Article

Failure of human rhombic lip differentiation underlies medulloblastoma formation

https://doi.org/10.1038/s41586-022-05215-w

Received: 8 December 2021

Accepted: 9 August 2022

Published online: 21 September 2022

Check for updates

Medulloblastoma (MB) comprises a group of heterogeneous paediatric embryonal neoplasms of the hindbrain with strong links to early development of the hindbrain[1–4]. Mutations that activate Sonic hedgehog signalling lead to Sonic hedgehog MB in the upper rhombic lip (RL) granule cell lineage[5–8]. By contrast, mutations that activate WNT signalling lead to WNT MB in the lower RL[9,10]. However, little is known about the more commonly occurring group 4 (G4) MB, which is thought to arise in the unipolar brush cell lineage[3,4]. Here we demonstrate that somatic mutations that cause G4 MB converge on the core binding factor alpha (CBFA) complex and mutually exclusive alterations that affect *CBFA2T2*, *CBFA2T3*, *PRDM6*, *UTX* and *OTX2*. *CBFA2T2* is expressed early in the progenitor cells of the cerebellar RL subventricular zone in *Homo sapiens*, and G4 MB transcriptionally resembles these progenitors but are stalled in developmental time. Knockdown of *OTX2* in model systems relieves this differentiation blockade, which allows MB cells to spontaneously proceed along normal developmental differentiation trajectories. The specific nature of the split human RL, which is destined to generate most of the neurons in the human brain, and its high level of susceptible EOMES^{+}KI67^{+} unipolar brush cell progenitor cells probably predisposes our species to the development of G4 MB.

MB comprises a group of malignant paediatric cerebellar embryonal neoplasms with extensive intertumoural and intratumoural heterogeneity[1,2]. Mutations in genes in the Sonic hedgehog (SHH) signalling pathway lead to SHH MB in the granule cell lineage[5–8]. Meanwhile, mutational activation of WNT signalling leads to WNT MB in the lower RL[9,10]. Although group 4 MB (G4 MB, 40% of patients) and group 3 MB (G3 MB, 20% of patients) constitute the majority of MB cases, and the most deaths, far less is known about their specific cellular origin or somatic driver events. G4 MB is proposed to arise in the unipolar brush cell (UBC) lineage, whereas G3 MB appears to arise from an earlier population of stem cells[3,4]. Activation of MYC is seen in many G3 tumours; however, the convergence and/or significance of various low-incidence driver mutations in G4 MB remains unclear[11].

The developing *H. sapiens* RL displays specific features compared to other mammals such as mice and macaques[12]. At around 11 post-conception weeks (PCW), the human RL splits into two molecularly and structurally distinct zones—the ventricular RL (RL^{VZ}) and the subventricular RL (RL^{SVZ})—that are separated by a vascular plexus[12]. The RL^{VZ} and RL^{SVZ} are transcriptionally distinct[13], with the RL^{VZ} primarily composed of stem cells and the RL^{SVZ} primarily composed of proliferative progenitor cells. Expression of the classical UBC marker gene *EOMES* (also known as *TBR2*) in the RL^{SVZ} implies human-specific aspects of UBC development[14]. The human-specific features of RL development necessitate analyses of the developing *H. sapiens* cerebellum to determine the true cellular origins of G4 MB.

We report new G4 MB mutations in genes that encode CBFA family proteins, including *RUNX1T1* (also known as *CBFA2T1*), *CBFA2T2* and *CBFA2T3*. These adaptor proteins form a large complex that recruits epigenetic modifiers and transcription factors (TFs) to chromatin[15–19]. Apical G4 MB tumour cells resemble a specific, human-expanded, EOMES^{+}KI67^{+} UBC progenitor cell population of the RL^{SVZ}, where members of the CBFA family are specifically expressed during human cerebellar development. We propose that the CBFA complex potentiates normal differentiation of EOMES^{+}KI67^{+} RL^{SVZ} progenitors, the failure of which results in G4 MB.

The mutational landscape of G3 and G4 MB

We transcriptionally profiled G3 and G4 MB samples (n = 545) using bulk RNA sequencing (RNA-seq)[2,11,20] (Fig. 1a). Recurrently mutated genes converged on epigenetic modifiers, the cell cycle and four gene families: *ELP*[21], *FANC*[11], *CHD*[22] and members of the CBFA polyprotein complex. Genes from these four gene families are proximally clustered on human chromosomes, particularly within regions affected by copy number aberrations (CNAs) in G4 MB (Extended Data Fig. 1a–e). The combined deletion of multiple physically proximate drivers suggested a reason for specific chromosome arm deletions in G4 MB (Extended Data Fig. 2a–c). We identified new G4 MB alterations that targeted the CBFA complex, including the histone methyltransferase *PRDM6* (ref. [16]), the histone demethylases *KDM6A* and *KDM2B* and the transcriptional corepressors *RUNX1T1*, *CBFA2T2* and *CBFA2T3*. This is consistent with a model in which CBFA driver mutations result in the failed differentiation of G4-MB-initiating progenitor cells, which accumulate to form G4 MB, an embryological remnant.

CBFA2T2 and *CBFA2T3* alterations are enriched in G4 MB

G4-MB-specific loss-of-function mutations that target *CBFA2T2* (Fig. 1b) were enriched in the NHR1 domain (Fig. 1c), which interacts with the SET and PR domains of PRDM proteins, including PRDM6 (ref. [16]). *CBFA2T2* mutations tended to occur independently of high

A list of authors and their affiliations appears online. ✉e-mail: mdt.cns@gmail.com

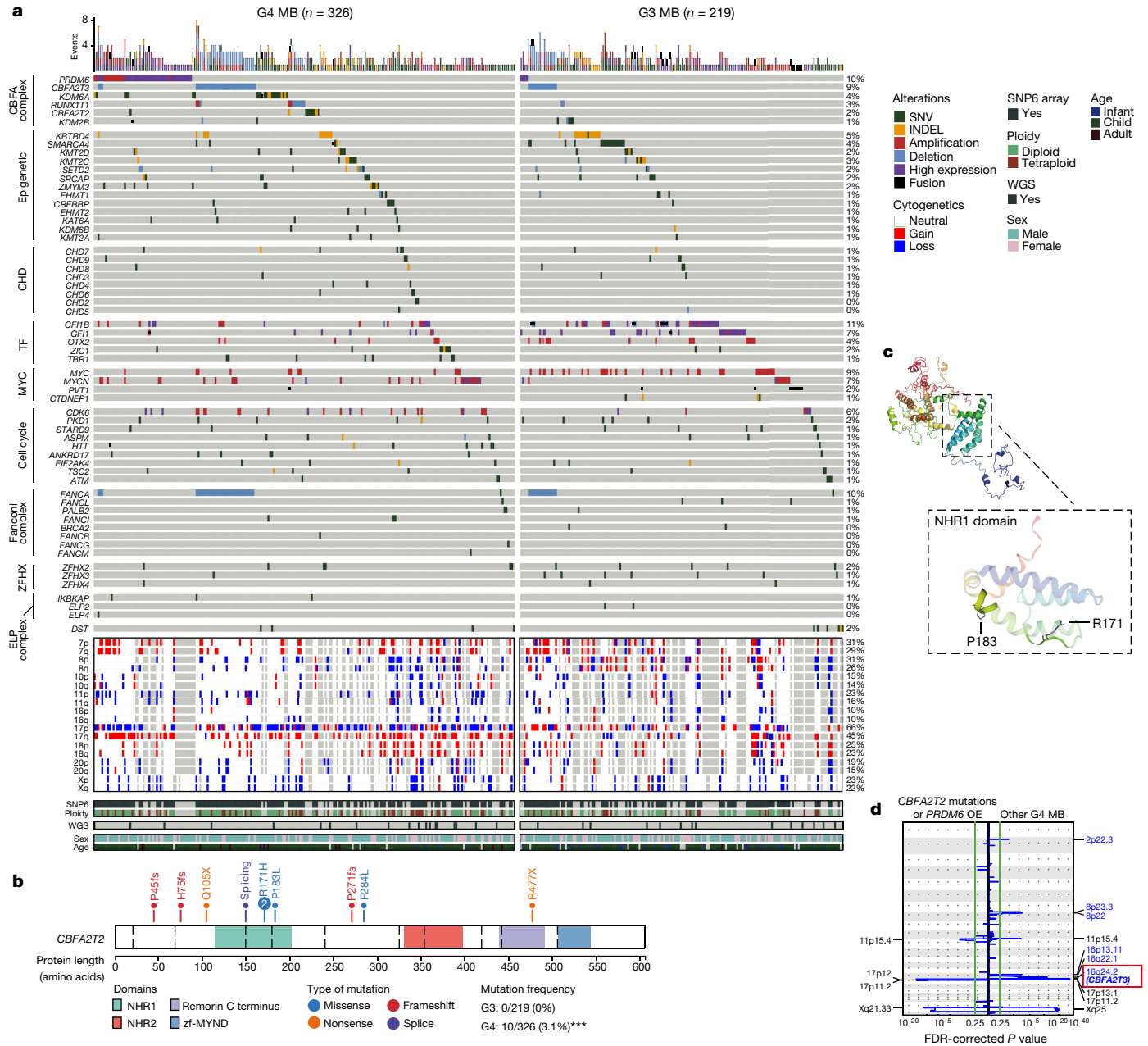


Fig. 1 | The landscape of oncogenic drivers in G3 and G4 MB. a, Oncoprint summary of mutations, CNAs, gene expression and gene fusions in G3 and G4 MB ($n = 545$ tumours). A total of 396 (73%) samples are shown; 173 samples without recurrent alterations detected are not shown. Percentage values (right) indicate the percentage of patients with the gene alteration. INDEL, insertion or deletion. **b**, Gene-level summary of *CBFA2T2* mutations in G4 MB. The R171H mutation was found in two patient tumours. **c**, Structural model of the NHR1 domain of CBFA2T2 protein, highlighting positions affected by missense mutations. The structure of the NHR1 domain has been previously determined (bottom), whereas the full protein structure was inferred using iTasser[39] (top). **d**, Comparison of significant focal deletions in $n = 206$ G4 MB, either with *CBFA2T2* mutations or *PRDM6* overexpression (OE), compared with tumours without *CBFA2T2* or *PRDM6* abnormalities. Significance was assessed using GISTIC 2.0 (ref.[40]) (false discovery rate (FDR) < 0.25).

PRDM6 expression, which indicated that there is a complementation group (Extended Data Fig. 3a). To uncover other members of this complementation group, we compared CNAs between G4 MB samples with *CBFA2T2* or *PRDM6* alterations, or neither (Fig. 1d). Focal chromosome 16q24 (*CBFA2T3*) deletions occurred in a mutually exclusive pattern with *CBFA2T2* and *PRDM6* mutations (Fig. 1d and Extended Data Fig. 3b). They also drove significantly reduced *CBFA2T3* expression (Extended Data Fig. 3c) and universally spanned *CBFA2T3* (Extended Data Fig. 3d–g). Although 16q deletions are detected in both G3 and G4 MB, they are rarely seen in SHH MB[20,23]. Indeed, both *CBFA2T2* and *CBFA2T3* showed significantly higher expression in SHH MB, which indicates that the role of the CBFA complex may differ between the MB cells of origin (Extended Data Fig. 3h).

Chromosome 16q contains 3 additional G4 MB recurrently mutated genes (*FANCA*, *ZFHX3* and *CHD9*), and mutations in these genes were mutually exclusive with 16q deletions (Fig. 2a). Tight genomic clustering of these G4 MB driver genes along human chromosome 16 explains the bias towards large deletions of *CBFA2T3* and the relative absence of somatic single-nucleotide variants (SNVs) in *CBFA2T3* (Fig. 2b). *CBFA2T2* and *CBFA2T3* interact with *GFI1B* and other epigenetic modifiers[19,24], and G4 MB with *GFI1B* enhancer hijacking events were mutually exclusive (Fig. 2c). Taken together, we propose that *CBFA2T2*, *PRDM6*, *CBFA2T3*, *GFI1* and *GFI1B* represent a new G4 MB complementation group.

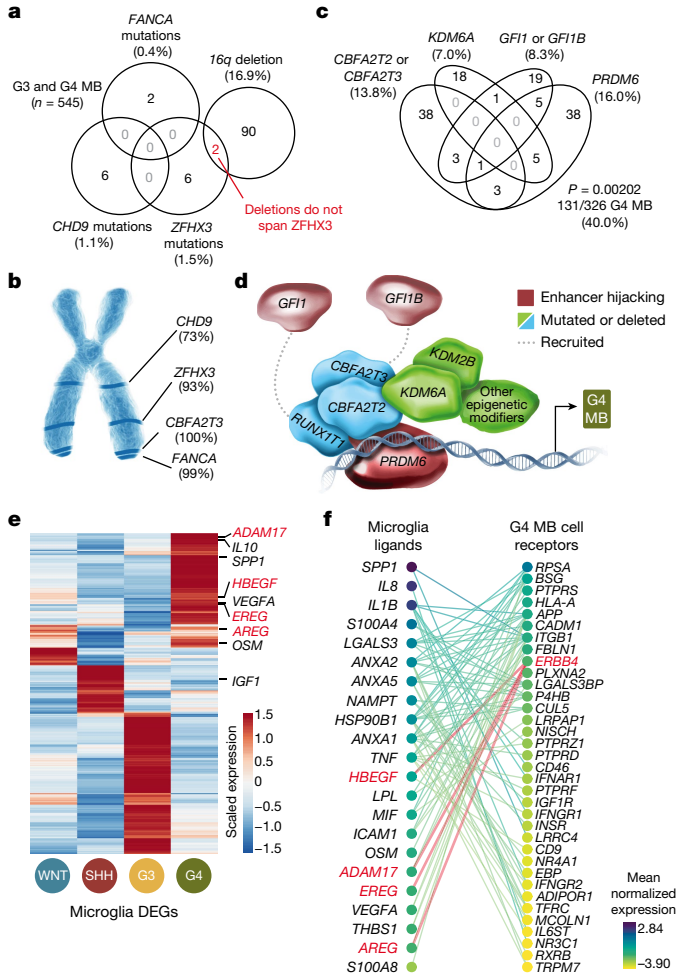


Fig. 2 | CBFA complex members are recurrently somatically mutated in G4 MB. a, Mutual exclusivity of somatic SNVs in the 16q genes *FANCA*, *ZFHX3* and *CHD9*. Deletions of 16q do not typically co-occur with mutations in these genes. G3 and G4 MB prefer deletions to simultaneously disrupt several tumour suppressor genes (TSGs). **b**, Locations of G4 MB TSGs on chromosome 16 and percentage of samples in which TSGs are deleted to haploinsufficiency in G3 and G4 MB samples with 16q deletions. **c**, Mutual exclusivity of *PRDM6* overexpression, *CBFA2T2* mutations, *CBFA2T3* deletions, *KDM6A* mutations and *GFI1* or *GFI1B* enhancer hijacking in G4 MB. Significance was assessed using the impurity test for mutual exclusivity implemented by DISCOVER[41]. **d**, Cartoon of known or suspected CBFA complex members. **e**, Expression of significant differentially expressed genes (DEGs) between microglia from each subgroup by scRNA-seq (MAST[42], FDR < 0.05). ERBB4 receptor ligands are highlighted in red. **f**, Predicted receptor–ligand interactions between microglia and tumour cells in G4 MB (CCInx[43]). Node colours represent mean normalized gene expression in each cell type, and edge colour represents the average of the node expression levels. Only significant DEGs from **e** were included, which demonstrate the microenvironment specificity of G4 MB.

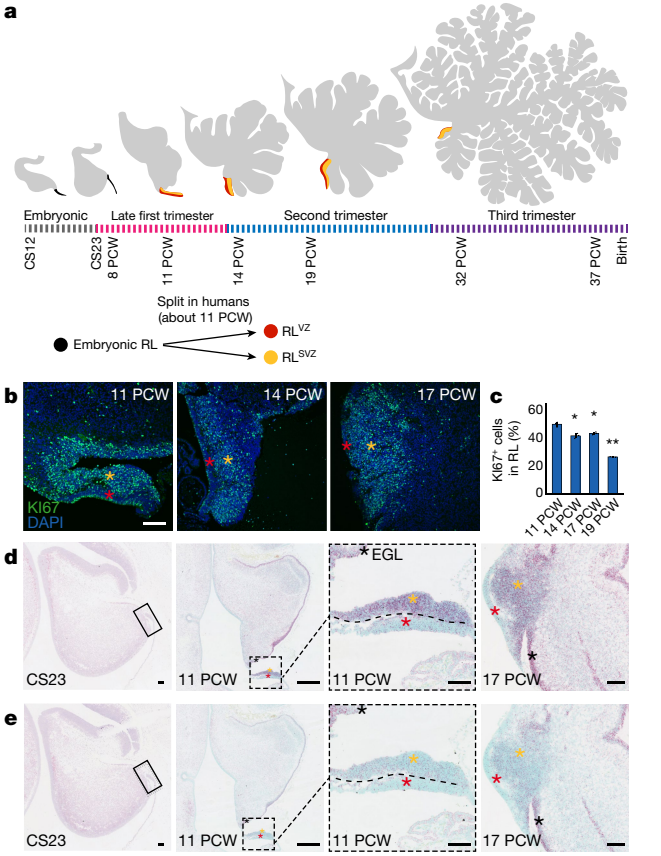


Fig. 3 | *CBFA2T2* and *CBFA2T3* define the RL VZ–SVZ boundary in the developing human cerebellum. a, Schematic summarizing *H. sapiens* cerebellar development. Before 11 PCW, the RL resembles that of *Mus musculus*. At around 11 PCW, the human RL splits into a ventricular and subventricular zone, which leads to human-specific cerebellar expansion. **b**, KI67 expression in the developing human RL at 11, 14 and 17 PCW. **c**, Per cent of human RL cells expressing KI67 across several developmental timepoints. Significance was assessed using paired two-tailed *t*-test compared with 11 PCW, $^{**}P < 0.005$, $^{*}P < 0.05$; 14 PCW, $P = 0.0072$; 17 PCW, $P = 0.012$; 19 PCW, $P = 0.0024$. $n = 3$ biological repeats per timepoint; error bars represent the s.e.m. **d**,**e**, In situ hybridization showing spatially resolved RNA expression of *CBFA2T2* (**d**) and *CBFA2T3* (**e**) in the developing human cerebellum at Carnegie stage 23 (CS23) and at 11 and 17 PCW. *CBFA2T2* and *CBFA2T3* expression is first observed at 11 PCW in the RL^{SVZ} but not the RL^{VZ}. Dashed lines denote approximate location of vasculature bed. Data presented in **b** are representative images from three independent experiments with similar results. Data in **d** and **e** were not performed in replicates. RL^{VZ}, RL^{SVZ} and EGL are denoted by red, yellow and black asterisks, respectively. Scale bars, 100 µm.

CBFA complex disruptions underlie G4 MB

As *CBFA2T2*, *PRDM6* and *CBFA2T3* form a complementation group, and are known to physically interact[16,19], we propose that the polyprotein CBFA complex contains additional G4 MB drivers. We performed in vitro TurboID[25] in the G3 MB cell line HDMB03 (Extended Data Fig. 4a,b), in which *MYC* is amplified. CBFA2T2 interacted with KDM6A, a known G4 MB driver gene (Extended Data Fig. 4c,d). Other notable interactors included RUNX1T1, KDM2B and SMARCA1. CBFA2T2 also interacted with GLI2, a recurrently amplified SHH MB oncogene. We combined our new CBFA2T2 protein interactions with known interactions between G3 and G4 MB driver genes (Extended Data Fig. 4e). Most G4 MB driver genes were within two steps of CBFA2T2, which suggests that they might be part of the CBFA complex. Alterations that disrupted the CBFA complex were found in at least 57% of G4 MB samples and 39% of G3 MB samples (Fig. 2d and Extended Data Fig. 5a).

A subgroup-specific analysis of microglial expression differences demonstrated stark variances in subgroup-specific microenvironments (Fig. 2e). G4 MB microglia, but not other subgroups, expressed the ERBB4 ligands *HBEGF* and *EREG* (Fig. 2f). Tumours that expressed high levels of ERBB4 ligands were less likely to have mutations in the CBFA complex (Extended Data Fig. 5b). These data provide a possible explanation for activation of ERBB4 in G4 MB, as previously demonstrated through unbiased proteomics[26]. As ERBB4 activation has been suggested to inhibit the activity of CBFA2T3 (ref. [27]), it appears that a subset of G4 MB tumours receive important and sustaining signals from their microenvironment that might be targetable for intervention.

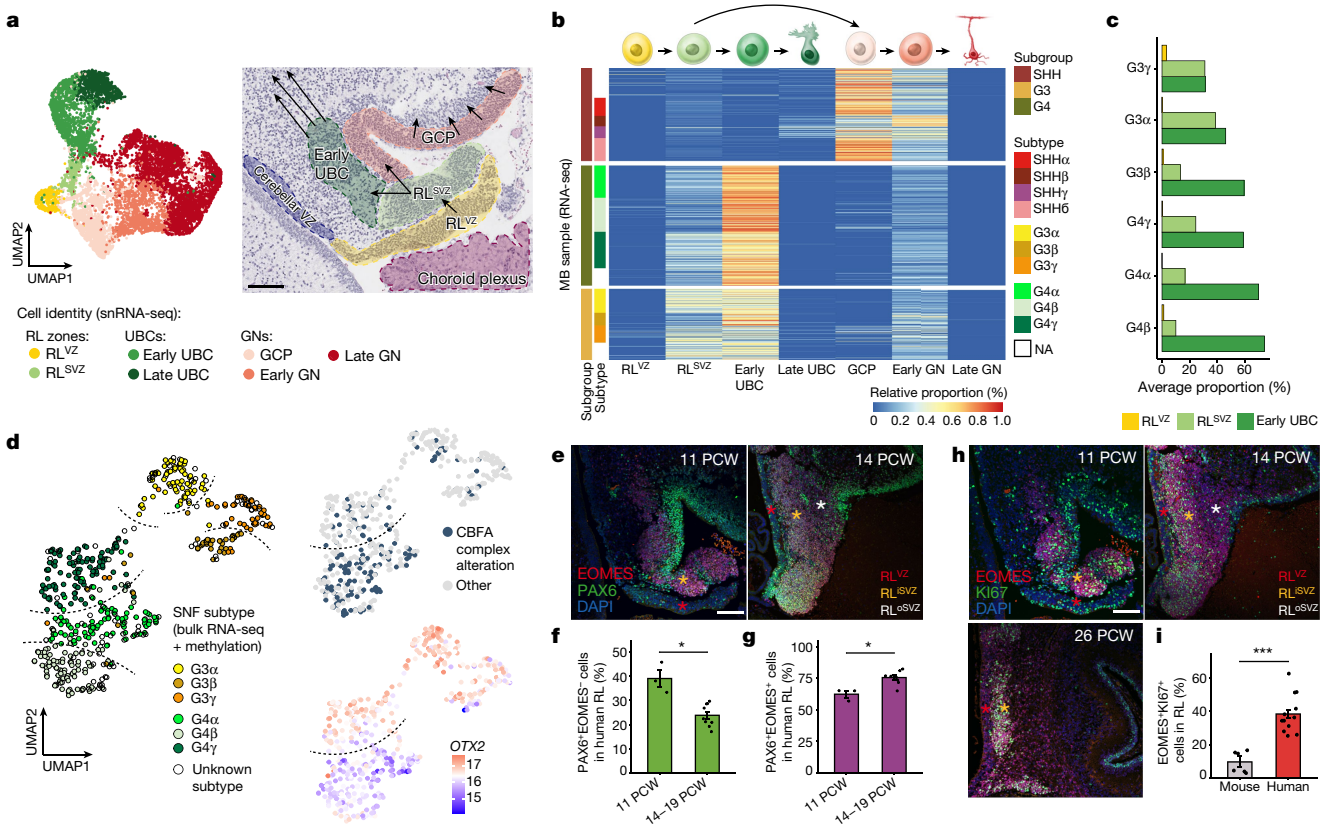


Fig. 4 | The expanded human RL is predisposed to G3 and G4 MB. a, Left, uniform manifold approximation and projection (UMAP) embedding of 9,208 single cells from the developing human cerebellum[13]. Right, Mid-sagittal H&E-stained section from a human RL cerebellum at 11 PCW highlighting the expected distribution of indicated cell types. **b**, Deconvolution of RNA-seq expression profiles from G3 ($n = 219$), G4 ($n = 326$) and SHH ($n = 250$) MB by cell types in **a**. **c**, Deconvolution proportions along the RL to UBC lineage in G3 and G4 MB subtypes. **d**, UMAP embedding of G3 and G4 MB RNA-seq samples ($n = 545$) showing subtype, *OTX2* expression and CBFA complex alterations. **e**, EOMES and PAX6 expression in human RL at 11 and 14 PCW. **f**,**g**, Per cent of RL cells expressing PAX6 only (**f**, $^*P = 0.036$) or both PAX6 and EOMES (**g**, $^*P = 0.017$). Significance assessed using unpaired two-tailed *t*-tests; $n = 3$ biological repeats per timepoint; error bars represent the s.e.m. **h**, EOMES and KI67 expression at 11, 14 and 26 PCW. **i**, EOMES$^+$KI67$^+$ population in humans and mice across timepoints. Significance assessed by two-sided Mann–Whitney *U*-test, $^{***}P = 0.000108$. $n = 3$ biological repeats per timepoint; error bars represent the s.e.m. Data presented in **e** and **h** (11 and 14 PCW) are representative images from three independent experiments with similar results. Data in **a** and **h** (26 PCW) were not performed in replicates. The RL^{VZ}, inner subventricular (RL^{iSVZ}) and outer subventricular (RL^{oSVZ}) zones are indicated with red, yellow and white asterisks, respectively. Scale bars, 100 µm.

Collectively, we propose a model in which driver genes of G3 and G4 MB converge to inhibit the physiological CBFA complex through various mechanisms (Extended Data Fig. 5c).

CBFA2T2 demarcates the human RL^{SVZ}

The role of the CBFA complex in cerebellar development is unknown. In humans, the RL is long-lived and seen throughout gestation and it undergoes a series of morphological and structural changes (Fig. 3a and Extended Data Figs. 5d and 6a). Notably, the human RL develops a substructure and expands into two distinct zones at 11 PCW: the RL^{VZ} and the RL^{SVZ} (Fig. 3a and Extended Data Fig. 5d). These zones are separated by a vascular plexus, which expresses *HBEGF* (Extended Data Fig. 6b,c). The RL^{VZ} was enriched for SOX2, whereas the RL^{SVZ} demonstrated strong cell division typical of progenitor cells (Fig. 3b). The developing RL became decreasingly proliferative with time (Fig. 3b,c and Extended Data Fig. 6d) until it disappeared around birth (Extended Data Fig. 6e).

CBFA2T2 was specifically expressed in the RL^{SVZ} starting at 11 PCW (Fig. 3d). At 14 and 17 PCW, *CBFA2T2* and *CBFA2T3* remain expressed in the RL^{SVZ} but not the RL^{VZ} (Fig. 3d,e and Extended Data Fig. 6f,g). At each timepoint, *CBFA2T2* and *CBFA2T3* were also expressed in the external granule layer (EGL), where *Cbfa2t2* and *Cbfa2t3* expression is observed in mice (Extended Data Fig. 6h,i). The EGL is composed of granule cell precursors (GCPs), the cell of origin for SHH MB. By contrast, the expression of both genes decreased along the ventral exit from the RL, where differentiating *LMX1A*$^+$ UBCs migrated away from the proliferative RL (Extended Data Fig. 6j). *CBFA2T2* and *CBFA2T3* expression was retained by GCPs (and SHH MB), but not UBCs, which suggests that these genes have a role in this cell fate decision during normal development.

G3 and G4 MB mirror embryonic human RL

CBFA2T2 expression in RL^{SVZ} progenitor cells indicates that G4 MB might arise from the RL^{SVZ} because of dysfunctional *CBFA2T2*. An analysis of single-nucleus RNA-seq (snRNA-seq) data from the developing human cerebellum[13] demonstrated the expected cell types and lineages descending from the apical RL^{VZ} (Fig. 4a and Extended Data Fig. 7a,b). *CBFA2T2* and *CBFA2T3* were highly expressed in the RL^{SVZ} and in GCP clusters (Extended Data Fig. 7a,c).

We performed single-cell RNA-seq (scRNA-seq) and integrative clustering on G3 MB (five patient samples), G4 MB (11 patient samples) and SHH MB (three patient samples) (Extended Data Fig. 7d). Non-tumour cells were identified through the expression of known marker genes and a paucity of CNA detection and were excluded from further analysis (Extended Data Fig. 7e–g). Using two distinct methods, a transcriptional comparison between the developing human cerebellum and MB cells

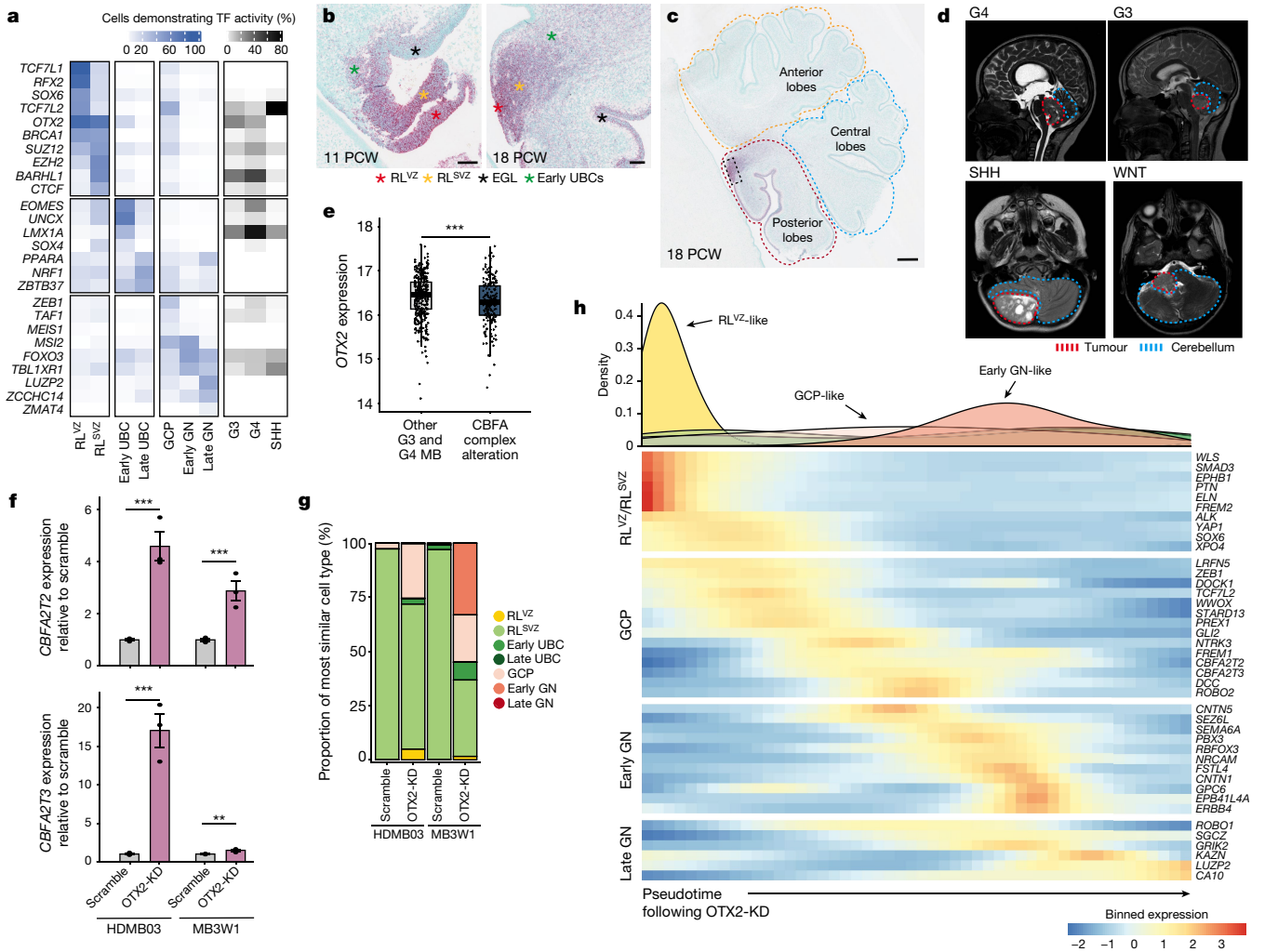


Fig. 5 | *OTX2* restrains RL differentiation through inhibition of the CBFA complex. a, Percentage of cells in normal cell types and MB cells that exhibit predicted TF activity (SCENIC[44]). **b**, *OTX2* expression in the developing human RL at 11 and 18 PCW. Scale bars, 100 µm. **c**, *OTX2* expression at 18 PCW. High expression is observed in the RL (black box) and the posterior lobes. Scale bar, 500 µm. **d**, Representative T1 enhanced or T2 MRI scans showing typical locations of each MB subgroup at initial diagnosis. Mid-sagittal scans are shown for G3 and G4 MB, axial scans for SHH and WNT MB tumours. **e**, *OTX2* expression in G3 and G4 MB samples by CBFA complex alterations (Fig. 4d). Significance was assessed using two-sided Mann–Whitney *U*-test, $^{***}P = 0.0000162$. $n = 545$ MB samples. Box plots show the median and interquartile range, and whiskers show the data range. Points outside this range are outliers and are plotted individually. **f**, *CBFA2T2* and *CBFA2T3* expression following OTX2-KD in G3 MB tumourspheres. Data shown as mean fold change ± s.e.m. compared to scramble, with $n = 3$ biological replicates. Significance was assessed using DESeq2 (ref. [45]) (FDR < 0.05), $^{***}P < 0.0005$, $^{**}P < 0.005$. *CBFA2T2*: HDMB03, $P = 1.20 \times 10^{-30}$; MB3W1, $P = 4.45 \times 10^{-15}$. *CBFA2T3*: HDMB03, $P = 2.90 \times 10^{-91}$; MB3W1, $P = 0.0055$. **g**, Single cells show increased similarity to normal differentiated cell types following OTX2-KD in G3 MB tumourspheres. **h**, Expression of GN lineage marker genes along pseudotime following OTX2-KD in MB3W1. Top, density of most similar developmental cell type (**g**) along pseudotime. Bottom, binned expression of GN differentiation genes (Fig. 4a). Data presented in **b** and **c** were not performed in replicates.

consistently revealed that apical G3 and G4 MB cells were most similar to the RL^{SVZ} (Fig. 4b and Extended Data Fig. 8a,b), whereas the more deadly G3 gamma subtype (G3γ)[2] displayed enrichment for the earlier RL^{VZ} (Fig. 4c and Extended Data Fig. 8c). SHH, G3 and G4 MB all displayed a differentiation block, with few cells mapping to late granule neurons (GNs) and late UBCs. G3 MB displayed the lowest similarity to normal cerebellar cells, overall (Extended Data Fig. 8d). Mutations in the CBFA complex were highly enriched in the *OTX2*-depleted subtypes of G4 MB (G4α and G4β), but not commonly found in G4γ, in which *OTX2* levels were high and *KBTBD4* mutations were found (Fig. 4d and Extended Data Fig. 8e). Whether the transcriptional differences between G4α, G4β and G4γ are due to them arising from discrete developmental cell states, compared to the effects of specific somatic mutations, will require additional consideration. *BARHL1* and *DDX31*, for which enhancers are known to drive aberrant *GFI1B* expression through enhancer hijacking[28], were expressed only in the RL^{SVZ}, whereas *GFI1B* itself was not expressed (Extended Data Fig. 8f–k). These data are consistent with a model in which G4 MB and some G3 MB tumours arise in the RL^{SVZ} owing to the specific human RL split.

Human-evolved predisposition to G3 or G4 MB

The RL^{SVZ} is composed of two distinct populations: nascent GCPs, which express PAX6, but not EOMES; and nascent UBCs, which express both PAX6 and EOMES (Fig. 4e and Extended Data Fig. 9a). At around 11 PCW, PAX6^{+}EOMES^{-} cells form a stream of cells between pockets of EOMES^{+} cells, which connect the tail-like RL with the EGL. However, this conspicuous subcompartmentalization was short-lived and no longer visible following RL internalization at 14 PCW. The frequency of PAX6^{+}EOMES^{-} cells decreased after 11 PCW, which suggests that the

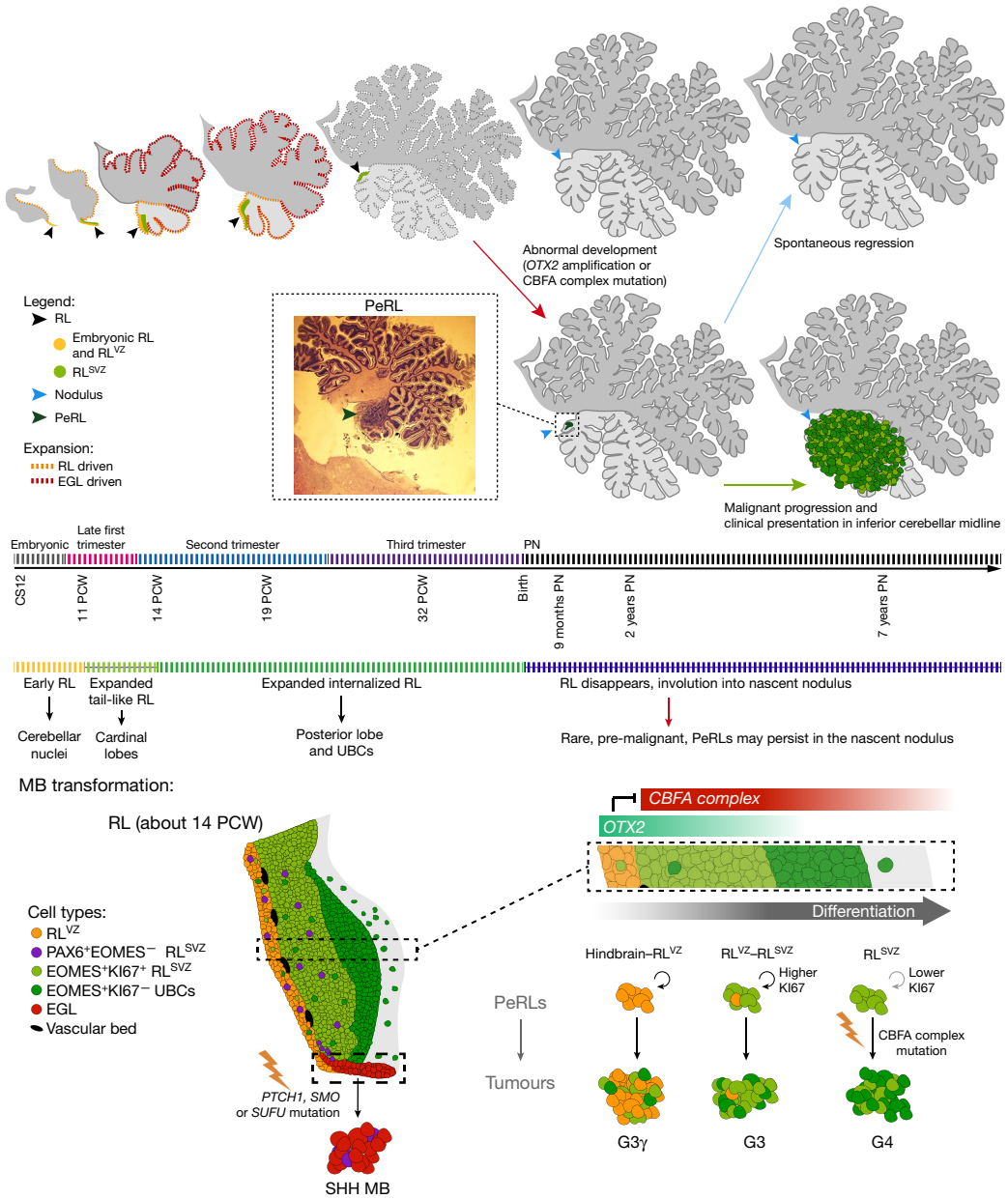


Fig. 6 | G3 and G4 MB initiate in utero from failed differentiation of components of the human RL. Proposed model of the initiation of SHH, G3 and G4 MB. Top, normal human cerebellum development is marked by changing roles and shrinking of the RL until it dissipates within the nodulus around or shortly after birth. Rarely, this process can be disrupted, leaving residual disorganized RL termed PeRL, which appears as a non-neoplastic malformation (dysplasia) in the nodulus. Bottom, following the split of the human RL around 11 PCW, EOMES$^+$KI67$^+$ RL^{SVZ} progenitors are generated, which, in the presence of chromosomal instability and/or driver alterations, fail to differentiate properly and give rise to PeRLs. In the presence of further genetic insult, PeRLs give rise to G3 and G4 MB in the inferior cerebellum. PN, postnatal.

early tail-like RL produces more GCPs relative to later stages (Fig. 4f). We propose that GCPs produced by the tail-like RL build the nascent EGL of all three cardinal lobes, and following internalization at 14 PCW, the RL seeds only the posterior lobe[12,29,30]. Consistently, the RL^{SVZ} shifts towards UBC production, as shown by an increase in EOMES$^+$ cell numbers after 14 PCW (Fig. 4g).

EOMES is currently thought of as a marker of post-mitotic UBCs[31,32], and previous comparisons of G4 MB to the developing mouse cerebellum suggested that G4 MB arises in post-mitotic EOMES$^+$ UBCs[3,4]. However, in humans, EOMES$^+$ cells in the RL^{SVZ} are predominantly proliferative, as shown by labelling for both EOMES and KI67 (Fig. 4h and Extended Data Fig. 9b,c). Furthermore, we identified the following additional features in the RL specific to humans compared to mice: (1) further compartmentalization of the RL^{SVZ} into an inner SVZ that contained EOMES$^+$KI67$^+$ cells and an outer SVZ that contained differentiated nascent or early UBCs (Fig. 4h, top right); (2) temporal expansion of the EOMES$^+$KI67$^+$ population late into gestation (Fig. 4h, bottom left, and Extended Data Fig. 9d); and (3) significant human-specific expansion of EOMES$^+$KI67$^+$ progenitors compared to mice (Fig. 4i and Extended Data Fig. 9e–h). Thus, there is a spatiotemporally expanded pool of MB-susceptible EOMES$^+$KI67$^+$ UBC progenitors throughout human gestation in comparison to mice, thereby providing a statistically larger risk for G3 and G4 MB to arise (Extended Data Fig. 9i).

OTX2 inhibits *CBFA2T2* to retain the RL state

We computationally inferred the activity of TFs in both the developing human cerebellum and our scRNA-seq MB samples (Fig. 5a). TFs highly

active in G3 and G4 MB were also highly active in the RL^{SVZ}. OTX2 activity was particularly high in G3 and G4 MB and the RL. *OTX2* is frequently amplified in both G3 and G4 MB[33,34] (Fig. 1a). *OTX2* was highly expressed in the RL^{VZ} and the RL^{SVZ}, and preceded *CBFA2T2* and *CBFA2T3* expression (Fig. 5b and Extended Data Fig. 10a,b). *OTX2* was also specifically expressed in the posterior cerebellar lobes (Fig. 5c). Consistent with our proposed cell of origin, G3 and G4 MB—but not SHH or WNT MB—always presented on the inferior surface of the cerebellum (Fig. 5d and Extended Data Fig. 10c–f). At 14 PCW, the RL was internalized into the nascent nodulus of the cerebellar flocculonodular lobe (Extended Data Fig. 6a). The location of the embryological remnant of the internalized RL explains the near universal localization of G3 and G4 MB in the inferior midline of the cerebellum (Fig. 5d). *OTX2* was highly expressed in G3 MB and G4γ, whereas *CBFA2T2*, *CBFA2T3*, *KDM6A*, *PRDM6* and other alterations to the CBFA complex tended to occur in $OTX2^{low}$ G4α and G4β MB (Fig. 4d). Indeed, *OTX2* expression was significantly reduced in G3 and G4 MB with alterations in the CBFA complex (Fig. 5e). This result suggests that either *OTX2* overexpression or CBFA complex dysfunction can impair RL^{SVZ} differentiation.

OTX2 bound the *CBFA2T2* but not the *CBFA2T3* promoter, suggesting transcriptional control of the CBFA complex (Extended Data Fig. 11a). We performed siRNA-mediated *OTX2* knockdown (OTX2-KD) on tumourspheres from two G3 MB lines harbouring *MYC* amplifications, which were profiled by bulk and snRNA-seq (Extended Data Fig. 11b–d and Supplementary Fig. 1). Expression of both *CBFA2T2* and *CBFA2T3* were significantly upregulated following OTX2-KD (Fig. 5f). Analysis of single OTX2-KD cells revealed a large increase in similarity to more differentiated cell types, which was dominated by GCPs and GNs (Fig. 5g). Pseudotime mapping of OTX2-KD cells revealed a sequential differentiation trajectory that mirrored normal GN development (Fig. 5h, Extended Data Fig. 11e–j and Supplementary Fig. 1). *CBFA2T2* and *CBFA2T3* were among the first genes to show increased expression following OTX2-KD. They were also highly expressed in GCP-like cells and preceded GN-like cells. This result provides further evidence for a role of OTX2 in RL^{SVZ} fate decisions, GN differentiation or both (Fig. 5h and Extended Data Fig. 11k). Previous studies have demonstrated that injection of significantly more OTX2-KD cells was required to generate G3 MB tumours in vivo, which may reflect the depletion of tumour-initiating cells through induced GN differentiation[33]. Overexpression of CBFA2T2 significantly reduced the number of live cells in tumoursphere culture, without affecting viability. This result provides evidence for a role for CBFA2T2 in the promotion of differentiation in G3 and G4 MB (Extended Data Fig. 11l–p and Supplementary Fig. 1).

Collectively, these data support a model in which *OTX2* maintains RL identity by inhibiting the CBFA complex until cells exit the RL and differentiate. Disruptions to the CBFA complex, or overexpression of *OTX2*, according to this model, result in failed RL^{SVZ} differentiation. The resulting ball of RL^{SVZ} progenitor cells are retained with the RL in the nascent nodulus of the developing cerebellum, where ongoing mitotic activity eventually results in a mass lesion diagnosed as G3 or G4 MB (Extended Data Fig. 11q).

Discussion

The RL gives rise to most of the neurons in the human brain, which provides a proximate explanation for why MB (cancer of the RL) is the most common embryonal neoplasm in humans and the most common malignant paediatric brain tumour. The physical splitting, interposed vascular plexus and dramatic expansion of the human RL, compared to that of mice and macaques, suggests that *H. sapiens* may be specifically and highly predisposed to the development of G4 MB. The teleological reasons and evolutionary benefits of the expansion of the human cerebellum are not immediately evident, although a predisposition to RL cancer appears to be a possible cost. Notably, the presence of atypical cell rests in the postnatal nodulus have been recognized since the 1940s and were already speculated to be a precursory lesion to MB on the basis of histological similarity[35,36]. However, without knowledge of the MB subgroups, the inconsistent anatomical presentation of MB in both the cerebellar hemispheres and the midline probably led to these hypotheses being forgotten. Further work in the 1960s, principally by Lucy Rorke, found an association between the presence of cerebellar heterotopias and various chromosomal trisomies, including trisomy of commonly gained chromosomes in G3 and G4 MB such as trisomy 17 (refs. [37,38]). We propose that these heterotopias in the nodulus are more aptly named persistent RL (PeRL), as they are probably neither permanent nor ectopic. Our findings add support to the 80-year-old hypothesis that PeRLs may represent a premalignant lesion for G3 and G4 MB, although presumably most PeRLs spontaneously regress in the absence of further genetic insult. Cancers arising from different anatomical regions of the RL, and different points in developmental time, account for much of the heterogeneity seen between MB subgroups (Fig. 6). Future studies using modern techniques will be necessary to further examine the probable link between PeRLs and MB.

The clustering of G4 MB driver genes along human chromosomes suggests a reason for the predilection of CNAs over SNVs in G4 MB. It also explains the specific patterns of genomic gain and loss and emphasizes that G4 MB is a disease of the human genome. G4 MB cells appear to be stalled in developmental time, with release of the developmental blockade resulting in the resumption of pseudo-normal developmental differentiation cascades. MB is a cancer of the human RL, with G4 MB probably arising from the embryological remnant of the RL^{SVZ} (that is, PeRLs) secondary to *OTX2* overactivity or CBFA complex failure. We propose that the detection of a PeRL in the postnatal period, through either imaging or biochemical detection of oncofetal antigens in the serum, could enable closer monitoring and potential early intervention to prevent the emergence of full-blown G3 or G4 MB.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-05215-w.

1. Northcott, P. A. et al. Medulloblastoma comprises four distinct molecular variants. *J. Clin. Oncol.* **29**, 1408–1414 (2011).
2. Cavalli, F. M. G. et al. Intertumoral heterogeneity within medulloblastoma subgroups. *Cancer Cell* **31**, 737–754.e6 (2017).
3. Vladoiu, M. C. et al. Childhood cerebellar tumors mirror conserved fetal transcriptional programs. *Nature* **572**, 67–73 (2019).
4. Hovestadt, V. et al. Resolving medulloblastoma cellular architecture by single-cell genomics. *Nature* **572**, 74–79 (2019).
5. Wechsler-Reya, R. J. & Scott, M. P. Control of neuronal precursor proliferation in the cerebellum by Sonic hedgehog. *Neuron* **22**, 103–114 (1999).
6. Wallace, V. A. Purkinje-cell-derived Sonic hedgehog regulates granule neuron precursor cell proliferation in the developing mouse cerebellum. *Curr. Biol.* **9**, 445–448 (1999).
7. Yang, Z.-J. et al. Medulloblastoma can be initiated by deletion of patched in lineage-restricted progenitors or stem cells. *Cancer Cell* **14**, 135–145 (2008).
8. Oliver, T. G. et al. Loss of patched and disruption of granule cell development in a pre-neoplastic stage of medulloblastoma. *Development* **132**, 2425–2439 (2005).
9. Gibson, P. et al. Subtypes of medulloblastoma have distinct developmental origins. *Nature* **468**, 1095–1099 (2010).
10. Jessa, S. et al. Stalled developmental programs at the root of pediatric brain tumors. *Nat. Genet.* **51**, 1702–1713 (2019).
11. Northcott, P. A. et al. The whole-genome landscape of medulloblastoma subtypes. *Nature* **547**, 311–317 (2017).
12. Haldipur, P. et al. Spatiotemporal expansion of primary progenitor zones in the developing human cerebellum. *Science* **366**, 454–460 (2019).
13. Aldinger, K. A. et al. Spatial and cell type transcriptional landscape of human cerebellar development. *Nat. Neurosci.* **24**, 1163–1175 (2021).
14. Juraschka, K. & Taylor, M. D. Medulloblastoma in the age of molecular subgroups: a review. *J. Neurosurg. Pediatr.* **24**, 353–363 (2019).
15. Tu, S. et al. Co-repressor CBFA2T2 regulates pluripotency and germline development. *Nature* **534**, 387–390 (2016).
16. Nady, N. et al. ETO family protein Mtgr1 mediates Prdm14 functions in stem cell maintenance and primordial germ cell formation. *eLife* **4**, e10150 (2015).

17. Tracey, L. J. et al. The pluripotency regulator PRDM14 requires hematopoietic regulator CBFA2T3 to initiate leukemia in mice. *Mol. Cancer Res.* **17**, 1468–1479 (2019).
18. Baulies, A. et al. The transcription co-repressors MTG8 and MTG16 regulate exit of intestinal stem cells from their niche and differentiation into enterocyte vs secretory lineages. *Gastroenterology* **159**, 1328–1341.e3 (2020).
19. Stadhouders, R. et al. Control of developmentally primed erythroid genes by combinatorial co-repressor actions. *Nat. Commun.* **6**, 8893 (2015).
20. Northcott, P. A. et al. Subgroup-specific structural variation across 1,000 medulloblastoma genomes. *Nature* **488**, 49–56 (2012).
21. Waszak, S. M. et al. Germline elongator mutations in Sonic hedgehog medulloblastoma. *Nature* **580**, 396–401 (2020).
22. Badodi, S. et al. Convergence of BMI1 and CHD7 on ERK signaling in medulloblastoma. *Cell Rep.* **21**, 2772–2784 (2017).
23. Skowron, P. et al. The transcriptional landscape of Shh medulloblastoma. *Nat. Commun.* **12**, 1749 (2021).
24. Parang, B. et al. The transcriptional corepressor MTGR1 regulates intestinal secretory lineage allocation. *FASEB J.* **29**, 786–795 (2015).
25. Branon, T. C. et al. Efficient proximity labeling in living cells and organisms with TurboID. *Nat. Biotechnol.* **36**, 880–887 (2018).
26. Forget, A. et al. Aberrant ERBB4–SRC signaling as a hallmark of group 4 medulloblastoma revealed by integrative phosphoproteomic profiling. *Cancer Cell* **34**, 379–395.e7 (2018).
27. Linggi, B. & Carpenter, G. ErbB-4 s80 intracellular domain abrogates ETO2-dependent transcriptional repression. *J. Biol. Chem.* **281**, 25373–25380 (2006).
28. Northcott, P. A. et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511**, 428–434 (2014).
29. Haldipur, P. et al. Evidence of disrupted rhombic lip development in the pathogenesis of Dandy–Walker malformation. *Acta Neuropathol.* **142**, 761–776 (2021).
30. Haldipur, P. et al. Expression of Sonic hedgehog during cell proliferation in the human cerebellum. *Stem. Cells Dev.* **21**, 1059–1068 (2012).
31. Englund, C. Unipolar brush cells of the cerebellum are produced in the rhombic lip and migrate through developing white matter. *J. Neurosci.* **26**, 9184–9195 (2006).
32. Leto, K. et al. Consensus paper: cerebellar development. *Cerebellum* **15**, 789–828 (2016).
33. Stromecki, M. et al. Characterization of a novel OTX2-driven stem cell program in group 3 and group 4 medulloblastoma. *Mol. Oncol.* **12**, 495–513 (2018).
34. Zagozewski, J. et al. An OTX2–PAX3 signaling axis regulates group 3 medulloblastoma cell fate. *Nat. Commun.* **11**, 3627 (2020).
35. Brzustowicz, R. J. Cell rests in the region of the fourth ventricle. *AMA Arch. Neurol. Psychiatry* **67**, 592 (1952).
36. Raaf, J. Relation of abnormal collections of cells in posterior medullary velum of cerebellum to origin of medulloblastoma. *Arch. Neurol. Psychiatry* **52**, 163 (1944).
37. Rorke, L. B., Fogelson, M. H. & Riggs, H. E. Cerebellar heterotopia in infancy. *Dev. Med. Child Neurol.* **10**, 644–650 (2008).
38. Yachnis, A. T., Rorke, L. B. & Trojanowski, J. Q. Cerebellar dysplasias in humans: development and possible relationship to glial and primitive neuroectodermal tumors of the cerebellar vermis. *J. Neuropathol. Exp. Neurol.* **53**, 61–71 (1994).
39. Yang, J. & Zhang, Y. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res.* **43**, W174–W181 (2015).
40. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
41. Canisius, S., Martens, J. W. M. & Wessels, L. F. A. A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence. *Genome Biol.* **17**, 261 (2016).
42. Finak, G. et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
43. Ximerakis, M. et al. Single-cell transcriptomic profiling of the aging mouse brain. *Nat. Neurosci.* **22**, 1696–1708 (2019).
44. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
45. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

Liam D. Hendrikse[1,2,3,100], Parthiv Haldipur[4,100], Olivier Saulnier[1,2,100], Jake Millman[4], Alexandria H. Sjoboen[4], Anders W. Erickson[1,2,5], Winnie Ong[1,2,5], Victor Gordon[6], Ludivine Coudière-Morrison[6], Audrey L. Mercier[7], Mohammad Shokouhian[8], Raúl A. Suárez[1,2], Michelle Ly[1,2,5], Stephanie Borlase[6], David S. Scott[1,2], Maria C. Vladoiu[1,2,5], Hamza Farooq[1,2,5], Olga Sirbu[1,2,3], Takuma Nakashima[9], Shohei Nambu[9], Yusuke Funakoshi[9], Alec Bahcheli[10,11], J. Javier Diaz-Mejia[12], Joseph Golser[4], Kathleen Bach[4], Tram Phuong-Bao[8], Patryk Skowron[1,2,5], Evan Y. Wang[1,2,3], Sachin A. Kumar[1,2,5], Polina Balin[1,2,5], Abhirami Visvanathan[1,2], John J. Y. Lee[1,2,5], Ramy Ayoub[3], Xin Chen[1,2], Xiaodi Chen[1,2], Karen L. Mungall[13], Betty Luu[1,2], Pierre Bérubé[14], Yu C. Wang[14], Stefan M. Pfister[15,16], Seung-Ki Kim[17], Olivier Delattre[18,19], Franck Bourdeaut[18,19], François Doz[18,20], Julien Masliah-Planchon[21], Wieslawa A. Grajkowska[22], James Loukides[1], Peter Dirks[1,2,10,23], Michelle Fèvre-Montange[24,25], Anne Jouvet[25,101], Pim J. French[26], Johan M. Kros[27], Karel Zitterbart[28], Sweneke D. Bailey[29,30], Charles G. Eberhart[31], Amulya A. N. Rao[32], Caterina Giannini[33], James M. Olson[34], Miklós Garami[35], Peter Hauser[35], Joanna J. Phillips[36,37], Young S. Ra[38], Carmen de Torres[39,102], Jaume Mora[39], Kay K. W. Li[40], Ho-Keung Ng[40], Wai S. Poon[41], Ian F. Pollack[42], Enrique López-Aguilar[43], G. Yancey Gillespie[44], Timothy E. Van Meter[45], Tomoko Shofuda[46], Rajeev Vibhakar[47], Reid C. Thompson[48], Michael K. Cooper[49], Joshua B. Rubin[50], Toshihiro Kumabe[51], Shin Jung[52], Boleslaw Lach[53,54], Achille Lolascon[55,56], Veronica Ferrucci[55,56], Pasqualino de Antonellis[55,56], Massimo Zollo[55,56], Giuseppe Cinalli[57], Shenandoah Robinson[58], Duncan S. Stearns[59], Erwin G. Van Meir[60], Paola Porrati[61], Gaetano Finocchiaro[61], Maura Massimino[61], Carlos G. Carlotti[62], Claudia C. Faria[63,64], Martine F. Roussel[65], Frederick Boop[65], Jennifer A. Chan[66], Kimberly A. Aldinger[4,67], Ferechte Razavi[68], Evelina Silvestri[69], Roger E. McLendon[70,71], Eric M. Thompson[71], Marc Ansari[72,73], Maria L. Garre[74], Fernando Chico[75], Pilar Eguía[75], Mario Pérezpeña[76], A. Sorana Morrissy[66,77,78], Florence M. G. Cavalli[79,80,81], Xiaochong Wu[1,2], Craig Daniels[1,2], Jeremy N. Rich[82], Steven J. M. Jones[13,83,84], Richard A. Moore[13], Marco A. Marra[13,83], Xi Huang[1,2,10], Jüri Reimand[3,10,11], Poul H. Sorensen[85,86], Robert J. Wechsler-Reya[87], William A. Weiss[36,88,89], Trevor J. Pugh[3,11,12], Livia Garzia[90], Claudia L. Kleinman[91,92], Lincoln D. Stein[10,93], Nada Jabado[94,95], David Malkin[3,96], Olivier Ayrault[7], Jeffrey A. Golden[97], David W. Ellison[98], Brad Doble[8], Vijay Ramaswamy[1,2,3,96], Tamra E. Werbowetski-Ogilvie[6,99,103], Hiromichi Suzuki[9,103], Kathleen J. Millen[4,103] & Michael D. Taylor[1,2,3,5,23,103] ✉

[1]The Arthur and Sonia Labatt Brain Tumor Research Centre, The Hospital for Sick Children, Toronto, Ontario, Canada. [2]Developmental and Stem Cell Biology Program, The Hospital for Sick Children, Toronto, Ontario, Canada. [3]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [4]Center for Integrative Brain Research, Seattle Children's Research Institute, Seattle, WA, USA. [5]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada. [6]Department of Biochemistry and Medical Genetics, University of Manitoba, Winnipeg, Manitoba, Canada. [7]PSL Research University, Université Paris Sud, Université Paris-Saclay, CNRS UMR 3347, INSERM U1021, Institut Curie, Orsay, France. [8]Department of Pediatrics and Child Health and Department of Biochemistry and Medical Genetics, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, Manitoba, Canada. [9]Division of Brain Tumor Translational Research, National Cancer Center Research Institute, Tokyo, Japan. [10]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [11]Computational Biology Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [12]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada. [13]Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia, Canada. [14]McGill University Genome Centre, McGill University, Montreal, Quebec, Canada. [15]Division of Pediatric Neurooncology, German Cancer Research Center (DKFZ), Heidelberg, Germany. [16]Department of Pediatric Oncology, Hematology, Immunology and Pulmonology, University Hospital Heidelberg, Heidelberg, Germany. [17]Department of Neurosurgery, Division of Pediatric Neurosurgery, Seoul National University Children's Hospital, Seoul, South Korea. [18]SIREDO Oncology Center (Pediatric, Adolescent and Young Adults Oncology), Institut Curie, Paris, France. [19]INSERM U830, Institut Curie, Paris, France. [20]Université Paris Cité, Paris, France. [21]Unit of Somatic Genetics, Institut Curie, Paris, France. [22]Department of Pathology, The Children's Memorial Health Institute, Warsaw, Poland. [23]Division of Neurosurgery, The Hospital for Sick Children, Toronto, Ontario, Canada. [24]INSERM U1028, CNRS UMR5292, Centre de Recherche en Neurosciences, Université de Lyon, Lyon, France. [25]Centre de Pathologie EST, Groupement Hospitalier EST, Université de Lyon, Bron, France. [26]Department of Neurology, Erasmus University Medical Center, Rotterdam, Netherlands. [27]Department of Pathology, Erasmus University Medical Center, Rotterdam, Netherlands. [28]Department of Pediatric Oncology, Masaryk University School of Medicine, Brno, Czech Republic. [29]Department of Surgery, Division of Thoracic and Upper Gastrointestinal Surgery, Faculty of Medicine, McGill University, Montreal, Quebec, Canada. [30]Cancer Research Program, Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada. [31]Departments of Pathology, Ophthalmology and Oncology, John Hopkins University School of Medicine, Baltimore, MD, USA. [32]Division of Pediatric Hematology/Oncology, Mayo Clinic, Rochester, MN, USA. [33]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA. [34]Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [35]2nd Department of Pediatrics, Semmelweis University, Budapest, Hungary. [36]Department of Neurological Surgery, University of California San Francisco, San Francisco, CA, USA. [37]Department of Pathology, University of California San Francisco, San Francisco, CA, USA. [38]Department of Neurosurgery, University of Ulsan, Asan Medical Center, Seoul, South Korea. [39]Developmental Tumor Biology Laboratory, Hospital Sant Joan de Déu, Esplugues de Llobregat, Barcelona, Spain. [40]Department of Anatomical and Cellular Pathology, The Chinese University of Hong Kong, Shatin, Hong Kong. [41]Department of Surgery, The Chinese University of Hong Kong, Shatin, Hong Kong. [42]Department of Neurological Surgery, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA. [43]Division of Pediatric Hematology/Oncology, Hospital Pediatría Centro Médico Nacional century XXI, Mexico City, Mexico. [44]Department of Neurosurgery, University of Alabama at Birmingham, Birmingham, AL, USA. [45]Pediatrics, Virginia Commonwealthy University, School of Medicine, Richmond, VA, USA. [46]Division of Stem Cell Research, Institute for Clinical Research, Osaka National Hospital, Osaka, Japan. [47]Department of Pediatrics, University of Colorado Denver, Aurora, CO, USA. [48]Department of Neurological Surgery, Vanderbilt Medical Center, Nashville, TN, USA. [49]Department of Neurology, Vanderbilt Medical Center, Nashville, TN, USA. [50]Departments of Neuroscience, Washington University School of Medicine in St Louis, St Louis, MO, USA. [51]Department of Neurosurgery, Kitasato University School of Medicine, Sagamihara, Japan. [52]Department of Neurosurgery, Chonnam National University Research Institute of Medical Sciences, Chonnam National University Hwasun Hospital and Medical School, Hwasun-gun, South Korea. [53]Department of Pathology and Molecular Medicine, Division of Anatomical Pathology, McMaster University, Hamilton, Ontario, Canada. [54]Department of Pathology and Laboratory Medicine, Hamilton General Hospital, Hamilton, Ontario, Canada. [55]Dipartimento di Medicina Molecolare e Biotecnologie Mediche (DMMBM), University of Naples Federico II, Naples, Italy. [56]CEINGE Biotecnologie Avanzate, Naples, Italy. [57]Department of Pediatric Neurosurgery, Santobono-Pausilipon Children's Hospital, Naples, Italy. [58]Division of Pediatric Neurosurgery, Case Western Reserve, Cleveland, OH, USA. [59]Department of Pediatrics–Hematology and Oncology, Case Western Reserve, Cleveland, OH, USA. [60]Department of Hematology and Medical Oncology, Emory University School of Medicine and Winship Cancer Institute, Atlanta, GA, USA. [61]Fondazione IRCCS Istituto Nazionale Tumori, Milan, Italy. [62]Department of Surgery and Anatomy, Faculty of Medicine of Ribeirão Preto, University of São Paulo, São Paulo, Brazil. [63]Division of Neurosurgery, Centro Hospitalar Lisboa Norte (CHULN), Hospital de Santa Maria, Lisbon, Portugal. [64]Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal. [65]Department of Tumor Cell Biology, St Jude Children's Research Hospital, Memphis, TN, USA. [66]Charbonneau Cancer Institute, University of Calgary, Calgary, Alberta, Canada. [67]Brotman Baty Institute for Precision Medicine, Seattle, WA, USA. [68]Assistance Publique Hôpitaux de Paris, Hôpital Necker-Enfants Malades, Paris, France. [69]Surgical Pathology Unit, San Camillo Forlanini Hospital, Rome, Italy. [70]Department of Pathology, Duke University, Durham, NC, USA. [71]Department of Neurosurgery, Duke University, Durham, NC, USA. [72]Cansearch Research Platform for Pediatric Oncology and Hematology, Faculty of Medicine, Department of Pediatrics, Gynecology and Obstetrics, University of Geneva, Geneva, Switzerland. [73]Division of Pediatric Oncology and Hematology, Department of Women, Child and Adolescent, University Geneva Hospitals, Geneva, Switzerland. [74]U.O. Neurochirurgia, Istituto Giannina Gaslini, Genova, Italy. [75]Department of Neurosurgery, Hospital Infantil de Mexico Federico Gomez, Mexico City, Mexico. [76]Instituto Nacional De Pediatría de México, Mexico City, Mexico. [77]Department of Biochemistry and Molecular Biology, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada. [78]Alberta Children's Hospital Research Institute, Calgary, Alberta, Canada. [79]INSERM U900, Institut Curie, Paris, France. [80]PSL Research University, Institut Curie, Paris, France. [81]CBIO-Centre for Computational Biology, PSL Research University, MINES ParisTech, Paris, France. [82]UPMC Hillman Cancer Center, Pittsburgh, PA, USA. [83]Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada. [84]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia, Canada. [85]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia, Canada. [86]Department of Molecular Oncology, BC Cancer Agency, Vancouver, British Columbia, Canada. [87]Tumor Initiation and Maintenance Program, NCI-Designated Cancer Center, Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA, USA. [88]Department of Neurology, University of California San Francisco, San Francisco, CA, USA. [89]Department of Pediatrics, University of California San Francisco, San Francisco, CA, USA. [90]Cancer Research Program, McGill University Health Centre Research Institute, Montreal, Quebec, Canada. [91]Department of Human Genetics, McGill University, Montreal, Quebec, Canada. [92]Lady Davis Research Institute, Jewish General Hospital, Montreal, Quebec, Canada. [93]Adaptive Oncology, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [94]Departments of Pediatrics and Human Genetics, McGill University, Montreal, Quebec, Canada. [95]The Research Institute of the McGill University Health Center, Montreal, Quebec, Canada. [96]Division of Haematology/Oncology, Department of Pediatrics, The Hospital for Sick Children, Toronto, Ontario, Canada. [97]Department of Pathology, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [98]Department of Pathology, St Jude Children's Research Hospital, Memphis, TN, USA. [99]CancerCare Manitoba Research Institute, Winnipeg, Manitoba, Canada. [100]These authors contributed equally: Liam D. Hendrikse, Parthiv Haldipur, Olivier Saulnier. [101]Deceased: Anne Jouvet. [102]Deceased: Carmen de Torres. [103]These authors jointly supervised this work: Tamra E. Werbowetski-Ogilvie, Hiromichi Suzuki, Kathleen J. Millen, Michael D. Taylor.

Methods

Data reporting

No statistical methods were used to predetermine sample sizes. The experiments were not randomized. Investigators were not blinded to allocation during experiments and outcome assessment, except in the case of performing immunohistochemistry (IHC)-based cell counts, for which the investigator performing the count was blinded to experimental conditions such as sample age and antibodies used.

Acquisition of patient samples and human tissue

Samples used to generate the bulk RNA-seq group were obtained from the Medulloblastoma Advanced Genomics International Consortium (MAGIC). Samples used to generate the scRNA-seq group were obtained in compliance with the ethical regulations of the Hospital for Sick Children and McGill University Health Centre. All patient material was collected after receiving informed consent, under approval and oversight by their respective internal review boards. Relevant clinical metadata and overlapping data types for the 819 primary MB samples used in this study are presented in Supplementary Table 2.

All human cerebellar samples used in this study were obtained using protocols approved by the Seattle Children's Research Institute's institutional review board. Samples used for histological analyses were collected by the Human Developmental Biology Resource located at University College London and at Newcastle University, United Kingdom, and the Birth Defects Research Laboratory at the University of Washington, United States. Third trimester samples were part of an archival collection at the Hôpital Necker-Enfants Malades in Paris, France. All samples were collected with previous patient consent and in strict accordance with institutional and legal ethical guidelines. Formalin-fixed paraffin-embedded sections of the cerebellum were collected at 4-µm thickness along the sagittal plane and mounted on Superfrost plus slides (Thermo Fisher Scientific). Slides were refrigerated to preserve antigenicity and to prevent RNA degradation.

Animals

Embryonic mouse tissue was collected in accordance with the guidelines laid down by the Institutional Animal Care and Use Committee of Seattle Children's Research Institute and with the National Institutes of Health guidelines on the care and use of laboratory animals and are in accordance with the applicable portions of the Animal Welfare Act. Mice were maintained on corncob bedding with a 10–14 h dark–light cycle. Ambient temperature and humidity were monitored and maintained within the recommended ranges by the Guide for the Care and Use of Laboratory Animals, 8th edition. CD1 mice were crossed, and the day of plug was taken as embryonic day 0.5 (E0.5). Embryos from both males and females were dissected out between E14.5 and E16.5, fixed in 4% paraformaldehyde for 1–2 h, washed in PBS and transferred to a solution of 30% sucrose overnight. Samples were then embedded in optimum cutting temperature matrix. Mid-sagittal cryosections of 11 µm thickness were collected.

Cell lines

Two G3 MB cell lines were used in this study: HDMB03 and MB3W1. HDMB03 was provided by T. Milde (described in ref. [46]) MB3W1 was provided by M. Wölfl (described in ref. [47]). No commonly misidentified cell lines were used in this study and the cell lines used were authenticated by STR profiling by ATCC in 2021. Cell lines were not tested for mycoplasma contamination.

Histology

Formalin-fixed paraffin-embedded sections were de-paraffinized in xylene and rehydrated in a gradient of ethanol before processing. Haematoxylin and eosin staining were carried out as previously described[30].

IHC

IHC was performed as previously described[12]. The following primary antibodies were used: KI67 (Agilent, M7240, mouse, 1:50; Thermo Fisher, MA5-14520, rabbit, 1:100); SOX2 (Thermo Fisher, PA1-094, rabbit, 1:200); PAX6 (BioLegend, 901301, rabbit, 1:300); TBR2 (*EOMES*) (Thermo Fisher, 14-4875-82, rat, 1:200); and GFAP (Agilent, Z0334, rabbit, 1:1,000). Fluorescent-dye-labelled secondary antibodies from Thermo Fisher were used at a dilution of 1:1,000 (anti-rabbit: Alexa Fluor 488, A-11034, goat; Alexa Fluor 568, A-11011, goat; anti-mouse: Alexa Fluor 488, A-11001, goat; Alexa Fluor 568, A-11004, goat). After secondary antibody incubation, sections were counterstained with DAPI (4′,6-diamidino-2-phenylindole) using Vectashield mounting medium (Vector laboratories, H-1200).

In situ hybridization

In situ hybridization assays were run using commercially available probes from Advanced Cell Diagnostics. Manufacturer-recommended protocols were used without modification. The following probes were used in the study: LMX1A (540661), OTX2 (484581), CBFA2T2 (410331), CBFA2T3 (406001), MKI67 (591771), HBEGF (431651), Cbfa2t2 (491601), Cbfa2t3 (434601) and Hbegf (437601). All sections were counterstained with haematoxylin or methyl green.

Microscopy

All slides from fluorescence immunohistochemical assays were imaged using a Zeiss LSM-Meta confocal microscope and ZEN 2009 software (Zeiss). A Nanozoomer Digital Pathology slide scanner (Hamamatsu) was used for brightfield microscopy. Barring minor adjustments of contrast and brightness, no additional image alteration was performed. Figures were prepared using Adobe Illustrator.

Cell counts

Cell counts were manually performed using ImageJ. For all counts, a minimum of three sections over three samples were used. For KI67, PAX6 and TBR2 (*EOMES*) counts, the total number of KI67 or TBR2 (*EOMES*) or $PAX6^+$ cells relative to the total DAPI count in the RL was determined.

CBFA2T2 TurboID

3×HA-TurboID-NLS_pCDNA3 was a gift from A. Ting (Addgene plasmid 107171). pCMV6-AC-CBFA2T2-GFP was purchased from Origene (RG202013). PCR-amplified TurboID and CBFA2T2 coding sequences were introduced into the KpnI/ClaI-linearized pCAG-H2B-mAG-P2A vector using NEBuilder HiFi assembly (New England Biolabs). Non-vector sequences in resultant plasmids pCAG-H2B-mAG-P2A-3×HA-TurboID-CBFA2T2 and pCAG-H2B-mAG-P2A-3×HA-TurboID were verified by Sanger sequencing. For TurboID experiments, 3×10^6 HDMB03 cells were seeded equally across a 6-well ultralow attachment plate and transfected with 12 µg of plasmid using Lipofectamine 3000 (Thermo Fisher) according to manufacturer's guidelines. Cells were grown in suspension for 72 h in StemPro NSC serum-free medium (Life Technologies). After 72 h of incubation, biotin-D (Sigma, B4639) was added to all wells to a final concentration of 500 µM, and the cells were incubated for 15 min. Tumourspheres were then collected, washed in 1×PBS, snap-frozen and stored at −80 °C until all runs were completed.

Replicate samples were thawed on ice and lysed with 1 ml of 1× RIPA (20 mM Tris-HCl, pH 7.5, 150 mM NaCl, 1 mM EDTA-Na_2, 1 mM EGTA, 1% NP-40; 1% sodium deoxycholate, 2.5 mM sodium pyrophosphate, 1 mM β-glycerophosphate, 1 mM Na_3VO_4 and 1× Halt protease/phosphatase inhibitor (ThermoScientific)) for 10 min. Samples were then sonicated twice with 10-s pulses at 30% load to shear DNA (Fisher Scientific Sonicator FB50 with microprobe). Lysed samples were clarified by centrifugation at 16,000*g* for 10 min at 4 °C. Each supernatant was quantified using the Bradford protein assay. Following quantification,

50 µl aliquots of streptavidin–sepharose bead slurry for each sample was washed once in 1 ml of 1× RIPA. Washed bead aliquots were then resuspended in 1 ml of 1× RIPA containing 2 mg of quantified protein supernatant and rotated overnight at 4 °C. Samples were then centrifuged at 1,000*g* for 5 min, and bead pellets were washed twice in 1 ml 2% SDS solution (in double-distilled H_2O) followed by three washes in 1 ml wash buffer (50 mM Tris-HCl, pH 7.4, and 8 M urea), rotating for 8 min at room temperature each time. Samples were then resuspended in storage buffer (285 µl of ammonium bicarbonate (50 mM) and 15 µl of 1 mM biotin) to saturate streptavidin binding and to prevent peptide recapture during on-bead digestion. Samples were stored on ice for analysis.

Mass spectrometry

Mass spectrometry (MS) data were obtained with an Orbitrap Exploris 480 instrument (Thermo Fisher Scientific) and analyses were performed at the Manitoba Centre for Proteomics and Systems Biology. All MS raw files were processed with Proteome Discoverer (v.2.20.388) and searched for tryptic peptides against the human UniProt protein database (date accessed: December 2020) using SEQUEST with the following standard Orbitrap settings: up to 2 missed cleavages were permitted, with a parent and fragment mass tolerance of 0.02 Da and 15 ppm, respectively. A fixed modification of cysteine carbamidomethylation was applied, and variable modifications including amino-terminal acetylation, deamidation (at N and Q), phosphorylation (at S, T and Y), oxidation (at M and W), ubiquitylation (at K), double oxidation (at M and W) and biotinylation (at K) were permitted. The results were filtered by 1% false discovery rates (FDRs) at both the protein and peptide levels. SAINTexpress (v.1.0.0) was then used to calculate the probability of each potential proximal protein interaction from background contaminants using default parameters[48,49]. Three replicates were used for control and experimental samples. A full list of CBFA2T2-interacting proteins can be found in Supplementary Table 3.

OTX2-KD

OTX2-KD in HDMB03 and MB3W1 G3 MB cells (2×10^5 cells per well) was performed as previously described[33,34]. In brief, *OTX2* was silenced using 30 nM Silencer Select siRNAs 9931 or 9932 (Life Technologies). A non-silencing (scramble) served as the negative control. For bulk RNA-seq, *OTX2* was knocked down in three independent biological replicates for each cell line, and silencing was confirmed by western blotting (OTX2, Abcam, ab21990, rabbit, at 1:500; β-actin, Sigma-Aldrich, A2228, mouse, at 1:1,000 was used as a loading control) 72 h following transfection. Total RNA was extracted from all samples using a Norgen RNA extraction kit (Norgen Biotek), and bulk RNA-seq was performed by StemCore laboratories at the Ottawa Hospital Research Institute. For snRNA-seq, the above was repeated but 4.5×10^5 cells per well for MB3W1 was used and only one replicate was performed. GN differentiation was validated by western blotting for RBFOX3 (NeuN) (Cell Signaling Technology, D4G4O, at 1:1,000).

CBFA2T2 overexpression

HDMB03 G3 MB cells were plated at a density of 2×10^5 cells per well (in a 6-well format) 24 h before infection. The following day, Precision LentiORF CBFA2T2 viral particles (OHS5898-219582524) or RFP control lentiviral particles (Horizon Discovery Biosciences) were added to the cells in OptiMEM (Life Technologies) at a multiplicity of infection of 1. Twenty-four hours after infection, OptiMEM was aspirated and replaced with EMEM/10% FBS. Positively transduced cells were stably selected in blasticidin starting at day 5. Increased CBFA2T2 expression was confirmed by quantitative PCR and immunoblotting (CBFA2T2, A303-593A-M, Bethyl Laboratories at 1:500; β-actin, Sigma-Aldrich, A2228, mouse, at 1:1,000 was used as a loading control).

Following stable selection and cell expansion, assessments of total cell number and viability were performed as previously described[34]. In brief, HDMB03 RFP control and CBFA2T2 overexpression cells were dissociated and cultured as tumourspheres in 6-well ultralow attachment plates at a density of 2×10^5 cells per well (in triplicate) in StemPro NSC serum-free medium (Life Technologies). The total number of cells was counted at day 5 and viability was assessed by trypan blue exclusion.

MRI

All MRI scans were performed on patients with MB from the Hospital of Sick Children, Canada. Tumours were subsequently subgrouped by methylation array.

Sanger sequencing

CBFA2T2 mutations were validated by Sanger sequencing (data not shown). In brief, nested PCR was performed using the following reagents: master mix (2.5 µl 10× PCR buffer (Invitrogen, 52724), 1 µl 50 mM $MgCl_2$ (Invitrogen, 52723), 0.5 µl 10mM dNTP (Bio Basic, DD0056), 0.5 µl F primer, 0.5 µl R primer, 2 µl loading dye, 0.2 µl Taq polymerase (Invitrogen, 100021278), 15.8 µl ddH_2O (Wisent, 809-115-CL)) and 2 µl DNA. Secondary PCR products were run on a 2% agarose gel with TAE buffer to confirm the expected product size. For each mutation, the band was excised, cleaned (with a Geneaid PCR Clean-up Kit, Geneaid, DFC100) and Sanger sequenced. In addition, we TA cloned each mutation using a TA Cloning kit (Thermo Fisher, 450641) followed by Mini prep (Geneaid Presto Mini Plasmid kit, Geneaid, PD100) and Sanger sequencing.

Bulk RNA-seq

Sample processing, mRNA library construction and sequencing. Samples were processed and sequenced as previously described[23].

Alignment. The hs37d5 reference genome FASTA (1000 Genomes Project Phase II) was appended to the C1_2 ERCC spike-in sequences used for C1 Fluidigm, as well as Caltech profile 3 spike-ins sequences by ENCODE. A STAR assembly was then built with this reference and GENCODE (v.19) gene annotations using the following parameter: --sjdbOverhang 124. RNA-seq library reads were then mapped with the built assembly using STAR (v.2.5.1b) and the following parameters: --outFilterMultimapNmax 20 -- alignSJoverhangMin 8 --alignMatesGapMax 200000 --alignIntronMax 200000 --alignSJDBoverhangMin 10 --alignSJstitchMismatchNmax 5 –1 5 5 --outSAMmultNmax 20 --twopassMode Basic. For alignment of OTX2-KD G3 cell lines, raw sequence data were processed with fastp (v.0.20.1)[50] to perform automatic adapter trimming and read-quality filtering, retaining reads with at least 60% of bases having $Q \geq 15$, and no more than 5 'N' bases. Transcripts from scramble and OTX2-KD samples were quantified using salmon (v.1.4.0)[51] against an index built from the GENCODE v.35 reference assembly with inclusion of genomic decoy sequences.

G3 and G4 MB subtype identification. Subtypes were determined using similarity network fusion (SNF) as previously described[2] using RNA-seq expression data in place of microarray expression data[23]. In brief, the full expression and methylation matrices were input into the SNF function of the R package SNFtool (v.2.3.0) with the following parameters: $K = 20$, $\alpha = 0.5$, $T = 100$. Spectral clustering was then performed, and the clusters obtained at $k = 6$ corresponded to the 6 G3 and G4 subtypes.

G3 and G4 MB subtype differentially expressed genes. Differential expression analysis was performed using the R package DESeq2 (v.1.28.1)[45], comparing each subtype to the other five, and controlling for subgroup as a covariate. Only genes with a FDR adjusted P value < 0.05 were considered in downstream analysis. Full lists of subtype-specific differentially expressed genes can be found in Supplementary Table 4.

RNA-seq mutation analysis. RNA-seq mutation calls were performed as previously described[23,52]. In brief, GATK (v.3.8.0)[53] was used to detect

variants that were then annotated with ANNOVAR (February 2016)[54]. SNPs present in the gnomAD database were discarded, and SNPs with frequencies greater than 0.01 in 1000 Genomes, dnSNP138, RADAR, the Exome Aggregation Consortium database, the NHLBI-ESP project, the Kaviar Genomic Variant Database, the Haplotype Reference Consortium database, the Greater Middle East Variome, the Brazilian Genomic Variants database and from an in-house SNP database (356 sequenced whole genomes) were discarded. Significantly mutated genes ($q < 0.05$) were identified using MutSigCV (v.1.41)[55] and OncodriveFML (v.2.3.0)[56]. A full list of mutations detected in G3 and G4 MB can be found in Supplementary Table 5.

Fusion calling and filtering. Gene fusions were called and filtered as previously described[23]. In brief, three fusion calling algorithms were run to maximize sensitivity: Star-fusion (v.0.8.0)[57], InFusion (v.0.7.3)[58] and Trans-Abyss (v.2.0.0)[59]. Only putative fusions called by at least two algorithms, not detected in normal, and occurring in known mutated genes were retained.

SNP 6.0 arrays

GISTIC analysis and determination of copy-number-responsive genes. Samples were pre-processed as previously described[23]. Processed SNP6 segment files were then input into GISTIC 2.0 (v.2.0.23)[40] and run with the following changes to default parameters: -conf 0.90 -broad 1 -ta 0.25 -td 0.3 -js 10 -rx 0 -brlen 0.7 -armpeel 1 -gcm extreme -genegistic. Significantly amplified and deleted regions were then extracted, manually inspected in IGV and categorized into either broad or focal events depending on whether the segment spanned >12 Mb or ≤12 Mb, respectively. To determine whether genes falling in significantly amplified or deleted regions showed concomitant expression changes, gene expression was categorized (where applicable) by amplification, neutral or loss. Kruskal–Wallis test was performed for greater than two conditions, whereas Mann–Whitney *U*-test was performed in the case of only two conditions. Significance was adjusted using post-hoc Dunn's test with Benjamini–Hochberg correction.

RNA-seq of MB tumours

Preparation of single-cell suspensions. Patient samples of tumours were collected at the time of surgical resection. Tumour tissue was mechanically dissociated followed by collagenase-based enzymatic dissociation as previously described[3].

Single-cell RNA library preparation and sequencing. Single-cell suspensions were assessed by trypan blue count. We aimed to load 10,000–15,000 cells per sample using a Chromium Controller in combination with Chromium Single Cell 3′ V3 and V3.1 Gel Bead and Chip kits (10X Genomics). Individual cells were partitioned into gel beads-in-emulsion followed by reverse transcription of barcoded RNA and cDNA amplification. Individual single-cell libraries with indices and Illumina P5/P7 adapters were generated using a Chromium Single Cell 3′ Library kit and a Chromium Multiplex kit. The libraries were sequenced on an Illumina Novaseq6000 sequencer.

Alignment of raw reads. Gene-level counts were obtained using the 10X CellRanger pipeline (v.3.1.0)[60]. In brief, raw base call files were de-multiplexed into fastqs using the mkfastq function. Fastq files were then aligned to the reference human genome hg19 v.3.0.0 (from 10X Genomics) using the count function to generate raw gene–barcode count matrices. Alignment quality control metrics for all scRNA-seq and snRNA-seq samples can be found in Supplementary Table 6.

Quality control and normalization. Quality control was performed at an individual sample level before sample integration and normalization. In brief, low-quality cells were determined and excluded from further analysis based on outlier mitochondrial content (indicative of cellular stress or damage) or gene counts using the R package Seurat (v.4.0.2)[61]. Genes expressed in fewer than ten cells were also removed. High-quality cells from each tumour sample were merged and normalized together using SCTransform (v.0.3.2)[62] using the parameter variable.features.n = 3,000 and regressing unwanted variance associated with mitochondrial content.

Clustering analysis and visualization. Single-cell clustering was performed by first using principal component analysis to determine statistically significant principal components, which were then used to construct a uniform manifold approximation and projection (UMAP) embedding in Seurat (v.4.0.2)[61]. Cell clusters were identified using Seurat's shared-nearest neighbour algorithm following modularity optimization using the Louvain algorithm with multilevel refinement.

Identification of cell types. Nontumour cell types present in the tumour single-cell samples were first identified by examining the expression of known cell-type-specific markers (Extended Data Fig. 7g). As expected, these clusters displayed the most overlap between different samples. To confirm that the remaining sample-specific cell clusters were composted of tumour cells, inferCNV (v.1.4.0)[63] was used, using the nontumour cells as a reference and using the following parameters: HMM_type = i6, noise_filter = 0.1, cut-off = 0.1 and sd_amplifier = 0.15 (Extended Data Fig. 7e). All tumour cell clusters were confirmed to have CNVs characteristic of their subgroup[11], with only one sample (SHH MB3862) having no detected CNVs (Extended Data Fig. 7f). Given the high proportion of tumour cells in other SHH MB samples compared to nontumour microenvironment cells, and the distinct clustering from microenvironment cells, we labelled these as tumour cells.

Cell-to-cell ligand–receptor analysis. To identify putative ligand–receptor interactions occurring between tumour cells and microenvironmental nontumour cells in the scRNA-seq data, we used CCInx (v.0.4)[43] as previously described[64]. In brief, scaled expression levels of both subgroup-specific ligands in microglia and receptors in G4 MB tumour cells were calculated using the function BuildGeneStatList, and interactions were determined using the built-in curated database of ligand–receptor interactions (https://baderlab.org/CellCellInteractions). Edges represent putative interactions and they are ordered and coloured by the magnitude of the average expression of the ligand and receptor.

snRNA-seq (human samples)

Sample collection and processing. Human cerebellar tissue from 9–21 PCW were previously processed and the data obtained[13]. Samples were aligned to the reference human genome hg19 as for the tumour samples.

Clustering analysis and visualization. Clustering analysis and visualization for the entire human cerebellum dataset were performed as previously stated[13]. For more discrete cell-type similarity mapping, a glutamatergic subset of the object was created by extracting cell clusters 02-RL, 03-GCP, 05-eCN/UBCs and 04-GN and reclustering by the same methods as described above for tumour scRNA-seq samples.

Pseudotime trajectory analysis. To recapitulate the expected developmental trajectories[13] in the glutamatergic cell subset, we used the trajectory inference method Slingshot (v.1.6.1)[65] as previously described[64]. In brief, a three-dimensional diffusion map embedding was constructed using the DiffusionMap function from the R package destiny (v.3.2.0)[66]. Slingshot analysis was then performed on the diffusion map to determine per-cell pseudotime estimates and mapped back to the UMAP embedding.

Identification of cell types. Re-clustering analysis of the glutamatergic cell clusters from the developing human cerebellum snRNA-seq dataset

revealed additional heterogeneity within each original cluster. The RL cells clustered into two distinct transcriptional clusters, consistent with previous histological findings[12,13]. We performed differential gene expression analysis using the FindAllMarkers methods with the following non-default parameters: only.pos = T, test.use = "MAST" (ref. [42]). We annotated these two clusters as RL^{VZ} and RL^{SVZ} based on the expression of characteristic markers of these distinct cell types[12,13] (Extended Data Fig. 7a). The eCN/UBC cluster also split into two distinct transcriptional clusters, which we annotated as 'early' and 'late' UBCs on the basis of the pseudotime results. The GN cluster split into several transcriptionally distinct clusters that we labelled as 'early' and 'late' GNs on the basis of the pseudotime results. A list of the top 50 differentially expressed gene markers of each cluster can be found in Supplementary Table 7.

Mapping transcriptional similarity between MB and the human cerebellum. To compare the transcriptional similarity between MB cells and the developing human cerebellum, we first used the SingleR package (v.1.2.4)[67]. A human cerebellum development reference was created using the function aggregateReference, with the parameter power = 0. We confirmed that the new reference was able to successfully self-classify the human cerebellar cells. This reference was then used to classify tumour cells from each MB sample by transcriptional similarity. This method was repeated with an additional step of removing genes associated with the cell cycle[61,68] from the expression matrices to confirm that there was no spurious cell-type correlation based on the MB cell cycle state. As a control, we found that SHH MB cells were most similar to GCPs, which is consistent with their presumed cell of origin[3,5,6,69].

Deconvolution analysis. As a parallel method to determine transcriptional similarity and in a much larger group of bulk RNA-seq MB samples, we used CIBERSORT[70] to deconvolute the relative proportions of each glutamatergic cell type in the developing human cerebellum. In brief, the CIBERSORT algorithm was called in R using the amritr package (v.0.1.0) with the parameters perm = 10, QN = F. As a control, we again found that the dominant cell-type proportion in SHH MB was GCPs, as expected.

TF activity analysis. The activity of specific TFs in each cell type was inferred in both the human development cells and the MB cells using the package pySCENIC (v.0.10.0)[44,71], implemented in python (v.3.7.6), with the parameter -min_genes 10. For each TF, the cellular activity was determined and binarized into an 'on' or 'off' state using the function binarize with default parameters.

snRNA-seq (G3 MB cell lines)

Preparation of single-nucleus suspensions. Nuclei were isolated from fresh, snap-frozen tumour tissues as previously described[72]. Frozen pellets were dounced in 1 ml of chilled lysis buffer (10 mM Tris-HCl (pH 7.4), 10 mM NaCl, 3 mM $MgCl_2$ and 0.05% NP-40 detergent) 5 times with a loose pestle, 10 times with a tight pestle and lysed for 10 min on ice. Chilled wash buffer (5 ml; 5% BSA, 0.04 U μl^{-1} RNase inhibitor and 0.25% glycerol) was added to the sample, passed through a 40-µm cell strainer and centrifuged at 500*g* for 5 min at 4 °C. After pelleting, the nuclei were resuspended in 5–10 ml of wash buffer. After two washes, single-nucleus suspensions were passed through a 20-µm cell strainer, pelleted and resuspended in PBS with 0.05% BSA. Library preparation and sequencing were then conducted as stated above for scRNA-seq of MB samples.

Alignment of raw reads. Alignment was performed as previously stated for human MB tumours, with the following differences. When using the CellRanger's (v.6.0.1) count method, the additional parameter --include-introns was specified as recommended for single-nucleus sequencing data.

Quality control, normalization and clustering. Quality control, normalization and clustering were performed as previously stated for human MB tumours, with the following difference. Following quality control, there were a number of cells with notably low numbers of features detected per RNA count, which was indicative of ambient RNA background or low complexity cells. These cells were filtered out on a per sample basis.

Mapping transcriptional similarity between cell lines and the human cerebellum. We observed differentiation in our cultures of OTX2-KD cells from both HDMB03 and MB3W1 cells. Using the above described method for comparing primary MB tumour cells to the human developing cerebellum, we used SingleR with the same reference dataset to classify individual cells from each scramble and OTX2-KD sample.

Identification of cellular differentiation state. In the OTX2-KD samples for both HDMB03 and MB3W1 cells, we noted a gradient of *OTX2* expression rather than a consistent decrease in expression. We suspected this could be due to inefficient OTX2-KD in certain cells. Using the differentially expressed genes (log fold change > 2, $P < 0.05$) obtained from bulk RNA-seq of OTX2-KD and scramble samples, we generated two gene signatures called 'scramble signature' and 'OTX2-KD signature', which represent gene expression programmes of the dominant cellular populations in each condition. Using the AddModuleScore function in Seurat, we scored the single cells from each condition for these signatures (Extended Data Fig. 11g) and found cluster-specific enrichment for the scramble signature, which probably indicated incomplete knockdown in a small subset of cells. If these cells indeed had inefficient OTX2-KD, we would expect these cells to retain features of wild-type cells, namely a lack of differentiation. To test this, we applied the CytoTRACE (v.0.3.3) algorithm, which estimates the differentiation state of single cells using the number of genes expressed per cell[73]. As expected, these putatively inefficient knockdown cells were inferred to be the least differentiated cells (Extended Data Fig. 11h).

Pseudotime trajectory analysis. We then conducted pseudotime trajectory analysis using Slingshot, as described for the tumour samples above; however, we specified the start of the lineage (or the root) as the cluster with the highest CytoTRACE scores. To identify genes for which expression was significantly correlated with the determined lineages, we first fit each gene with a negative binomial general additive model as implemented in the R package TradeSeq (v.1.2.01)[74]. We then tested whether these gene-level models were correlated with inferred pseudotime using the associationTest function with the parameter lineages = T, and filtered out genes with FDR corrected *P* values of >0.05. Finally, we overlapped these significantly lineage-associated genes with significant cell-type marker genes from the human developing cerebellum dataset and used the predictSmooth function to bin the expression of these genes along OTX2-KD and normal human cerebellar development pseudotime.

Survival analysis

Overall survival for each of the subtypes of G3 and G4 MB was analysed using the Kaplan–Meier method. *P* values were determined using the log-rank test. Survival analysis was conducted using the R package survival (v.3.2-11)[75].

CBFA2T2 protein structure model

The structure of the NHR1 domain of CBFA2T2 was accessed from SWISS-MODEL (accession number O43439). This structure was previously inferred using homology to the known structure of the RUNX1T1 NHR1 domain[76]. The full protein structure was predicted using i-TASSER[39]. The optimal CBFA2T2 prediction was selected for visualization based on which model's NHR2 domain exhibited the highest degree of similarity to the known SWISS-MODEL structure. This model was then visualized and coloured in PyMOL (v.2.4.2). However, the high degree of disorder in CBFA2T2 makes the predicted model imprecise and is provided only for context of the NHR1 domain.

Mutual exclusivity analysis

Mutual exclusivity of various alterations was statistically assessed using the R package DISCOVER (v.0.9.3)[41]. All *P* values derived from multiple testing were corrected using built-in FDR.

Protein–protein interaction network

To determine the extent of protein–protein interactions that are known to occur between G3 and G4 MB driver genes (Fig. 1a), and with new CBFA2T2 interacting proteins (TurboID), we used STRING analysis (v.11.5)[77]. In brief, gene lists of driver genes and genes encoding CBFA2T2-interacting proteins were queried for known interactions. The resulting network was further filtered for only interactions with experimental evidence and a confidence score of at least 0.250. The network was then imported into Cytoscape[78,79] for visualization. Each node corresponds to a protein and was coloured by the relative frequency that the corresponding gene is altered (that is, mutated, deleted, amplified, enhancer hijacked, among others) across G3 and G4 MB. Edges correspond to known protein–protein interactions and their weight is proportional to the STRING confidence score. The nodes of CBFA2T2 prey proteins as determined by TurboID were assigned a diamond shape. To reduce complexity, not all edges between diamond-shaped nodes and CBFA2T2 were drawn. Nodes were then manually arranged to group proteins in the same gene family; therefore, edge length is arbitrary. However, the degree of connectivity between groups guided the organization of the nodes.

Illustrations

Oncoprint landscape figures were generated in R (v.4.0.2) using the ComplexHeatmap (v.2.4.3) library[80]. Gene mutation summary lollipop figures were generated using ProteinPaint[81].

Reporting summary

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

Data availability

The bulk and scRNA-seq data generated from MB tumour samples in this study have been deposited in the European Genome-Phenome Archive (EGA) database under the accession code EGAS00001005826. The bulk and scRNA-seq data generated from the G3 MB cell lines HDMB03 and MB3W1 in this study have been deposited in the Gene Expression Omnibus (GEO) database under the access codes GSE189238 and GSE200791, respectively. The published MB bulk and scRNA-seq data referenced in this study are available in the EGA database under the accessions EGAD00001006305, EGAD00001004435, EGAD00001004958, EGAS00001003170 and EGAS00001003368. The referenced GTEx normal cerebellum RNA-seq controls were acquired from the NCBI public repository phs000424.v6.p1. The Affymetrix SNP 6.0 data referenced during the study are available in the GEO database under the accession GSE37385. The whole-genome sequencing data referenced during the study are available in the EGA database under the accessions EGAD00001003125 and EGAD00001004347. The Illumina 450k methylation data referenced during the study are available in the GEO database under the accession GSE85218. The G3 tumoursphere ChIP-seq data referenced during the study are available in the GEO database under the accession GSE132269. There were multiple databases used for annotation and filtering referenced in this study. These include the Exome Aggregation Consortium/gnomAD (https://gnomad.broadinstitute.org/downloads), the NHLBI-ESP project (https://esp.gs.washington.edu/drupal/), the Kaviar Genomic Variant Database (http://db.systemsbiology.net/kaviar/#:-:text=Kaviar%20Genomic%20Variant%20Database%20%7C%20SNP,and%20frequency%20of%20observed%20variants.), the Haplotype Reference Consortium (http://www.haplotype-reference-consortium.org/), the Greater Middle East Variome (http://igm.ucsd.edu/gme/), the Brazilian Genomic Variants Database (http://abraom.ib.usp.br/), RADAR (http://rnaedit.com/), GENCODE (v.19) (https://www.gencodegenes.org/human/release_19.html), the hs37d5 reference genome (https://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/phase2_reference_assembly_sequence/), ERCC spike-in sequence (https://www.encodeproject.org/files/ENCFF908UQN/) and Caltech profile 3 spike-in sequence (https://www.encodeproject.org/references/ENCSR193ZXE/). snRNA-seq data from the developing human cerebellum were obtained through correspondence from ref. [13] and are available through the Human Cell Atlas (https://www.covid19cellatlas.org/aldinger20), the UCSC Cell Browser (https://cbl-dev.cells.ucsc.edu) or from Database of Genotypes and Phenotypes (dbGaP) (accession number phs001908.v2.p1). Bulk RNA-seq data from the developing human cerebellum were obtained through correspondence from ref. [12] and are available through the dbGaP (accession number phs001908.v2.p1). Source data are provided with this paper.

Code availability

No custom code was used in this study. Only open-source algorithms were used, and their applications are detailed in the Methods. Further details on how these algorithms were used are available from the corresponding author upon request.

46. Milde, T. et al. HD-MB03 is a novel group 3 medulloblastoma model demonstrating sensitivity to histone deacetylase inhibitor treatment. *J. Neurooncol.* **110**, 335–348 (2012).
47. Dietl, S. et al. MB3W1 is an orthotopic xenograft model for anaplastic medulloblastoma displaying cancer stem cell- and group 3-properties. *BMC Cancer* **16**, 115 (2016).
48. Choi, H. et al. SAINT: probabilistic scoring of affinity purification–mass spectrometry data. *Nat. Methods* **8**, 70–73 (2011).
49. Teo, G. et al. SAINTexpress: improvements and additional features in significance analysis of INTeractome software. *J. Proteomics* **100**, 37–43 (2014).
50. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
51. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
52. Suzuki, H. et al. Recurrent noncoding U1 snRNA mutations drive cryptic splicing in SHH medulloblastoma. *Nature* **574**, 707–711 (2019).
53. Auwera, G. A. et al. From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).
54. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
55. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
56. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.* **17**, 128 (2016).
57. Haas, B. J. et al. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol.* **20**, 213 (2019).
58. Okonechnikov, K. et al. InFusion: advancing discovery of fusion genes and chimeric transcripts from deep RNA-sequencing data. *PLoS ONE* **11**, e0167417 (2016).
59. Robertson, G. et al. De novo assembly and analysis of RNA-seq data. *Nat. Methods* **7**, 909–912 (2010).
60. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
61. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
62. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
63. Patel, A. P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
64. Khazaei, S. et al. H3.3 G34W promotes growth and impedes differentiation of osteoblast-like mesenchymal progenitors in giant cell tumor of bone. *Cancer Discov.* **10**, 1968–1987 (2020).
65. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
66. Angerer, P. et al. destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics* **32**, 1241–1243 (2016).
67. Aran, D. et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).
68. Kowalczyk, M. S. et al. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res.* **25**, 1860–1872 (2015).
69. Vanner, R. J. et al. Quiescent Sox2$^+$ cells drive hierarchical growth and relapse in Sonic hedgehog subgroup medulloblastoma. *Cancer Cell* **26**, 33–47 (2014).

70. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
71. Van de Sande, B. et al. A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat. Protoc.* **15**, 2247–2276 (2020).
72. Nagy, C. et al. Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. *Nat. Neurosci.* **23**, 771–781 (2020).
73. Gulati, G. S. et al. Single-cell transcriptional diversity is a hallmark of developmental potential. *Science* **367**, 405–411 (2020).
74. Van den Berge, K. et al. Trajectory-based differential expression analysis for single-cell sequencing data. *Nat. Commun.* **11**, 1201 (2020).
75. Therneau, T. A Package for Survival Analysis in R. R package version 3.4-0, https://CRAN.R-project.org/package=survival (2022).
76. Wei, Y. et al. A TAF4-homology domain from the corepressor ETO is a docking platform for positive and negative regulators of transcription. *Nat. Struct. Mol. Biol.* **14**, 653–661 (2007).
77. Szklarczyk, D. et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605–D612 (2021).
78. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment Map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS ONE* **5**, e13984 (2010).
79. Shannon, P. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
80. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
81. Zhou, X. et al. Exploring genomic alteration in pediatric cancer using ProteinPaint. *Nat. Genet.* **48**, 4–6 (2016).
82. Penas, C. et al. Time series modeling of cell cycle exit identifies Brd4 dependent regulation of cerebellar neurogenesis. *Nat. Commun.* **10**, 3028 (2019).
83. Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* **35**, W193–W200 (2007).

Acknowledgements M.D.T. is supported by the NIH (R01NS106155, R01CA159859 and R01CA255369), The Pediatric Brain Tumor Foundation, The Terry Fox Research Institute, The Canadian Institutes of Health Research, The Cure Search Foundation, the Matthew Larson Foundation (IronMatt), b.r.a.i.n.child, Meagan's Walk, the SWIFTY Foundation, The Brain Tumor Charity, Genome Canada, Genome BC, Genome Quebec, the Ontario Research Fund, Worldwide Cancer Research, the V-Foundation for Cancer Research and the Ontario Institute for Cancer Research through funding provided by the Government of Ontario. M.D.T. is also supported by a Canadian Cancer Society Research Institute Impact grant, a Cancer Research UK Brain Tumor Award and by a Stand Up To Cancer (SU2C) St Baldrick's Pediatric Dream Team Translational Research Grant (SU2C-AACR-DT1113) and SU2C Canada Cancer Stem Cell Dream Team Research Funding (SU2C-AACR-DT-19-15) provided by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research, with supplementary support from the Ontario Institute for Cancer Research through funding provided by the Government of Ontario. Stand Up to Cancer is a programme of the Entertainment Industry Foundation administered by the American Association for Cancer Research. M.D.T. is also supported by the Garron Family Chair in Childhood Cancer Research at the Hospital for Sick Children and the University of Toronto. K.J.M. is supported by NIH-R01-NS080390, R01-NS095733 and R37 NS095733. T.E.W.-O. and B. Doble are supported by the Canadian Institutes of Health Research (CIHR) and the CancerCare Manitoba Foundation. H.S. is supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (21K21001) and National Cancer Center Research and Development Funds (2021-A-1). C.L.K is supported by CIHR (PJT-156086), salary awards from Fonds de Recherche du Québec-Santé (FRQS) and NSERC (RGPIN-2016-04911). P. Haldipur is supported by NIH-R21 NS117848, Brain and Behavior Young Investigator Award. We also thank the staff at Human Developmental Biology Resource London and Newcastle for their technical help and support; B. Pakuts for support with graphic design and illustrations; the Fernando Goldsztein Family for supporting this research; and L. Rorke for thoughtful discussions on cerebellar dysplasia.

Author contributions L.D.H., P. Haldipur and O. Saulnier designed, performed and analysed the majority of the experiments in this study. J.M., A.H.S. and J.G. contributed to in situ hybridization and cell count analyses. A.W.E. contributed to bioinformatics analysis and SCENIC workflow. W.O. and D.S.S. contributed to cell line preparation for single-nucleus sequencing. W.O. and M.C.V. contributed to primary tissue isolation, single-cell preparation and sequencing. V.G. and L.C.-M. contributed to TurboID experiments and OTX2-KD experiments. A.L.M. contributed to GCP isolation and design and overexpression of the DN-CBFA2T3 construct. M.S. and T.P.-B. contributed to fusion CBFA2T2-TurboID construct design. M.L., R.A.S. and D. S. Scott contributed to *CBFA2T2* mutation validation. S.B. contributed to western blot validations. H.F. contributed to SNP6 copy number analyses. T.N. and S.N. contributed to mutation analysis. A.B. contributed to pathway analysis. J.J.D.-M. contributed to CNV analysis of the scRNA-seq data. K.B. contributed to analysis and model design. P.S. and E.Y.W. contributed to RNA-seq fusion calling and analysis. O. Sirbu, S.A.K., P.B., A.V., J.J.Y.L., R.A., Xin Chen, Xiaodi Chen and K.L.M. contributed to experimental, statistical, bioinformatics analysis and/or interpretation. B. Luu, P. Bérubé, Y.C.W., A.S.M. and X.W. contributed to RNA-seq library preparation. S.M.P., N.J., S.-K.K., F.D., O.D., F.B., J.M.-P., W.A.G., J.L., P.D., M.F.-M., A.J., P.J.F., J.M.K., K.Z., S.D.B., C.G.E., A.A.N.R., C.G., J.M.O., M.G., P. Hauser, J.J.P., Y.S.R., C.d.T., J. Mora, K.K.W.L., H.-K.N., W.S.P., I.F.P., E.L.-A., G.Y.G., T.E.V.M., T.S., R.V., R.C.T., M.K.C., J.B.R., T.K., S.J., B. Lach, A.L., V.F., P.d.A., M.Z., G.C., S.R., D. S. Stearn, E.G.V.M., P.P., G.F., M.M., C.G.C., C.C.F., M.F.R., F. Boop, J.A.C., R.E.M., E.M.T., M.A., M.L.G., F.C., P.E. and M.P. provided patient tumour material and helped design the study. K.A.A., A.S.M., F.M.G.C., J.N.R., S.J.M.J., R.A.M., M.A.M., X.H., J.R., P.H.S., R.J.W.-R., W.A.W., T.J.P., L.G., C.L.K., L.D.S., N.J., D.M. and O.A. helped analyse data and provided expert advice. F.R. and E.S. provided late-gestation human samples. X.W. and C.D. provided reagents, equipment and expert advice. D.W.E. and J.A.G. provided expert neuropathology analysis and images. T.E.W.-O. and B.D. contributed to and provided expert advice on TurboID and OTX2-KD experiments. V.R. provided MRI images and provided expert advice on their interpretation. H.S., T.E.W.-O., K.J.M. and M.D.T. jointly supervised the project. L.D.H., P.H. and M.D.T. prepared the figures and wrote the manuscript.

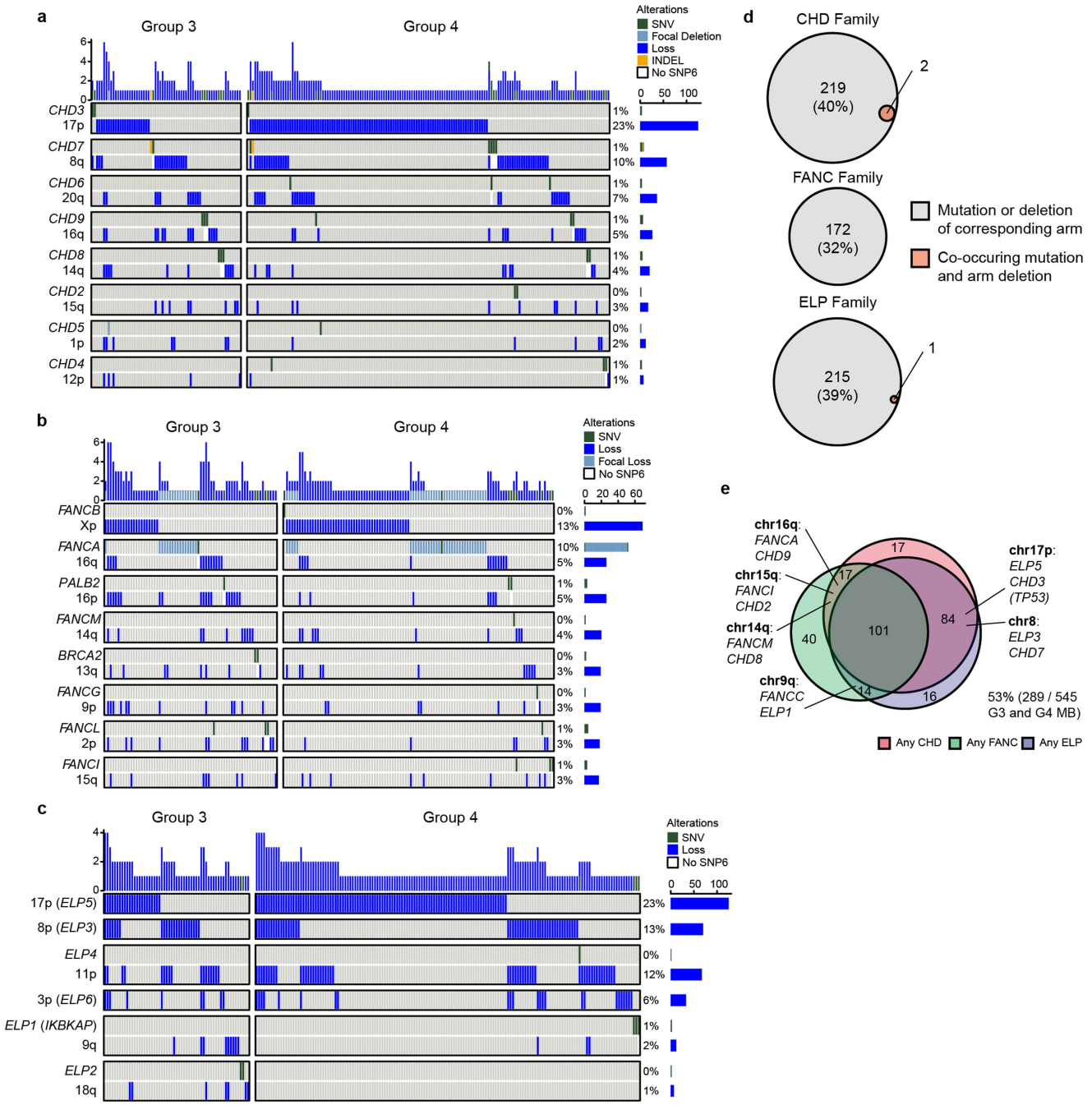
Competing interests The authors declare no competing interests.

Additional information

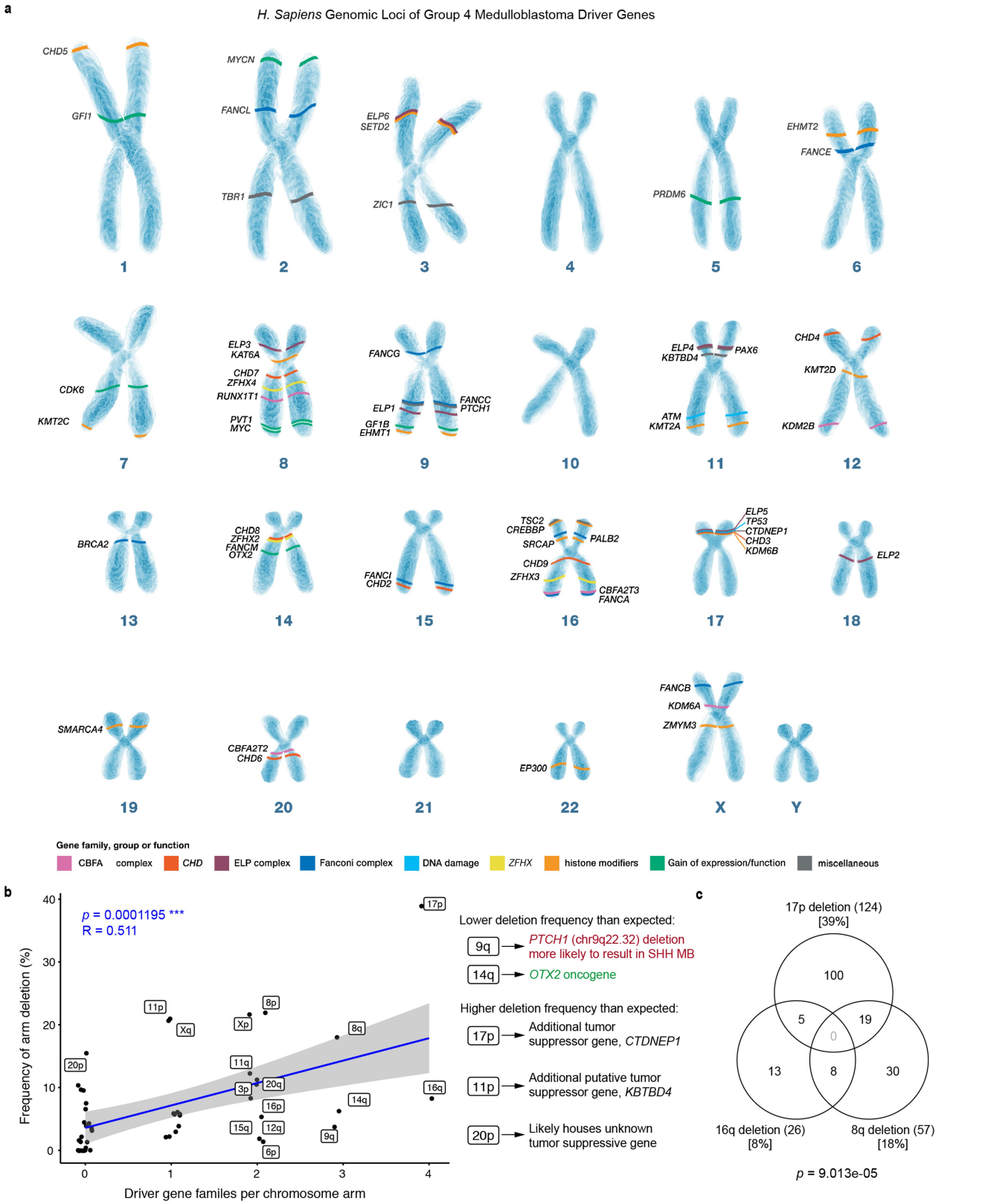
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-022-05215-w.

Correspondence and requests for materials should be addressed to Michael D. Taylor.

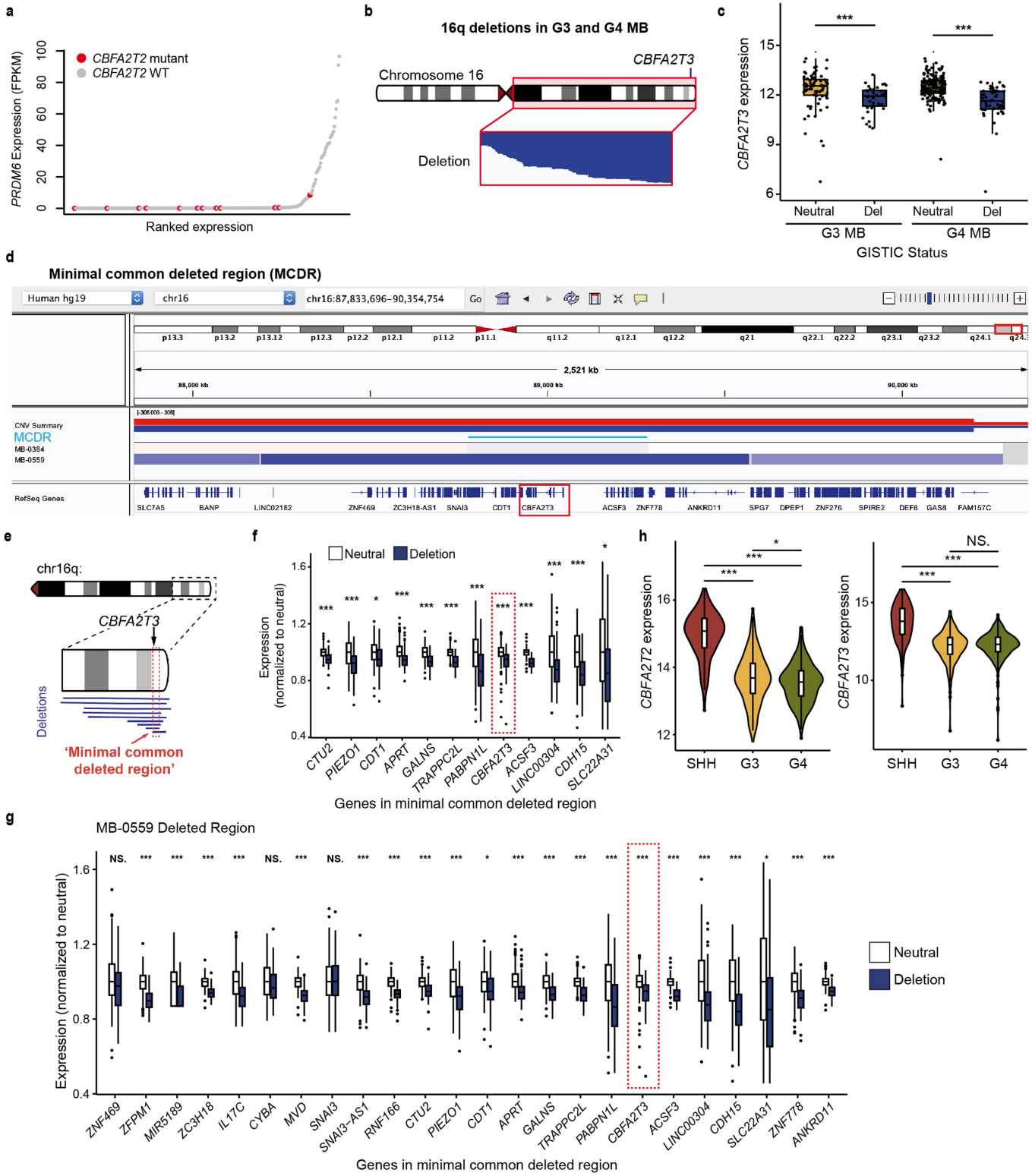
Peer review information *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work.



Extended Data Fig. 1 | Mutations and copy number losses targeting the same genes are mutually exclusive in G3 and G4 MB. a, b, c, Oncoprint summarizing gene mutations and copy number losses in their corresponding genomic loci (chromosome arm) for CHD (**a**), FANC (**b**) and ELP (**c**) driver families. Mutations are less frequent than copy number losses but tend to occur independently suggesting they are targets of the deletions. **d**, Mutual exclusivity of mutation and copy number loss events targeting genes in the CHD, FANC, and ELP families. **e**, Overlap between events targeting CHD, FANC, or ELP genes. Most G4 MB tumours are haploinsufficient for genes in at least two families through a single deletion event.

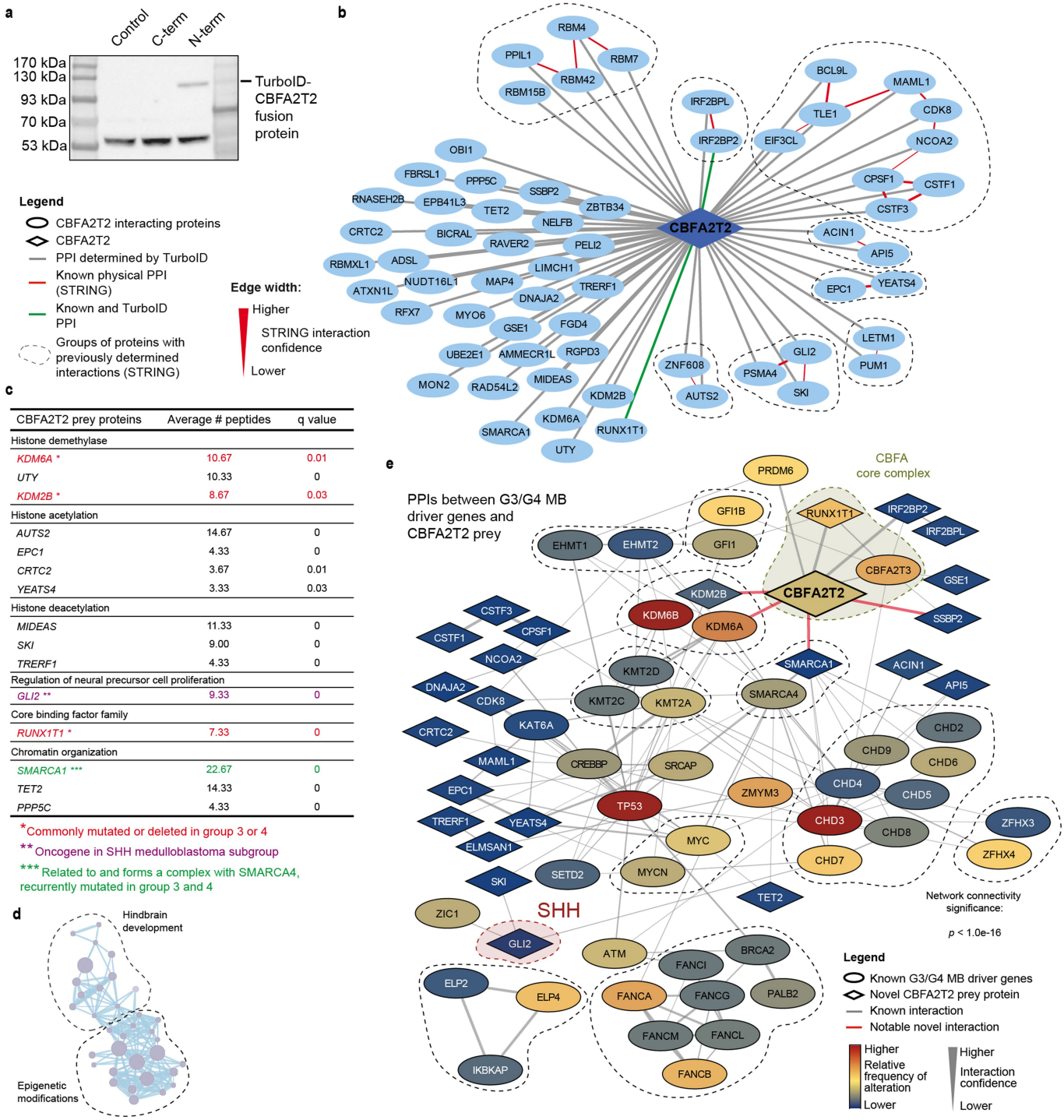


Extended Data Fig. 2 | Clustering of G4 MB driver genes in the human genome predisposes humans to develop MB. a, Cartoon of the *Homo sapiens* genome with the locations of known and newly identified G3 and G4 MB candidate driver genes demonstrating clustering of genes at locations known to be deleted in G3 and G4 MB. **b**, Frequency of whole chromosome arm loss is significantly correlated with the number of driver gene families—as detailed in (**a**)—contained on the arm. Significance was assessed by a two-sided linear regression model; grey shaded area denotes the 95% confidence interval. **c**, Mutual exclusivity of copy number losses of chromosome arms 17p, 16q, and 8q. Significance was assessed using the impurity test for mutual exclusivity, implemented in the R package DISCOVER[41].

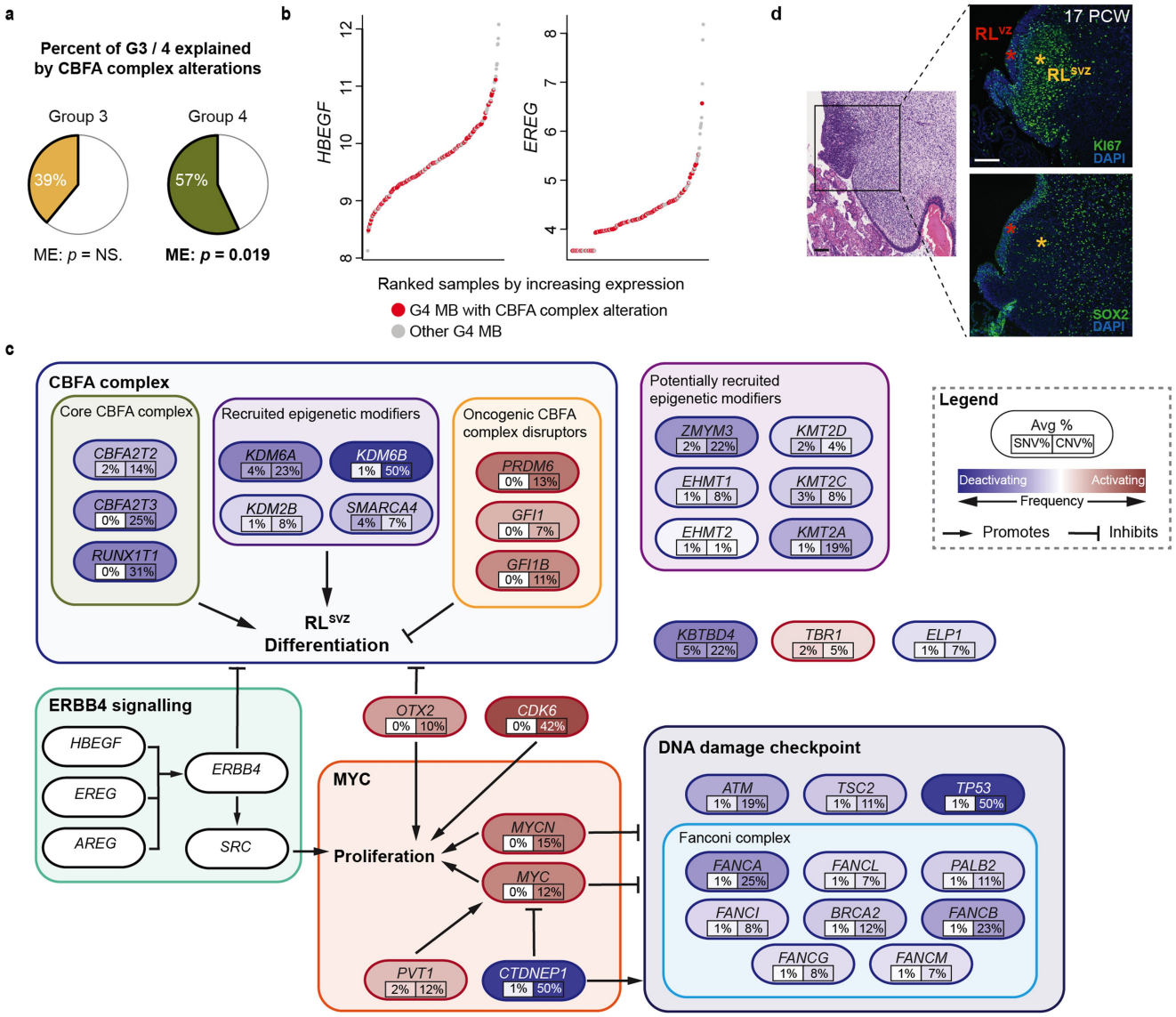


Extended Data Fig. 3 | See next page for caption.

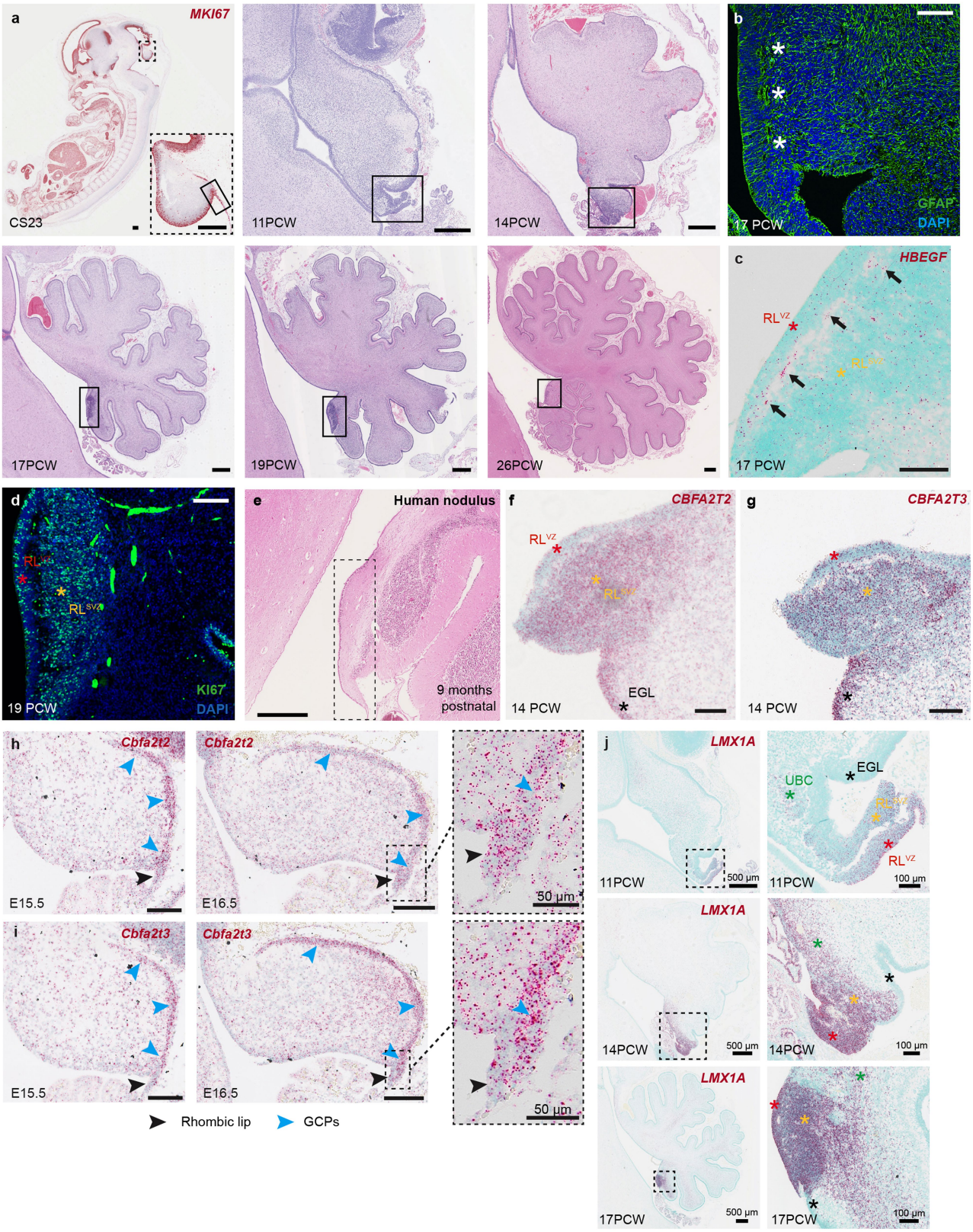
Extended Data Fig. 3 | *CBFA2T3* is a G4 MB tumour suppressor gene. **a**, *PRDM6* expression in *CBFA2T2* mutant (red) and *CBFA2T2* WT (grey) G3 and G4 MB samples demonstrates that enhancer hijacking mediated *PRDM6* expression is largely limited to *CBFA2T2* WT cases. **b**, Density of regions of chromosomal gain and loss along human chromosome 16q in G3 and G4 MB cases, demonstrating that deletions are biased towards the telomeric end of 16q, the location of known drivers, particularly *CBFA2T3*. **c**, *CBFA2T3* expression differences between samples with and without *CBFA2T3* deletions, split by subgroup. Statistical significance was assessed using Kruskal-Wallis rank-sum test (FDR < 0.05), *** $p < 0.0005$, G3, $p = 2.88e^{-05}$; G4, $p = 2.60e^{-09}$. G3, $n = 112$; G4, $n = 206$. *CBFA2T3* is a copy-number responsive tumour suppressor gene in G4 MB. **d**, IGV analysis showing focal deleted region in two G4 MB samples MB-0364 and MB-0559. MB-0364, which is the minimal common deleted region (MCDR) on 16q in G3 and G4 MB, though does not quite achieve statistical significance in the GISTIC analysis. MB-0559 is the MCDR achieving statistical significance in GISTIC analysis. *CBFA2T3* is identified with a red box. **e**, Cartoon illustrating the MCDR concept. **f**, Expression differences between copy neutral or hemizygously deleted G3 and G4 MB samples for genes within the MB-0364 MCDR on chr16q24.3. Statistical significance was assessed using two-sided Mann-Whitney U tests with FDR adjustment, * $p < 0.05$, *** $p < 0.0005$. Deletion, $n = 86$; Neutral, $n = 232$. **e**, Expression differences between copy neutral or hemizygously deleted G3 and G4 MB samples for genes within the MB-0559 MCDR on chr16q24.3. Statistical significance was assessed using two-sided Mann-Whitney U tests with FDR adjustment, * $p < 0.05$, *** $p < 0.0005$. Deletion, $n = 86$; Neutral, $n = 232$. A full list of p values for genes presented in **(f)** and **(g)** can be found in Supplementary Data Table 1. **h**, *CBFA2T2* (left) and *CBFA2T3* (right) expression in SHH, G3, and G4 MB by bulk RNAseq. Statistical significance was assessed by Kruskal-Wallis rank-sum test (FDR < 0.05), * $p < 0.05$, *** $p < 0.0005$. For *CBFA2T2*: SHH-G3, $p = 2.29e^{-47}$; SHH-G4, $p = 4.42e^{-73}$; G3-G4, $p = 0.035$. For *CBFA2T3*: SHH-G3, $p = 7.10e^{-42}$; SHH-G4, $p = 1.13e^{-46}$; G3-G4, $p = 0.61$. G3, $n = 219$; G4, $n = 326$; SHH, $n = 250$. While *CBFA2T2* and *CBFA2T3* are recurrently targeted and have low expression in G3 and G4 MB, high expression of both genes and an absence of alterations are observed in SHH MB. *CBFA2T2* and *CBFA2T3* likely have different roles in SHH MB compared to G3 and G4 MB. For **c**, **f**, **g**, and **h** box plots show the median and interquartile range, and whiskers show the data range. Points outside this range are outliers and are plotted individually.



Extended Data Fig. 4 | The CBFA polyprotein complex contains multiple known and novel G4 MB driver genes. a, Western blot showing successful expression of the TurboID-CBFA2T2 fusion protein when the TurboID construct is fused to the N-terminal of CBFA2T2, but not to the C-terminal. **b**, Protein-protein interaction (PPI) network of novel CBFA2T2 interacting proteins. Each node represents a protein and edges between the proteins represent known or novel PPIs. Edges in red represent known interactions between CBFA2T2 interacting proteins, and edges in green represent known interactions with CBFA2T2 that were recapitulated in our TurboID screen. Proteins are grouped with dashed lines if they contain known interactions between each other. **c**, Significant CBFA2T2 prey proteins enriched in each indicated biological process. GLI2 is a SHH oncogene and has been recently shown to maintain GCP proliferation and identity, implicating the CBFA complex[82]. **d**, Enrichment map of biological processes (GO:BP) enriched in CBFA2T2 prey proteins by TurboID. Each node represents a significantly enriched pathway and edges represent shared genes between nodes. Nodes are grouped and labelled with a common biological theme. Significance was assessed using G:Profiler[83] with FDR correction. **e**, Protein-protein interaction (PPI) network of CBFA2T2 TurboID proteins and G3/G4 MB driver genes (Fig. 1a) using STRING[41]. Edges between CBFA2T2 and diamond-shaped nodes are not drawn for simplicity. Connectivity significance was assessed by STRING, $p < 0.1e^{-16}$.

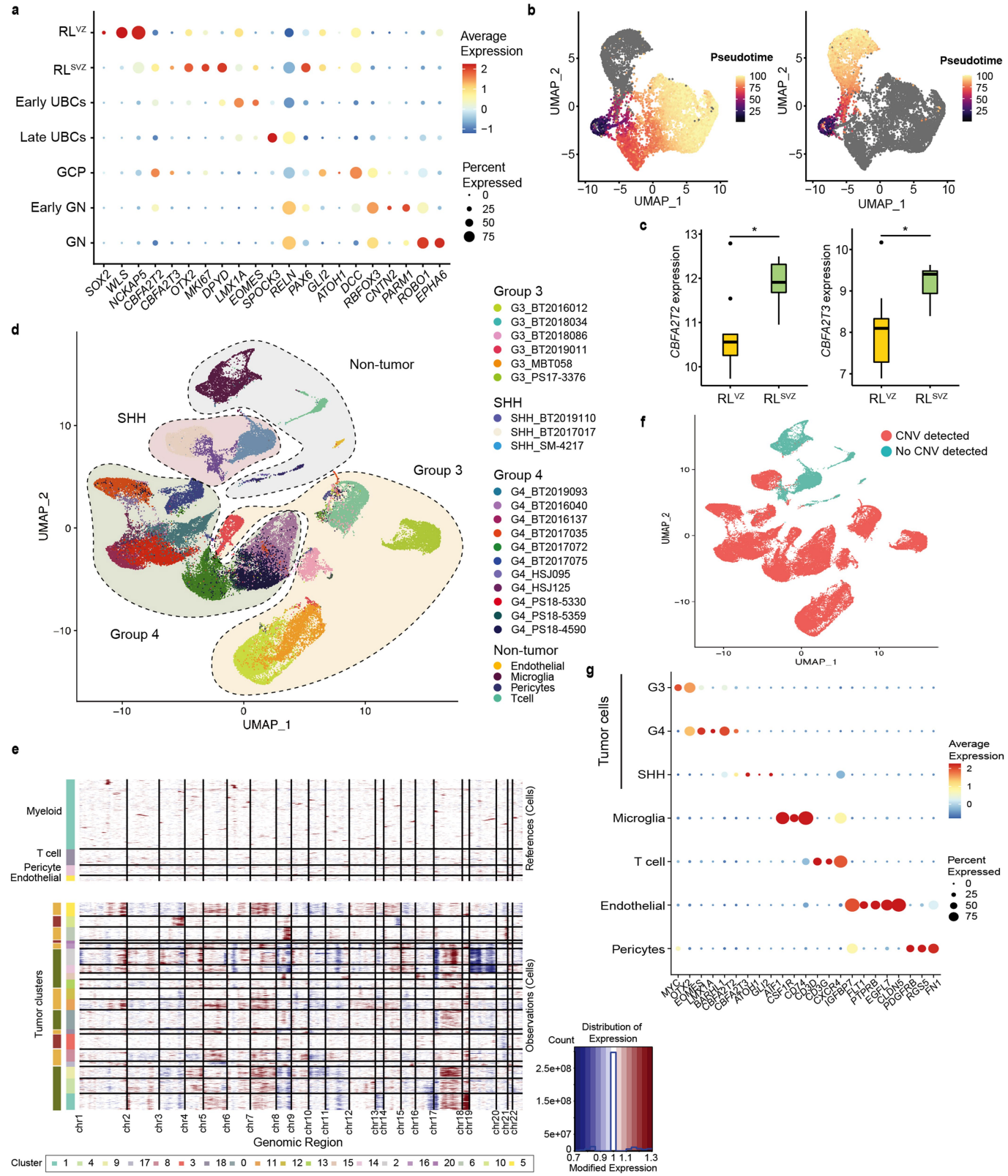


Extended Data Fig. 5 | Disruption to the CBFA complex explains most G4 MB tumours. a, Per cent of G3 and G4 MB in our cohort ($n = 545$) explained by alterations in genes connected to *CBFA2T2* with a known or novel PPI (one step in the network presented in Extended Data Fig. 4e). Significance assessed using the impurity test for mutual exclusivity implemented in the R package DISCOVER[41]. **b**, Ranked expression of *HBEGF* (left) and *EREG* (right) in G4 MB ($n = 326$). Points are coloured by the presence (red) or absence (grey) of known CBFA complex alterations. Samples with the highest expression of *HBEGF* and *EREG* typically do not have CBFA complex alterations, suggesting an alternate mechanism of CBFA complex inhibition. Data presented in **a** were not performed in replicates. **c**, Summary of disrupted pathways in G3 and G4 MB. Altered genes are grouped by pathway and labelled with alteration frequency. **d**, (Left) H&E–stained midsagittal section from 17 PCW human cerebellum. (Right) Fluorescence immunohistochemistry (IHC) showing *KI67* and *SOX2* expression in the human RL compartments. Data presented in **d** were not performed in replicates. RL^{VZ} and RL^{SVZ} are denoted by red and yellow asterisks, respectively. Scale bars: 100 µm.



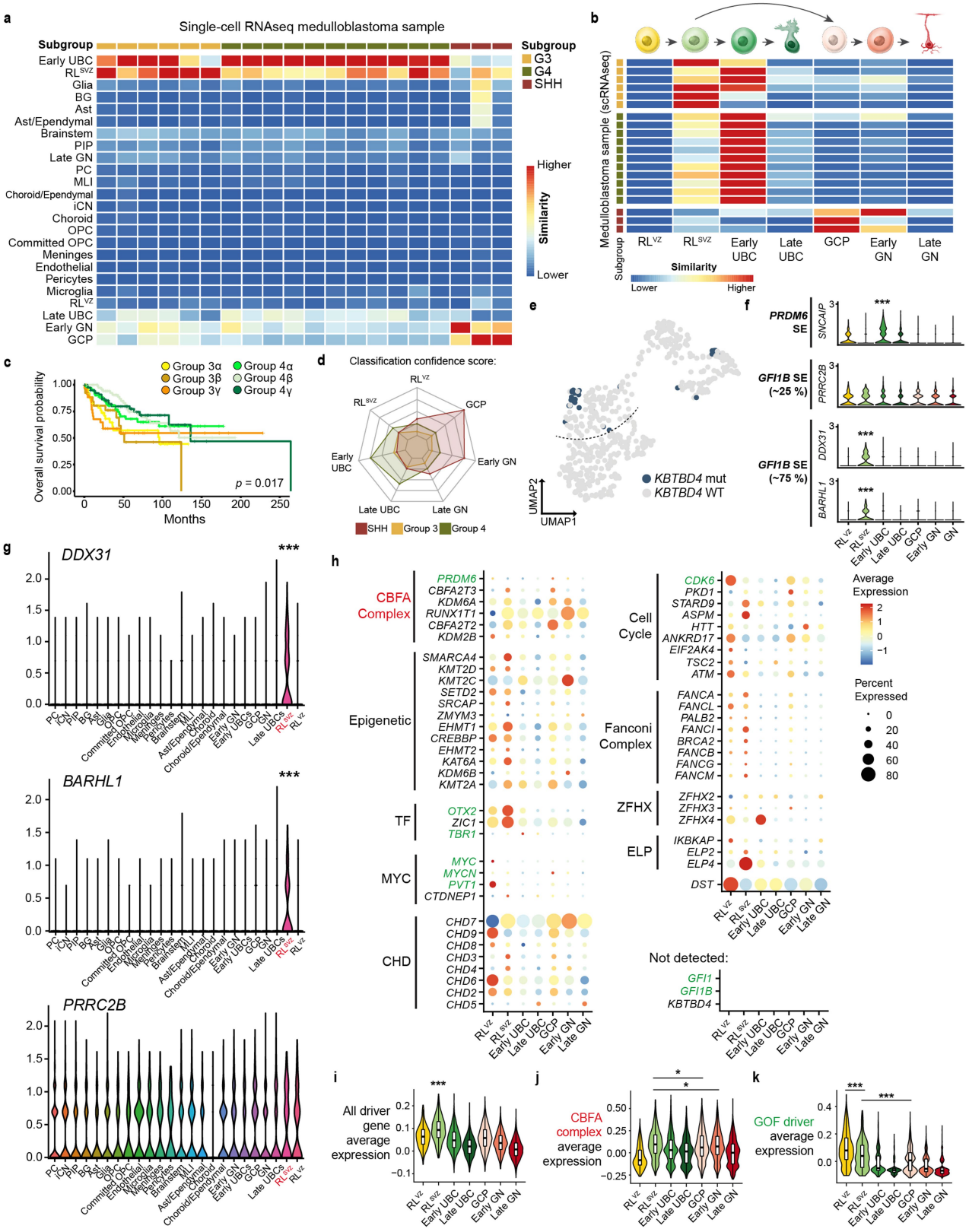
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | *LMX1A* expression distinguishes the two downstream lineages of the RL^{SVZ}. a, (Top left) *In situ* hybridization (ISH) showing *MKI67* expression. In-set highlights the developing cerebellum, and the RL is indicated by the black box. (Other images) Hematoxylin and eosin (H&E)–stained midsagittal sections of the developing human cerebellum. In each, the rhombic lip is indicated by the black box. Scale bars: 500 µm. **b**, GFAP expression in the developing human RL at 17 PCW. Scale bar: 100 µm. The RL^{VZ} and RL^{SVZ} are physically divided by a vascular plexus, as indicated with white asterisks. **c**, ISH showing spatially resolved RNA expression of *HBEGF* in the developing human cerebellum at 17 PCW. Scale bar: 50 µm. *HBEGF* foci are enriched along the RL vascular plexus. **d**, KI67 expression in the developing human RL at 19 PCW. Scale bar: 100 µm. **e**, H&E–stained midsagittal sections of the 9-month postnatal human cerebellum. Scale bar: 500 µm. The RL is only present during gestation and disappears around birth. **f, g**, ISH showing spatially resolved RNA expression of *CBFA2T2* (**f**) and *CBFA2T3* (**g**) in the developing human cerebellum at 14 PCW. Scale bars: 100 µm. **h, i**, ISH showing spatially resolved RNA expression of *Cbfa2t2* (**h**) and *Cbfa2t3* (**i**) in the developing mouse cerebellum at E15.5 (Left) and E16.5 (Right). Scale bars: 100 µm. We do not observe a similar expression pattern of either gene in the mouse RL as we do in the human RL, and note an enrichment of expression in the EGL, similar to humans. **j**, ISH showing spatially resolved RNA expression of *LMX1A* in the developing human cerebellum at 11, 14, and 17 PCW. *LMX1A* is highly expressed in both the RL^{VZ} and RL^{SVZ}, but *LMX1A* expression is only retained in UBCs migrating away from the RL and is completely absent in GCPs that migrate to the EGL. Data presented in **d** is a representative image from three independent experiments with similar results, data in remaining panels were not performed in replicates.



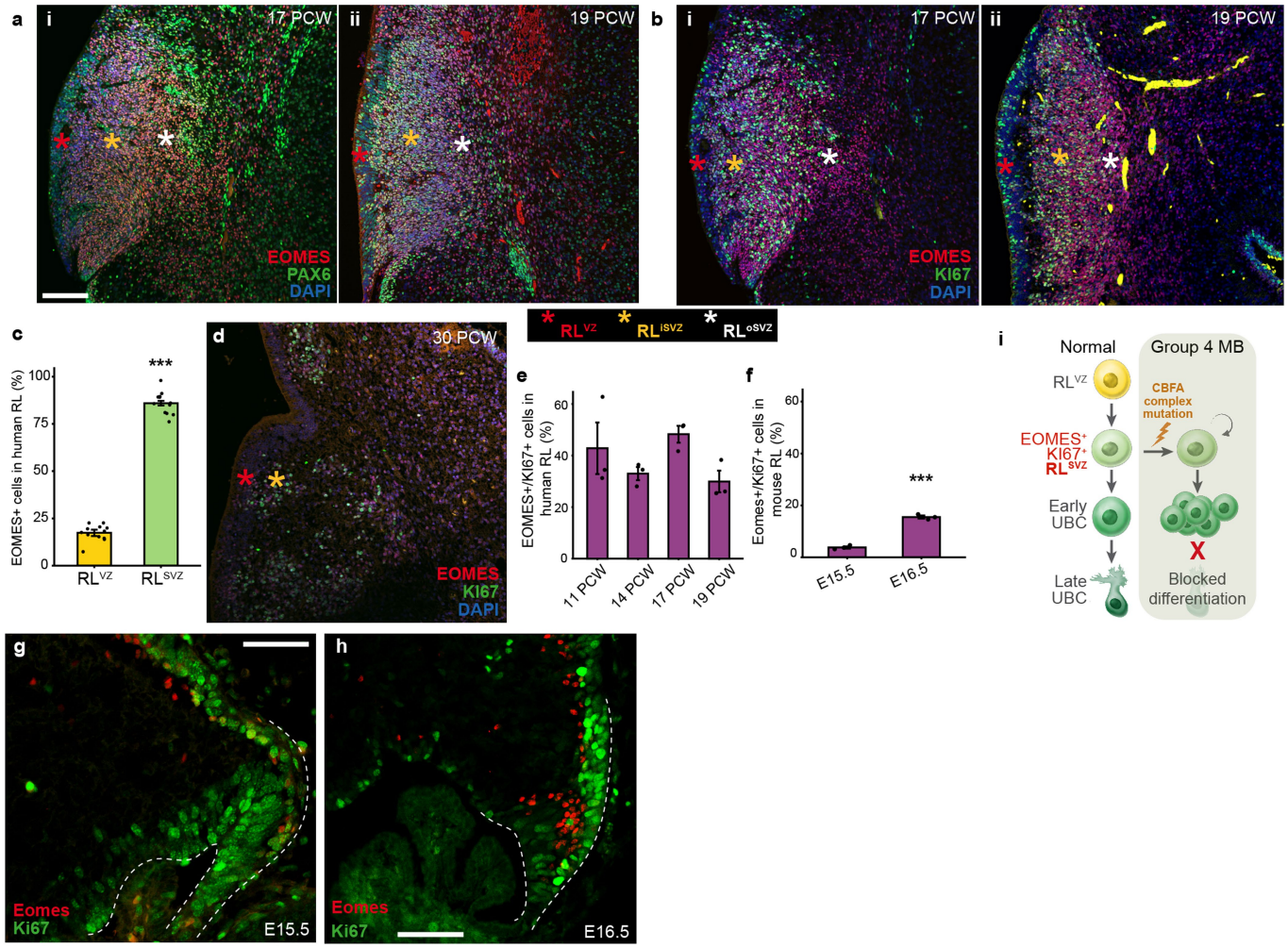
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Characterization of single cells used in transcriptional mapping between MB and human cerebellum development. a, Dot plot showing expression of characteristic marker genes across RL glutamatergic cell types in the developing human cerebellum[13]. **b**, UMAP embeddings coloured by pseudotime inferred from Slingshot[65], where the direction of pseudotime is from dark to light colours, for the granule cell lineage (Left) and the UBC lineage (Right). **c**, Expression of *CBFA2T2* (Left) and *CBFA2T3* (Right) in each zone of the developing human RL by bulk RNAseq[12]. Statistical significance was assessed using a two-sided Mann-Whitney U test, * $p < 0.05$; *CBFA2T2*, $p = 0.0078$; *CBFA2T3*, $p = 0.0056$. $n = 9$ biological samples, per zone, acquired between 9 and 19 PCW. Box plots show the median and interquartile range, and whiskers show the data range. Points outside this range are outliers and are plotted individually. **d**, UMAP embedding of 63,296 single cells derived from G3 ($n = 6$), G4 ($n = 11$), and SHH ($n = 3$) MB scRNAseq samples. Clusters of transcriptionally similar cells are colored and labeled by tumour sample or annotated cell type for non-tumour cells. **e**, Copy number variations detected in single cells inferred using inferCNV[63]. (Top) Reference non-tumour cells are devoid of copy number variations. (Bottom) Tumour cell clusters were enriched for copy number variations characteristic of the sample subgroup. Cells containing CNVs were assigned as tumour cells for downstream analysis. **f**, UMAP embedding as in **(d)** coloured by the detection of copy number variations. **g**, Dot plot showing expression of characteristic marker genes of SHH, G3, G4 MB, and the non-tumour cell types identified. For **a**, **g**, colour indicates average expression and size of each dot indicates the per cent of cells in that cluster that express the genes.

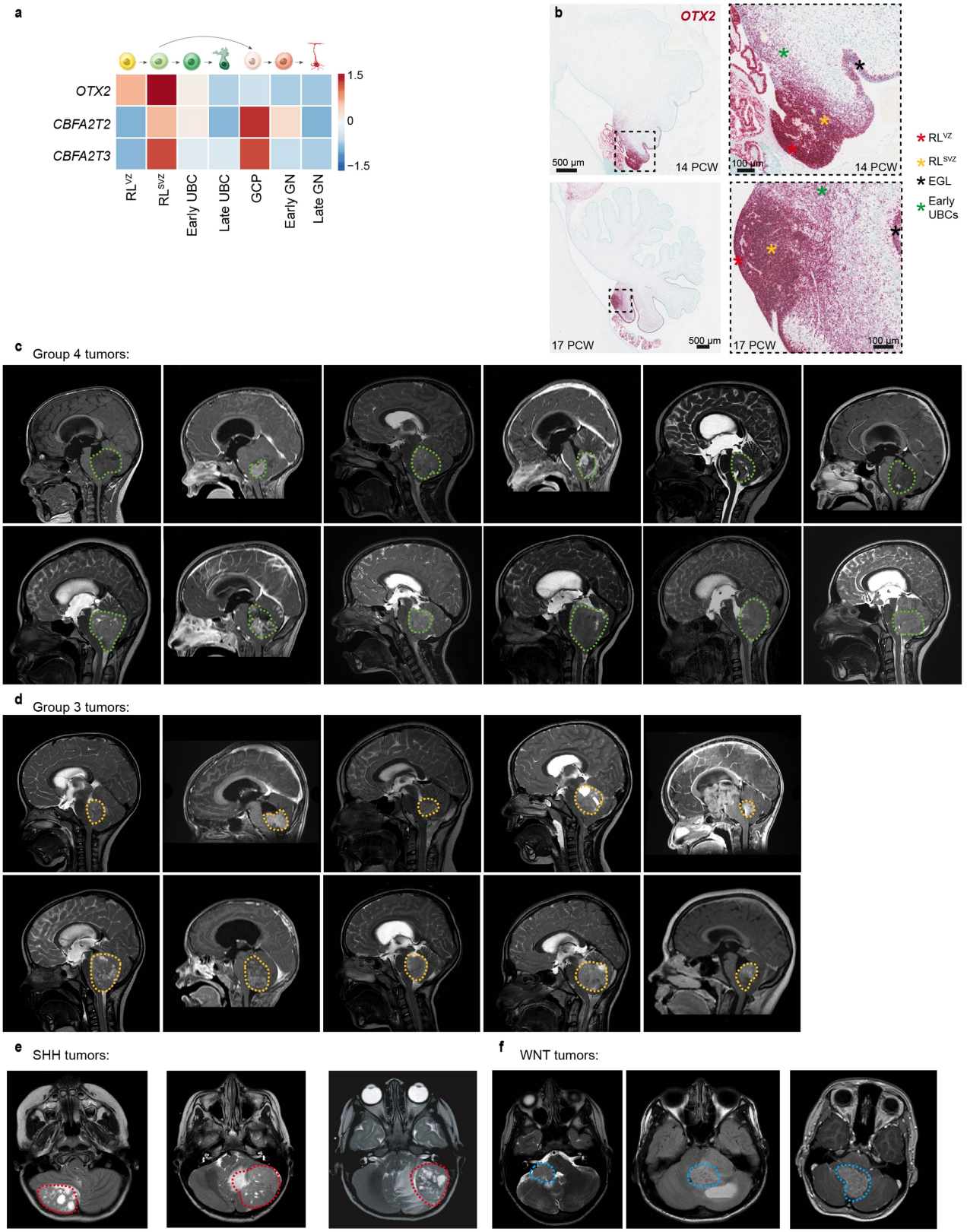


Extended Data Fig. 8 | See next page for caption.

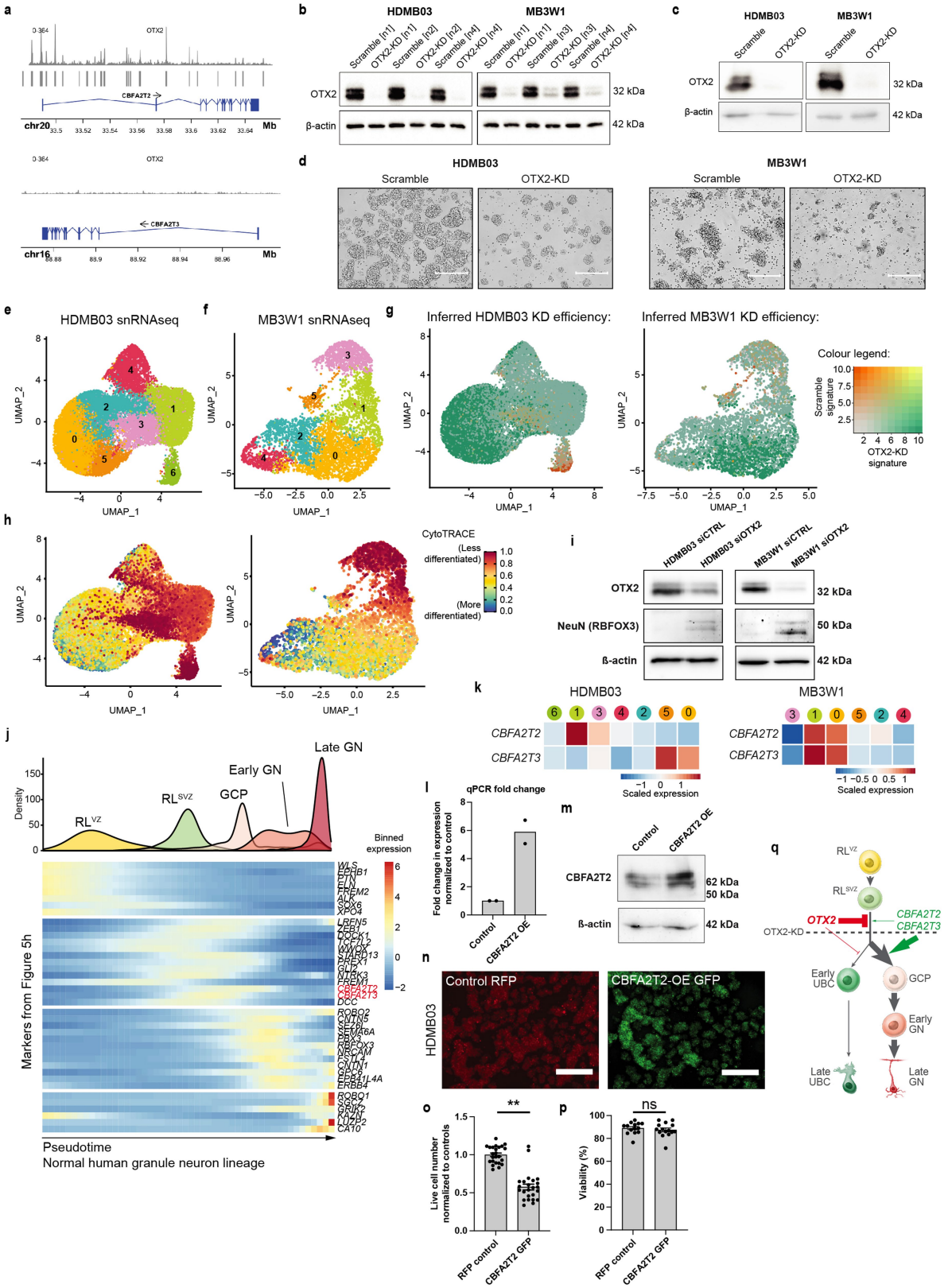
Extended Data Fig. 8 | G3 and G4 MB resemble specific components of the human RL, whose differentiation is stalled in time. a, SingleR[67] classification of tumour cells from G3 ($n = 6$), G4 ($n = 11$), and SHH ($n = 3$) MB scRNAseq samples, by comparison to the entire developing human cerebellum[13]. As expected, MB cells are most similar to glutamatergic cells. **b**, SingleR classification of tumour cells from **(a)** by comparison to glutamatergic cell types. **c**, Kaplan-Meier plot showing overall survival of G3 and G4 MB subtypes in the current dataset. Significance assessed using a log-rank test. Censored cases, +. **d**, Relative confidence of per cell classifications, calculated as the average similarity score per subgroup **(b)**, minus median similarity scores from other subgroups per cell type. **e**, UMAP embedding of $n = 545$ G3 and G4 MB bulk RNAseq samples, coloured by presence of *KBTBD4* mutations. $G4_\chi$ which are impoverished for CBFA complex mutations, and display high *OTX2* expression, are also enriched for *KBTBD4* mutations. **f**, Expression of super enhancer (SE) genes in the developing human cerebellum snRNA-seq data. These gene promoters have been demonstrated to promote transcription of *PRDM6* (*SNCAIP*) and *GFI1B* (*DDX31/BARHL1* and *PRRC2B*) in G3 and G4 MB secondary to enhancer hijacking events[11,28]. Significance was assessed using a two-sided Wilcoxon Rank Sum test with FDR correction, *** $p < 0.0005$. *SNCAIP*, $p = 3.39e^{-261}$; *DDX31*, $p = 4.27e^{-71}$; *BARHL1*, $p = 4.91e^{-40}$. $n = 9{,}208$ cells. **g**, Expression of *DDX31*, *BARHL1*, and *PRRC2B* across all cell types in the developing human cerebellum. *DDX31* and *BARHL1* exhibit correlated expression specific to the RL^{SVZ}, while *PRRC2B* is non-specifically expressed. Significance was assessed using a two-sided Wilcoxon Rank Sum test with FDR correction, *** $p < 0.0005$. *DDX31*, $p = 3.66e^{-113}$; *BARHL1*, $p = 6.26e^{-191}$. $n = 59{,}608$ cells. **h**, Expression of G3 and G4 MB driver genes (from Fig. 1a) in the developing human cerebellum snRNA-seq data. **i, j, k**, Average expression of all G3 and G4 MB driver genes **(i)**, CBFA complex genes **(j)**, and gain of function (GOF) driver genes **(k)** in the developing human cerebellum snRNA-seq data. Significance was assessed using a two-sided Wilcoxon Rank Sum test, * $p < 0.05$, *** $p < 0.0005$. $n = 9{,}208$ cells. **(i)**, $p = 1.5e^{-13}$; **(j)** GCP, $p = 0.0085$; Early GN, $p = 0.047$; **(k)** RL^{SVZ}, $p = 1.9e^{-05}$; GCP, $p = 3.2e^{-13}$. For **i**, **j**, and **k**, box plots show the median and interquartile range, and whiskers show the data range. Points outside this range are outliers and are plotted individually.



Extended Data Fig. 9 | Human EOMES^{+ve} RL cells are mitotically active UBC progenitors. a, EOMES and PAX6 expression in the developing human RL at 17 (**i**) and 19 (**ii**) PCW. Scale bars: 100 µm. **b**, EOMES and KI67 expression in the developing human RL at 17 (**i**) and 19 (**ii**) PCW. Scale bars as in **a**. Proliferating EOMES^{+ve} UBC progenitors are common across all developmental timepoints assessed. **c**, EOMES+ cells in the human RL zones. The RL^{SVZ} contains significantly more EOMES+ cells that the RL^{VZ}. Significance was assessed using an unpaired two-tailed t-test, *** $p = 1.048e^{-18}$. $n = 3$ biological repeats, per N = 4 time points; error bars, SEM. **d**, EOMES and KI67 expression in the developing human RL at the late timepoint 30 PCW. Scale bar as in **(a)**. Proliferating EOMES^{+ve} UBC progenitors can be found across fetal development, though at reduced frequency at later time points as KI67 expression is reduced (Fig. 3c). **e**, Quantification of the number of EOMES+/KI67+ cells in the human RL across various developmental timepoints. All comparisons to 11 PCW were non-significant using two-tailed unpaired t-tests; 14PCW, $p = 0.43$; 17PCW, $p = 0.65$; 19PCW, $p = 0.33$. $n = 3$ biological repeats per timepoint; error bars, SEM. EOMES+/KI67+ UBC progenitors are a long-lived and dominant population of the RL, rather than a transient state preceding differentiation. **f**, Quantification of the number of Eomes+/Ki67+ cells in the mouse RL across various developmental timepoints. Significance was assessed using an unpaired two-tailed t-test, *** $p = 0.00015$. $n = 3$ biological repeats per timepoint; error bars, SEM. Eomes+/Ki67+ UBC progenitor cells are a rare population in the mouse RL. **g, h**, Eomes and Ki67 expression in the mouse RL at E15.5 **(g)** and E16.5 **(h)**. The RL boundaries are indicated with white dashed lines. Scale bars: 50µm. Eomes+ UBCs are rarely Ki67+. Data presented in **a, b, g, h** are representative images from three independent experiments with similar results, data in **d** were not performed in replicates. **i**, Oncogenic divergence of RL^{SVZ} progenitors from normal initiate G4 MB.



Extended Data Fig. 10 | The location of G3 and G4 MB tumours coincides with *OTX2* expression and supports an RL cell of origin. a, Scaled *OTX2*, *CBFA2T2*, and *CBFA2T3* expression by scRNAseq. **b**, Expression of *OTX2* at 14 and 17 PCW by ISH in the developing human cerebellum. **c**, T1 enhanced or T2 mid-sagittal MRI images of G4 MB ($n = 12$) tumours at initial diagnosis. **d**, T1 enhanced or T2 mid-sagittal MRI images of G3 MB ($n = 10$) tumours at initial diagnosis. Both G3 and G4 MB tumours present exclusively in the *OTX2*+ inferior cerebellum. **e**, Axial T1 enhanced, T2 or FLAIR images of SHH MB ($n = 3$) at initial diagnosis. SHH tumours occur in the cerebellar hemispheres, consistent with an EGL cell of origin. **f**, Axial T1 enhanced, T2 or FLAIR images of WNT MB ($n = 3$) at initial diagnosis. Data presented in **b** were not performed in replicates.



Extended Data Fig. 11 | See next page for caption.

Extended Data Fig. 11 | *OTX2* knockdown promotes G3 MB differentiation through intermediate upregulation of *CBFA2T2* and *CBFA2T3*. **a**, *OTX2* ChIP-seq[34] peaks are enriched at *CBFA2T2* gene locus, but not *CBFA2T3*. **b**, OTX2 protein expression is reduced following OTX2-KD. Samples were used for bulk RNA sequencing. Beta actin used as a loading control. **c**, OTX2 protein expression is reduced following OTX2-KD. Samples were used for single-nucleus RNA sequencing. Beta actin used as a loading control. **d**, Representative images of primary tumourspheres in OTX2-KD and scramble conditions for both HDMB03 and MB3W1 cultures. Scale bar: 300 µm. **e, f**, Unbiased clustering of single nuclei following OTX2-KD in HDMB03 and MB3W1 G3 MB cells lines **(c)**. **g**, Average expression of gene signatures derived from bulk RNAseq on OTX2-KD from HDMB03 and MB3W1 **(b)**. Cells that are more orange than green indicate cells with higher expression of genes characteristic of the unchanged G3 MB cell lines, and vice-versa. Orange cells are likely cells where OTX2-KD was inefficient. **h**, Differentiation score as determined by CytoTRACE[73]. Less differentiated cells are indicated in red and more differentiated cells are indicated in blue. The results support a model where cluster 6 in HDMB03 and cluster 3 in MB3W1 represent inefficient OTX2-KD cells that retain the most similarity to WT tumour cells. **i**, RBFOX3 (NeuN) protein expression is increased following OTX2-KD in both HDMB03 and MB3W1, validating GN differentiation following OTX2-KD. **j**, Expression of genes significantly correlated with granule neuron differentiation along pseudotime in normal human RL development. (Top) Density of cell along pseudotime in the granule neuron lineage (Extended Data Fig. 6b, **left**). (Bottom) Binned gene expression of markers derived from the developing human cerebellum snRNA-seq dataset (Fig. 4a). The stepwise expression of granule neuron genes observed when OTX2 is knocked down in G3 MB (Fig. 5i) strikingly mirrors that of normal granule neuron differentiation, suggesting that G3 and G4 MB arise from failed normal differentiation rather than alternate hypotheses, such as trans- or de-differentiation. **k**, *CBFA2T2* and *CBFA2T3* expression in HDMB03 (Left) and MB3W1 (Right). *CBFA2T2* expression is strongly upregulated in cells where the CytoTRACE score drops below 0.8, and *CBFA2T3* follows. The results suggest *CBFA2T2* and then *CBFA2T3* are upregulated early in response to efficient OTX2-KD. **l**, *CBFA2T2* expression change in response to CBFA2T2 overexpression (CBFA2T2-OE) in HDMB03 by qPCR. **m**, CBFA2T2 protein expression is increased following CBFA2T2-OE. β-actin used as a loading control. **n**, Representative images of primary tumourspheres in CBFA2T2-OE GFP and Control RFP conditions. Scale bar: 600 µm. **o, p**, Live cell number (**o**) and viability (**p**) in response to CBFA2T2-OE. Live cell number is significantly reduced in response to CBFA2T2-OE, while viability is unchanged. Data are normalized to their respective controls and presented showing points from $n = 3$ technical replicates per N = 8 or N = 5 biological replicates, for live cell number and viability, respectively. Error bars indicate SEM. Significance assessed using a two-tailed paired t-test on biological replicates, $^{**}p = 0.0047$. **q**, *OTX2* restrains differentiation of RL progenitors through CBFA complex inhibition. Data presented in **d, n** are representative images from 4 and 8 independent experiments, respectively, with similar results, data in **b, c, i, m** were not performed in replicates. For gel source data, see Supplementary Fig. 1.

nature portfolio

Corresponding author(s): Michael D. Taylor

Last updated by author(s): Jun 13, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Bulk RNA sequencing was performed using the Illumina HiSeq 2000 or 2500 platform and aligned to the hs37d5 reference using STAR (v2.5.1b) software. Single-cell/nucleus RNA sequencing was performed using the Illumina Novaseq6000 platform, fastqs were generated and aligned using CellRanger (v3.1.0) against the hg19 v3.0.0 reference. For single-nucleus RNA sequencing processing, CellRanger (v6.0.1) was used to include the additional parameter; --include-introns. Mass spectrometric data were obtained with an Orbitrap Exploris 480 instrument, raw output files were processed with Proteome Discoverer (v2.20.388) against the human Uniprot protein database (Dec, 2020).
Data analysis	Bulk RNA sequencing data was analyzed using R (v4.0.2) or bash environments implementing the following published algorithms which are further described and cited in the methods section. R packages: SNFtool (v2.3.0), DESeq2 (v1.28.1), DISCOVER (v0.9.3), and survival (v3.2-11). Bash environment: GATK (v3.8.0), ANNOVAR (Feb 2016), MutSigCV (v1.41), OncodriveFML (v2.3.0), Star-fusion (v0.8.0), InFusion (v0.7.3), Trans-Abyss (v2.0.0). SNP6 data was analyzed using GISTIC 2.0 (v2.0.23) in a bash environment. Single-cell/nucleus RNA sequencing data was analyzed in R (v4.0.2) or python (v3.7.6), implementing the following published algorithms which are further described and cited in the methods section. R packages: Seurat (v4.0.2), SCTransform (v0.3.2), inferCNV (v1.4.0), CCInx (v0.4), Slingshot (v1.6.1), destiny (v3.2.0), SingleR (v1.2.4), amritr (v0.1.0), CytoTRACE (v0.3.3), TradeSeq (v1.2.01). Python libraries: pySCENIC (v0.10.0). Known protein-protein interactions were accessed from STRING database (v11.5). CBFA2T2 structural prediction was performed using i-TASSER (https://zhanggroup.org/I-TASSER/, accessed 12-Jun-2021) and visualized using PyMOL (v2.4.2). Oncoprints were generated in R using ComplexHeatmap (v2.4.3).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The bulk and single-cell RNAseq data generated from MB tumor samples in this study has been deposited in the European Genome-Phenome Archive (EGA) database under the accession code EGAS00001005826. The bulk and single-cell RNAseq data generated from the G3 MB cell lines HDMB03 and MB3W1 in this study has been deposited in the Gene Expression Omnibus (GEO) database under the access codes GSE189238 and GSE200791, respectively. The published MB bulk and single-cell RNA-seq data referenced in this study is available in the EGA database under the accessions EGAD00001006305, EGAD00001004435, EGAD00001004958, EGAS00001003170, and EGAS00001003368. The referenced GTEx normal cerebellum RNAseq controls were acquired from the NCBI public repository phs000424.v6.p1. The Affymetrix SNP 6.0 data referenced during the study are available in the Gene Expression Omnibus (GEO) under the accession GSE37385. The whole-genome sequencing data referenced during the study are available in EGA under the accessions EGAD00001003125 and EGAD00001004347. The Illumina 450k methylation data referenced during the study are available in GEO under the accession GSE85218. The G3 tumorsphere ChIP-seq data referenced during the study is available in GEO under the accession GSE132269. There were multiple databases used for annotation and filtering referenced in this study. These include the Exome Aggregation Consortium / gnomAD [https://gnomad.broadinstitute.org/downloads], the NHLBI-ESP project [https://esp.gs.washington.edu/drupal/], the Kaviar Genomic Variant Database [http://db.systemsbiology.net/kaviar/#:~:text=Kaviar%20Genomic%20Variant%20Database%20%7C%20SNP,and%20frequency%20of%20observed%20variants.], the Haplotype Reference Consortium [http://www.haplotype-reference-consortium.org/], the Greater Middle East Variome [http://igm.ucsd.edu/gme/], the Brazilian Genomic Variants Database [http://abraom.ib.usp.br/], RADAR [http://rnaedit.com/], GENCODE (v19) [https://www.gencodegenes.org/human/release_19.html], the hs37d5 reference genome [https://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/phase2_reference_assembly_sequence/], ERCC spike-in sequence [https://www.encodeproject.org/files/ENCFF908UQN/], and Caltech profile 3 spike-in sequence [https://www.encodeproject.org/references/ENCSR193ZXE/]. Single-nucleus RNA sequencing data from the developing human cerebellum was obtained through correspondence from Aldinger et al., 2021, and is also available through the Human Cell Atlas [https://www.covid19cellatlas.org/aldinger20], the UCSC Cell Browser [https://cbl-dev.cells.ucsc.edu], or from Database of Genotypes and Phenotypes (dbGaP) (accession number: phs001908.v2.p1). Bulk RNA sequencing data from the developing human cerebellum was obtained through correspondence from Haldipur et al., 2019, and is available through the dbGaP (accession number: phs001908.v2.p1).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For sequencing, sample size was determined by the availability of the human samples and availability of previously published data. Human cerebellar development time points were chosen based on availability and relevance of important developmental events. Due to the scarcity of these samples, an n of 3 biological replicates were obtained for statistical quantification. Mouse samples were acquired at time points with relevance to corresponding human time points. An n of 3 biological replicates were used for mouse experiments for consistency with availability of human samples and due to the observed consistency between the 3 replicates (i.e. very low standard deviation between the replicates allowing high statistical confidence).
Data exclusions	All of the data acquired was utilized for analysis.
Replication	Technical replication of sequencing was not performed for individual patient medulloblastoma tumours, however our cohorts are comprised of a large number of patient samples with demonstrated consistencies. Where possible, samples from human and mouse development were collected in 3 biological replicates to improve reproducibility. Bulk RNA sequencing of OTX2-KD samples from HDMB03 and MB3W1 were performed in triplicate to improve reproducibility and allow for statistical comparisons. All attempts at replication were successful. Single-nucleus RNA sequencing of OTX2-KD samples were not performed in replicate as the data was very consistent with the bulk RNAseq data and between the two cell lines. Furthermore, statistical comparisons were not required to reach our conclusions.
Randomization	All available Group 3 and Group 4 medulloblastoma patients available from Medulloblastoma Advanced Genomics International Consortium (MAGIC) were selected for this study. Developing human cerebellum samples used in this study were randomly selected by HDBR and BDRL from their respective collections, based solely upon the requested developmental stage. Downstream allocation of these samples into ISH or IHC experiments was also random, based solely on developmental stage. Similarly, developing cerebellums from mice were randomly assigned to downstream ISH or IHC experiments based on their developmental stage.
Blinding	In our exploratory genomics analyses, investigators were not blinded to the subgroups and/or subtypes nor clinical metadata, because knowledge of these conditions was required for thoughtful analyses, however the same analysis scripts were used for all the tumour samples. For all IHC counts, the individual performing the count was blinded to all experimental variables, including but not limited to, antibodies used and sample age. Live cell counts and viability for in vitro studies with group 3 MB cell lines were machine automated and thus did not require blinding. Medulloblastoma MRIs were performed prior to subgroup determination, thus blinded to subgroup.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Primary antibodies used in the study were as follows: KI67 (Agilent, M7240, mouse, 1:50; Thermofisher, MA5-14520, rabbit, 1:100), SOX2 (Thermofisher, PA1-094, Rabbit, 1:200), PAX6 (Biolegend, 901301, rabbit, 1:300), TBR2 (EOMES) (Thermofisher, 14-4875-82, Rat, 1:200), GFAP (Agilent, Z0334, rabbit, 1:1000), OTX2 (Abcam, ab21990, rabbit, 1:500), CBFA2T2 (Bethyl Laboratories, A303-593A-M, Rabbit,1:500), RBFOX3 (NeuN) (Cell Signaling Technology, D4G4O, Rabbit, 1:1000), Beta-actin (Sigma-Aldrich, A2228, mouse, 1/1000). Secondary antibodies used in the study were as follows. Anti-rabbit: Alexa Fluor 488, Thermofisher, A-11034, Goat, 1:1000 and Alexa Fluor 568, Thermofisher, A-11011, Goat, 1:1000. Anti-mouse: Alexa Fluor 488,Thermofisher, A-11001, Goat, 1:1000 and Alexa Fluor 568, Thermofisher, A-11004, Goat, 1:1000.

Validation

KI67 (M7240) is intended for diagnostic IHCs and certified manufacturing facilities guarantee full quality control. Additionally, extensively validated in our previous publications Haldipur et al., Science, 2019, Haldipur et al., Acta Neuropathol, 2021.
KI67 (MA5-14520) has Manufacturer Advanced Verification: This Antibody was verified by Cell treatment to ensure that the antibody binds to the antigen stated. Manufacturer tested applications include IHC, IF, WB, and Flow. Cited in at least 705 publications which include the following recent studies: Low et al., iScience, 2021, Kalucka et al., Cell, 2020.
SOX2 (PA1-094) has Manufacturer Advanced Verification: This Antibody was verified by Relative expression to ensure that the antibody binds to the antigen stated. Manufacturer tested applications include IHC, IF, WB, Flow, IP and Cut&Run. Cited in 34 publications which include the following recent studies: Krishnan et al., Front Surg, 2021, Ram et al., Front Surg, 2017, and validated in studies from our own lab, Haldipur et al., Science, 2019, Haldipur et al., Acta Neuropathol, 2021.
PAX6 (901301) is manufacturer tested for applications including IHC and WB. Each lot of this antibody is quality control tested by Western blotting and formalin-fixed paraffin-embedded immunohistochemical staining of brain tissue. Cited in 223 publications which include the following recent studies: Haag D, et al., Cancer Cell, 2021, Padmanabhan, et al., Cell Stem Cell, 2021, and validated in studies from our own lab, Haldipur et al., Science, 2019, Haldipur et al., Acta Neuropathol, 2021, Aldinger et al., Nat Neurosci, 2021.
TBR2 (14-4875-82) is manufacturer tested for WB and IHC applications. Cited in 71 publications which include the following recent studies: Hu et al., Nat Commun, 2021, Ha et la., J Dev Biol, 2020, and validated in studies from our own lab, Haldipur et al., Science, 2019, Haldipur et al., Acta Neuropathol, 2021.
GFAP (Z0334) is manufacturer optimized and validated for IHC applications, with over 2,635 citations and previously validated in studies from our own lab, including Haldipur et al., Science, 2019, Haldipur et al., Acta Neuropathol, 2021.
OTX2 (ab21990) manufacturer tested applications include IHC, IF, ChIP, WB, and IP, and extensively validated in our previous publications for both gain and loss of function experiments (Kaur et al., Dis Model Mech, 2015; Stromecki et al., Mol Oncol, 2018; Zagozewski et al., Nat Commun, 2020).
CBFA2T2 (A303-593A-M) manufacturer validated for use in WB and IP. All Bethyl Laboratories® antibodies are validated based on 6 pillars (independent antibodies, complementary assays, orthogonal characteristics, biological characteristics, protein OE/epitope tags, genetic strategies) to meet strict performance standards. Multiple pillars are used in a complementary fashion to validate each antibody. Citations include the following recent publications: Yamamoto et al., J Cell Sci, 2020, Luo et al., Biochem Biophys Res Commun, 2020.
RBFOX3 (NeuN) (D4G4O) manufacturer validated for WB, IHC and IF according to Cell Signaling Technologies' antibody performance guarantee which, for immunoblotting, may include KD/KO, validation, validation across several cell lines/tissues, etc. Cited in at least 112 publications which include the following recent studies: Xiang et al., J Neuroinflammation, 2022, Verdone et al., Sci Rep, 2022, Yu et al., Commun Biol, 2022.
Beta-actin (A2228) is widely used and validated in WB and immunocytochemistry applications. Extensively used in previous publications from our labs including Stromecki et al., Mol Oncol, 2018; Zagozewski et al., Nat Commun, 2020. Additional recent citations include the following: Xian et al., Autophagy, 2020, Yao et al., J Clin Invest, 2019.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

HDMB03 was kindly provided by Dr. Till Milde (Milde T, et al., Journal of Neuro-oncology, 2012)
MB3W1 was kindly provided by Dr. Matthias Wölfl (Dietl S, et al., BMC cancer, 2016)

Authentication	Cell lines were authenticated by STR profiling by ATCC recently in 2021.
Mycoplasma contamination	Cell lines were not tested for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used in this study.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Cerebella from CD1 mice were collected between E14.5 and E16.5. Both male and female mice were used with no discrimination.
Wild animals	No wild animals were used in this study.
Field-collected samples	No field collected samples were used in this study.
Ethics oversight	IACUC Protocol No. IACUC00006, Institutional Animal Care and Use Committee (IACUC), of Seattle Children's Research Institute, Seattle, WA, USA.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Samples used for the bulk RNAseq cohort were collected at diagnosis after informed consent was obtained from subjects as part of the Medulloblastoma Advanced Genomics International Consortium (MAGIC). Research participants were patients with confirmed diagnosis of medulloblastoma at local centers. All cases used in this study were primary, treatment-naive Group 3 or Group 4 medulloblastomas. Samples were predominantly classified as Group 4 (326 Group 4, 219 Group 3), and all subtypes were represented (68 Group 3 alpha, 36 Group 3 beta, 49 Group 3 gamma, 93 Group 4 alpha, 98 Group 4 beta, 105 Group 4 gamma, 96 not profiled by methylation array). Patients were predominantly male (365 males, 149 females, 31 unknown) and the median age was 7.1 years old. All histologies were represented, though samples predominantly displayed classic histology (290 classic, 55 large cell/anaplastic, 32 desmoplastic, 9 MBEN, and 159 unknown). Samples used for the single-cell RNAseq cohort were obtained prospectively following surgical resection. Samples used in this study were those having subsequent diagnoses as Group 3 or Group 4 medulloblastomas and were all treatment-naive primary tumours. Samples were predominantly classified as Group 4 (7 Group 4, 4 Group 3), subtypes were not determined. Patients were predominantly male (9 male, 1 female, 1 unknown) and the median age was 8.7 years old. Developing human cerebellum samples were obtained from the Human Developmental Biology Resource (HDBR) and Birth Defects Research Laboratory (BDRL) tissue repositories. Due to the sensitivity of the samples, population characteristics were not recorded.
Recruitment	Tumour samples used to generate the bulk RNAseq cohort were obtained from the Medulloblastoma Advanced Genomics International Consortium (MAGIC). Samples used to generate the single-cell RNA sequencing cohort were prospectively obtained following surgical resection from Hospital for Sick Children or McGill University Health Centre. There is no bias of recruitment as patients are not pre-screened.
Ethics oversight	Medulloblastoma samples were obtained in compliance with the ethical regulations of the Hospital for Sick Children (REB 0020020238 and REB1000055059 approved by the Research Ethics Board of The Hospital for Sick Children) and McGill University Health Centre (REB MCH003-26 approved by the Research Ethics Board of McGill University Health Centre). All patient material was collected after receiving informed consent, under approval and oversight by their respective internal review boards. All human cerebellar samples used in this study were obtained using protocols approved by the Seattle Children's Research Institute IRB. Samples used for histological analysis were collected by the Human Developmental Biology Resource (HDBR) located at University College London, and Newcastle University, United Kingdom and Birth Defects Research Laboratory (BDRL) at the University of Washington. Third trimester samples were part of an archival collection at the Hôpital Necker-Enfants Malades in Paris, France. All samples were collected with previous patient consent and in strict accordance with institutional and legal ethical guidelines.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Michael D. Taylor  June 14, 2022