medulloblastoma

Piyush Joshi[1,2,3#], Tamina Stelzer[1,3,4#], Konstantin Okonechnikov[1,2#], Ioannis Sarropoulos[5,6,] Mari Sepp[6], Mischan V. Pour-Jamnani[1,2], Anne Rademacher[7], Tetsuya Yamada-Saito[6], Céline Schneider[6], Julia Schmidt[6], Philipp Schäfer[6], Kevin Leiss[6], Michele Bortolomeazzi[8], Jan-Philipp Mallm[8], Patricia B.G. da Silva[1,3], Britta Statz[1,2], Andrea Wittmann[1,2], Kathrin Schramm[1,9], Mirjam Blattner-Johnson[1,9], Petra Fiesel[1,10], Barbara Jones[1,9,11], Till Milde[1,11,12,13], Kristian Pajtler[1,2,11], Cornelis M. van Tilburg[1,11,12,13], Olaf Witt[1,11,12,13], Karsten Rippe[7], Andrey Korshunov[1,10,14], David T.W. Jones[1,9], Volker Hovestadt[15], Paul A. Northcott[16], Supat Thongjuea[1,2], Natalie Jäger[1,2], Henrik Kaessmann[6], Stefan M. Pfister[1,2,11§], Lena M. Kutscher[1,3§]

Abstract

Resolving the molecular mechanisms driving childhood brain tumors will uncover tumor-specific vulnerabilities and advance mechanism-of-action-based therapies. Here we describe a continuum of cell-states in Group 3/4 medulloblastomas, the most frequent and fatal cerebellar embryonal tumor subgroups, based on the differential activity of transcription-factor-driven gene networks derived using a comprehensive single-nucleus multi-omic medulloblastoma atlas. We show that Group 3/4 tumor diversity stems from enriched cell-states along four molecular identity axes: photoreceptor, MYC, precursor, and unipolar brush cell-like. We identified a potential role of *PAX6* in driving dual Group 3- and Group 4-like tumor trajectories in subtype VII tumors. Our study demonstrates how oncogenic events together with lineage determinants drive Group 3/4 tumor identity away from their original source in the cerebellar unipolar brush cell lineage.

Introduction

Scarcity of accurate models of medulloblastoma, a highly heterogeneous and malignant childhood tumor group arising in the cerebellum (*1-3*), has hindered the development of effective mechanism-of-action-based treatment strategies. Advances in molecular profiling in the last decade have characterized medulloblastoma into four major subgroups: WNT, SHH, Group 3 and Group 4 (*4*). Group 3 and 4 medulloblastomas (hereafter referred together as Group 3/4 tumors), which are further categorized into eight molecular subtypes (I-VIII) encompassing pure Group 3 (II, III, IV), mixed (I, V, VII), to pure Group 4 (VI, VIII) tumors (*5*), together represent the most common and lethal cohort. Despite their prevalence, our knowledge of the tumor heterogeneity and underlying regulatory networks in Group 3/4 tumors is limited, and this lack of understanding has hampered the development of mechanism-of-action-based therapies that could improve patient survival at lower rates of collateral damage (*6*).

Recent transcriptomic studies comparing Group 3/4 tumor gene expression programs to those of developing

1. Hopp Children's Cancer Center (KiTZ), Heidelberg, Germany
2. Division of Pediatric Neurooncology, German Cancer Research Center (DKFZ) and German Cancer Consortium (DKTK), Heidelberg, Germany
3. Developmental Origins of Pediatric Cancer Junior Research Group, German Cancer Research Center (DKFZ), Heidelberg, Germany.
4. Faculty of Biosciences, University of Heidelberg, Germany
5. Wellcome Sanger Institute, Cambridge, United Kingdom
6. Center for Molecular Biology of Heidelberg University (ZMBH), DKFZ-ZMBH Alliance, Heidelberg, Germany
7. Division of Chromatin Networks, German Cancer Research Center (DKFZ) and Bioquant, Heidelberg, Germany
8. Single-cell Open Lab, German Cancer Research Center (DKFZ), Heidelberg, Germany
9. Division of Pediatric Glioma Research (B360), German Cancer Research Center (DKFZ), Heidelberg, Germany
10. CCU Neuropathology, German Cancer Research Center (DKFZ) and German Cancer Consortium (DKTK), Heidelberg, Germany
11. Department of Pediatric Oncology, Hematology & Immunology, Heidelberg University Hospital, Heidelberg, Germany
12. CCU Pediatric Oncology, German Cancer Research Center (DKFZ) and German Cancer Consortium (DKTK), Heidelberg, Germany
13. National Center for Tumor Diseases (NCT), Heidelberg, Germany
14. Department of Neuropathology, Institute of Pathology, Heidelberg University Hospital, Heidelberg, Germany
15. Department of Pediatric Oncology, Dana Farber Cancer Institute, Boston, MA, USA
16. Department of Developmental Neurobiology, St Jude Children's Research Hospital, Memphis, TN, USA

# These authors contributed equally to this study
§ These authors jointly supervised this study, l.kutscher@kitz-heidelberg.de, s.pfister@kitz-heidelberg.de

human cerebellum have hinted that these tumors likely arise from upper rhombic lip-derived unipolar brush cell (UBC) progenitors (*7-9*). However, it is still unclear how the heterogeneous Group 3/4 biology can be derived from and explained by the linear UBC differentiation process, and which regulatory networks drive malignant transformation. In this study, we generated and analyzed single-nucleus multi-omic data of 38 Group 3/4 medulloblastoma samples to provide unparalleled insight into the molecular mechanisms explaining similarities and differences within Group 3/4 medulloblastoma. We focused on differential activity of transcription-factor regulated gene regulatory networks (TF-GRNs), a set of genes comprising putative downstream targets of the TF along with the TF itself, and identified four molecular axes of identity of Group 3/4 medulloblastoma development. We show that the spectrum of Group 3/4 subtypes can be attributed to the continuum of cell-states along these axes, which are connected through a shared regulatory landscape. We further identified that the intermediate nature of subtype VII tumors is due to the co-existence of Group 3- and Group 4-like tumor trajectories arising from bi-potent precursor cells in single tumors. Our findings provide the mechanistic framework to explain Group 3/4 medulloblastoma biology in the context of its normal developmental origin, opening new avenues to explore and test novel medulloblastoma treatment strategies and to faithfully model the different disease subtypes.

RESULTS

Group 3/4 medulloblastoma multi-omic atlas

Group 3/4 medulloblastomas appear as a separable, yet continuous group of tumors when their transcriptomic programs (bulk RNA-Seq samples, Fig. 1A; Fig. S1A-C; Table S1) (8, 10-14) are visualized in a low dimensional space, such as tSNE (t-distributed Stochastic Neighbor Embedding) or UMAP (Uniform Manifold Approximation and Projection). This result suggests the existence of a gradient of biology that connects their distinct molecular characteristics. Consequently, subtype-specific metagene programs are also enriched in other subtypes of the same subgroup (Fig. 1B; Fig. S1D-G; Table S2). For example, Group 4 subtypes VI, VII and VIII demonstrate enrichment of the same signature, *Sig_g* (Fig. 1B). These observations align with the previously proposed model that places the continuum of medulloblastoma biology on a bipolar Group 3 vs Group 4 axis (Fig. S1D) (*10*). However, using diffusion trajectory analysis to identify potential directions of the metagene programs, we discerned that both Group 3 and Group 4 subtypes have their own linear axis of separation (Fig. 1C; Fig. S1H-J), suggesting that a multi-axial spectrum exists within Group 3/4 biology.

We hypothesized that the conserved biology across closer subtypes is driven by the same underlying molecular programs, as defined by TF-GRNs, while separable subtypes are regulated by distinct TF-GRNs. To determine the molecular programs that define this multi-axial tumor biology, we generated multi-omic single-nucleus (interchangeable with "single-cell" for simplicity) data for a cohort of 38 Group 3/4 patient samples encompassing all eight Group 3/4 molecular subtypes (total nuclei = 355,295; total samples = 38: 32 samples with both RNA and ATAC profiles from same nuclei, 1 sample with both RNA and ATAC profiles from different nuclei, 5 samples with RNA profiles only; Fig. 1D-G; Fig. S2A-H; Fig. S3A-F; Table S3). Expectedly, transcriptomic and chromatin accessibility profiles showed sample-specific cell-clusters (Fig. S2E; Fig. S3F), with samples from the same molecular subtype located closer on the UMAP (Fig. 1F).

To integrate the tumor data such that tumor cells exhibiting similar molecular biology, but distinct levels of gene expression, cluster together, we transformed gene expression data into molecular program enrichment profiles. We focused on TFs with highly variable expression in our tumor atlas, to obtain the TF-GRN sets driving inter-tumor heterogeneity and continuity across Group 3/4 tumors, and employed a two-step approach. Firstly, for each of the above identified TFs, we defined a TF-GRN in a tumor sample using *SCENIC+* based analysis (*15*), by identifying genes with correlated expression to that TF and filtering for targets with putative binding sites for the candidate TF in target-associated cis-regulatory elements (CREs). We then converted the gene expression matrix into TF-GRN score matrix using *AUCell* (*16*), and obtained TF-GRNs that are differentially active in the tumor clusters of the sample. Secondly, to integrate the multi-omic data, we selected TFs associated with intra-tumor heterogeneity across multiple samples and obtained a conserved TF-GRN for each of the selected TFs based on recurrent TF-target associations. We then obtained TF-GRN scores for each of the selected TFs in the tumor cells of the integrated data and used this TF-GRN score matrix for further analysis, such as to generate

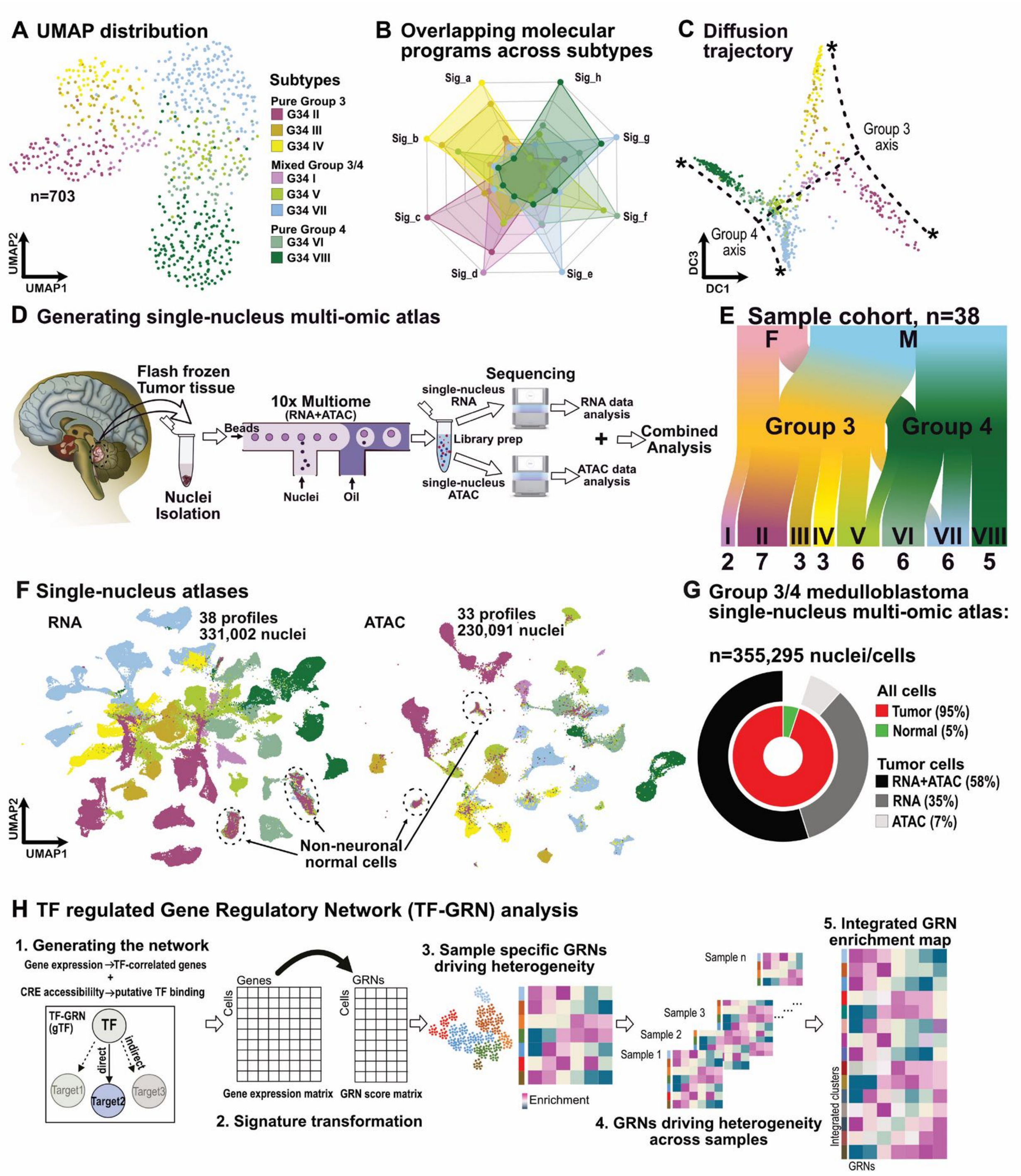


Fig. 1. Overlapping heterogeneity defines the molecular continuity among G3/4 medulloblastoma.
A. UMAP distribution of Group 3/4 medulloblastoma (n=703, bulk-RNA-Seq) samples on the transcriptomic landscape colored by subtype identity. **B**. Spider plot of scaled enrichment of metagene signatures across subtypes. **C**. Diffusion trajectory of Group 3/4 tumors on the transcriptomic landscape. Gradient of subtype identity along the Group 3 and Group 4 axes is shown by dotted lines. **D**. Experimental design for generating single-nucleus multi-omic data from patient-derived tumor samples. **E**. Sample metadata of our Group 3/4 medulloblastoma single-nucleus multi-omic study cohort. F, female. M, male. **F**. UMAP distribution of snRNA-Seq (left) and snATAC-Seq (right) data colored by subtype identity. Non-neuronal cells are encircled. **G**. Graphical summary of data modalities of single-nuclei comprising the Group 3/4 multi-omic atlas. snATAC-Seq nuclei from MB248 (n=3,194 nuclei) are excluded in the chart. **H**. Graphical representation of *SCENIC+* based TF-GRN approach to integrate snRNA-Seq and snATAC-Seq data for the identification of the regulatory signatures driving intra-tumor heterogeneity. Conserved TF-GRNs across samples provide insights into the continuous heterogeneity observed within Group 3/4 medulloblastoma.

an integrated TF-GRN enrichment map (Fig. 1H and see *Methods* for additional details).

Gene regulatory networks driving Group 3/4 identity

Using scaled enrichment of area under the curve (AUC) scores for a set of TF-GRNs (n=108, Table S4) selected from TF-GRNs active across tumor samples, we integrated tumor cells based on their shared biology (Fig. 2A,B; Fig. S4A-H). This integrated tumor cell atlas displayed four axes on the diffusion map, which we labeled as photoreceptor-like (PR_t, t=tumor), MYC-enriched, Precursor-like and UBC-like (UBC_t, t=tumor), based on the known function of associated TFs and the enrichment of molecular programs in the annotated cells, as described below. Cells belonging to Group 3 vs Group 4 tumors differentially contributed to these four axes (Fig. 2B).

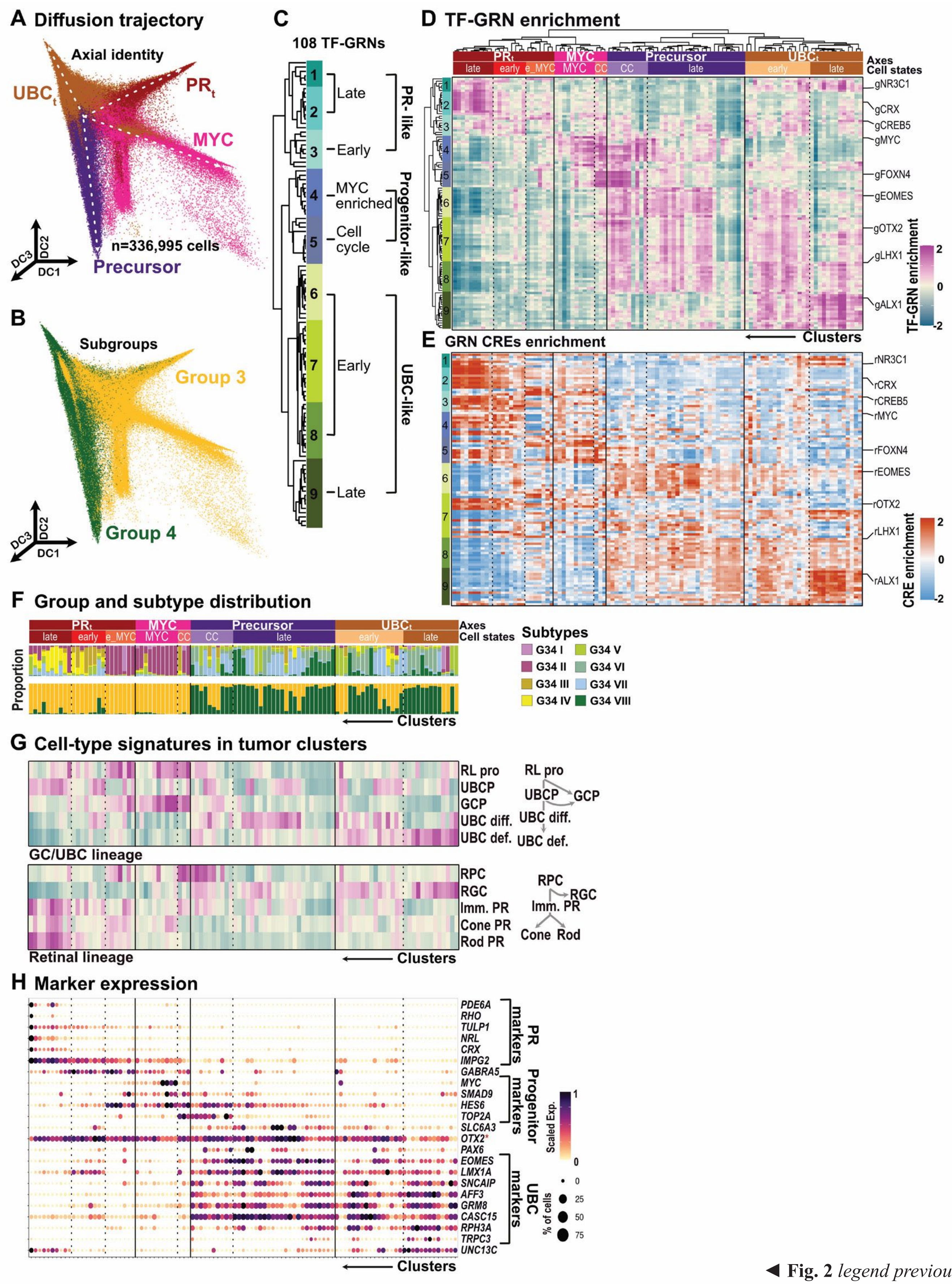
To further molecularly define these four axes, we first clustered the tumor cells (Fig. S4B) and identified the TF-GRNs enriched in each cluster. We grouped the identified 108 TF-GRNs into nine groups by hierarchical clustering to identify co-enriched programs (Fig. 2C; Fig. S5A-L; Table S4). TF-GRN programs 1 (representative GRN: gNR3C1), 2 (gCRX) and 3 (gCREB5) included well-known regulators of the photoreceptor lineage (17) (Fig. S5D-F; Fig. S6A-G). TF-GRN programs 4 (gMYC) and 5 (gFOXN4) were enriched for cell-cycle and progenitor- associated TFs (Fig. S5G,H) (*18, 19*). Similarly, TF-GRN programs 6 (gEOMES), 7 (gOTX2), 8 (gLHX1) and 9 (gALX1) included well-known regulators of early and late UBC development (Fig. S5I-L) (*20, 21*). Using hierarchical clustering, we grouped tumor clusters exhibiting similar program enrichment along the identified axes and subdivided these groups into tumor cell-states based on co-enrichment of molecular programs defining more than one axis (Fig. 2D; Fig. S7A-K; Table S5). For example, while all clusters in the MYC axis were enriched for TF-GRN program 4, tumor cells in the MYC_CC states were also co-enriched for TF-GRN program 5 (Cell cycle), the TF-GRN program these cells share with the cell cycling Precursor states (Prec_CC). We also investigated the differential enrichment of CREs associated with these TF-GRNs, which showed similar enrichment profiles (Fig. 2E). TF-GRNs and associated open chromatin regions mostly showed co-enrichment patterns, except in tumors cells along the PR_t-axis, where progenitor-like programs (4 and 5) were turned down while the associated CREs remained comparatively accessible as in undifferentiated MYC-axis clusters (Fig. 2D,E), a phenomenon shared with normal human rod photoreceptors (Fig. S6F,G).

We observed that the tumor cells' subgroup and subtype identity was also distinctively associated with the four axes. PR_t and MYC axes were almost uniquely populated by tumor cells from Group 3 samples, while the Precursor and UBC_t axes were predominantly populated by tumor cells from Group 4 samples (Fig. 2F). At the subtype level, subtypes III and IV exhibited PR_t axis cell-states, while subtype II was enriched in the MYC axis (Fig. 2F). Group 4 associated subtypes, VI and VIII, showed almost exclusive association with Precursor and UBC_t states. Interestingly, subtypes I, V and VII, which have intermediate Group 3/4 identity, were distributed across the four axes.

Group 3/4 tumors originate from the cerebellar

Fig. 2. Four axes of Group 3/4 medulloblastoma identity.
A. 3D diffusion map of Group 3/4 tumor cells obtained from TF-GRN enrichment colored by axial identity. Dotted line indicate axial trajectories. **B**. 3D diffusion map of Group 3/4 tumor cells colored by group identity. **C**. Hierarchical clustering of the 108 TF-GRNs based on co-enrichment in tumor clusters. **D**. Differential enrichment of TF-GRN score across tumor cell-clusters in the integrated data. **E**. Differential enrichment of activity of constituent CREs of the TF-GRN-sets across cell-clusters in the integrated data. **F**. Subtype and subgroup identity of cells comprising the cell-cluster in the integrated atlas. Each bar represent a cluster's proportional tumor subtype (top) or subgroup (bottom) composition. **G**. Enrichment of cell-state signatures in the cerebellar granule cell/unipolar brush cell lineage and retinal photoreceptor lineages in the cell-cluster of the integrated Group 3/4 medulloblastoma atlas. UBCP cells were labelled as GCP/UBCP in the original atlas (*20*) but termed as UBCP here for simplicity. RL pro, rhombic-lip progenitor. GCP, granule cell progenitor. UBCP, UBC progenitor. UBC diff., differentiating UBC. UBC def., defined UBC. RPC, retinal progenitor. RGC, retinal ganglion cell. Imm. PR, immature photoreceptor. Cone PR, cone photoreceptor. Rod PR, rod photoreceptor. **H**. Marker gene expression distribution in the integrated atlas. Photoreceptor, progenitor or UBC cell-states marker genes are annotated as such. *OTX2* (marked with asterisk) is a marker gene for both photoreceptor and UBC lineages. Dot size indicates the proportion of cells in a cluster expressing a gene, and color denotes mean expression scaled across cluster per gene.

Fig. 2 next page ▶



◀ **Fig. 2** *legend previous page*

rhombic lip (*7*-*9*) and also show enrichment of photoreceptor programs (*7, 9, 22*). To identify the tumor cells resembling the cell-states in cerebellar rhombic lip or retinal lineage, we investigated the enrichment of these lineage programs (Table S6) in tumor clusters (Fig. 2G). Briefly, the PR_t axis is linked to normal UBC progenitor and normal retinal photoreceptor program, while UBC_t is characterized by an enrichment of normal differentiating and differentiated UBC programs, as well as the enrichment of normal retinal ganglion cell signature: a retinal lineage associated with *EOMES* expression (Fig. S6D) (*23*). The MYC axis showed enrichment of normal rhombic lip and granule cell progenitor (GCP) signatures. The Precursor axis showed enrichment of normal differentiating UBCs, with normal retinal progenitor signature enriched in cell cycling Precursors (Prec_CC). Expression patterns of marker genes of retinal photoreceptor (e.g. *CRX*, *NRL*), cell cycling progenitor (e.g. *TOP2A*) and cerebellar UBC lineage (e.g. *EOMES*, *LMX1A*) further validated our axial and cell-states annotation (Fig. 2H).

In summary, the differential TF-GRN activity enrichment map defines the continuum of biology of the eight subtypes of Group 3/4 medulloblastoma along the four axes of molecular identity.

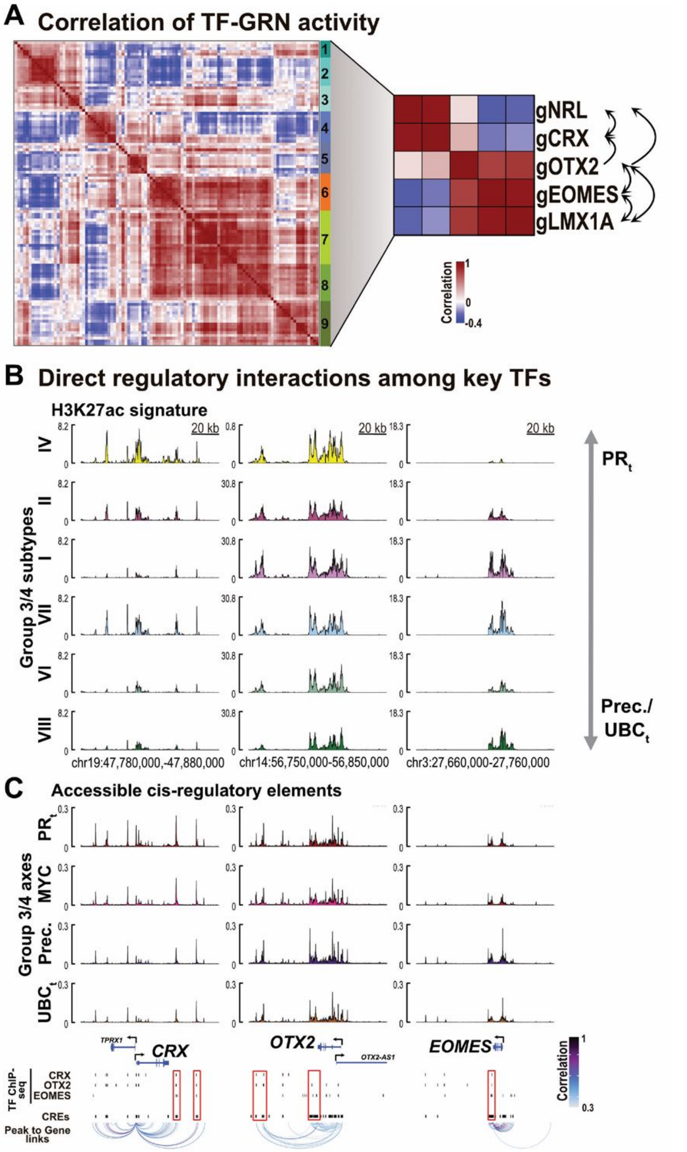


Fig. 3. Mutually repressive PR_t and UBC_t associated TF-GRNs drive Group 3 and Group 4 identity apart

A. Pearson correlation analysis of TF-GRN activity in the integrated single-cell tumor data. PR_t, MYC and UBC_t associated TF-GRNs show high anti-correlation. Inset shows correlation between key GRNs: gNRL, gCRX, gOTX2, gEOMES and gLMX1A. Arrow shows putative direct interaction between TF pairs based on *SCENIC+* analysis and arrow head denotes the target of the interaction. **B**. H3K27ac ChIP-Seq (*11*) shows distinct enhancer signature enrichment at *CRX*, *OTX2*, and *EOMES* loci across Group 3/4 subtypes. Subtypes are arranged from pure high PR_t (top) to high Precursor (Prec.)/UBC_t (bottom) phenotype. **C**. Integrated snATAC-Seq reveals differential accessibility of CREs at *CRX*, *OTX2*, and *EOMES* loci across Group 3/4 medulloblastoma axial identities (this study). ChIP-Seq peaks for OTX2, CRX (*24*) and EOMES (*25*) overlap CREs positively associated with expression of key genes: *CRX* (left), *OTX2* (middle) and *EOMES* (right). Interaction arcs depict representative peak to gene links colored by correlation of peak accessibility and gene expression. Red boxes depict putative CREs involved in cross-regulations for each gene.

Mutually repressive TF-GRN interactions drive Group 3 versus Group 4 separation

To investigate the transition between the four axial identities, we investigated the correlation between the TF-GRNs activity. We hypothesized that co-expressed TF-GRNs will show high positive correlation, and mutually exclusive, potentially repressive interactions between TF-GRNs will be negatively correlated (Fig. S8A). Broadly, PR_t, MYC and UBC_t associated TF-GRNs were negatively correlated, and Precursor and UBC_t TF-GRNs were positively correlated (Fig. 3A). Particularly, gCRX/gNRL (PR_t axis) and gEOMES/gLMX1A (Precursor/UBC_t axis) were highly negatively correlated (Fig. 3A, inset). gMYC activity was likewise negatively correlated with gEOMES/gLMX1A, and not-correlated to gCRX/gNRL (Fig. S8B). These anti-correlative relationships suggest a mutual exclusivity between PR_t, MYC and UBC_t axial identities: individual tumor cells cannot have two or more of these identities simultaneously.

We next focused on TF-GRNs that drive the PR_t and UBC_t separation. We hypothesized that this separation results from the mutual repression of TF-GRN programs associated with PR_t and UBC_t. To confirm direct regulatory

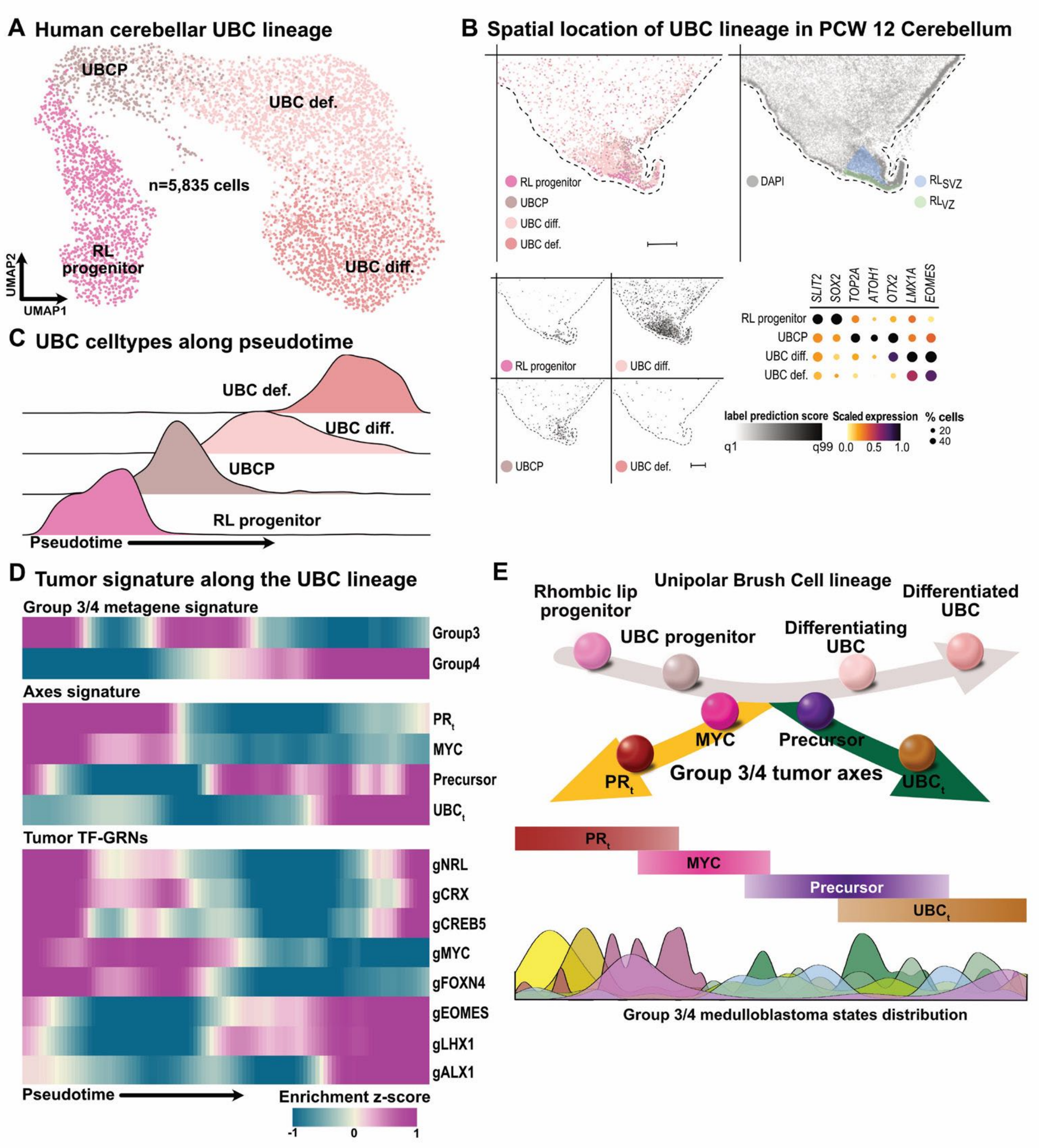


Fig. 4. Group 3/4 identity programs are enriched in distinct stages of the developing cerebellar UBC lineage.
A. UMAP representation of cells in the human cerebellar UBC lineage (*20*) in the transcriptomic space. RL progenitor, rhombic-lip progenitor. UBCP, UBC progenitor. UBC diff., differentiating UBC. UBC def., defined UBC. **B**. Mapping of the UBC lineage cell-states in the fetal human cerebellum (12 pcw) by alignment of the multiplexed single-molecule fluorescent *in situ* hybridization FISH data with 11 pcw snRNA-Seq data (*20*). Region of a coronal section containing the rhombic lip is shown. Top left: cell's estimated rhombic-lip (RL) progenitor, UBC progenitor (UBCP), differentiating (diff.) or defined (def.) UBC cell-identity colored by state. Top right: DAPI-stained section with the rhombic lip ventricular zone (RLvz; green) and sub-ventricular zone (RL_{SVZ}; blue) highlighted. Bottom left: cell-label prediction score (scale capped at the 1st and 99th quantiles). Bottom right: Expression of key markers across the labelled segments. Dot size indicates the proportion of segments expressing a gene, and color shows the mean expression level normalized to segment area and scaled per gene. Scale bars, 250 μM. **C**. Cell-type density variations along predicted pseudotime within the UBC lineage. **D**. Differential enrichment of Group 3 vs Group 4 metagene signature (top); Group 3/4 axial identity signature (middle), and selected TF-GRNs activity (bottom) in the UBC lineage along pseudotime. **E**. Proposed model of Group 3/4 medulloblastoma identity bifurcation. Tumor cell-states derived before UBC identity specification drive Group 3 tumor identity. Tumor cells that turn on UBC specification program become Group 4 tumors.

interactions between *CRX/NRL* and *EOMES/LMX1A*, the key TF regulators of PR_t and UBC_t states, respectively, we examined enhancer regions around *CRX, NRL, OTX2, EOMES*, and *LMX1A* gene loci using our snATAC-Seq atlas. By analyzing and overlaying co-localization of active enhancers (H3K27ac signal) in Group 3/4 tumors (*11*), accessibility of identified CREs (this study), and the binding sites of CRX and OTX2 in the human retina (*24*) and of EOMES in human embryonic stem cells (*25*), we identified potentially functional enhancers regulating cross-talk between these key TFs (Fig. 3B,C; Fig. S8C,D).

These findings suggest a mutual repression between CRX and EOMES, indicated by the potential binding of CRX and EOMES to each other's functional enhancers and their anti-correlated gene expression (Fig. 3C). OTX2, on the other hand, potentially directly regulates expression of *CRX/NRL* and *EOMES/LMX1A* (Fig. 3A, inset). Altogether, these data suggest that the presence of OTX2 provides a permissive environment for tumor cells to differentiate along both the PR_t and UBC_t lineage, while the mutually repressive interaction between *CRX/NRL* and *EOMES/LMX1A* drives the trajectories apart.

The UBC lineage exhibits Group 3/4 specific programs at distinct time-points

We (*7*) and others (*8, 9*) have previously shown that the Group 3/4 transcriptomic program is best matched to the UBC lineage of the developing human cerebellum. We hypothesized that if tumor cells are arrested in the cell-state space of normal UBC development, the tumor axial or TF-GRN programs would be differentially enriched during normal UBC development, allowing us to determine the putative stages in which these tumor cells are arrested. Therefore, we extracted the cells belonging to the developing UBC lineage from our previously generated snRNA-Seq atlas of the developing human cerebellum (Fig. 4A) (*20*). We further estimated the spatial locations of the UBC cell-states in the 12 post-conception week (PCW 12) human cerebellum based on our multiplexed single-molecule *in situ* hybridization dataset (Molecular Cartography, Resolve Biosciences) (Fig. 4B) (*20*). Investigating the differential enrichment of the identified gene-sets along the UBC lineage (Fig. 4C), we observed that a Group 3 specific metagene signature, PR_t/MYC axial programs, and TF-GRNs driving PR_t/MYC axial identities are enriched in rhombic lip or UBC progenitor cell-states (Fig. 4D; Fig. S9A-C). Conversely, a Group 4 specific metagene signature, Precursor/UBC_t axial program, and TF-GRNs driving Precursor/UBC_t tumor identities are enriched in differentiating and differentiated UBC (*i.e.* defined UBC) cell-states (Fig. 4D; Fig. S9A-C). Together, this nearly mutually exclusive enrichment pattern of Group 3 and Group 4 regulatory networks along the UBC lineage suggests that the coarse Group 3 vs Group 4 separation occurs at the point of UBC identity specification (Fig. 4E). The spatial location of UBC progenitors and differentiating UBCs, in the rhombic lip sub-ventricular zone (RL_{SVZ}), a region with proliferative capacity (*8, 9*), further confirms the source of Group 3/4 medulloblastomas in the cerebellar rhombic lip, as reported by others (*8, 9*).

PAX6 expression drives dual PRt-UBCt lineage identity in subtype VII tumors

To investigate molecular drivers of Group 3/4 tumor identities, we extrapolated the TF-GRNs obtained from our single-cell multi-omic atlas to a larger bulk RNA-Seq dataset of Group 3/4 medulloblastoma samples (n=703) (*8, 10-14*). We obtained the relative enrichment profiles of the above identified TF-GRNs in the tumor bulk transcriptomic data and observed a high correspondence between cell-state enrichment patterns across subtypes, similar to our mutli-omic atlas results (Fig. S10A-H). tSNE analysis of the bulk data based on TF-GRN enrichment scores showed that Group 3/4 tumors can also be divided into four major axes at the bulk level that correlated with enrichment of specific axial signatures (Fig. 5A).

Overlaying the status of common genetic driver events (*8, 14*) in Group 3/4 medulloblastomas suggested a causal relation between the driver event and the resultant phenotype (Fig. S10A-H). Briefly, predominantly *MYC*-driven subtype II tumors, with documented *MYC* amplification or *PVT1-MYC* fusion (Fig. S10B,I) and high *MYC* expression (Fig. S10J), showed enrichment of the MYC and early PR_t axial-signature (Fig. S10B). *SNCAIP* duplication associated with *PRDM6* activation (*14*) in subtypes VI, VII and VIII tumors drove tumors towards the UBC_t axis (Fig. S10F-I). While *GFI1B* rearrangements were distributed across subtypes, *GFI1B*-driven subtype I and II tumors typically exhibited a mixed PR_t-UBC_t identity as observed from co-enrichment of associated TF-GRNs (Fig. S10A,B,I).

We next focused on intermediate Group 3/4 tumors, which primarily belonged to subtypes I, V and VII, and exhibited a lower Group 3/Group 4 classification score (Fig. 5B; Fig. S10K,L). The intermediate identity of subtype V is possibly due to lack of enrichment of late-PR_t or late-UBC_t

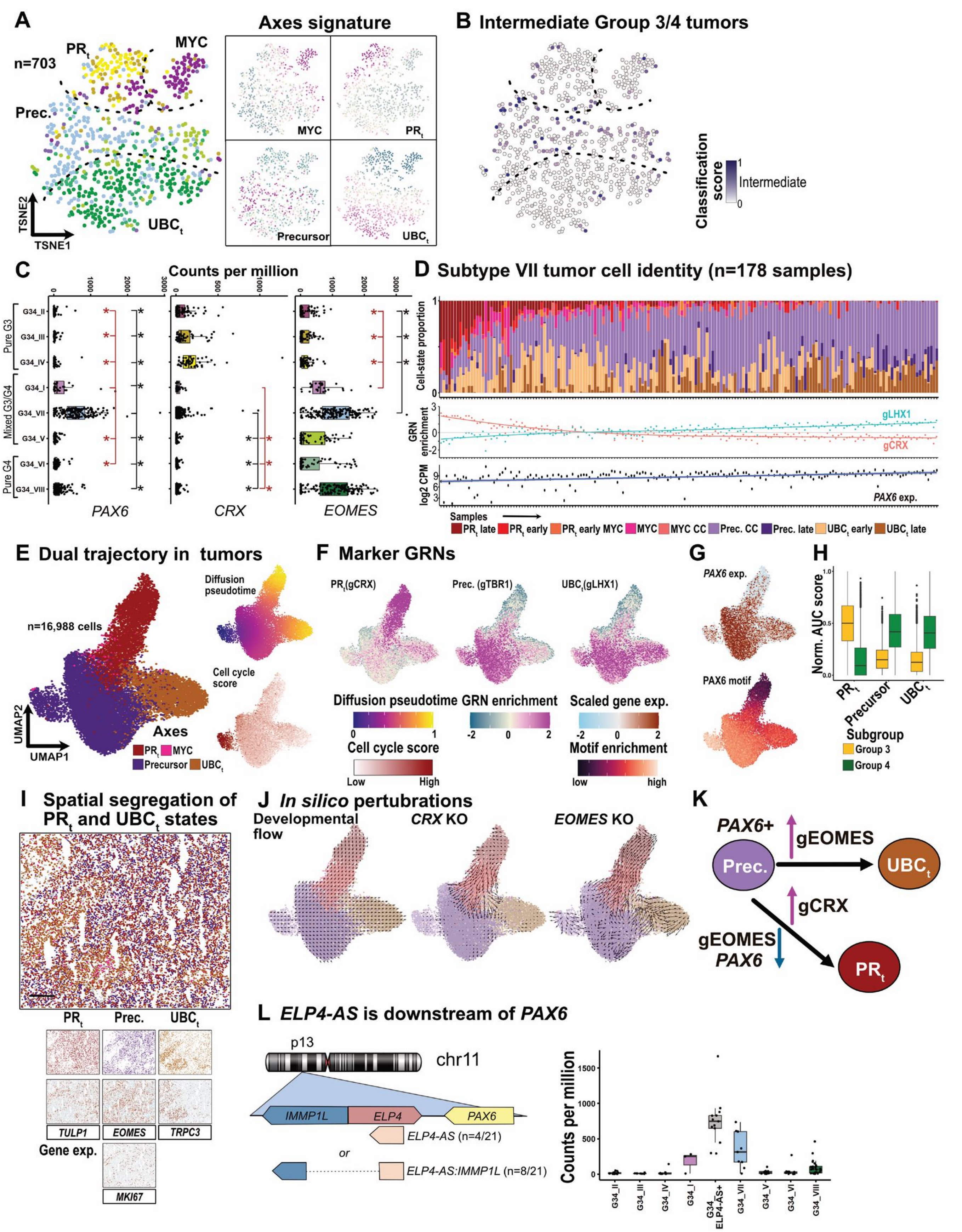


Figure 5 *legend next page* ►

Fig. 5. ***PAX6*** **drives dual Group 3 and Group 4-like trajectory in subtype VII medulloblastomas.**
A. tSNE distribution of Group 3/4 medulloblastoma bulk-RNA-Seq samples (n=703) on the TF-GRN enrichment space. Relative enrichment of axial signatures (middle) and marker TF-GRNs (right) on the tSNE landscape. PR_t, Photoreceptor (tumor)-like. Prec., Precursor. UBC_t, UBC (tumor)-like. **B**. Intermediate methylation classification score (1- abs(G3 score- G4 score)) on the tSNE map. Dashed lines highlights presumptive separation among bulk axes. **C** Boxplot distribution of *PAX6*, *CRX* and *EOMES* expression in bulk RNA-Seq samples (n=703) across subtypes. Expression in individual samples is shown as dots. Asterisk denotes log-fold change > 1 and adjusted p-value < 0.001 for pairwise comparisons. Black, pair-wise comparisons to subtype VII tumors, Red, pair-wise comparisons to subtype I tumors. **D**. Predicted deconvoluted axial cell-state identity in subtype VII samples arranged in order of increasing UBC_t identity (increasing gLHX-gCRX score). Each bar represents a sample's proportional composition of tumor cell-states after removing predicted normal neuronal cell fraction. Middle panel shows gCRX and gLHX1 AUC scores per sample, and the bottom panel illustrates *PAX6* (log2 counts per million) expression in each sample. Fitted linear model for *PAX6* expression along the sample order: R2=0.1643. p-value =1.99e-08. **E**. UMAP distribution of a subtype VII tumor (MB129) cells in the TF-GRN space, annotated by axial identities. Panels on right show predicted diffusion pseudotime (top) and cell-cycle score (bottom). **F**. Scaled enrichment of marker TF-GRNs in MB129 tumor cells is shown on the UMAP. **G**. UMAP distribution of scaled *PAX6* expression (top) and scaled PAX6 motif enrichment (bottom) in MB129 tumor cells. **H**. Relative enrichment of Group 3 and Group 4 metagene signature in MB129 tumor cells annotated with PR_t, Precursor and UBC_t axial identity. **I**. Top: Spatial distribution of cells annotated as per axial identities. Middle: Cells assigned to PR_t, Precursor (Prec.) and UBC_t identities. Bottom: Expression of marker genes, *TULP1*, *EOMES* and *TRPC3*, for axial identities PR_t, Precursor and UBC_t, respectively. *MKI67* expression denotes cell-cycling cells. Scale bar, 200 μM. **J**. Developmental trajectory of cell-states in MB129 (left). *In silico* loss of *CRX* (middle) and *EOMES* (right) inhibits and promotes the acquisition of PR_t cell-states, respectively. Tumor cells are colored as per cell-state identities. Arrows show predicted local trajectory of cells. **K**. Axial-state transition model suggesting presence of *PAX6* expression leads to differentiation of tumor cells along the bifurcating PR_t and UBC_t trajectory. **L.** Frequency of novel lncRNA (*ELP4-AS*) and a novel spliced form (*ELP4-AS:IMMP1L*) in subtype VII tumors of the ICGC cohort. Boxplot distribution of *PAX6* expression in Group 3/4 tumors (ICGC cohort). Samples with *ELP4-AS* transcription (with or without splicing to *IMMP1L*, n= 13/104) (grey box) are grouped separately from rest of the tumors, which are grouped as per subtype identity.

TF-GRN programs (1-2 and 9, respectively) that defines either Group 3- or Group 4-like identity, respectively (Fig. S10E). Conversely, intermediate subtypes I and VII could be ascribed to the co-enrichment of PR_t and UBC_t associated TF-GRN programs (1-3 and 6-9, respectively) in the same tumor (Fig. S10A,G).

In the integrated multi-omic atlas, subtype VII tumor cells were distributed along the PR_t-to-UBC_t axis. We confirmed this observation using bulk RNA-Seq data, where ~31% (55/178) of subtype VII tumors exhibited co-enrichment of PR_t and UBC_t TF-GRN programs and ~7% (12/178) showed predominance of PR_t TF-GRN programs (Fig. S10G). In bulk tumors, we identified that the TF *PAX6*, a key regulator in retinal and UBC lineage specification and differentiation (*26, 27*), was highly expressed in subtype VII tumors (Fig. 5C; Table S7; Table S8). Further, subtype VII tumors expressed the Group 3-associated *CRX* (Fig. 5C) and *NRL* (Fig. S11A) at significantly higher levels when compared to Group 4 subtypes (VI and VIII). On the other hand, they also expressed the Group 4-associated *EOMES* (Fig. 5C) and *LMX1A* (Fig. S11A) at significantly higher levels when compared to pure Group 3 subtypes (II, III and IV). Therefore, the intermediate identities of subtype VII likely arise from this co-expression of TFs typically associated with core networks regulating tumor-cell specification along the PR_t or UBC_t axes, respectively. Increased *PAX6* expression was also positively correlated to an increased proportion of Precursor and UBC_t cell-states, suggesting that *PAX6* drives tumor identity from the PR_t axis towards the UBC_t axis (Fig. 5D; Fig. S11B-G).

To investigate whether the intermediate Group 3/4 identity arises from co-expression of dual lineage factors in the same cells or instead results from the presence of two distinct lineages in separate cells in the same tumor, we focused our analysis on three (out of six) subtype VII samples (MB26, MB292 and MB129, ICGC cohort) from our atlas that showed a co-enrichment of PR_t and UBC_t-associated TF-GRN programs (1-3 and 6-9) (Fig. S12A). These samples also exhibited a distinct dual PR_t (gCRX) and UBC_t (gLHX1) trajectory arising from a common Precursor pool (gTBR1) in all the three samples (Fig. 5E,F; Fig. S12B-E). In all three samples, *PAX6* expression and PAX6 motif enrichment was high in the Precursor cells and almost completely absent in the PR_t tumor cells (Fig. 5G; Fig. S12F,G). *PAX6* expression further correlated to Precursor/UBC_t markers and anti-correlated to PR_t markers (Fig. S12H). Axial compartments in these tumors exhibited a mutually inverse enrichment of Group 3 (PR_t cells) and

Group 4 (Precursor/UBC$_t$ cells) tumor programs, confirming the intermediate nature of these tumor samples (Fig. 5H; Fig. S12I,J). Individual tumors recapitulated the axial TF-GRN activity pattern as observed in the integrated Group 3/4 multi-omic atlas (Fig. S12K-P; Fig. 2B), albeit without the MYC states due to absence of *MYC* expression. We then investigated the spatial distribution of tumor cell-states in two (out of three) of these intermediate tumors, in which we had appropriate tissue available. This spatial analysis showed that PR$_t$ and UBC$_t$ cells were spatially resolved (Fig. 5I; Fig. S13A-C), suggesting spatial compartmentalization of axial-states within intermediate subtype VII samples.

Altogether, the presence of the divergent PR$_t$/Precursor/UBC$_t$ tumor states in individual tumors suggests a shared regulatory landscape connecting these states, and that the TF-GRN interactome driving heterogeneity across Group 3/4 tumors also drives the intermediate identity of individual tumors.

To further test if the dual lineage in these intermediate tumors arises from mutual repression of PR$_t$- and UBC$_t$-associated TF-GRNs, as proposed earlier (Fig. 3D), we computationally knocked-down *CRX* and *EOMES* in individual tumors using CellOracle (*28*). *In silico* loss of *CRX* inhibited specification of PR$_t$ trajectory and loss of *EOMES* inhibited acquisition of UBC$_t$ states while pushing cells toward PR$_t$ identity (Fig. 5J; Fig. S14A-B). This data suggests that indeed potential mutual repression between key PR$_t$ and UBC$_t$ TF-GRNs drives tumor to acquire either Group 3 or Group 4 identity, and UBC specification is indeed the developmental time point that separates Group 3 and Group 4 (Fig. 4E). The absence of PAX6 TF-GRN in our *SCENIC+* analysis prevented us from performing *in silico PAX6* knock-down. However, based on *PAX6* expression and motif enrichment, together with the known dual function of *PAX6* in retinal and rhombic lip development (*26, 27*), we propose that *PAX6* expression in the Precursor pool maintains a bi-potent state that facilitates both the PR$_t$- and UBC$_t$-identity within the same tumor sample, but not in the same tumor cells (Fig. 5K).

Genetic aberrations that could explain this sustained subtype-specific *PAX6* expression, such as small variants, copy number aberrations or structural variants, have not been identified to date. Therefore, we searched for potential somatic aberrations underlying this aberrant expression, using bulk RNA-Seq data of subtype VII tumors from the ICGC cohort (*11-14*). We identified a previously unknown non-coding transcript downstream of the *PAX6* locus, and antisense to the *ELP4* gene (termed here as: *ELP4-AS*, Fig. 5L; Table S9). Expression of this novel transcript positively correlated with PAX6 expression (Fig. 5L; 12/21 of subtype VII samples and 1/4 of subtype I sample). We also identified samples where *ELP4-AS* was spliced to the downstream *IMMP1L* gene (*ELP4-AS:IMMP1L*, Fig. 5L), resulting in a putative chimeric lncRNA in ~57% (8/13 tumors) of *ELP4-AS*+ cases (Fig. S15A). All the 12 subtype VII tumors harboring *ELP4-AS* expression showed intermediate (n=4) or a predominantly Precursor/UBC$_t$ (n=8) identity, further alluding that a mechanism driving *PAX6* upregulation drives tumors toward Group 4-like tumor states (Fig. S15B).

DISCUSSION

Despite advances in identifying a unified rhombic lip origin of Group 3/4 medulloblastoma (*8, 9*), the causes of the underlying heterogeneity within this group remain unknown. Our single-cell multi-omic atlas unravels the molecular underpinnings driving Group 3/4 subtype-specific biology, while also addressing the continuity among these subtypes. Master regulators of retinal lineages, such as OTX2, CRX and PAX6, together with TFs driving UBC differentiation, such as BARHL1, LMX1A and EOMES, are among the known modulators of regulatory circuits driving Group 3/4 medulloblastoma heterogeneity (*11, 29*). Our analysis delineates the TF-interaction network that connects these master regulators to drive divergent tumor states in the same regulatory landscape; we also propose the regulatory logic that determines the transition across these states. We show that the presence of photoreceptor signature in Group 3/4 medulloblastoma, first reported in 1991 by Kramm *et al.* (*22*) is due to aberrant activation of a *CRX*-driven photoreceptor-specification cascade, as also suggested by Garancher *et al.* (*29*). Additionally, we show that the broad separation of Group 3 and Group 4 medulloblastoma stems from the failure of Group 3 tumors to attain *EOMES/LMX1A*-driven UBC identity and thus are propelled towards an alternative *CRX/NRL*-driven photoreceptor identity through the remodeling of the UBC progenitor (RLsvz) regulatory network. We suggest expression of key master regulators including *OTX2* and *PAX6* in the UBC progenitors prime this state to acquire divergent retinal photoreceptor lineage, in the case of stalled UBC specification. Further, we propose that, apart from arising at distinct stages/states during UBC lineage differentiation (*8, 9*), the mutual repression between CRX/NRL- and EOMES/LMX1A-driven GRNs contributes to the mutual exclusion of Group 3 and Group 4 tumor identities.

Our study identifies the connecting links between the oncogenic events and underlying lineage determinants that drive tumor identity away from the normal cerebellar UBC lineage, and induce aberrant retinal photoreceptor-lineage identity. Our data opens up an exciting possibility whereby a *GFI1B/PAX6*-driven tumor model, which shows co-enrichment of typically mutually exclusive PR_t and Precursor/UBC_t associated TF-GRN programs, can be tuned by modulating the TF activity to obtain pure Group 3- or Group 4-like tumors. Such model(s) would represent the spectrum of Group 3/4 heterogeneity and further improve our understanding of mechanisms that drive Group 3 or Group 4 identity and pinpoint underlying therapeutic vulnerabilities. A deeper understanding of lineage specification in Group 3/4 medulloblastoma could further identify yet unknown genetic or regulatory determinants of tumor identity.

Author Contributions

Conceptualization: PJ, SMP, LMK
Data acquisition: PJ, MS, AR, CS, JS, PBGdS, BS
Data analysis: PJ, TS, KO, MS, MVPJ
Methodology: PJ, TS, KO, IS, MS, TY-S, KL, VH, PAN, ST
Resources: IS, MS, PS, KL, MB, JPM, KS, MBJ, PF, BJ, TM, KP, CMvT, OW, KR, AK, DTWJ
Funding acquisition: SMP, LMK
Project administration: SMP, LMK
Supervision: HK, ST, NJ, SMP, LMK
Writing – original draft: PJ, TS, KO, SMP, LMK
Writing – review & editing: PJ, TS, KO, IS, MS, TM, DTWJ, VH, PAN, ST, NJ, HK, SMP, LMK.
All authors have read and approved the manuscript.

Conflict of interest

CMvT: Alexion, Bayer and Novartis (Advisory boards)
SMP and DTWJ: Heidelberg Epignostix (Founders)

Data and code availability:

Processed snRNA-Seq and snATAC-Seq data generated in this study are available on GEO under the identifier GSE253557 and GSE253573, respectively. Raw data will be provided after data transfer agreement. Scripts use in data processing are available on github: github.com/piyushjo15/G34MBGRN

Acknowledgments

The study was funded in part by European Research Council (ERC Advanced grant to S.M.P), the European Horizon 2020 Programme - ‘iPC - individualizedPaediatricCure’ (Grant 826121 to S.M.P). We thank the Fernando Goldsztein Family and the Cure Group 4 Consortium for their continued effort in fighting Group 4 medulloblastoma. The INFORM program is financially supported by the German Cancer Research Center (DKFZ), several German health insurance companies, the German Cancer Consortium (DKTK), the German Federal Ministry of Education and Research (BMBF), the German Federal Ministry of Health (BMG), the Ministry of Science, Research and the Arts of the State of Baden-Württemberg (MWK BW), the German Cancer Aid (DKH), the German Childhood Cancer Foundation (DKS), RTL television, the aid organization BILD hilft e.V. (Ein Herz für Kinder) and the generous private donation of the Scheu family. We also would like to express our sincere thanks to Carsten Maus, Erjia Wang (Genomics and Proteomics Core Facility, DKFZ) and Lena Weiser, Gregor Warsow (Omics IT and Data Management Core Facility, DKFZ) for their highly dedicated support in data management and processing. We thank Kimberly Siletti for critical reading of the manuscript.

References

1. V. Hovestadt et al., Medulloblastomics revisited: biological and clinical insights from thousands of patients. Nat Rev Cancer 20, 42-56 (2020).
2. M. F. Roussel, J. L. Stripay, Modeling pediatric medulloblastoma. Brain Pathol 30, 703-712 (2020).
3. D. P. Ivanov, B. Coyle, D. A. Walker, A. M. Grabowska, In vitro models of medulloblastoma: Choosing the right tool for the job. J Biotechnol 236, 10-25 (2016).
4. Central Nervous System Tumours: WHO Classification of Tumours., (International Agency for Research on Cancer, Lyons (France), ed. 5, 2021), vol. 6.
5. T. Sharma et al., Second-generation molecular subgrouping of medulloblastoma: an international meta-analysis of Group 3 and Group 4 subtypes. Acta Neuropathol 138, 309-326 (2019).
6. D. R. Ghasemi, G. Fleischhack, T. Milde, K. W. Pajtler, The Current Landscape of Targeted Clinical Trials in Non-WNT/Non-SHH Medulloblastoma. Cancers (Basel) 14, (2022).
7. K. Okonechnikov et al., Mapping pediatric brain tumors to their origins in the developing cerebellum. Neuro Oncol 25, 1895-1909 (2023).
8. L. D. Hendrikse et al., Failure of human rhombic lip differentiation underlies medulloblastoma formation. Nature 609, 1021-1028 (2022).

9. K. S. Smith et al., Unified rhombic lip origins of group 3 and group 4 medulloblastoma. Nature 609, 1012-1020 (2022).
10. D. Williamson et al., Medulloblastoma group 3 and 4 tumors comprise a clinically and biologically significant expression continuum reflecting human cerebellar development. Cell Rep 40, 111162 (2022).
11. C. Y. Lin et al., Active medulloblastoma enhancers reveal subgroup-specific cellular origins. Nature 530, 57-62 (2016).
12. D. T. Jones et al., Dissecting the genomic complexity underlying medulloblastoma. Nature 488, 100-105 (2012).
13. P. A. Northcott et al., Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. Nature 511, 428-434 (2014).
14. P. A. Northcott et al., The whole-genome landscape of medulloblastoma subtypes. Nature 547, 311-317 (2017).
15. C. Bravo González-Blas et al., SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. Nat Methods 20, 1355-1367 (2023).
16. S. Aibar et al., SCENIC: single-cell regulatory network inference and clustering. Nat Methods 14, 1083-1086 (2017).
17. P. Lyu et al., Gene regulatory networks controlling temporal patterning, neurogenesis, and cell-fate specification in mammalian retina. Cell Rep 37, 109994 (2021).
18. S. Li et al., Foxn4 controls the genesis of amacrine and horizontal cells by retinal progenitors. Neuron 43, 795-807 (2004).
19. O. Zaytseva, N. H. Kim, L. M. Quinn, MYC in Brain Development and Cancer. Int J Mol Sci 21, (2020).
20. M. Sepp et al., Cellular development and evolution of the mammalian cerebellum. Nature, (2023).
21. A. McDonough et al., Unipolar (Dendritic) Brush Cells Are Morphologically Complex and Require Tbr2 for Differentiation and Migration. Front Neurosci 14, 598548 (2020).
22. C. M. Kramm, H. W. Korf, M. Czerwionka, W. Schachenmayr, W. J. de Grip, Photoreceptor differentiation in cerebellar medulloblastoma: evidence for a functional photopigment and authentic S-antigen (arrestin). Acta Neuropathol 81, 296-302 (1991).
23. C. A. Mao et al., Eomesodermin, a target gene of Pou4f2, is required for retinal ganglion cell and optic nerve development in the mouse. Development 135, 271-280 (2008).
24. T. J. Cherry et al., Mapping the cis-regulatory architecture of the human retina reveals noncoding genetic variation in disease. Proc Natl Acad Sci U S A 117, 9001-9012 (2020).
25. A. K. Teo et al., Pluripotency factors regulate definitive endoderm specification through eomesodermin. Genes Dev 25, 238-250 (2011).
26. V. Oron-Karni et al., Dual requirement for Pax6 in retinal progenitor cells. Development 135, 4037-4047 (2008).
27. J. Yeung, T. J. Ha, D. J. Swanson, D. Goldowitz, A Novel and Multivalent Role of Pax6 in Cerebellar Development. J Neurosci 36, 9057-9069 (2016).
28. K. Kamimoto et al., Dissecting cell identity via network inference and in silico gene perturbation. Nature 614, 742-751 (2023).
29. A. Garancher et al., NRL and CRX Define Photoreceptor Identity and Reveal Subgroup-Specific Dependencies in Medulloblastoma. Cancer Cell 33, 435-449.e436 (2018).
30. C. M. van Tilburg et al., The Pediatric Precision Oncology INFORM Registry: Clinical Outcome and Benefit for Patients with Very High-Evidence Targets. Cancer Discov 11, 2764-2779 (2021).
31. B. Kaminow, D. Yunusov, A. Dobin, STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. bioRxiv, 2021.2005.2005.442755 (2021).
32. M. Alvarez et al., Enhancing droplet-based single-nucleus RNA-seq resolution using the semi-supervised machine learning classifier DIEM. Sci Rep 10, 11019 (2020).
33. M. D. Young, S. Behjati, SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. Gigascience 9, (2020).
34. S. Yang et al., Decontamination of ambient RNA in single-cell RNA-seq with DecontX. Genome Biol 21, 57 (2020).
35. S. L. Wolock, R. Lopez, A. M. Klein, Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Syst 8, 281-291.e289 (2019).
36. A. T. Lun, D. J. McCarthy, J. C. Marioni, A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Res 5, 2122 (2016).
37. L. Haghverdi, A. T. L. Lun, M. D. Morgan, J. C. Marioni, Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol 36, 421-427 (2018).
38. J. D. Welch et al., Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. Cell 177, 1873-1887.e1817 (2019).
39. J. M. Granja et al., ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nat Genet 53, 403-411 (2021).
40. P. Angerer et al., destiny: diffusion maps for large-scale single-cell data in R. Bioinformatics 32, 1241-1243 (2016).
41. T. Stuart et al., Comprehensive Integration of Single-Cell Data. Cell 177, 1888-1902.e1821 (2019).
42. P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559 (2008).
43. A. Dobin et al., STAR: ultrafast universal RNA-seq aligner.

Bioinformatics 29, 15-21 (2013).

44. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15, 550 (2014).

45. Y. Lin et al., scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. Nat Biotechnol 40, 703-710 (2022).

46. T. Chu, Z. Wang, D. Pe'er, C. G. Danko, Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. Nat Cancer 3, 505-517 (2022).

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

48. S. Uhrig et al., Accurate and efficient detection of gene fusions from RNA sequencing data. Genome Res 31, 448-460 (2021).

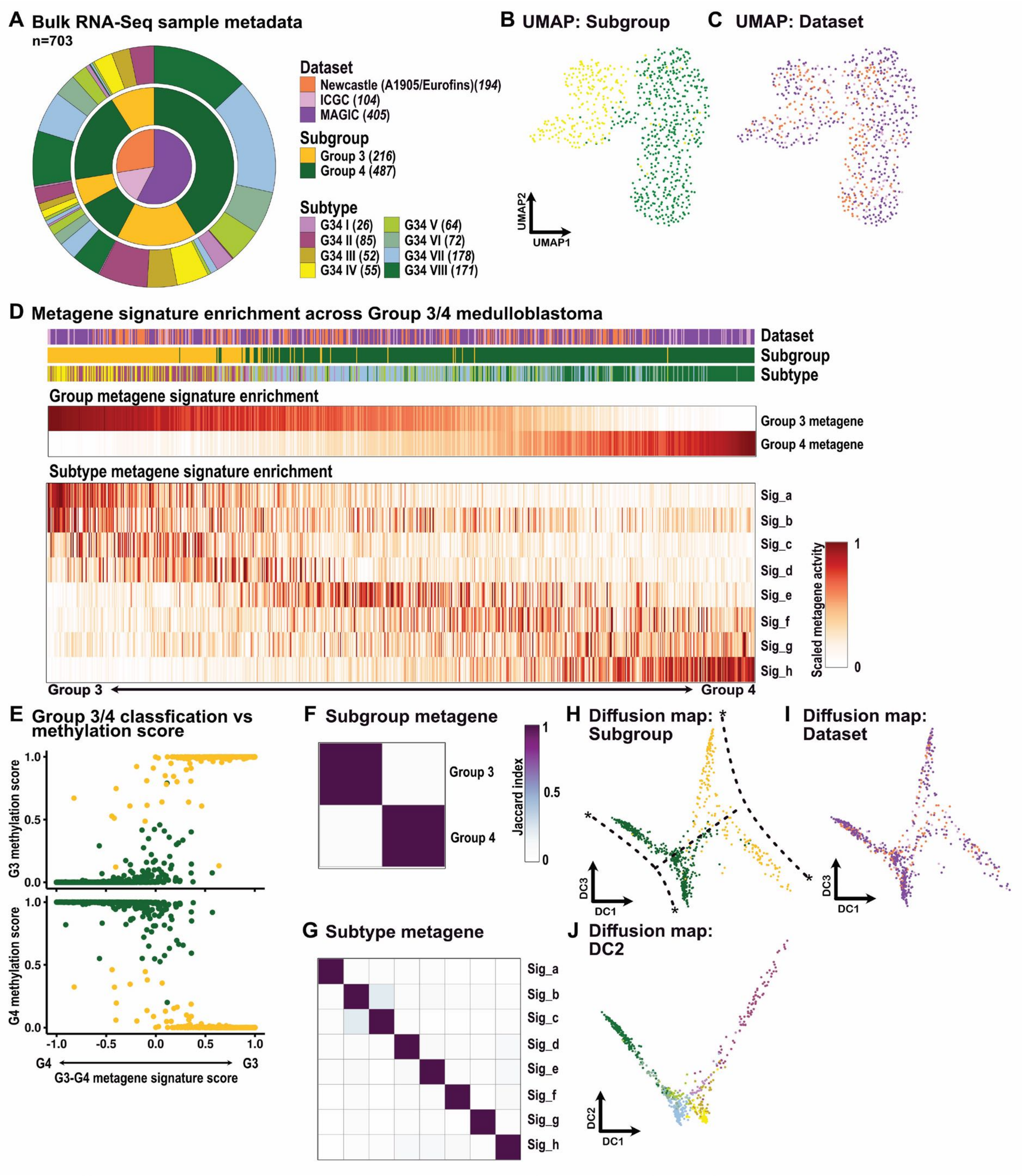
49. F. Sahm et al., Meningiomas induced by low-dose radiation carry structural variants of NF2 and a distinct mutational signature. Acta Neuropathol 134, 155-158 (2017).

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

51. V. Petukhov et al., Cell segmentation in imaging-based spatial transcriptomics. Nat Biotechnol 40, 345-354 (2022).

52. T. Biancalani et al., Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. Nat Methods 18, 1352-1362 (2021).

53. D. R. Ghasemi et al., Compartments in medulloblastoma with extensive nodularity are connected through differentiation along the granular precursor lineage. Nat Commun 15, 269 (2024).



Supplemental Figure S1 *legend next page* ►

◀*cont. from previous page*

Fig. S1: Group 3/4 medulloblastoma bulk RNA-Seq data analysis and metagene signatures. (A) Group 3/4 medulloblastoma bulk-RNA-Seq metadata collated from three sources: ICGC, MAGIC and Newcastle (*8, 10-14*). Numbers of samples per category are depicted in parenthesis. **(B and C)** UMAP distribution of tumor samples on the transcription program landscape colored by group (B) and dataset (C) identity. **(D)** Scaled subgroup and subtype-specific metagene score (NMF component value) per sample. Samples are arranged on a Group 3 –Group 4 metagene score scale. **(E)** Per sample methylation-based Group 3 (top) and Group 4 (bottom) classification score (y axis) vs Group 3 – Group 4 metagene score (x axis). Tumor samples are colored as per subgroup identity. **(F and G)** Jaccard similarity between subgroup (F) and subtype (G) specific metagene gene-sets. For each metagene gene-set top 100 genes ranked by contribution per metagene were used. **(H and I)** Diffusion map of samples colored as per subgroup (H) and dataset (I) identity. **(J)** Diffusion map with samples colored as per subtype identity. DC2 is shown instead of DC3.

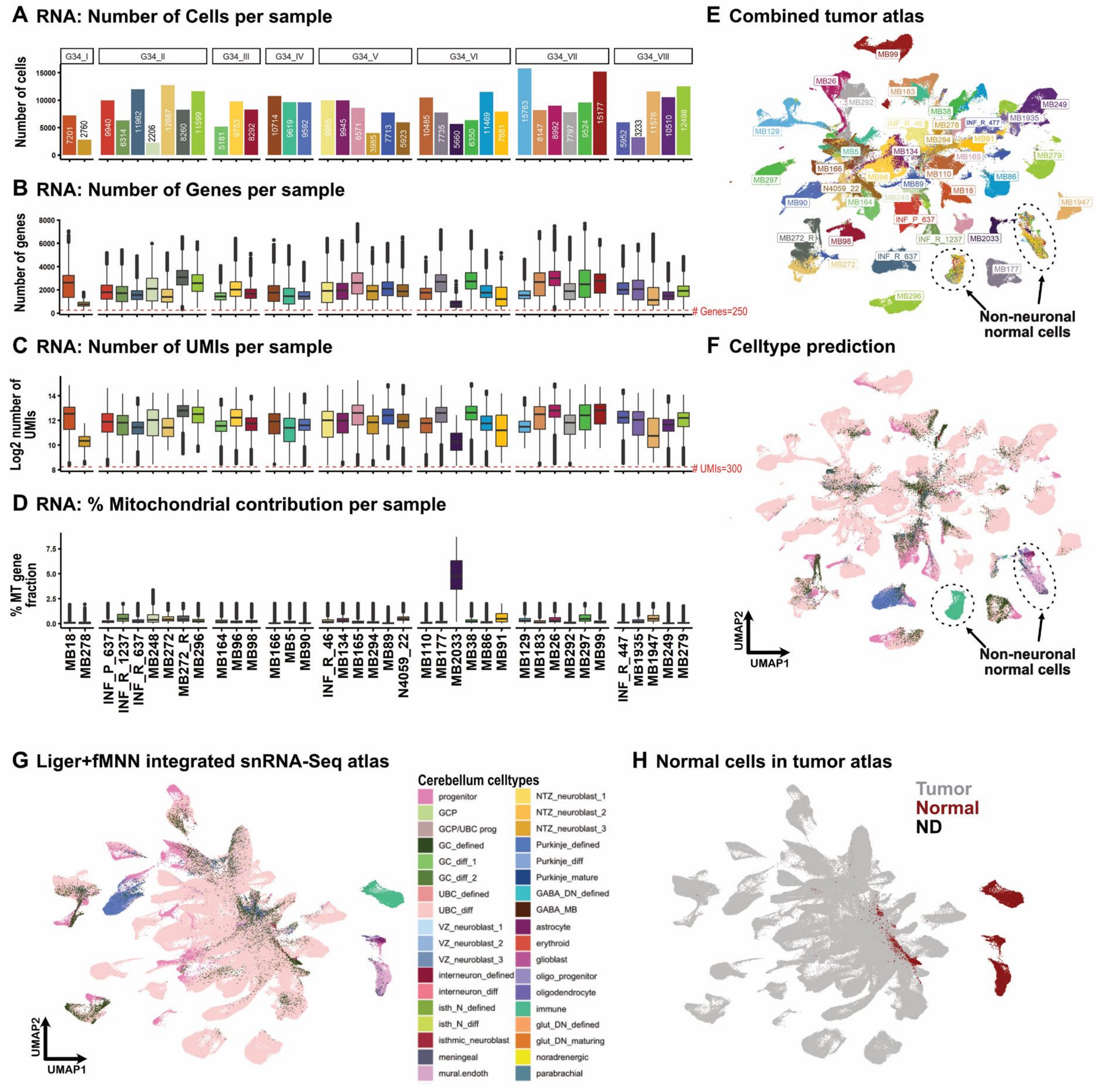


Fig. S2: Group 3/4 tumor single-nucleus RNA-Seq (snRNA-Seq) data quality control (QC) metrics. (A) Number of cells per sample in snRNA-Seq data post QC filtering. (**B-D**) Per sample distribution of number of genes (B), unique molecular identifiers (UMIs) (C), and fractional mitochondrial gene contribution (D). Dotted line shows cut-off at 250 Genes (B) and 300 UMIs (C). **(E and F)** UMAP distribution of cells in the merged snRNA-Seq data (without batch correction) colored by sample identity (E) and predicted cell-type labels using reference cerebellum data (*20*) (F). Non-neuronal normal cells are encircled. **(G and H)** UMAP distribution of LIGER-fMNN batch-corrected snRNA-Seq data colored by predicted cell-type label (G) and identified non-tumor cells (Normal and not-determined/ ND) (H).

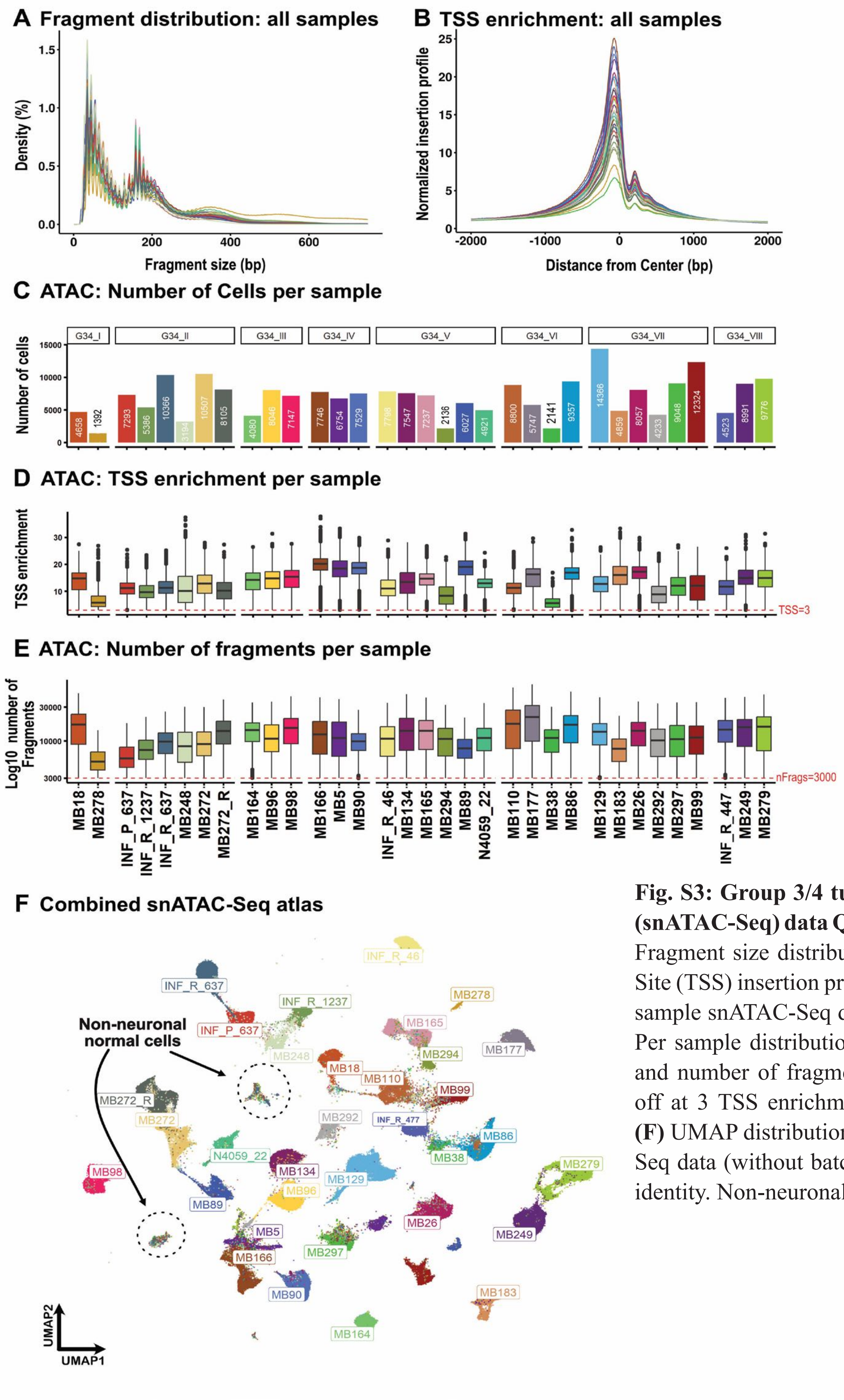


Fig. S3: Group 3/4 tumor single-nucleus ATAC-Seq (snATAC-Seq) data QC metrics. (A and B) Per sample Fragment size distribution (A) and Transcription Start Site (TSS) insertion profile (B). **(C)** Number of cells per sample snATAC-Seq data post QC filtering. **(D and E)** Per sample distribution of TSS enrichment score (D), and number of fragments (E). Dotted line shows cut-off at 3 TSS enrichment (D) and 3000 fragment (E). **(F)** UMAP distribution of cells in the merged snATAC-Seq data (without batch correction) colored by sample identity. Non-neuronal normal cells are encircled.

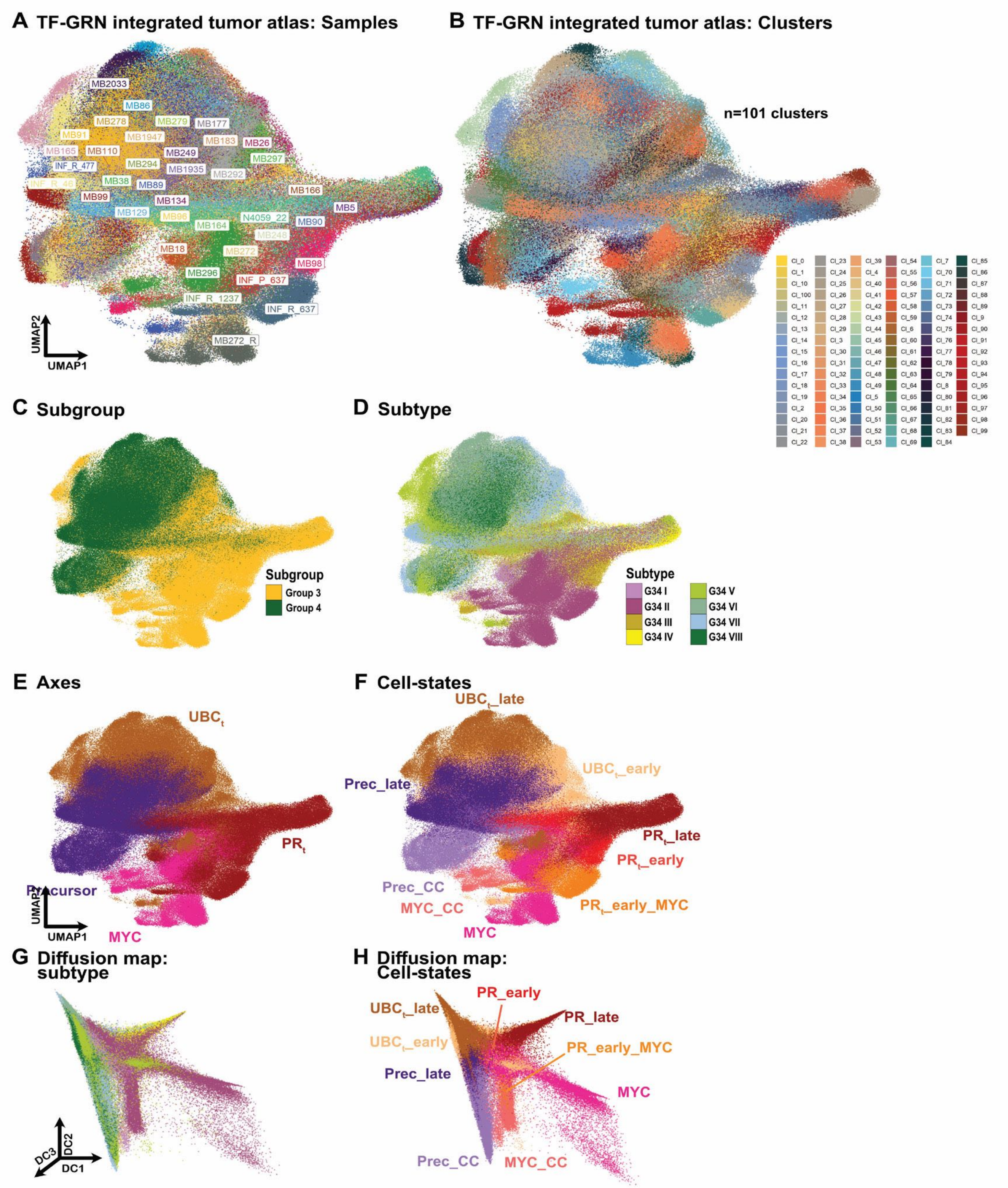
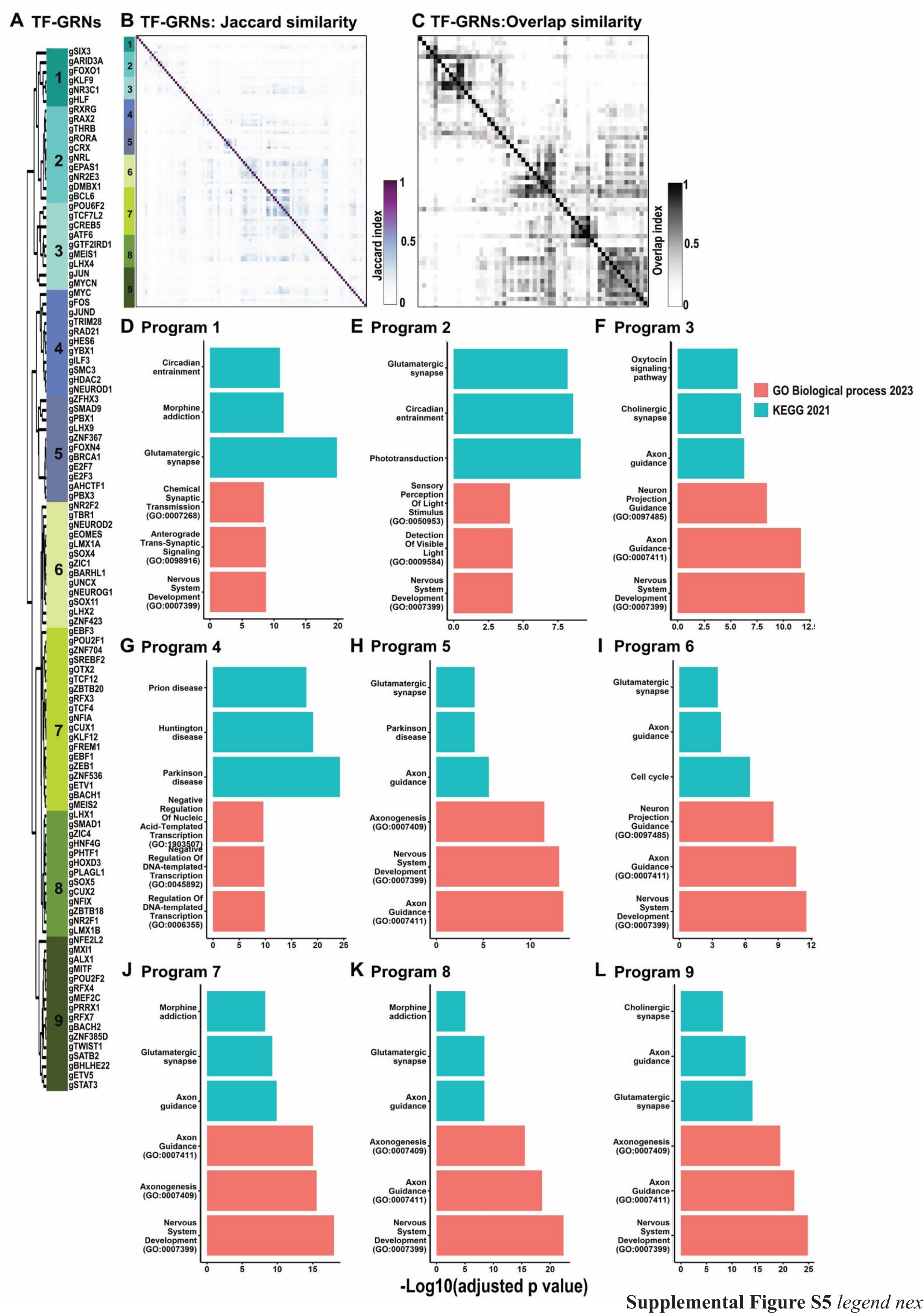


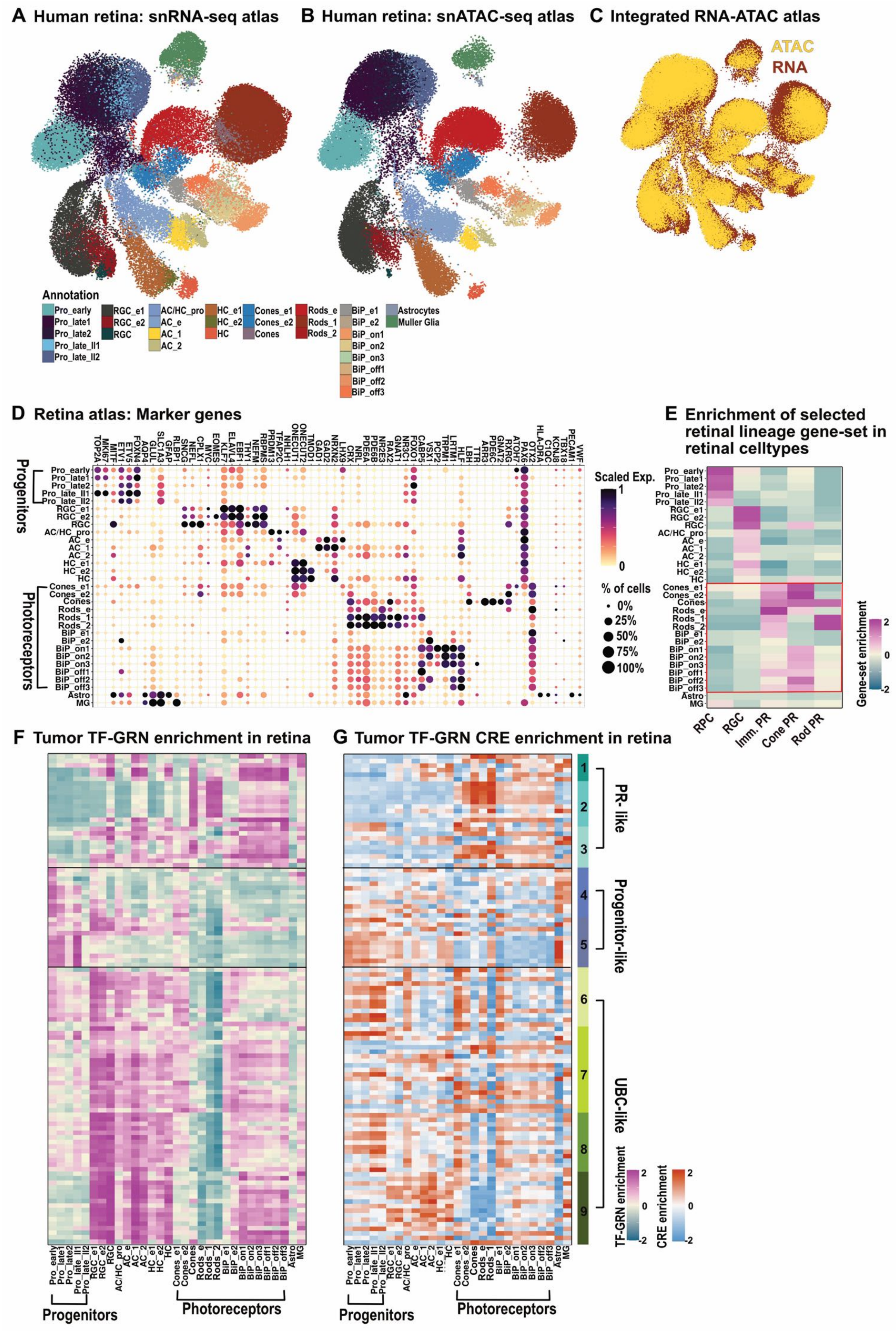
Fig. S4: UMAP and diffusion map distribution of tumor cells based on TF-GRN AUC scores. (A-F) UMAP distribution of tumor cells colored by sample (A), KNN-leiden clusters (B), subgroup (C), subtype (D), axial (E) and cell-state (F) identities in the TF-GRN enrichment space. **(G and H)** 3D diffusion map distribution of tumor cells colored by subtype (G) and cell-state identity (H) in the TF-GRN enrichment space.



Supplemental Figure S5 *legend next page* ►

◀*cont. from previous page*

Fig. S5: TF-GRNs driving Group 3/4 medulloblastoma axial identities. (A) TF-GRNs (n=108) denoted by the name of regulatory TF. TF-GRNs are arranged by their order in the hierarchical clustering. Colored column bars represent TF-GRN groups, referred to as TF-GRN programs (1-9). **(B and C)** Jaccard similarity (B) and overlap similarity (C) indices for TF-GRN sets. **(D-L)** Top three gene ontology (GO) Biological process and KEGG terms (sorted by –(Log10(adjusted p-value)) associated with genes contributing to each TF-GRN programs 1-9 (D-L).



Supplemental Figure S6 *legend next page* ►

◀*cont. from previous page*

Fig. S6: Human developing retina multi-omic atlas. (A and B) UMAP distribution of developing human retina snRNA-Seq data (A) and snATAC-Seq data (B) (*17*) colored by cell-type annotation in the retinal TF-GRN enrichment space. **(C)** Overlap of snRNA-Seq and snATAC-Seq data colored by data modality. **(D)** Expression of selected marker genes in the annotated retinal cell-types. **(E)** Enrichment of selected retinal lineage marker gene-sets in retinal cell-types. Red box encircles retinal photoreceptor cell-types. **(F)** Relative enrichment of Group 3/4 medulloblastoma tumor TF-GRNs in the retina snRNA-Seq atlas. **(G)** Relative enrichment of tumor TF-GRNs associated cis-regulatory elements (CREs) in the retina snATAC-Seq atlas.

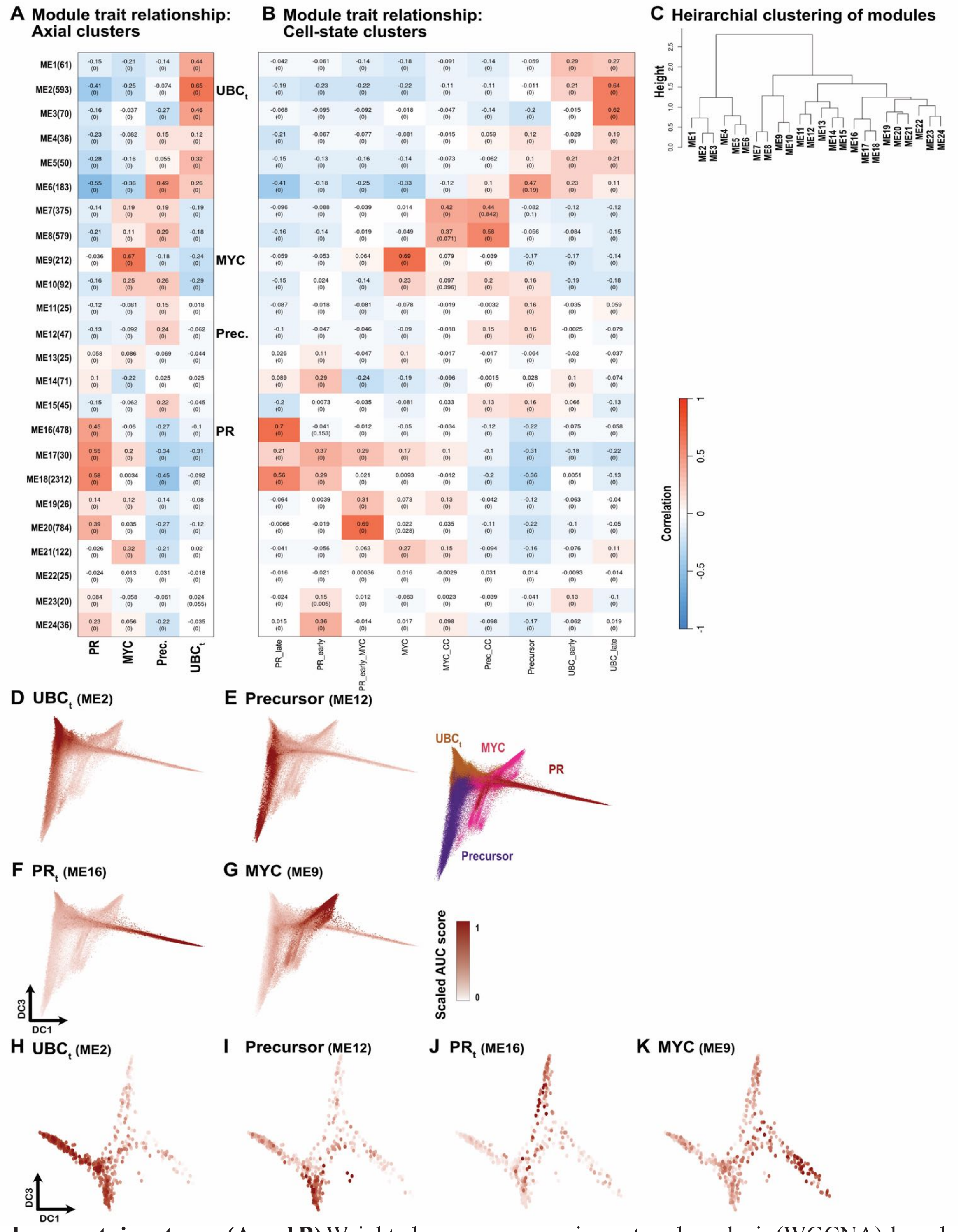


Fig. S7: Axial gene-set signatures. (A and B) Weighted gene co-expression network analysis (WGCNA)-based module-trait relationship correlation heatmap. Rows are modules identified from WGCNA analysis. Columns are tumor cells clustered based on annotated axial identities (A) or cell-state identities (B). Correlation and associated p-value (in brackets) for each module-trait combination are noted in each cell (A). Representative module per axis marked with the axis name on the left in (A). **(C)** Hierarchical clustering of modules. **(D-G)** Scaled enrichment of signature axial gene-set module for UBC$_t$ (D), Precursor (E), PR$_t$ (F), and MYC (G) axes on the integrated tumor diffusion map. Diffusion map with cells colored by axial identities at the bottom for reference. **(H-K)** Scaled enrichment of signature axial gene-set module for UBC$_t$ (G), Precursor (H), PR$_t$ (I), and MYC (J) axes on the bulk-RNA-Seq diffusion map.

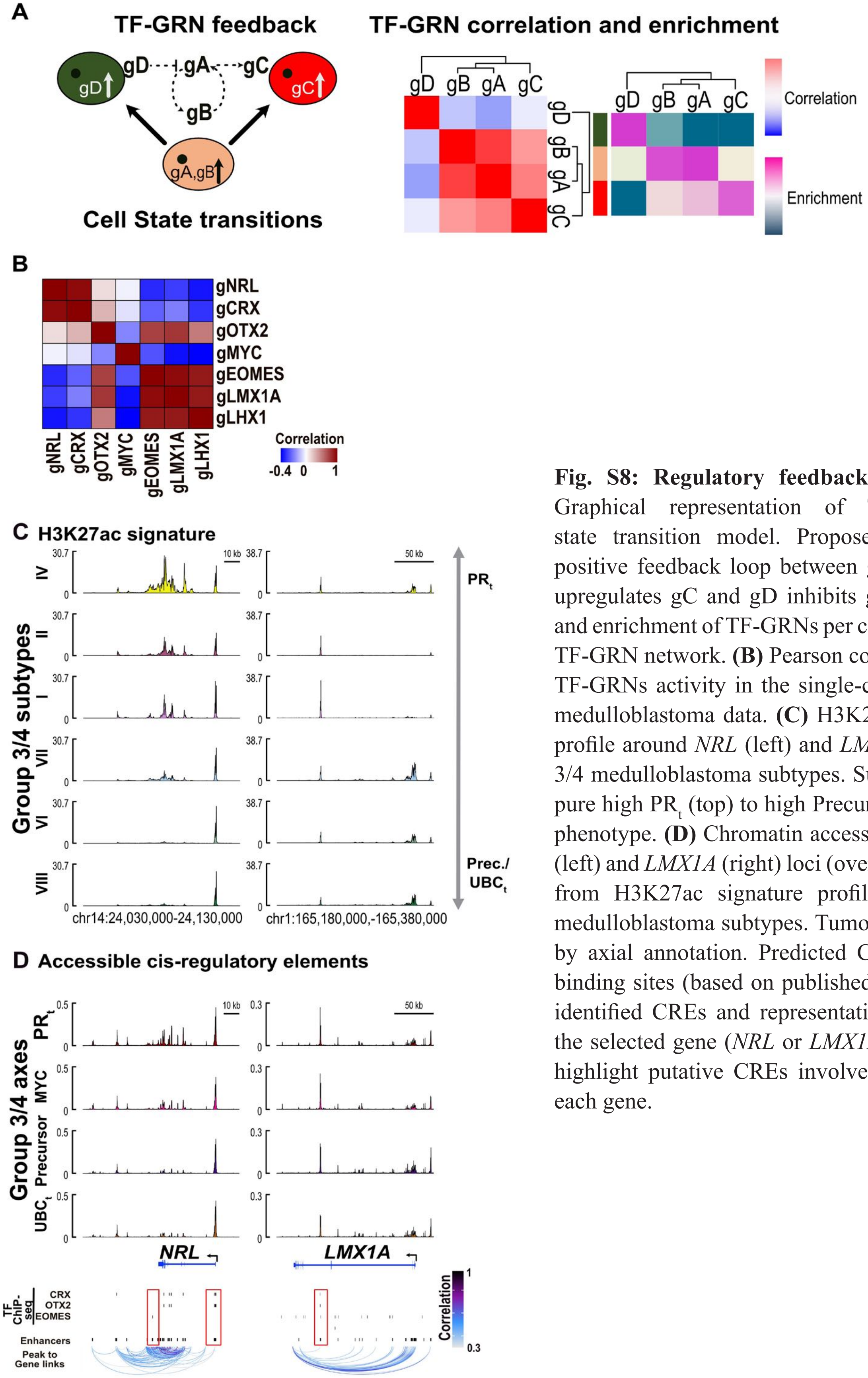


Fig. S8: Regulatory feedback among TF-GRNs. (A) Graphical representation of TF-GRN directed cell-state transition model. Proposed TF-GRN interactions: positive feedback loop between gA and gB. gA positively upregulates gC and gD inhibits gA. Expected correlations and enrichment of TF-GRNs per cell-state from the proposed TF-GRN network. **(B)** Pearson correlation between selected TF-GRNs activity in the single-cell multi-omic Group 3/4 medulloblastoma data. **(C)** H3K27ac ChIP-Seq (*11*) signal profile around *NRL* (left) and *LMX1A* (right) loci in Group 3/4 medulloblastoma subtypes. Subtypes are arranged from pure high PR_t (top) to high Precursor (Prec.)/UBC_t (bottom) phenotype. **(D)** Chromatin accessibility profile around *NRL* (left) and *LMX1A* (right) loci (overlapping region as selected from H3K27ac signature profile in (C)) in Group 3/4 medulloblastoma subtypes. Tumor cells were pseudobulked by axial annotation. Predicted CRX, OTX2 and EOMES binding sites (based on published ChIP-Seq data) (*24, 25*), identified CREs and representative peak-to-gene links for the selected gene (*NRL* or *LMX1A*) shown below. Red box highlight putative CREs involved in cross-regulations for each gene.

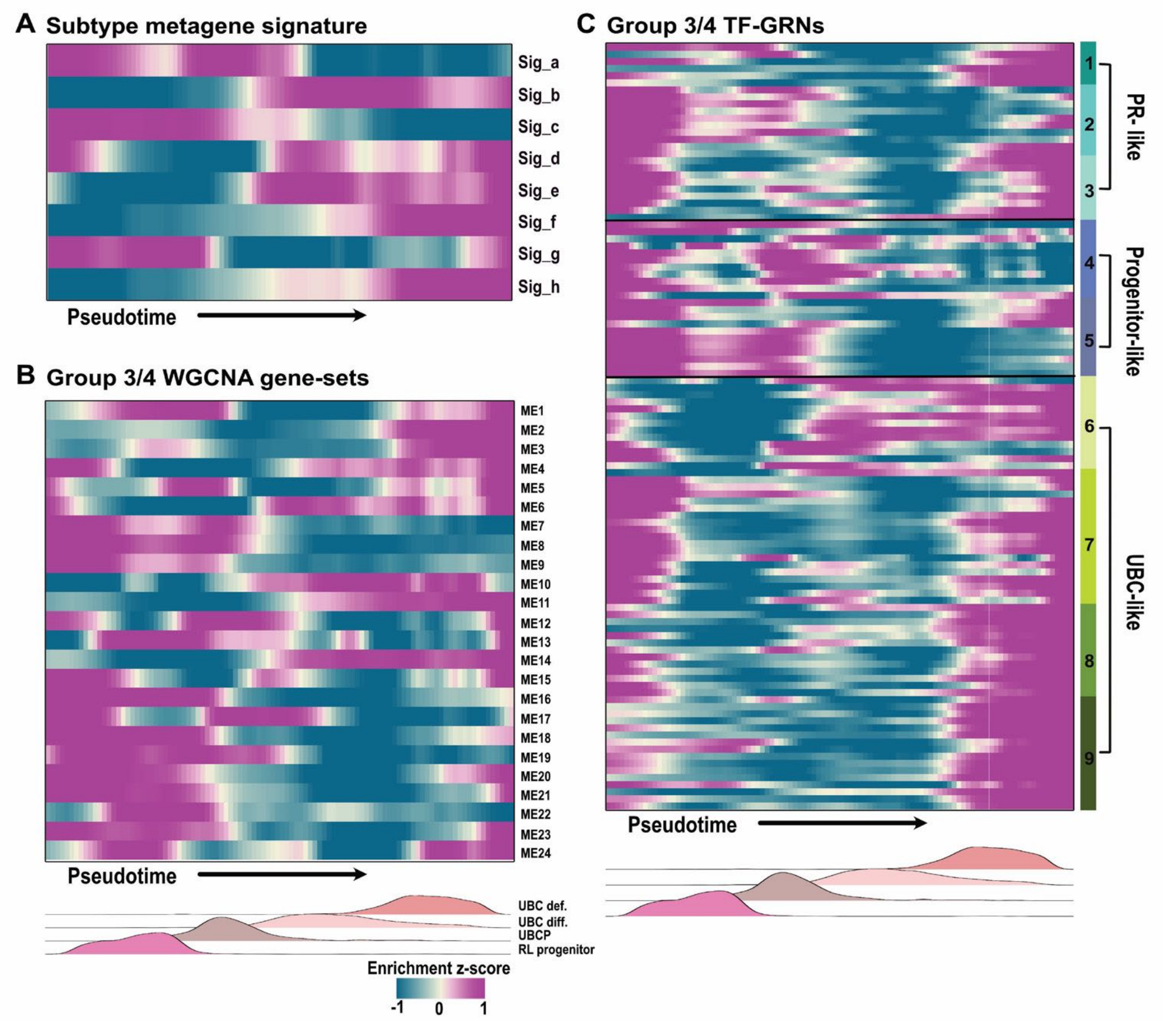
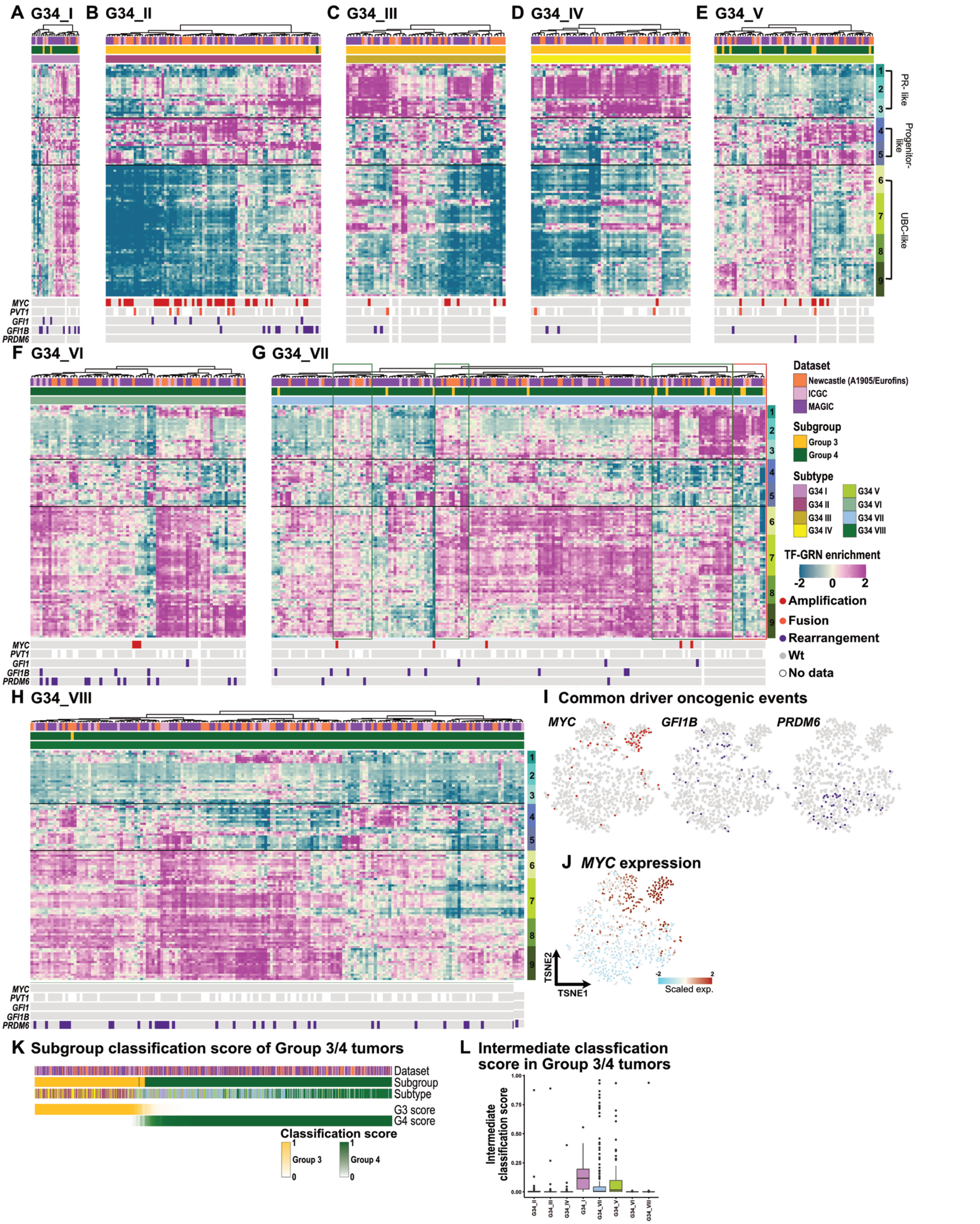


Fig. S9: Signature gene-set enrichment in cerebellar UBC lineage. (A-C) Relative enrichment of subtype-specific metagene signature (top 100 genes ranked by contribution) (A), Group 3/4 medulloblastoma weighted gene co-expression network analysis (WGCNA) module sets (B), and Group 3/4 medulloblastoma TF-GRNs (C) in the cerebellar UBC lineage. Density distribution of cerebellar UBC-lineage along pseudotime at the bottom for reference.



Supplemental Figure S10 *legend next page* ►

◀*cont. from previous page*

Fig. S10: Bulk-RNA-Seq tumor TF-GRN enrichment. (A-H) TF-GRNs enrichment heatmap for subtype I (A), II (B), III (C), IV (D), V (E), VI (F), VII (G), VIII (H). *MYC* (amplification), *PVT1* (fusion), *GFI1*, *GFI1B* and *PRDM6* (rearrangements) events shown at the bottom of each strip. Samples with dual enrichment of PR_t- and UBC_t-associated TF-GRN programs in subtype VII samples are encircled in green box. Samples with enrichment of only PR_t-associated TF-GRN programs are encircled in red box. (G). **(I)** Distribution of samples with documented genomic alterations in *MYC*, *GFI1B* and *PRDM6* on the tSNE landscape. **(J)** Scaled expression of *MYC* on the tSNE landscape. **(K)** Group 3/4 tumor arranged on a Group 3 – Group 4 methylation classification score. Methylation classification score for each subgroup is on a scale of 0-1. **(L)** Boxplot distribution of intermediate classification score (1-abs(G3 score – G4 score)). Samples are grouped by subtype identity.

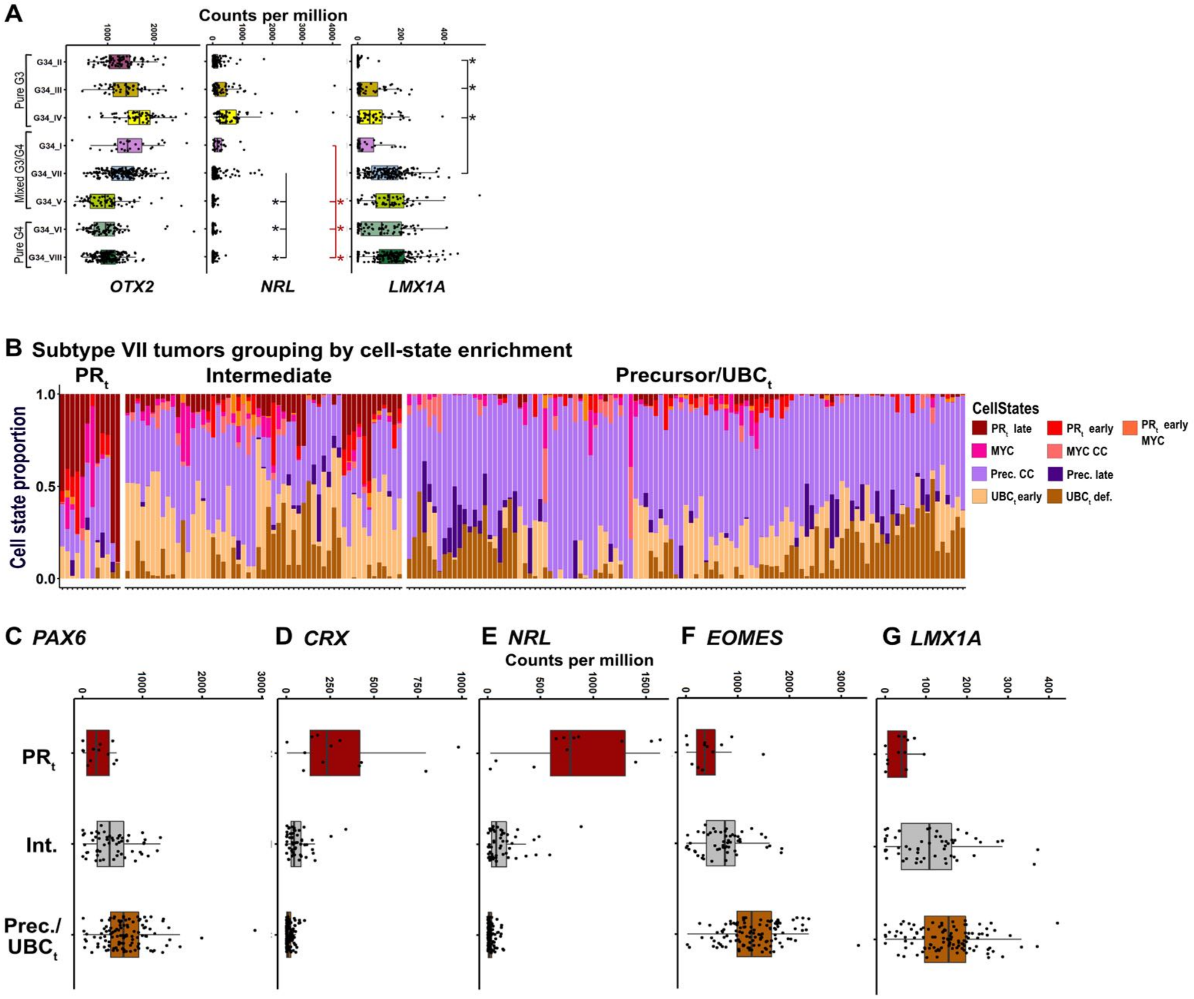
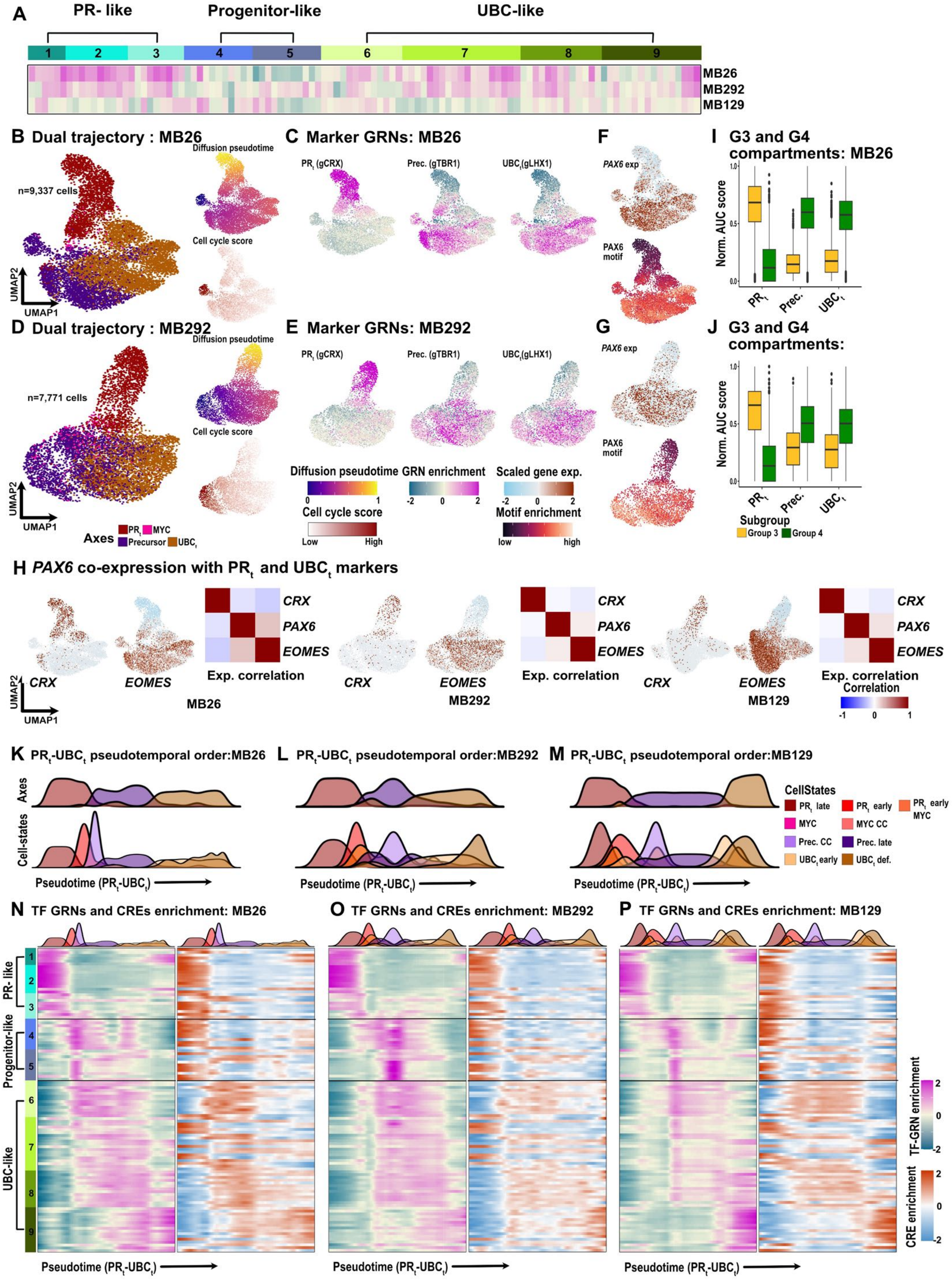


Fig. S11: Increased *PAX6* expression drives tumors toward UBC$_t$ states. (A) Expression of *OTX2* (left), *NRL* (middle) and *LMX1A* (right) in bulk tumor samples across eight subtypes. Statistically significant upregulation of genes is shown by dashed lines and asterisk (log-fold change >0.5 and adjusted p-value < 0.01). Black, subtype VII tumors compared with II-IV tumors. Red, subtype I tumors compared with II-IV tumors. **(B)** Predicted composition of axial cell-states in subtype VII samples (as shown in Fig. 5D) split by PR$_t$, Intermediate or Precursor/UBC$_t$ annotation. PR$_t$ annotated samples exhibited enrichment of PR$_t$-associated TF-GRN programs (red box, Fig. S10G), Intermediate samples exhibited dual enrichment of PR$_t$ and UBC$_t$ associated TF-GRN programs (green boxes, Fig S10G), Precursor/UBC$_t$ annotated samples are rest of the samples. **(C-G)** Expression distribution of *PAX6* (C), *CRX* (D), *NRL* (E), *EOMES* (F) and *LMX1A* (G) in PR$_t$, Intermediate and Precursor (Prec.)/UBC$_t$ annotated subtype samples, as in (B). Dots represent individual samples. Outliers not shown.



Supplemental Figure S12 *legend next page* ►

◀*cont. from previous page*

Fig. S12: Intermediate nature of *PAX6*+ subtype VII samples. (A) Enrichment of TF-GRNs in bulk RNA-Seq data of selected subtype VII samples, MB129, MB292 and MB26. (**B**) UMAP distribution of tumor cells for MB26. Cells are colored as per axial identities. Panels on the left shows diffusion pseudotime (top) and cell cycle score (bottom). **(C)** Enrichment of marker TF-GRNs gCRX (PR_t), gTBR1 (Prec./Precursor) and gLHX1 (UBC_t) shown on the MB26 UMAP. **(D)** UMAP distribution of tumor cells for MB292. Cells are colored as per axial identities. Panels show diffusion pseudotime (top) and cell cycle score (bottom). **(E)** Enrichment of marker TF-GRNs shown on the MB292 UMAP. **(F-G)** Scaled *PAX6* expression (top) and PAX6 motif enrichment (bottom) on the MB26 (F) and MB292 (G) UMAP. **(H)** Scaled expression of *CRX* and *EOMES* along with their Pearson correlation with *PAX6* in MB26 (left), MB292 (middle) and MB129 (right). Enrichment of marker TF-GRNs shown on the MB292 UMAP. **(I-J)** Scaled Group 3 and Group 4 metagene AUC score in cells labeled as PR_t, Precursor or UBC_t in MB26 (I) and MB292 (J). **(K-M)** Tumor cell density along the PR_t-to-UBC_t trajectory pseudotime for MB26 (K), MB292 (L) and MB129 (M). **(N-P)** TF-GRN enrichment (left) and TF-GRN associated CREs enrichment (right) along the predicted PR_t-to-UBC_t trajectory pseudotime for MB26 (N), MB292 (O) and MB129 (P).

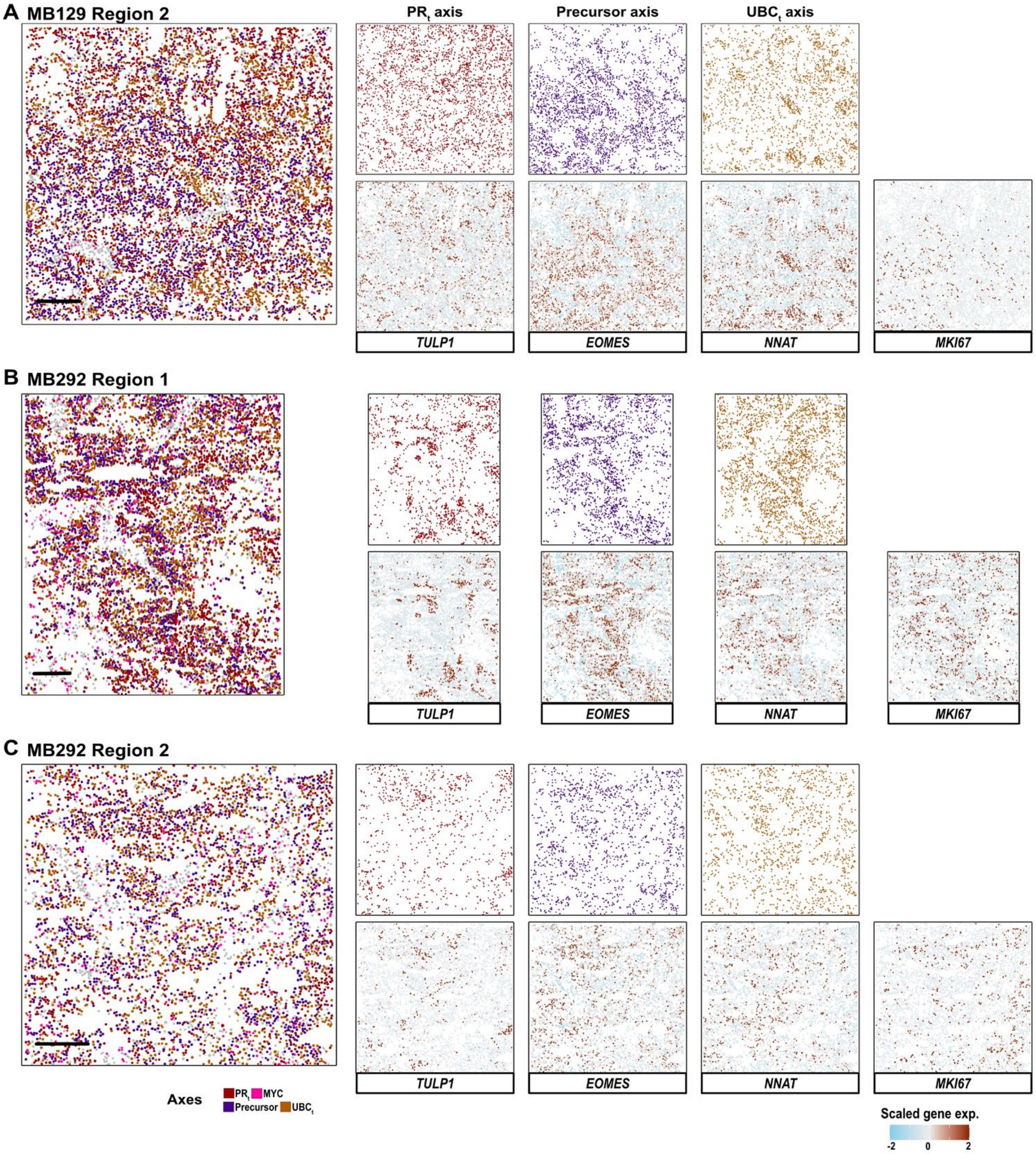


Fig. S13: Spatial compartmentalization in subtype VII samples. (A-C) Spatial transcriptomic data for MB129 region 2 (A), MB292 region 1 (B) and MB292 region 2 (C). Cells are colored by predicted axial annotation. Top panels show spatial location of cells labeled as PR$_t$-, Precursor- and UBC$_t$- like tumor cells. Bottom panels show scaled expression of marker genes *TULP1* (PR$_t$), *EOMES* (Precursor) and *NNAT* (UBC$_t$). Expression of *MKI67* represent cell-cycling tumor cells. Scale bars, 200 µM.

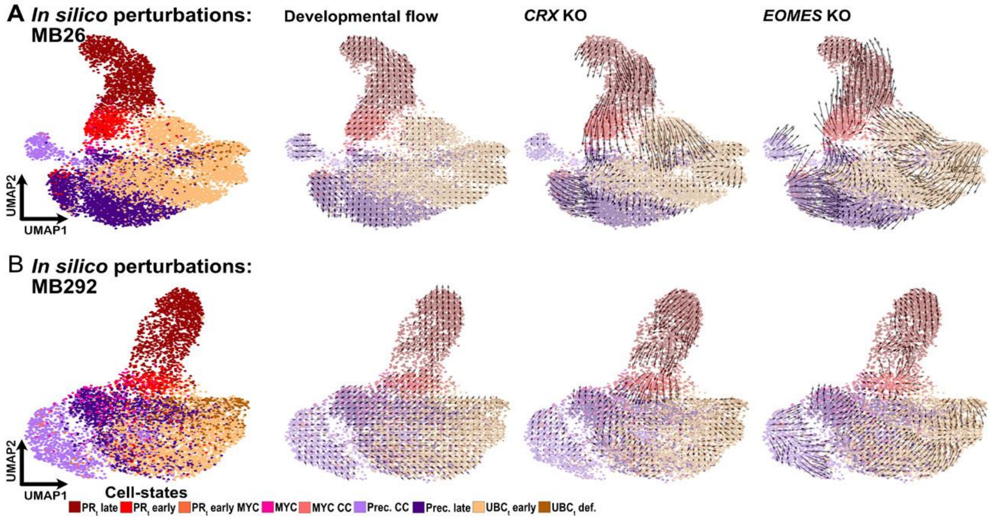


Fig. S14: *In silico* perturbation of CRX and EOMES GRN. (A and B) *In silico* knock-out (KO) of *CRX* and *EOMES* in MB26 (A) and MB292 (B). Cells are colored by cell-states. Arrows show predicted local trajectory of cells in control, *CRX* KO and *EOMES* KO simulations.

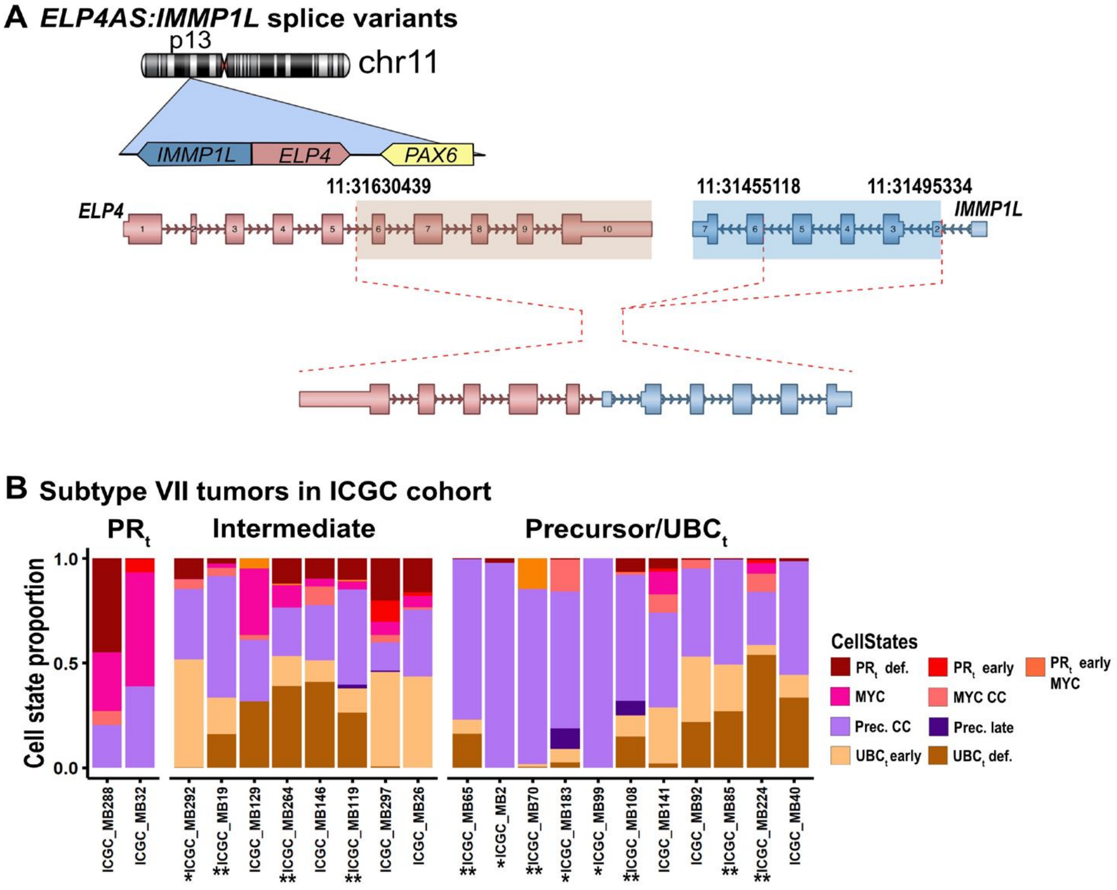


Fig. S15: *ELP4-AS* expression is associated with enrichment of UBC$_t$ states. (A) Summary figure depicting splice events between *ELP4-AS* and *IMMP1L*. **(B)** Predicted composition of axial cell-states in subtype VII samples belonging to the ICGC cohort (as shown in Fig. 5D) split by PR$_t$, Intermediate or Precursor/UBC$_t$ annotation. Single asterisk (*) denotes presence of *ELP4-AS* and doublet asterisks (**) denotes presence of *ELP4-AS:IMMP1L*.

Supplemental Table Legends

Table S1. Metadata for samples included in the bulk RNA-Seq data.
Table S2. Metagene sets from NMF analysis of bulk RNA-Seq samples. Top 100 genes ranked by contribution are shown.
Table S3. Metadata for samples included in the single-nucleus multi-omic atlas.
Table S4. 108 TF-GRN sets obtained from integrated tumor data analysis.
Table S5. Gene module sets obtained from the weighted gene co-expression network analysis (WGCNA).
Table S6. Signature gene set for selected cell-types in GC/UBC and retinal lineages.
Table S7. Log-fold change and adjusted p-value for selected genes in subtype I/VII pair-wise comparisons.
Table S8. Top 250 differentially expressed genes in subtype I and subtype VII tumor in pair-wise comparisons.
Table S9. ICGC samples with *ELP4-AS* or *ELP4-AS:IMMP1L* transcripts.

Materials and Methods

Sample selection

Target tumor tissue samples were collected from published studies (ICGC (*14*) and INFORM (*30*) cohorts). No statistical methods were used to pre-determine the sample size. Experiments were not randomized, and investigators were not blinded to tumor sample characteristics and experiment outcome.

per cluster to obtain the TF-GRN enrichment heatmap. The scaled TF-GRN matrix (clusters x TF-GRNs) was hierarchically clustered to obtain groups of co-enriched TF-GRNs (annotated as TF-GRN programs) and groups of tumor cluster exhibiting similar TF-GRN activity (annotated as tumor axes and cell-states). We also used *addmodulescore()* (*Seurat*, *R*) (*41*) to calculate activity scores for each of the identified 108 TF-GRN sets in the combined tumor cell data and used this score to calculate Pearson correlation between TF-GRNs.

Single-nucleus multi-omic sequencing

Flash frozen tumor samples were processed to extract nuclei as described (*20*). Tumor samples were finely cut into pieces using a surgical blade on dry ice. Cut tissue was homogenized in the homogenization buffer (for details of reagents, *20)* by trituration or douncing with a micropestle. Cellular debris was removed by centrifugation at 100g for 1 min, followed by nuclei pelleting from the supernatant at 500g for 5 min. Pelleted nuclei were washed once in the homogenization buffer before pelleting again at 500g for 5 min. Washed nuclei were re-suspended in 1x Nuclei buffer (10x Genomics) and filtered through a 40µm filter to remove the left-over debris. Nuclei concentration was estimated by counting nuclei on Countess II FL Automated Cell Counter (Thermo Fisher Scientific) with Hoechst DNA dye and propidium iodide for nuclei staining. Extracted nuclei were processed using Chromium Single Cell Multiome ATAC + Gene expression kit and Chromium Controller instrument (10x Genomics) as per manufacturer's recommendations. One sample, MB248, was processed with Chromium Next GEM Single Cell 3' v3.1 and ATAC v1.1 kits, as per manufacturer's recommendation. 15,000-20,000 nuclei were loaded per channel along with the Multiome/3'/ATAC gel bead. DNA and cDNA libraries were prepared as described in respective kit protocols. Libraries were quantified using Qubit Fluorometer (Thermo Fisher Scientific) and profiled using Fragment Analyzer. GEX and ATAC libraries were sequenced using NextSeq2000 to recommended lengths and depth. If the ATAC library was not of good quality, we still used the obtained RNA-Seq library if that was found to be of sufficient quality. RNA-Seq and ATAC-Seq datasets were further analyzed separately.

Single-nucleus RNA sequencing (snRNA-Seq) data processing

De-multiplexed reads were aligned to human genome assembly GRCh38 (v. p13, release 37, gencodegenes.org). Genome version associated comprehensive gene annotation (PRI) was customized by filtering to transcripts with the following biotypes: protein coding, lncRNA, IG and TR gene and pseudogene as recommended by *cellranger mkgtf* wrapper. Reads were aligned using STARsolo (*31*) with parameters: --soloType *CB_UMI_Simple* --soloFeatures *Gene GeneFull* --soloUMIfiltering *MultiGeneUMI* --soloCBmatchWLtype *1MM_multi_pseudocounts* --soloCellFilter *None* --outSAMmultNmax *1* --limitSjdbInsertNsj *1500000*. For overlapping genes where intronic alignment recovered low counts, exonic alignment counts were used. Predicted cells were separated from debris using *diem* pipeline (*R*) (*32*). Cells with mitochondria fraction > 1 median absolute deviation (MAD) above the mean or above 2% (whichever is greater), and number of detected genes greater than 6600 were filtered out. We further removed cells with an intronic fraction (number of reads aligned to intron/total number of reads aligned to exon+intron) less than 25%. Filtered cells were then corrected for background signature using *SoupX* (*R*) (*33*) and *celda* (*decontXcounts()*, *R*) (*34*) pipeline. Finally, putative doublets identified by *scrublet* (*Python*) (*35*) for snRNA-Seq data and those identified from snATAC-Seq data (see below, *Single-nucleus ATAC sequencing data processing*) were removed. Filtered gene expression matrices were normalized using the *scran* (*R*) (*36*) approach. A list of 1,500 highly variable genes (HVG) per sample was also obtained after removing mitochondrial (prefix: MT-) and ribosomal genes (prefixes: RPS, RPL, MRPS, MRPL). HVG from all the samples were combined, and sex-chromosome-specific genes (chr X and Y) were further removed to obtain a set of combined sample HVG gene-set for the single-cell cohort. Post-identification of "normal" cells (described below, Single-cell annotation), a list of 1,500 HVG was re-calculated from each sample and a combined tumor HVG gene-set was obtained from their union after filter sex chromosome specific genes.

Tumor single-cell annotation

We used a published single-nucleus developing human cerebellum atlas (*20*) as a reference to identify putative cell-identities of each tumor cell, particularly to identify non-tumor cells, such as endothelial, immune or glial cell-types. Normalized gene expression matrices from reference and target (tumor samples) were subsetted to the intersection of HVGs (5,000 genes from reference, combined sample HVG from single-cell tumor data) and cosine scaled (*cosineNorm()*, *batchelor*, *R*) (*37*). A *LinearSVC* model (*sklearn.svm, Python*) was first calibrated using *CalibratedClassifierCV(method='isotonic')* (*sklearn.calibration, Python*) using the reference data and then the fitted model was used to assign best matching cell identities to tumor cells. Cells that were identified as immune, mural/endothelial, astrocytes or oligodendrocytes were assigned as "normal" cells. Additionally, cells identified as cerebellar granule neurons (GC-defined) but appeared as a distant cluster on UMAP, separated from the bulk of tumor cells, were also assigned as "normal". These normal cells were removed for the integrated tumor data analysis.

Integration of snRNA-Seq data

We integrated all tumor samples together with and without batch-correction (across tumor samples) using LIGER (*R*) (*38*). Normalized gene-expression matrices from individual samples were subsetted to the combined sample HVG set, followed by cosine scaling. The scaled expression matrices were then used as an input for integrative NMF factorization using the function optimizeALS*(k=50, max.iters=100000)*. The obtained factors were then batch corrected using the fastMNN approach (*reducedMNN()*, *batchelor*, *R*). Corrected and uncorrected factors were used to obtain UMAP embedding of the integrated snRNA-Seq data. The batch corrected factors were further used to cluster cells using KNN (*sklearn.neighbors*, *kneighbors_graph(n_neighbors=11, metric='cosine', include_self=True), Python*) and leiden clustering (*leidenalg*, *lfind_partition*(), *Python*).

Single-nucleus ATAC sequencing (snATAC-Seq) data processing

ATAC-Seq reads were aligned to GRCh38 using Cellranger's *cellranger arc* wrapper and processed downstream using *ArchR* (*R*) (*39*). Briefly, fragment files obtained post alignment were converted into arrow files (*createArrowFiles()*) using custom gene annotation (same annotation as used for snRNA-Seq analysis) with a cut-off Transcription Start Site (TSS) enrichment of 3 and minimum 3000 fragments per cell. Putative doublets were identified by calculating a doublet score per cell (*addDoubletScores()*) and filterRatio of 1 (*filterDoublets()),* and were removed along with doublets identified in the snRNA-Seq processing. Cells with high fragment counts, 2x MAD above mean, were further removed. Filtered cells were then clustered and a final QC was done by removing clusters that exhibited comparatively low TSS enrichment and number of fragments per cell, along with lack of enrichment of known marker genes, obtained from the integrated snRNA-Seq data analysis. Cell clusters were also assigned putative "normal" identity if they were enriched for markers for immune, mural/endothelial, astrocyte or oligodendrocyte lineage, based on predicted gene-scores.

Integrating snRNA-Seq and snATAC-Seq data

Out of the 38 samples in the single-cell cohort, 32 were obtained from the multi-omic approach, with only a single tumor sample, MB248, that had snRNA-Seq and snATAC-Seq data from separate experiments. From here onwards, we only used tumor cell data in snATAC-Seq and hence any cell identified as "normal" based on snRNA-Seq or snATAC-Seq processing were removed. For multi-omics data, the majority of cells had both snRNA-Seq and snATAC-Seq data, but as per-sample snRNA-Seq and snATAC-Seq data was processed separately, variable number of cells were obtained per sample that passed QC parameters in one modality (snRNA-Seq or snATAC-Seq) but not in the other. To maximize data for downstream processing, we did not remove these cells from either data set, snRNA-Seq or snATAC-Seq, but imputed the missing RNA counts (normalized logcounts) for cells in the snATAC-Seq data of the same sample. Before imputing, snATAC-Seq data clusters that had RNA counts for less than 50% of cells or total number of cells with RNA counts was less than 100 were removed due to lack of a proper reference in these clusters. The imputed RNA count was then obtained from a weighted sum of normalized logcounts of 5 nearest neighbors (*sklearn.neighbors.NearestNeighbors()*). For sample MB248, snRNA-Seq and snATAC-Seq data were integrated using *addGeneIntegrationMatrix()* (*ArchR*).

Post integration, a joint dimensionality reduction of snRNA-Seq and snATAC-Seq data was obtained per sample. Using *addCombinedDims()* (*ArchR*), we combined Latent Semantic Indexing (LSI) based factorization of snATAC-Seq data to singular value decomposition (SVD) based factorization of snRNA-Seq data, excluding dimensions that had a correlation of greater than 0.75 to sequencing depth. The joint dimensional reduction was used to identify clustering (referred to as *combined_cluster*) and UMAP representation of the combined ATAC-RNA data.

Per sample peak calling in the snATAC-Seq data

Peaks were called per-sample on the tumor cells grouped by *combined_cluster* annotation. First a minimum of 40 cells and a maximum of 500 cells per group, with a sampling ratio of 0.8, were used to generate pseudobulk replicates via *addGroupCoverages()*. Then peaks were identified using MACS2 caller with a reproducibility of 2 via *addReproduciblePeakSet()*. Rest of the parameters used were defaults as defined in the function definition.

Creating a cisTopic object per sample

In order to prepare data for *SCENIC+* pipeline (*Python*) (*15*), the "peaks by cells" matrix (referred to as peak matrix here onwards) obtained from the *ArchR* analysis was converted to cisTopic object (*pycisTopic*, *Python*) to obtain topics and differentially accessible regions (DARs), which represent candidate enhancers for *SCENIC+* analysis. Peak matrix was reduced to 50 topics (*run_cgs_models(), pycisTopic*), obtained topics were binarized into region sets by '*otsu'* method and selection of top 3,000 regions per topic. DARs were identified by first identifying highly variable features (HVF), based on the log-normalized peak matrix, and then identifying marker regions using a cut-off adjusted p-value less than 0.05 and Log2FC greater than 0.5. If no marker regions were identified, then lower thresholds (Log2FC <0.1 and adjusted p-value <0.5) were used.

Creating motif-enrichment dictionary

Candidate enhancer regions identified from topic analysis and DARs were then assessed for motif-enrichment leading to creation of cistromes, an object associating transcription factors (TFs) to potential target regions. We used *run_pycisTarget()* wrapper from *SCENIC+*, along with motif-ranking, motif-score and motif-annotation provided by the Aertslab for GRCh38 (*15*) to obtain the TF-region cistromes per sample. Default settings were used for the function with the exception of *run_without_promoters* = *True*. Further, only TFs that were present in the combined tumor HVG set were selected for further processing.

Gene regulatory network identification

We used the *SCENIC+* approach for the multi-omic data to identify TF-associated gene regulatory networks (TF-GRNs) per sample. To identify tumor TF-GRNs, we first removed cells that were assigned as "normal" identity in snRNA-Seq or snATAC-Seq data processing. For each sample, we used snRNA-Seq data (after converting it into anData object), snATAC-Seq data (as cisTopic object) and motif-enrichment dictionary (obtained from pycisTarget) to create a *SCENIC+* object. Additionally, we provide a TF adjacency matrix with correlation values from a separate run of *pyscenic* (*Python*) (*16*) using *'genie3'* method (-m flag). *SCENIC+* first identified region-to-gene linkage for identified enhancers and their target genes and then assigned TF-to-gene links by associating TF that are enriched in the enhancers found linked to target genes. In the final step, *SCENIC+* uses region-to-gene and TF-to-gene links to identify regulons (TF-to-region-to-gene links) that are among the top ranked based on importance scores and assigns positive or negative regulatory relationships based on the correlation between the TF and assigned target gene. *SCENIC+* outputs a list of possible regulons with putative activation or repression relationships. For our analysis, we focused on positive TF-target interactions, represented as '+_+' in *SCENIC+*.

TF-GRNs selection and compilation

For each sample, a set of active TF-GRNs was identified using *SCENIC+* approach as described above. For each of the TF-GRNs, an "Area Under the Curve" (AUC)-based enrichment score (*AUCell_run()*, *AUCell*, *R*) (*16*) was calculated for all the tumor cells using log normalized RNA counts (including the imputed counts). From the identified TF-GRNs, GRNs associated with heterogeneity were identified based on the differential enrichment of TF-GRN AUC scores across combined_cluster annotation using Wilcox-rank test (*findmarkers(), scran*, *R*). The top three marker TF-GRNs per cluster per sample were used as representative of differentially active GRNs for that sample. After identifying such sets of TF-GRNs for each sample, we combined the obtained gene-sets as follows: 1) we selected TFs that were found to be associated with differentially active TF-GRNs in at least two samples, and then 2) for each of these selected TFs, we filtered target genes that were identified as linked to the TF in more than 20% of the samples where the TF was found to be active, with the association being present in at-least three samples. TF-GRN sets with sizes of less than 15 genes (including the TF) were also removed. In this way, we identified a conserved set of TF-gene links that were biologically replicated while reducing the number of associated genes by increasing the number of replicates required for the TFs that were widely used. This resulted in 108 TF-GRNs (Supplementary Table S4).

Integrating tumor RNA data across samples using TF-GRN enrichment scores

We obtained the AUC enrichment score for each of the

TF-GRN gene-sets (n=108) for all of the tumor cells using *AUCell_run(aucMaxRank=0.1*nGenes, normAUC=TRUE) (AUCell)*. The resulting enrichment score matrix was factorized using NMF (rank=25) and the obtained NMF factors were used for clustering (KNN-leiden) the integrated tumor data (resulting in 101 clusters), and obtained UMAP embedding and diffusion plots (*destiny*, *R*) (*40*). The TF-GRN AUC score matrix was scaled across cells and averaged per cluster to obtain the TF-GRN enrichment heatmap. The scaled TF-GRN matrix (clusters x TF-GRNs) was hierarchically clustered to obtain groups of co-enriched TF-GRNs (annotated as TF-GRN programs) and groups of tumor cluster exhibiting similar TF-GRN activity (annotated as tumor axes and cell-states). We also used *addmodulescore()* (*Seurat*, *R*) (*41*) to calculate activity scores for each of the identified 108 TF-GRN sets in the combined tumor cell data and used this score to calculate Pearson correlation between TF-GRNs.

Integrating snATAC-Seq data across samples

ArchR generated arrows files across tumors were merged to obtain a combined *ArchR* object. The merged *ArchR* object was factored using *addIterativeLSI(iterations=5, clusterParams = list(resolution = c(0.1, 0.2, 0.4, 0.8), sampleCells = 20000, n.start = 10), varFeatures = 100000, dimsToUse = 1:100, totalFeatures = 500000)* and obtained factors were used to calculate joint UMAP representation of the snATAC-Seq data. The merged ArchR object was then subsetted to tumor cells to identify peaks in the integrated data. Similar to peak identification in individual samples, first the integrated data was pseudobulked by tumor cell clusters (as identified in *Integrating tumor data using TF-GRN enrichment scores*) using *addGroupCoverages(maxCells = 1000, minReplicates = 5, maxReplicates = 15, maxFragments = 50 * 10^6)*. Peaks were called using *addReproduciblePeakSet(reproducibility = "2")*. Frequency of the identified peak's activity per tumor cluster was calculated by dividing the number of cells in a cluster in which the peak was detected by the total cluster population. Peaks that showed less than 3% frequency in all the tumor clusters were filtered out to obtain a robust peak set.

TF-GRN cis-regulatory elements (CREs) activity in tumor cells

Cis-regulatory elements (CREs) associated with a candidate TF and its identified target genes were combined to obtain a non-overlapping region set that defined the putative functional binding regions of that TF. For each TF-GRN, the obtained CREs were filtered to those CREs that overlapped with the above identified robust peak set (see *Integrating snATAC-Seq data across samples*), which together represented a pseudo-peak for that TF-GRN. A TF-GRN x tumor cluster pseudo-peak counts matrix was obtained by summing the peak counts of the associated CREs per tumor cluster. This matrix was divided by sum of column values, scaled to 10,000, and finally log2 transformed to obtain a normalized CRE activity matrix. The normalized CRE activity matrix was scaled across rows to obtain the CRE enrichment heatmap.

Subtype VII tumor sample trajectory analysis

TF-GRN AUC score for the integrated tumor data was subsetted by sample and used to obtain UMAP representation and diffusion map based pseudotime. Tumor cells with snATAC-Seq data were used to obtain PR_t to UBC_t trajectory using *addTrajectory()*(*ArchR*). The obtained trajectory was used to calculate TF-GRN and associated CRE enrichment signatures across pseudotime.

Weighted gene co-expression network analysis (WGCNA)

We used the combined logcounts and final annotation for the tumor data to identify a set of genes that showed axes or cell-state correlated activity using WGCNA (*R*) (*42*). A normalized gene expression matrix was subsetted to a combined tumor HVG set. For the WGCNA run, softPower was set to 9 and minimum module size was set to 20. A total of 24 modules were identified. Modules showing highest correlation with axial identities were selected as representative for the respective annotation.

In silico gene knock-out

CellOracle (*R*) (*28*) approach was used to perform *in silico* perturbations for *CRX* and *EOMES* in subtype VII tumor single-cell data. A sample-specific TF-GRN network, identified through the *SCENIC+* analysis, was provided as an input in the form of a TF-target dictionary. For each sample, raw gene expression counts, PCA calculated using *runPCA()* (*scran*), and TF-GRN AUC score-based diffusion pseudotime were used. Expression of *CRX/EOMES* was set to 0 to perform *in silico* loss-of-function analysis.

Bulk tumor RNA-Seq data processing

The bulk RNA-Seq data was collated from published studies for three cohorts: ICGC (*11-14*), MAGIC (*8*) and Newcastle (*10*). Except for the ICGC cohort, processed read count matrices were used for MAGIC and Newcastle samples. For samples belonging to the ICGC cohort, raw reads were aligned to human genome assembly GRCh38 (v. p13, release 37, gencodegenes.org), using STAR aligner (*43*). RNA-Seq samples belonging to individual cohorts were normalized separately using *DESeq2* (*R*) (*44*). Intersection of genes among the top 5,000 HVGs per cohort were used for subsetting data for NMF factorization. NMF factorization was performed using *sklearn.composition. NMF (init="nndsvd", max_iter=100000)*. NMF rank 2 and 8 were used to obtain subgroup- and subtype-associated latent factors or metagene signatures. To obtain the gene-set associated with each latent factor/metagene signature, the top 100 genes ranked by contribution to that factor were used. The obtained NMF latent factors (rank=8) were used for UMAP, tSNE, and Diffusion map projection of the bulk data. Differentially active genes in subtype I or VII tumors were obtained from pairwise comparison using *lfcShrink(type="ashr")* (*DeSeq2*).

Gene-set AUC scores for bulk RNA-Seq data

AUC enrichment scores of the TF-GRN gene-sets or WGCNA identified modules were calculated for each of the bulk tumor samples using *AUCell_run()*. AUC scores for tumor samples were scaled for each cohort (ICGC, MDT, and Newcastle) separately and then merged. Scaled TF-GRN AUC scores were used to obtain tSNE representation of the bulk-RNA-Seq tumor data on the TF-GRN enrichment space.

Human retina single-cell multi-omic atlas data processing

Processed filtered snRNA-Seq and snATAC-Seq data for the developing human retina were obtained from GSE183684 (*17*). snRNA-Seq and snATAC-Seq data were processed similar to tumor data. snRNA-Seq data was integrated together without batch-correction using NMF factorization (rank 25) and clustered using KNN-leiden approach. Obtained clusters were annotated based on marker gene expression (*17*). For *SCENIC+* analysis, snRNA-Seq data was converted to anDATA format, and snATAC-Seq data was converted into cisTopic format followed by processing with *pycisTarget* to obtain cistromes, as described for the tumor data. Processed data was then used as input for *SCENIC+* pipeline to obtain active regulons per sample. The top three TF-GRNs per "combined_cluster" for each of the samples were obtained based on differential AUC score enrichment (Wilcoxon test, *findMarkers*(), *scran*). TF-GRNs identified per sample were combined with a minimum requirement of the TF being associated with differentially expressed GRNs in at least two samples and the target gene being associated with the TF in at least 20% of the samples, with a minimum of three samples. Finally, TF-GRNs with a size of less than 15 genes were removed. Similar to the tumor data, the AUC score was calculated for each of the retinal lineage cells using *AUCell_run(aucMaxRank=0.1*nGenes, normAUC=TRUE).*

Integration of snRNA-Seq and snATAC-Seq data was obtained using scJoint (*45*). Normalized logcounts were used for snRNA-Seq and predicted gene scores (*addGeneScoreMatrix(), ArchR*) were used for snATAC-Seq. Gene expression matrices were subsetted to top 5000 HVGs across the integrated snRNA-Seq data excluding mitochondrial, ribosomal and sex chromosomal genes. scJoint based predicated labels for ATAC cells were used to annotate integrated snATAC-Seq data. For each of the ATAC cells, TF-GRN activity was imputed from the weighted sum of TF-GRN AUC score of the five nearest RNA cells obtained based on scJoint generated embedding for snRNA-Seq and snATAC-Seq data. Calculated TF-GRN AUC scores for snRNA-Seq data and imputed AUC scores for snATAC-Seq data were used to obtain joint representation of snRNA-Seq and snATAC-Seq data on the TF-GRN enrichment space. AUC scores from the snRNA-Seq data were used to obtain NMF model (rank=25), then the obtained model was used to factorize both the RNA-Seq and ATAC-Seq AUC score matrices. Post factorization, RNA and ATAC factor matrices were merged for a combined UMAP embedding.

Post-integration peak calling was done on the snATAC-Seq data by grouping cells based on scJoint predicted labels. A sample ratio of 0.8, 2 minimum replicates and 8 maximum replicates were used to obtain pseudo-bulks, followed by calling peaks by *MACS2* caller using a reproducibility of 2. A robust peak-set was obtained by removing peaks that were detected in less than 3% of cells in all the clusters.

Tumor TF-GRN gene-set enrichment score was obtained using *AUCell_run()* and retina logcounts gene expression matrix. TF-GRN AUC scores were scaled across cells and averaged by cluster to obtain a tumor TF-GRN gene-set enrichment heatmap.

Tumor TF-GRN CREs were intersected to obtain overlapping regions in the retinal robust peak matrix. Peak counts for all the CREs associated with each TF-GRN were summed to obtain TF-GRN by retina cluster matrix. The pseudo-bulked peak matrix was divided by column sums, scaled to 10,000, log2 transformed and finally scaled across clusters to obtain tumor TF-GRN CRE enrichment heatmap.

Deconvolution of bulk RNA-Seq tumor data

Bulk RNA-Seq data was deconvoluted using *BayesPrism* (*R*) (*46*) separately for each cohort (ICGC, MDT, Newcastle). Tumor data with cell-state annotation was combined with non-neuronal cells from single-nucleus human cerebellum data (*20*) to create the reference for deconvolution. An intersection of combined single-cell multi-omic atlas derived tumor HVG set and the top 7,500 HVGs from the bulk tumor cohort was used to subset the gene expression matrices of the reference and target data. Estimated proportion for each of the reference cell-state were obtained for each of the tumor sample, and combined estimate of the non-neuronal cells were removed to obtain the proportional composition of tumor cells in terms of the reference cell-states as annotated in the integrated Group 3/4 medulloblastoma atlas.

ChIP-Seq data analysis

Published H3K27Ac ChIP-Seq data (*11*) was aligned to GRCh38 using *bowtie2 (47)*. Duplicated, unmapped and multi-mapped reads were marked and removed using *sambamba*. Deduplicated alignment bam files were sorted using *sambamba* and indexed using *samtools*. Obtained alignment was normalized using *bamCoverage –normalizeUsing CPM –binSize 20 smoothLength 60 –extendedReads 150 –bl hg38.blacklist.v2.bed* (*deepTools*, *Python*) and converted into bigwig format. Enhancer signal for a subtype was obtained from averaged normalized signal of the constituting samples using *wiggletools*. *wigToBigWig* was used to convert obtained *Wig* files to bigwig and followed by conversion to *BedGraph* format using *bigWigtoBedGraph* tool. Bed files for human OTX2 (GSE137311), CRX (GSE137311) and EOMES (GSE26097) binding regions were obtained from Remap (https://remap2022.univ-amu.fr/). Track plots were prepared by *SparK* (https://github.com/harbourlab/SparK).

Identification of *ELP4-AS* and *ELP4-AS:IMMP1L*

Novel long non-coding RNA transcript, *ELP4-AS*, was identified using *StringTie* based *de novo* transcriptome assembly using the ICGC cohort RNA-Seq data. The spliced variant of *ELP4-AS* with downstream *IMMP1L* was identified using *Arriba* toolkit based on the RNA-Seq data (*48*). Presence of *ELP4-AS* and novel splicing transcript was confirmed by RT-qPCR in individual samples. Presence of fusions at genome level was also investigated using WGS data and SOPHIA algorithm (*49*).

Gene-set AUC scores in the unipolar brush cell (UBC) lineage

In published human cerebellar snRNA-Seq data (*20*), the UBC lineage was defined as composed of the following cell-types: rhombic-lip progenitor (RL progenitor), bi-potent GC/UBC progenitor (GCP/UBCP, annotated as UBCP in current study), differentiating UBC (UBC diff.) and defined (or differentiated) UBC (UBC def.). Normalized gene expression counts for the 5,835 cells representing the UBC lineage were extracted from the combined cerebellum atlas. Gene expression matrix was further subsetted to the top 1,000 HVGs and factorized using *optimizeALS*() (*LIGER*, rank 15). Obtained iNMF factors were batch corrected using *reducedMNN()* and obtained corrected-iNMF factors were used to generate the UMAP representation of the UBC lineage. Obtained UMAP factors were used to calculate *slingshot*() (*slingshot, R*) (*50*) based pseudo-temporal lineage order with RL progenitor as the starting point. Cells were binned into 100 distinct bins based on pseudotime. AUC score for the NMF metagenes, TF-GRNs and WGCNA was calculated for each of the cells in the UBC lineage and scaled across cells. Scaled gene-set scores were smoothened using *loess*() along pseudotime, averaged per bin, followed by scaling across bins.

Multiplexed single molecule *in situ* hybridisation (smFISH) data analysis

For 12-week post-conception human cerebellum spatial mapping was performed using published processed smFISH dataset generated using the Molecular Cartography (Resolve Biosciences) and smFISH probeset targeting 100 genes (*20*). The dataset contains information on segmentation as defined by Baysor (*51*), and independently imputed cell type and state/subtype labels together with their prediction scores as estimated by Tangram (*52*). For tumor data, tumor samples were processed using a tumor specific set of target genes (*53*). Tumor cell identities were imputed at the cell-state level using Tangram using sample-specific snRNA-Seq data as reference.