Title: Rates and patterns of clonal oncogenic mutations in the normal human brain

Authors: Javier Ganz[1,2,3,4,†], Eduardo A. Maury[1,2,3,4,5,†], Basheer Becerra[1,2,3,4], Sara Bizzotto[1,2,3,4], Ryan N. Doan[1], Connor J. Kenny[1,2,3,4], Taehwan Shin[1,2,3,4], Junho Kim[1, 4], Zinan Zhou[1,2,3,4], Keith L. Ligon[4,6], Eunjung Alice Lee[1, 4*], Christopher A. Walsh[1,2,3,4,*]

Affiliations:
[1] Division of Genetics and Genomics, Manton Center for Orphan Disease, Boston Children's Hospital; Departments of Pediatrics, Harvard Medical School, Boston, MA 02115, USA.
[2] Howard Hughes Medical Institute, Boston Children's Hospital, Boston, MA 02115, USA.
[3] Departments of Neurology, Harvard Medical School, Boston, MA 02115, USA.
[4] Broad Institute of MIT and Harvard, Cambridge, MA, 02142, USA.
[5] Bioinformatics & Integrative Genomics Program and Harvard/MIT MD-PHD Program, Harvard Medical School, Boston, MA, USA.
[6] Department of Oncologic Pathology, Dana-Farber Cancer Institute; Department of Pathology, Brigham & Women's Hospital; Center for Patient Derived Models, Dana-Farber Cancer Institute, Department of Pathology, Boston Children's Hospital; Harvard Medical School, Boston, MA, USA.
[†] These authors contributed equally to this work.
[*] Corresponding authors.

Running Title: Oncogenic mutations in non-diseased human brain.

Corresponding authors: Christopher A. Walsh, 3 Blackfan St, CLS-15064, Boston, MA 02115, Phone (617) 919-2923, Fax (617) 919-2010, email Christopher.Walsh@childrens.harvard.edu, and Eunjung Alice Lee, 3 Blackfan St, CLS-15020, Boston, MA 02115, Phone (617) 919-1589, Fax (617) 919-2923, email Ealee@childrens.harvard.edu

Competing interests: The authors declare no conflict of interests.

Words (7708), Pages (21), Main figures (6), Supplementary figures (6), Supplementary tables (2).

Abstract

While oncogenic mutations have been found in non-diseased, proliferative non-neural tissues, their prevalence in the human brain is unknown. Targeted sequencing of genes implicated in brain tumors in 418 samples derived from 110 individuals of varying ages, without tumor diagnoses, detected oncogenic somatic single-nucleotide variants (sSNVs) in 5.4% of the brains, including *IDH1* R132H. These mutations were largely present in subcortical white matter and enriched in glial cells, and surprisingly, were less common in older individuals. A depletion of high-allele frequency sSNVs representing macroscopic clones with age was replicated by analysis of bulk RNAseq data from 1,816 non-diseased brain samples ranging from fetal to old age. We also describe large clonal copy number variants, and that sSNVs show mutational signatures resembling those found in gliomas, suggesting that mutational processes of the normal brain drive early glial oncogenesis. This study helps understand the origin and early evolution of brain tumors.

Statement of Significance:

In the non-diseased brain, clonal oncogenic mutations are enriched in white matter and are less common in older individuals. We revealed early steps in acquiring oncogenic variants, which are essential to understanding brain tumor origins and building new mutational baselines for diagnostics.

Introduction

Tumors result from the clonal expansion of cells due to the presence of somatic driver gene mutations in stem cells (1). Even in normal individuals, self-renewing tissues such as skin, blood, esophagus, endometrium, bladder, colon, and liver harbor cancer-associated somatic mutations, which increase with age (2-11). Recent work examining post-mortem brains from a small cohort of 14 non-diseased-aged individuals found mutations in cancer-associated genes, though none of the identified mutant alleles have a known role in oncogenesis (12). Consequently, the prevalence of oncogenic driver mutations in the non-diseased human brain remains largely unknown.

Primary gliomas and other brain tumors are considered to occur mostly within the white matter (WM) (13,14). WM is highly enriched in glial cells (approximately 75% oligodendrocyte-lineage cells, 20% astrocytes, and 5% microglia) (15), while grey matter (GM) represents a combination of neurons and glial cells (15). One factor impeding assessing the contribution of clonal oncogenic mutations in the brain is the abundance of non-proliferating neurons concentrated in the GM (15). In contrast, brain-derived glial cells, including oligodendrocyte precursor cells (OPC) (16) and astrocytes to a lower extent (17), retain the ability to proliferate in the postnatal brain. Proliferation is a significant source of somatic mutations (18,19), thus, prior conventional bulk analyses with mixtures of neurons and glial cells may have low sensitivity to discover oncogenic variants if they are present in glial cells which represent just a fraction of the total cells (15).

To overcome this challenge, we performed targeted deep-sequencing of white matter (WM) areas to identify clonal somatic oncogenic variants and compared these to adjacent grey matter from the same brain region (Fig. 1, Fig. S1). We also investigated 1,816 non-diseased brain RNA-seq datasets obtained from two independent cohorts, Genotype–Tissue Expression (GTEx) (20) and BrainVar (21), representing different brain regions and ages, with methods that allow identification of clonal somatic variants. With our approach, we observe that the normal human brain harbors sSNVs and large somatic copy number variants (sCNVs) with oncogenic potential, suggesting glial susceptibility to acquire or further expand existing variants. In contrast with other tissues (3-8), the burden of clonal mutations representing macroscopic clones (VAF>7%) (22) in the brain does not detectably increase with age, so that the mutations are less common in older individuals. The patterns of nucleotide substitution for these sSNVs resemble those previously reported in brain tumors, suggesting that the mutational processes that give rise to brain tumors pre-exist in normal tissue.

Results

Experimental scheme

We analyzed a total of 418 samples derived from 110 individuals spanning different ages (0-108 years) (Table S1A), with no history of neuro-oncological or other neurological diagnoses. We designed molecular inversion probes (MIPs) (23) that target all exons and adjacent intronic sequence (to capture splice site mutations) of 121 genes directly associated with brain tumors and other cancer types (Fig. 1, Table S1B,C). Our panel represented multiple pathways implicated in disease and different classes of proto-oncogenes and tumor suppressor genes. First, to evaluate the presence and accumulation of oncogenic variants during aging in the normal brain, for each of the 110 subjects, we analyzed data from at least two different brain regions and one non-brain sample (Table S1A). We primarily focused on the brain's frontal lobe since it is the most prevalent location for malignant tumor emergence, followed by temporal, parietal, and occipital (24). Brain samples consisted of hippocampus (HC), cerebellum (CER), and prefrontal cortex that was subdivided into grey (CXG) and white matter (CXW) (Fig. S1A). When no clear anatomical distinction between CXW and CXG was possible due to tissue size or frosting, the sample was labeled CX. Second, to test whether sampling more brain locations from one individual would increase our sensitivity for oncogenic variants, we also evaluated 91 samples derived from 17 different organs and the entire left hemisphere from one 17-year-old individual (UMB1465) (Fig. 1, Table S1A).

Identification of variants from deep-targeted sequencing

Somatic mutations associated with cancer can be either driver mutations, promoting clonal expansions in some cases, or passenger mutations with a less clear effect. To investigate the potential clonal variant accumulation in cancer genes within the non-diseased brain, we focused on low-allele-frequency variants with oncogenic potential. We define oncogenic variants as 1) previously reported pathogenic variants in cancer, or 2) predicted to be damaging by *in silico* analysis in a known cancer-driver gene (25). We generated deep-targeted sequencing data (average coverage of 590x per sample across targeted regions) (Fig. S1B) and used the CLCbio

Low Frequency Variant Detection algorithm (QIAGEN) to call somatic variants with a probabilistic error model to account for sequencing errors (see methods). No significant difference in sequencing coverage across ages or tissue types was observed (Fig. S1C,D).

We obtained a total of 51,036 raw calls (Table S1D) and conservatively filtered out variants to obtain high-confidence mutations (see methods), focusing only on variants with 0.5 – 15% allele frequency (VAF) due to germline contamination at higher VAFs (Fig. S1E). Experimental sensitivity analyses using 165 spike-in somatic mutations (108 heterozygous and 57 homozygous SNPs) showed that this computational approach and filtration was optimized for low false-positive rates and achieved specificities of >99% with sensitivities comparable to other studies (12) at different mosaic fractions (Fig. S1F). 35 variants (average depth of 1086x, median VAF of 1.86%) passed our filtering criteria, and of those, 28 were unique, while 7 were seen in different samples within the same individual (Table S1E). One of these was discovered to be a germline event during validation. We used ultra-deep Ion Torrent sequencing (MIPP-Seq) (26) to validate 19 candidate mosaic variants, including all the 13 variants predicted to affect protein structure and 6 random synonymous, intronic, and promoter variants. With an average 92,757x per site, we achieved a validation rate of 89% with a high correlation of the VAFs between discovery and validation sequencing (r^2=0.93, Pearson coefficient, Fig. 2A, Table S1F).

Presence of brain tumor associated oncogenic variants in normal brain

Among the validated variants, 12 occurred in cancer driver genes. These variants were predicted damaging and pathogenic by multiple algorithms (see methods) using similar criteria to a recent study (27). Nine of these 12 variants were found in the brain, and 3 in non-brain tissue. The brain-specific variants were all exonic and had a median VAF of 1.2% (Fig. S1G). These variants distributed similarly between proto-oncogenes and tumor suppressor genes (25,28), and had the highest score in our predicted pathogenicity scale (NSYND3, see methods) (27) (Fig. 2B). Importantly, the variants we detected did not occur in genes most frequently mutated in clonal hematopoiesis that were included in our panel, indicating that these variants are not likely derived from blood contaminants, as observed in a recent study (12). All of our identified genes were among the most frequently mutated genes in lower-grade gliomas (LGG) and glioblastoma (GBM), but not medulloblastoma (25) (Fig. 2C-D). Of the 9 brain-specific variants, 6 were previously reported pathogenic mutations in cancer (COSMIC, Clinvar, HGMD, and ICGC) (Fig. 2B). Among these variants, *IDH1*, *PTPN11*, *NF1*, and *PTEN* gene variants are of particular interest due to their high prevalence and established pathogenic effects in brain tumors.

For each subject, we evaluated germline variants with predicted deleterious impact to assess possible interactions with somatic mutations or double-hit events (somatic + germline). We identified 240 germline variants (median VAF 50.3±3.3%) predicted to be damaging, and of those, 51 were unique over 41 individuals (see methods, Table S1G). We did not detect biallelic double-hit events in our cohort, though our method would be unable to detect mosaic LOH or deletions (29), and none of these variants had been previously related to disease, as expected from a non-diseased cohort (Fig. 2B, Table S1G). We found no enrichment in predicted damaging germline variants (Fisher exact test p=0.621) among the individuals with somatic oncogenic mutations.

Oncogenic brain variants are not detectably more common in older individuals.

We detected 7 previously reported oncogenic variants over 6 different brains out of 110 total; therefore, 5.4% of the evaluated brains carried a reported pathogenic variant in a cancer-driver gene. We did not find evidence that older (>30 years) individual brains were enriched with known oncogenic variants in cancer genes (Fig. 2E) at the level of mosaicism detectable with this approach, which contrasts with many other tissues that show accumulation of variants with age (3-8). On the contrary, we observed a depletion of oncogenic variants in the CXW samples of individuals older than 30 y/o without cancer diagnosis (6/52 vs. 0/42, Fisher exact test, p=0.025; and 5/43 vs 0/42 maintaining only one sample of the oversampled individual, p=0.029). This observation was also true using 19 predicted pathogenic and non-pathogenic brain variants (13/114 vs. 6/143, Fisher exact test p=0.025) from our filtered call set (89% validation rate). We interpret these observations as suggestive of lack of age-related increase in oncogenic variants in the normal brain. Importantly, this behavior was further confirmed in two orthogonal datasets presented in the following sections.

Oncogenic brain variants are more prevalent in the white matter.

Oncogenic brain variants exhibited increased occurrence in WM (6 mutations in 94 CXW samples tested), compared with no variants in adjacent gray matter samples (92 CXG samples) (Fisher exact test p=0.028) (Fig. 2F). There were no significant differences in sequencing coverage across tissue types that might explain this difference (Fig. S1D). CX samples, which we could not distinguish as WM or GM, also had 0 variants (53 samples). HC samples came second after CXW, with 2 detected variants over 69 samples tested (ratio 0.03, Fisher exact test CXW vs. HC p=0.4, CXG vs. HC p=0.182). For CER, we observed 1 mutation (1/10, ratio 0.1), but given the small sample size, we cannot derive conclusions on this region.

IDH1 R132H mutation is enriched in glial cells.

The most prevalent somatic mutation in glioma is *IDH1* R132H, present in more than 70% of World Health Organization (WHO) grade II and III astrocytomas and oligodendrogliomas (30). We observed three mutations in *IDH1* in two different individuals. The first individual (UMB1465, 17y/o) had two *IDH1* reported mutations, R132H and R100Q, detected in cortical white matter from distinct distant regions of the brain, corresponding to PFC and primary motor area, respectively (Fig. 3A). *IDH1* R100Q has been infrequently reported in gliomas, and its contribution to oncogenesis is still being determined (31,32). The second individual (UMB 5621, 37y/o) bore the *IDH1* R132H mutation in the HC at a VAF of 0.8%. Interestingly, the R132H mutation observed in the PFC of UMB1465 was called in two adjacent WM samples of the same brain section with different VAFs (0.9 and 5.0%, Fig. 2B). For this individual, we sequenced >50 additional brain samples derived from the entire left hemisphere and most organs (Fig. 1 Table S1), and we did not detect *IDH1* R132H in any of these with at least 0.1%VAF at >3500x coverage (Fig. 3B). The difference in VAFs between the two adjacent R132H-bearing samples (5% vs. 0.9%) may reflect the distance from the center of the mutant clone (Fig. 3B). Mutations with a 5% VAF are often shared between tissues or within an organ (33), and therefore one would not expect such relatively high VAF variants to be restricted to one organ sub-region. Since heterozygous *IDH1* mutations confer a proliferative advantage in human astrocytes and remodel

chromatin into a neural progenitor-like state (34,35), the sharply different VAFs of this mutation are most consistent with a clonal event with enhanced proliferation, present in a ≈5 mm^3 sample within the white matter. This mutation's high VAF and focal nature suggest that it was acquired very late in development or postnatally, then further amplified by glial proliferation. Similar mechanisms can be proposed for the *PTPN11*, *PTEN*, and *NF1* mutations we found (36-38).

To investigate the lineage of origin of the *IDH1* R132H mutation, we evaluated the presence of this mutation in neuronal and non-neuronal nuclei (see methods). NEUN-positive nuclei were analyzed using single nucleus RNA sequencing (scRNA-seq) to confirm the identity of the sorted cells, demonstrating that this population was all neuronal nuclei, broadly sub-classified into excitatory and inhibitory neurons, without any further contamination (Fig. 3C). As expected, the NEUN-negative fraction lacked neurons, and represented a mixture of glial cells, including OPCs, astrocytes, oligodendrocytes, and a small fraction of microglial cells (Fig. 3C). NEUN-positive and NEUN-negative populations showed gene expression profiles consistent with neurons and glia (oligodendrocyte, astrocyte, and OPCs), respectively (Fig. 3D). The remaining sample containing the *IDH1* R132H mutation (individual UMB1465) was subjected to nuclear sorting and genotyping using digital droplet PCR (ddPCR) targeting the R132H mutation. Our results demonstrate that the *IDH1* R132H mutation was enriched in the NEUN-negative population (Fig. 3E), consistent with our initial observation of its presence only in WM, and suggesting a glial localization of this mutation. Interestingly, the R132H mutation appears to be detectable also at very low levels in the NEUN-positive population as well, though we cannot rule out low-level contamination in generating this signal or an early event in neural precursors. The shared presence of the R132H mutation in neurons and glia would suggest a congenital origin of this mutation since cortical neurons are virtually all generated prenatally.

Clonal brain variants at high allele fraction do not increase with age

To confirm the existence of oncogenic variants, we evaluated two independent cohorts of non-diseased human brains, including 1,640 and 166 bulk RNA-seq samples obtained from the GTEx project (20-79y)(20) and BrainVar database (fetal-19y)(21), respectively, also representing different brain regions (Fig. S2A,B). RNA-MuTect (22) was used to identify somatic variants (Fig. S2C) and suspected RNA editing bases were removed (A>G, T>C). RNA-MuTect sensitivity was tested in normal samples with coextracted DNA and RNA data, and was able to detect DNA mutations with allele fractions of >7% in the RNA, in cases where the gene was sufficiently highly expressed (22), and they define these as macroscopic clones. In the GTEx cohort, we found a total of 590 variants, including 325 missense, of which 62 variants (19%) overlap with the exact amino acid change reported in COSMIC (CMC v92), and 27 others are disruptive (splice site or nonsense, 5 with COSMIC overlap, 19%). In BrainVar, we found a total of 746 variants, including 493 missense (56 with COSMIC overlap, 11%) and 70 disruptive (5 with COSMIC overlap, 7%) (Table S2A-D). Within variants with COSMIC overlap, we identified reported variants in cancer driver genes (VAF ≤10%) associated with brain tumors, such as *DDX3X* and *MAX*, with *DDX3X* being mutated in 8% of medulloblastomas(25). We also detected predicted pathogenic variants (with a high score of pathogenicity NSYND3 and LOF) in other brain tumor driver genes highly represented in LGGs (*FGFR1, PDGFRA, MTOR*), but these exact base substitutions were not previously reported. Interestingly, we observed several variants in

PDGFRA in both GTEx and BrainVar (VAF ≤10%) datasets suggesting that this gene, which is of importance in all gliomas, is frequently mutated (25).

We used the BrainVar (n=166, fetal-19y) and GTEx (n=1640, 20-79y) mutation calls to further investigate our initial observation about the lack of age-related accumulation of detectable clonal variants in the brain with a less biased approach, including all expressed genes rather than only those previously implicated in brain tumors. By integrating both datasets and modeling the mutation counts per sample using mixed effect negative binomial regression, while adjusting for standardized RNA integrity score, standardized ischemic time, and standardized total mapped reads, we observed depletion of all (mean ratio=0.241, 0.149-0.391 95% CI, p=3.5e-09), predicted pathogenic (mean ratio= 0.28, 0.155-0.506 95% CI, p=1.7e-05), and disruptive variants (mean ratio=0.366, 0.2-0.66 95% CI, p=0.001) with age (Fig. 4A-C, Fig. S2D-F, Fig. S3), a discovery consistent with our panel findings. The negative association with age was also significant by analyzing all mutations from each dataset independently (BrainVar p=0.00029, GTEx p=0.043), and also in GTEx for pathogenic and disruptive mutations (Fig. S3A-E). For BrainVar, the regression model was not able to converge due to the low number of pathogenic and disruptive mutations. As a control, we did not observe a similar significant depletion for the T>C variants, removed as potential RNA editing events, in the combined dataset (mean ratio=0.765, p=0.282) and also in each cohort independently (Fig. S3F,G). Furthermore, analysis of all variants from both cohorts after only filtering known RNA editing sites in databases, also showed a significant decrease with age (mean ratio=0.411, p=5.1e-06), demonstrating that T>C removal does not affect our observed aging trend (Fig. S3H). The effect with age may vary for different brain sub-regions. Among 13 evaluated regions, cortex showed a nominal depletion of disruptive and pathogenic variants with age (p=0.035 and p=0.032, respectively), while cerebellum also showed significant depletion for pathogenic variants p=0.01 (and nominal for disruptive p=0.07) (Fig S4A). Since we observed a general negative association when combining GTEx and BrainVar, and none of the evaluated brain regions showed a significant increase with age, we conclude that there is no age-related increase of clonal somatic mutations that reach the level of detection of this method in normal brain.

To learn more about how different brain region compare to each other in their mutational burden, we used the negative binomial model to rank regions based on comparing region-specific mutation incidence rate to the overall mutation incidence rate (Fig. S4B). When assessing all variants, only caudate (basal ganglia) had nominally fewer mutations than average (p=0.039), while cerebellum exhibited a very significant increase in the relative mutation count (p= 4.2e-08) (Fig. S4B). The same pattern was true when assessing pathogenic and non-pathogenic variants, with the exception that in the cortex, pathogenic mutations also showed a nominal increase in the relative mutation burden compared to the overall brain mean (p=0.024) with hypothalamus showing a decrease in overall mutation count (p=0.034) (Fig. S4B).

Interestingly, we found one outlier sample (6-year-old female) in the BrainVar dataset with a high mutation load (139 mutations called) including 23 mutations overlapping reported COSMIC events, suggesting a potential pre-neoplastic clonal expansion. The only drivers detected were *GPC5* and *FAT3*, though neither of these have been directly associated with brain

tumors. However, these types of mutations might increase proliferative fitness leading to subsequent mutation accumulation as reflected in the high mutational load of this sample.

Somatic copy number alterations in non-diseased brain

Since somatic copy number alterations (sCNVs) are the most frequent and important driver events in the oncogenesis of multiple brain tumors, we next assessed the prevalence of sCNVs in 1,636 brain samples across 253 subjects from the GTEx (v7) consortium. We used a recently developed algorithm called superFreq(39) that leverages allele frequency information across germline heterozygous sites and read depth to identify sCNVs from RNA-seq data. While this algorithm was designed for cancer samples, it can provide lower-bound estimates of the sCNV landscape in normal tissues. The initial raw call-set consisted of 1,242 variants across 213 samples (Table S2E). Due to the noisy nature of RNA-seq data, we implemented a stringent filtering strategy (see methods). Briefly, we removed variants that overlapped fewer than 100 genes, so that the precision for those events is expected to be 80-90% (39), and we filtered variants whose log-fold-change (LFC) and clonality were too noisy to be reliably estimated via visual inspection. The final call-set was 37 sCNV across 20 subjects, consisting of 15 gains, 13 copy-number neutral loss of heterozygosity (CN-LOH), and 9 losses (Fig. 5A). The mosaic fraction of these events ranged from 13.4% to 48.0%, with a median of 29.8%. From this sample size we estimated the percentage of normal adult individuals to have at least 1 sCNV in a brain sample to be 7.9% (95% CI: 4.90-11.94) (Fig. 5B). No sCNV rate differences were detected between brain regions and no evidence of age-related change was observed (Fig. 5A). We also analyzed 147 subjects from BrainVar and obtained 262 initial calls, of which 7 remained after filtering (2 losses, 1 gain and 4 CN-LOH), across 5 prenatal subjects (Fig. 5A, Table S2E), suggesting that 4.8% (95% CI: 1.11-7.78) of young brain samples have at least 1 sCNV detectable with this approach.

Some of the sCNV overlap reported events in glioma and other brain tumors. Among the copy number losses in both datasets (n=11), we observed four events in chromosome (Chr) 22q, four in Chr19, two in Chr. 16, and one in Chr. 2q (Fig.5A). LOH22q has been reported in brain tumors including astrocytoma (9-30%), GBM (24%) and meningioma (65%) (40). In two out of the four 22q events, we detected loss of *NF2* and *SMARCB1* (Fig. 5C and Fig. S5), which are highly involved in meningioma (40) and atypical teratoid rhabdoid tumors(41). Events in Chr 19 were characterized by one 19q-arm loss and two 19p-arm losses. 19q LOH and loss events have been frequently reported in olidendrogliomas (100%), astrocytomas (30-40%), and GBM (30%) (42). We observed CN-LOH events overlapping CDKN2A and SMARCA4, two important genes in brain tumors(40), but the effect of these are less clear. Among copy number gains (n=16), we detected five events in Chr. 6q, all of them gaining half of the q-arm, which includes relevant genes such as *MYB*, involved in pediatric gliomas (43). Chr. 1 had four partial gains of the q-arm, and 1q gains were reported in high-grade gliomas (44). We also detected partial and whole gains in Chr. 12p, 13q, 15q, 17q and 18.

Clonal variants in normal brain share mutational mechanisms seen in brain tumors

To understand specific mutagenic processes underlying the accumulation of clonal point mutations in the brain compared to other tissues, we performed mutational signature analyses

using clonal sSNVs obtained from a recent study using the GTEx database (22). Using this dataset allowed us to compare brain spectra with other organs using variants called with a consistent pipeline across tissues. We performed our analyses using various VAF cut-offs of 5%-40% and obtained consistent results throughout this range (Fig. 6, Fig. S6A). We focused on sSNVs with a VAF of less than 15% since they are most likely to reflect the sSNVs we targeted in our panel data. We found that the estimated mutational signatures from normal brains were similar to those from brain tumors. Normal brain sSNVs statistically decomposed into several signatures (SBS 39, 5, 23, 1, 30, and 2), each reported in COSMIC as present in brain tumors (Fig. 6) and consistent with previous findings (45). Our analysis confirmed that the mutational signatures found in normal brains are indeed enriched for brain cancer signatures (Permutation test, p=0.00018). The brain tumor signature enrichment was not observed in any other non-brain tissue tested using a Bonferroni corrected p-value of 0.01, except for pancreas (Permutation test, pancreas p=0.00236, skin p=0.07895, and heart p=0.4), (Fig. 6, Fig. S6B). To validate our analysis, we processed skin sSNVs from the same study (22) and found that those signatures were enriched for skin cancer signatures (p=0.00025) (Fig. 6), consistent with previous findings (2). Normal pancreas sSNVs also showed a good correlation with those found in pancreatic cancer (Permutation test, p=0.0075) (Fig. S6B). The significant overlap of pancreas sSNVs with brain cancer signatures suggest similar mutational processes in these tissues, perhaps reflecting similarities in transcriptional and developmental programs (46). Brain mutational signatures reflect a combination of processes, including replication and transcription-induced mutations and their respective repair mechanisms. Interestingly, the COSMIC signature SBS1 observed in normal brain and in all tumor-types is associated with cell division and proliferation (18), reflecting developmental processes or postnatal glial proliferation.

Discussion

In this study, we observed oncogenic variants in the brain of individuals without diagnosed cancer at a rate higher than the brain tumor prevalence (24), indicating that the mere presence of these events in the brain is not equivalent to clinical progression to cancer. This may have diagnostic implications since knowing the occurrence of oncogenic variants in normal tissue may help establish baselines for more accurate diagnosis.

Evaluation of 1,816 normal brain samples from two orthogonal studies allowed us to independently confirm the existence of oncogenic variants in the normal human brain, in concordance with a previous study (47). Although we do not see the same pattern of affected genes in our panel data (DNA-seq) and the BrainVar and GTEx-based analysis (RNA-seq), such as *NF1* and *IDH1* recurrences, this difference may relate to limitations of RNA-based mutation calling, such as tissue-specific and expression-bias (WM enriched or GM enriched), coverage, and removal of RNA-editing bases (22). We also describe that sSNVs mutational signatures associated with brain tumors can be observed in normal brains, reflecting transcription and replication-induced mutations. Our data suggest that many signatures previously reported in brain tumors include many passenger mutations present in the normal brain and are not necessarily all tumor-specific or strictly associated with malignancy. Based on our data, we believe that replication-induced mutations are likely a result of pre-natal development or post-natal glial proliferation in concordance with previous etiological factors contributing to brain cancer (1,19).

The contribution of signatures we see in normal tissue and brain tumors are different likely due to tissue sampling differences and because during tumor development particular mutational mechanisms, such as SBS1, can diverge from those observed in the low-proliferating normal brain.

We adapted SuperFreq (39) for cloud computing to evaluate sCNVs in 1,783 normal brain samples, which to our knowledge comprises the largest normal brain cohort examined in this context. We report large chromosomal alterations in line with previous studies in single neurons (48-50) with some overlapping events reported in brain tumors such as 22q and 19q deletions (40). We only focused on large events to improve calling precision, limiting our discovery of smaller events. Gains were more frequent than losses and losses mostly affected Chr 22 and 19, while gains most commonly involved Chr 6 and 1. sCNV events occurred at surprisingly high frequency in our cohorts (7.9% GTEx, 4.8% BrainVar) with a median mosaic fraction of 29.8% (14.9% VAF). Despite the high level of mosaicism of these sCNVs, they were often not shared between multiple brain regions, suggestive of restricted events arising during development or postnatally due to local clonal expansion. Given the low number of sCNV events we found, we cannot draw conclusions about any regional or aging trends. Our estimated rates of frequency in the cohort and mosaic fraction are reasonable compared to those found in a recent study using bulk whole-genome sequencing of postmortem brains (51). However, larger samples sizes and more sensitive techniques will be needed to more definitively determine rates of sCNV.

All the clonal oncogenic sSNVs found in the white matter were detected in younger individuals in our targeted panel (<30 years), and we failed to find evidence of an age-related accumulation of oncogenic events. Given the relatively young age of the subjects, and postmortem nature of the data, we do not know whether those same individuals may develop cancer in the future from those mutant clones. In our panel data we found a surprising lack of age-related oncogenic variant accumulation in the brain, which differs from findings in other tissues (3-8). It is worth noting that our targeted-sequencing approach evaluated only oncogenic variants, which differs from those studies in blood, skin, and esophagus, among others, that evaluated all mutations in known cancer genes (2,4-9). Our BrainVar and GTEx analysis allowed us to look for all mutations in brain expressed genes, thus resembling more closely previous studies. In this case, we also found a stable number, or even depletion, of all clonal, disruptive and predicted pathogenic variants with age considering all brain samples from all ages. These results confirm our panel data finding and further contrast the mutational dynamics between the brain and other tissues. Two recent reports using GTEx data also included brain samples within a broader study. The first report indicates that some brain regions may have high correlation between mutations and age, even more than sun-exposed skin or blood, while other regions seem to be negatively associated (47). Nonetheless, all the evaluated brain regions do not achieve a high level of statistical support (showing a FDR>0.1) and hence are not inconsistent with a lack of age-related increase. In the second study (45), they report negative values of age-correlation with brain mutations, but these are also insignificant, supporting a lack of age-related increase and a trend consistent with our findings.

The relative stability of oncogenic brain mutations with age, combined with the ability of RNAseq analysis to detect only those clonal mutations with high mosaic fraction, suggests that some oncogenic mutations at a young age may be congenital. The modest reduction of oncogenic mutations with age may then either reflect postnatal elimination of mutant clones from an individual, perhaps by immune surveillance, or postnatal elimination of individuals carrying mutant clones from the healthy cohort. Indeed, we evaluated pre-natal and post-natal brain samples and found that during brain development, the mutation count and the frequency of non-diseased individuals with mutations is highest prenatally and then declines with age. Since the brain is an organ with low overall proliferation in postnatal stages, oncogenic clonal expansions over time can directly result in disease.

Both of our methods only have sensitivity to detect clones with relatively large mosaic fraction. While our targeted sequencing approach has a similar sensitivity to other methods (12), detecting ultra-low clonal events remains challenging. For example, the sensitivity for events with 0.5% VAF is ~10%; hence our rates may be underestimated. Similarly, the RNAseq approach only detects macroscopic clones with VAF>7% (22). Therefore, we cannot by any means rule out, and in fact it seems plausible, that “micro” somatic variants at lower mosaic fraction may indeed show age-related accumulation at levels below our sensitivity to detect them.

All the pathogenic variants found in the cerebral cortex occurred in the WM. Two scenarios may explain this observation: 1) These are derived from active glial proliferation or 2) sub-cortical WM is closer to the ventricles, and clones arising there can reach the WM more readily than the GM. Also, GM harbors large numbers of neurons and these may further dilute such mutations, which might only be detectable by deeper sequencing. A follow-up study of white matter of varying depths could test the second scenario. Future studies evaluating cancer variants in non-diseased brains should evaluate large cohorts of WM samples and GM to a higher depth.

We detected the *IDH1* R132H variant to be enriched in the non-neuronal population, which is 60% OPCs, and this is the most highly proliferative endogenous cell-type of the brain (52). Brain tumors are thought to originate mainly from progenitor cells in neurogenic niches (53); however, the effect of oncogenic mutations in progenitors, such as OPCs residing in non-neurogenic niches such as the cortical WM, remains elusive. OPCs can produce tumors and have been identified in several reports as the most common cell of origin for gliomas (54-57). Thus, the *IDH1* R132H mutation detected in our glial fraction may constitute an early event in a pathogenic progression towards infiltrating glioma. In our case, we did not find more than one mutation per sample, but our limits of sensitivity would likely preclude the identification of emerging sub-clones.

While our study represents a comprehensive survey of those sSNVs and sCNVs identifiable at high allele frequency from fetal to the old ages, a universe of events at lower mosaic fraction remains to be explored. Until now the differential mutational burden between WM and GM remained largely unexplored, and this proved to be critical for discovering oncogenic variants in a normal brain. While our findings also provide important information of

early processes in the acquisition of oncogenic events in the brain, future studies addressing the accumulation of somatic variants in single glial cells may provide another layer of information to continue dissecting early mechanisms of brain oncogenesis.

Acknowledgments: JG was supported by a Basic Research Fellowship from the American Brain Tumor Association BRF1900016 and by the NCI-Brain Cancer SPORE grant P50CA165962. EAM is supported by the Harvard/MIT MD-PhD MSTP program (T32GM007753), the Biomedical Informatics and Data Science Training Program (T15LM007092), and the Ruth L. Kirschstein NRSA F31 Fellowship (F31MH124292). BB is supported by the NHGRI T32 Training Grant (5T32HG002295-18). KLL is supported by the NIH (R01CA188228; R01CA215489; P50CA165962), Pediatric Brain Tumor Foundation and the PLGA Foundation. EAL is supported by the NIH (K01 AG051791; DP2 AG072437) and Suh Kyungbae Foundation. CAW is supported by the NINDS (R01 NS032457), the NIMH (U01 MH106883), and the NIA (R01AG070921). EAL and CAW are supported by the Allen Discovery Center program through The Paul G. Allen Frontiers Group. CAW is an Investigator of the Howard Hughes Medical Institute. We thank W. Bainter and K. Stafstrom for performing the Ion Torrent sequencing. J. Neil and the University of Maryland Brain Bank for helping with the tissue collection. We thank the donors and their families for their invaluable donations for the advancement of science, and the Boston Children's Hospital IDDRC Molecular Genetics Core Facility supported by NIH award U54HD090255 from the National Institute of Child Health and Human Development. We specially thank Diane Shao and Mike Miller for scientific discussion, as well as Sean Hill, Michael Lodato, and Sonia Kim for valuable technical help and discussion. We also thank Nenad Sestan, Christoffer Flensburg, Hunter Fraser, Gad Getz and Keren Yizhak and collaborators for facilitating access to their datasets and methodological advice. Illustrations were created with BioRender.com.

Author contributions: JG conceived and lead the project under the guidance of CAW; EAM lead statistical and computational analyses under the guidance of CAW and EAL; JG, RND and TS designed MIP panel; JG and SB performed the experiments; CK optimized MIP captures; ZZ, JK and EAM contributed to the MIP panel sensitivity analysis; JG, SB, EAM and RND analyzed MIPs captures and validation data; JG, BB and EAM analyzed and interpreted GTEx and BrainVar data; EAM and BB performed statistical analyses; KLL contributed to data interpretation; EAL suggested the spike-in benchmarking and supervised computational analyses; JG, EAM, EAL, and CAW wrote the manuscript.

Materials and Methods

Study design

We analyzed a total of 418 samples derived from 110 individuals with no history of neuro-oncological or other brain diseases, spanning different ages (0-108 years) (Table S1). We used MIPs (23) to evaluate mutations in 121 genes implicated in brain tumors, other cancers, and focal cortical dysplasia (Fig. 1, Table S1). For each of the 110 subjects, we analyzed at least two different brain regions and one non-brain sample (Table S1). Brain samples consisted of hippocampus (HC), cerebellum (CER) and prefrontal cortex that was subdivided into grey (CXG)

and white matter (CXW) (Fig. S1A). When no clear distinction between CXW and CXG was possible, the sample was labeled as CX. Furthermore, we also evaluated 91 samples derived from 17 different organs and the entire left hemisphere from one 17-year-old individual (UMB1465) (Fig. 1, Table S1). We also evaluated the presence of somatic mutations in two large independent cohort of 1,640 and 167 brain samples obtained from the Genotype–Tissue Expression (GTEx) project(20) and BrainVar(21), respectively, using RNA-MuTect (22) (Table S2A-B). GTEx provided samples from 13 different brain regions, and BrainVar provided samples from the DLPFC (mainly from Brodmann area 46) or from the frontal cerebral wall (for donors younger than 10 post-conception weeks).

Variant Calling

Sample FASTQs were first subjected to a local realignment step using CLCbio (QIAGEN). Variant calling was performed using CLCbio- Low Frequency Variant Detection mode that relies on statistical models for evaluating the sequencing error rate based on parameters defined by each batched analysis (http://resources.qiagenbioinformatics.com/manuals/clccancerresearchworkbench/200/index.php?manual=Low_Frequency_Variant_Detection.html). An error model is assumed and estimated for each nucleotide quality score. Error model parameters are all estimated from the data set being analyzed, so will adapt to the sequencing technology used and the characteristics of the particular sequencing run. Samples with average coverage below 396x (1 standard deviation from the mean) were not considered for the analysis. Resulting called variants were filtered out if they had a maximum allele frequency on the population greater than 0.1% (Gnomad)(58), occurred in more than 3 individuals from our cohort, the call quality score was less than 200, found in homopolymeric regions greater than 1 in length, were covered by less than two different MIPS, had less than 12 alternate reads covering the variant, a reference read depth less than 200, variant was present in SNP clusters, were not in a targeted region and was not predicted to have functional impact on the protein function (see below, pathogenic classification). Only variants with VAF between 0.5-15% were analyzed.

For germline calling we evaluated variants between 40% and 60% AF, present in less than 4 individuals to avoid common variants, a site coverage of at least 200x, call quality score 200, covered with more than one MIP and not present in SNP clusters or found in homopolymeric regions greater than one in length. In addition, since in most cases we have multiple tissues from the same individual, the germline variants were required to be present in at least 2 samples to be considered for our analysis. Only variants with potential impact on protein function were included (see below, pathogenic classification) (Table S1).

Pathogenic classification

Pathogenic classification of damaging missense variants was performed following a method reported in a previous study (27), categorizing the predicted pathogenicity relied on 6 different prediction algorithms (SIFT51, PolyPhen2_HDIV52, PolyPhen2_HVAR, MutationTaster53, MutationAssessor54, and LRT55) (59-61), damaging status and conservation sites. For example, NSYND3 the highest pathogenic score, was given if a variant was predicted to

be pathogenic by at least 5/6 of the prediction tools above, considered damaging by CADD, DANN, or FATHMM and affected a conserved site (42,62,63).

GTEx and BrainVar mutation calling and signatures

We implemented RNA-MuTect (22) in Terra's Google Cloud Platform to call somatic mutations in 1,640 bulk RNA-seq brain samples retrieved from the Genotype–Tissue Expression (GTEx) project (release v7) and 167 bulk RNA-seq brain samples retrieved from BrainVar(21). RNA-MuTect was run using both the provided DNA panel-of-normals (PoN) based on ~7000 TCGA normal samples and the provided RNA PoN based on a panel of ~6500 GTEx samples. The threshold for the minimum number of reads supporting the alternative allele was set to 4 as recommended by the pipeline authors (22).All A>G and T>C variants were removed from the callset to reduce false positives from RNA editing artifacts using the maftools package (release 3.11) in R (64) from all analysis unless otherwise specified. Variants greater than 40% VAF were removed from all analysis unless otherwise specified. Additionally, variants were annotated as oncogenic if the protein change was present in the Cancer Mutation Census (v92)(65). BrainVar sample "HSB498" was removed from all analysis as an outlier due to abnormally high mutation count and presence of several oncogenic mutations. Additionally, BrainVar variants present in more than one sample (based on exact cDNA change) were removed from all analysis.

For age trend analyses, statistical regression was performed by fitting a mixed-effects negative binomial model on all sample mutation counts using the *lme4* package (version 1.1-23) in R. For regression on the GTEx samples, the fixed effects included age (years) while adjusting for standardized RNA integrity score, standardized ischemic time, and standardized total mapped reads. GTEx donor ID and brain sub-region were set as a random effect to adjust for donor-specific and region-specific variation. To compare the aging trend for each brain region, the brain region was removed as a random effect and instead the model was fit for each brain region separately. To compare the average mutation count for each brain region, the brain region was changed to a fixed effect where each brain region variable compares its region-specific incidence rate with the overall incidence rate. For regression on the BrainVar samples, the fixed effects included stage (prenatal or postnatal) and sex (male or female) while adjusting for standardized RNA integrity score and standardized total mapped reads. To perform regression on the GTEx and BrainVar samples together, only the GTEx samples from the Frontal Cortex (BA9) were included.

For mutational signature analysis, we used variants from different organs obtained from a recent published study (22). We performed analyses using VAF filtering thresholds of 5-40%, which were relatively consistent. Mutations were filtered to retain only variants with less than 15%VAF (to match the variants we targeted with DNA sequencing) and suspected RNA editing bases were removed (A>G, T>C). List of mutations were analyzed using Mutalisk software (66). Enrichment analysis was performed with a permutation test, where we identified the number of mutational signatures present in a particular cancer type according to the COSMIC signature database (artifact signatures were not included), and then obtained random uniform samples from the total of 50 signatures with replacement and counted the number of signatures that were related to the given cancer type. The number of signatures sampled at each simulation was determined by

the number of signatures that were estimated to contribute to a particular mutational spectrum by Mutalisk.

Somatic copy alteration calling of normal brains

We identified putative sCNVs from non-diseased human brain samples of the GTEx consortium v7(20) using superFreq(39) and BrainVar(21). The algorithm of superFreq employs an error-propagation framework to leverage information from B-allele frequencies and read depth to identify sCNVs from RNA-seq data. Due to the large sample size (n ~ 1,700) samples, which were stored in the cloud, we adapted the superFreq workflow to be run in the cloud. The code to deploy the superFreq workflow as well as usage instruction were made publicly available through a github repository (https://github.com/emauryg/superFreq_cloud). Based on discussions with the developers of superFreq, we used 10 random samples as quality control reference samples for each cohort. These samples are used to remove variation in the data that originate from technical variability. The GTEX and BrainVar cohorts were run separately.

The raw call-set was then filtered to obtain a final call-set that minimized potential false-positives. We focused our analyses to the autosomes, and filtered sCNVs that overlapped less than 100 genes. Based on validation studies of the original superFreq manuscript, events that overlapped at least 100 genes had a high precision of 80-90%, and a recall of 60%. We further filtered variants with predicted breakpoints in the MHC (6: 27486711-33448264, GRCh37) and KIR (9: 54574747-55504099) regions. We also filtered out events that had a clonality of >0.80, since these events would be more likely to be germline events in the non-cancer setting. Lastly, we filtered out variants that did not pass visual inspection based on diagnostic plots.

Statistics

Statistical analyses were performed as described in the main text with un-corrected p-values, and the p-value necessary for significance after multiple hypothesis testing was provided for comparison where relevant. The calculations were done with custom scripts in the R computing language (http://www.r-project.org).

Data and materials availability

Tables of the data are included in supplementary data. Cloud pipeline for RNA MuTect is available at https://github.com/CodingBash/rna_mutect_cloud. Cloud pipeline for superFreq is available at https://github.com/emauryg/superFreq_cloud. Other materials, including analysis scripts are available through the authors upon reasonable request.

Further methodological descriptions can be found in the supplementary information.

References

1. Tomasetti C, Vogelstein B, Parmigiani G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. Proc Natl Acad Sci U S A **2013**;110(6):1999-2004 doi 10.1073/pnas.1221068110.

2. Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science **2015**;348(6237):880-6 doi 10.1126/science.aaa6806.
3. Genovese G, Kahler AK, Handsaker RE, Lindberg J, Rose SA, Bakhoum SF, *et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N Engl J Med **2014**;371(26):2477-87 doi 10.1056/NEJMoa1409405.
4. Jaiswal S, Fontanillas P, Flannick J, Manning A, Grauman PV, Mar BG, *et al.* Age-related clonal hematopoiesis associated with adverse outcomes. N Engl J Med **2014**;371(26):2488-98 doi 10.1056/NEJMoa1408617.
5. Moore L, Leongamornlert D, Coorens THH, Sanders MA, Ellis P, Dentro SC, *et al.* The mutational landscape of normal human endometrial epithelium. Nature **2020**;580(7805):640-6 doi 10.1038/s41586-020-2214-z.
6. Lac V, Nazeran TM, Tessier-Cloutier B, Aguirre-Hernandez R, Albert A, Lum A, *et al.* Oncogenic mutations in histologically normal endometrium: the new normal? J Pathol **2019**;249(2):173-81 doi 10.1002/path.5314.
7. Yokoyama A, Kakiuchi N, Yoshizato T, Nannya Y, Suzuki H, Takeuchi Y, *et al.* Age-related remodelling of oesophageal epithelia by mutated cancer drivers. Nature **2019**;565(7739):312-7 doi 10.1038/s41586-018-0811-x.
8. Martincorena I, Fowler JC, Wabik A, Lawson ARJ, Abascal F, Hall MWJ, *et al.* Somatic mutant clones colonize the human esophagus with age. Science **2018**;362(6417):911-7 doi 10.1126/science.aau3879.
9. Lee-Six H, Olafsson S, Ellis P, Osborne RJ, Sanders MA, Moore L, *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. Nature **2019**;574(7779):532-7 doi 10.1038/s41586-019-1672-7.
10. Lawson ARJ, Abascal F, Coorens THH, Hooks Y, O'Neill L, Latimer C, *et al.* Extensive heterogeneity in somatic mutation and selection in the human bladder. Science **2020**;370(6512):75-82 doi 10.1126/science.aba8347.
11. Brunner SF, Roberts ND, Wylie LA, Moore L, Aitken SJ, Davies SE, *et al.* Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. Nature **2019**;574(7779):538-42 doi 10.1038/s41586-019-1670-9.
12. Keogh MJ, Wei W, Aryaman J, Walker L, van den Ameele J, Coxhead J, *et al.* High prevalence of focal and multi-focal somatic genetic variants in the human brain. Nat Commun **2018**;9(1):4257 doi 10.1038/s41467-018-06331-w.
13. Laug D, Glasgow SM, Deneen B. A glial blueprint for gliomagenesis. Nat Rev Neurosci **2018**;19(7):393-403 doi 10.1038/s41583-018-0014-3.
14. Bohman LE, Swanson KR, Moore JL, Rockne R, Mandigo C, Hankinson T, *et al.* Magnetic resonance imaging characteristics of glioblastoma multiforme: implications for understanding glioma ontogeny. Neurosurgery **2010**;67(5):1319-27; discussion 27-8 doi 10.1227/NEU.0b013e3181f556ab.
15. von Bartheld CS, Bahney J, Herculano-Houzel S. The search for true numbers of neurons and glial cells in the human brain: A review of 150 years of cell counting. J Comp Neurol **2016**;524(18):3865-95 doi 10.1002/cne.24040.
16. Yeung MS, Zdunek S, Bergmann O, Bernard S, Salehpour M, Alkass K, *et al.* Dynamics of oligodendrocyte generation and myelination in the human brain. Cell **2014**;159(4):766-74 doi 10.1016/j.cell.2014.10.011.
17. Ge WP, Jia JM. Local production of astrocytes in the cerebral cortex. Neuroscience **2016**;323:3-9 doi 10.1016/j.neuroscience.2015.08.057.
18. Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, *et al.* Clock-like mutational processes in human somatic cells. Nat Genet **2015**;47(12):1402-7 doi 10.1038/ng.3441.

19. Tomasetti C, Li L, Vogelstein B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. Science **2017**;355(6331):1330-4 doi 10.1126/science.aaf9011.
20. Consortium GT. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science **2020**;369(6509):1318-30 doi 10.1126/science.aaz1776.
21. Werling DM, Pochareddy S, Choi J, An JY, Sheppard B, Peng M, *et al.* Whole-Genome and RNA Sequencing Reveal Variation and Transcriptomic Coordination in the Developing Human Prefrontal Cortex. Cell Rep **2020**;31(1):107489 doi 10.1016/j.celrep.2020.03.053.
22. Yizhak K, Aguet F, Kim J, Hess JM, Kubler K, Grimsby J, *et al.* RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. Science **2019**;364(6444) doi 10.1126/science.aaw0726.
23. Hardenbol P, Baner J, Jain M, Nilsson M, Namsaraev EA, Karlin-Neumann GA, *et al.* Multiplexed genotyping with sequence-tagged molecular inversion probes. Nat Biotechnol **2003**;21(6):673-8 doi 10.1038/nbt821.
24. Ostrom QT, Gittleman H, Liao P, Vecchione-Koval T, Wolinsky Y, Kruchko C, *et al.* CBTRUS Statistical Report: Primary brain and other central nervous system tumors diagnosed in the United States in 2010-2014. Neuro Oncol **2017**;19(suppl_5):v1-v88 doi 10.1093/neuonc/nox158.
25. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, *et al.* IntOGen-mutations identifies cancer drivers across tumor types. Nat Methods **2013**;10(11):1081-2 doi 10.1038/nmeth.2642.
26. Doan RN, Miller MB, Kim SN, Rodin RE, Ganz J, Bizzotto S, *et al.* MIPP-Seq: ultra-sensitive rapid detection and validation of low-frequency mosaic mutations. BMC Medical Genomics **2021**;14(1):47 doi 10.1186/s12920-021-00893-3.
27. Doan RN, Lim ET, De Rubeis S, Betancur C, Cutler DJ, Chiocchetti AG, *et al.* Recessive gene disruptions in autism spectrum disorder. Nat Genet **2019**;51(7):1092-8 doi 10.1038/s41588-019-0433-8.
28. Chakravarty D, Gao J, Phillips S, Kundra R, Zhang H, Wang J, *et al.* OncoKB: A Precision Oncology Knowledge Base. JCO Precision Oncology **2017**(1):1-16 doi 10.1200/po.17.00011.
29. Tischfield JA. Loss of heterozygosity or: how I learned to stop worrying and love mitotic recombination. Am J Hum Genet **1997**;61(5):995-9 doi 10.1086/301617.
30. Yan H, Parsons DW, Jin G, McLendon R, Rasheed BA, Yuan W, *et al.* IDH1 and IDH2 mutations in gliomas. N Engl J Med **2009**;360(8):765-73 doi 10.1056/NEJMoa0808710.
31. Avellaneda Matteo D, Grunseth AJ, Gonzalez ER, Anselmo SL, Kennedy MA, Moman P, *et al.* Molecular mechanisms of isocitrate dehydrogenase 1 (IDH1) mutations identified in tumors: The role of size and hydrophobicity at residue 132 on catalytic efficiency. J Biol Chem **2017**;292(19):7971-83 doi 10.1074/jbc.M117.776179.
32. Gupta R, Flanagan S, Li CC, Lee M, Shivalingham B, Maleki S, *et al.* Expanding the spectrum of IDH1 mutations in gliomas. Mod Pathol **2013**;26(5):619-25 doi 10.1038/modpathol.2012.210.
33. Lodato MA, Woodworth MB, Lee S, Evrony GD, Mehta BK, Karger A, *et al.* Somatic mutation in single human neurons tracks developmental and transcriptional history. Science **2015**;350(6256):94-8 doi 10.1126/science.aab1785.
34. Koivunen P, Lee S, Duncan CG, Lopez G, Lu G, Ramkissoon S, *et al.* Transformation by the (R)-enantiomer of 2-hydroxyglutarate linked to EGLN activation. Nature **2012**;483(7390):484-8 doi 10.1038/nature10898.
35. Turcan S, Rohle D, Goenka A, Walsh LA, Fang F, Yilmaz E, *et al.* IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. Nature **2012**;483(7390):479-83 doi 10.1038/nature10866.

36. Gutmann DH, Loehr A, Zhang Y, Kim J, Henkemeyer M, Cashen A. Haploinsufficiency for the neurofibromatosis 1 (NF1) tumor suppressor results in increased astrocyte proliferation. Oncogene **1999**;18(31):4450-9 doi 10.1038/sj.onc.1202829.
37. Liu KW, Feng H, Bachoo R, Kazlauskas A, Smith EM, Symes K, *et al.* SHP-2/PTPN11 mediates gliomagenesis driven by PDGFRA and INK4A/ARF aberrations in mice and humans. J Clin Invest **2011**;121(3):905-17 doi 10.1172/JCI43690.
38. Endersby R, Baker SJ. PTEN signaling in brain: neuropathology and tumorigenesis. Oncogene **2008**;27(41):5416-30 doi 10.1038/onc.2008.239.
39. Flensburg C, Oshlack A, Majewski IJ. Detecting copy number alterations in RNA-Seq using SuperFreq. Bioinformatics **2021** doi 10.1093/bioinformatics/btab440.
40. von Deimling A, Fimmers R, Schmidt MC, Bender B, Fassbender F, Nagel J, *et al.* Comprehensive Allelotype and Genetic Analysis of 466 Human Nervous System Tumors. Journal of Neuropathology & Experimental Neurology **2000**;59(6):544-58 doi 10.1093/jnen/59.6.544.
41. Hasselblatt M, Isken S, Linge A, Eikmeier K, Jeibmann A, Oyen F, *et al.* High-resolution genomic analysis suggests the absence of recurrent genomic alterations other than SMARCB1 aberrations in atypical teratoid/rhabdoid tumors. Genes Chromosomes Cancer **2013**;52(2):185-90 doi 10.1002/gcc.22018.
42. Cooper GM, Stone EA, Asimenos G, Program NCS, Green ED, Batzoglou S, *et al.* Distribution and intensity of constraint in mammalian genomic sequence. Genome Res **2005**;15(7):901-13 doi 10.1101/gr.3577405.
43. Tatevossian RG, Tang B, Dalton J, Forshew T, Lawson AR, Ma J, *et al.* MYB upregulation and genetic aberrations in a subset of pediatric low-grade gliomas. Acta Neuropathol **2010**;120(6):731-43 doi 10.1007/s00401-010-0763-1.
44. Jeuken JW, Sprenger SH, Boerman RH, von Deimling A, Teepen HL, van Overbeeke JJ, *et al.* Subtyping of oligo-astrocytic tumours by comparative genomic hybridization. J Pathol **2001**;194(1):81-7 doi 10.1002/path.837.
45. Muyas F, Zapata L, Guigo R, Ossowski S. The rate and spectrum of mosaic mutations during embryogenesis revealed by RNA sequencing of 49 tissues. Genome Med **2020**;12(1):49 doi 10.1186/s13073-020-00746-1.
46. Arntfield ME, van der Kooy D. beta-Cell evolution: How the pancreas borrowed from the brain: The shared toolbox of genes expressed by neural and pancreatic endocrine cells may reflect their evolutionary relationship. Bioessays **2011**;33(8):582-7 doi 10.1002/bies.201100015.
47. Garcia-Nieto PE, Morrison AJ, Fraser HB. The somatic mutation landscape of the human body. Genome Biol **2019**;20(1):298 doi 10.1186/s13059-019-1919-5.
48. Cai X, Evrony GD, Lehmann HS, Elhosary PC, Mehta BK, Poduri A, *et al.* Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. Cell Rep **2014**;8(5):1280-9 doi 10.1016/j.celrep.2014.07.043.
49. McConnell MJ, Lindberg MR, Brennand KJ, Piper JC, Voet T, Cowing-Zitron C, *et al.* Mosaic copy number variation in human neurons. Science **2013**;342(6158):632-7 doi 10.1126/science.1243472.
50. Chronister WD, Burbulis IE, Wierman MB, Wolpert MJ, Haakenson MF, Smith ACB, *et al.* Neurons with Complex Karyotypes Are Rare in Aged Human Neocortex. Cell Rep **2019**;26(4):825-35 e7 doi 10.1016/j.celrep.2018.12.107.
51. Sherman MA, Rodin RE, Genovese G, Dias C, Barton AR, Mukamel RE, *et al.* Large mosaic copy number variations confer autism risk. Nat Neurosci **2021**;24(2):197-203 doi 10.1038/s41593-020-00766-5.

52. Huang W, Bhaduri A, Velmeshev D, Wang S, Wang L, Rottkamp CA, *et al.* Origins and Proliferative States of Human Oligodendrocyte Precursor Cells. Cell **2020**;182(3):594-608 e11 doi 10.1016/j.cell.2020.06.027.
53. Lee JH, Lee JE, Kahng JY, Kim SH, Park JS, Yoon SJ, *et al.* Human glioblastoma arises from subventricular zone cells with low-level driver mutations. Nature **2018**;560(7717):243-7 doi 10.1038/s41586-018-0389-3.
54. Assanah M, Lochhead R, Ogden A, Bruce J, Goldman J, Canoll P. Glial progenitors in adult white matter are driven to form malignant gliomas by platelet-derived growth factor-expressing retroviruses. J Neurosci **2006**;26(25):6781-90 doi 10.1523/JNEUROSCI.0514-06.2006.
55. Zong H, Parada LF, Baker SJ. Cell of origin for malignant gliomas and its implication in therapeutic development. Cold Spring Harb Perspect Biol **2015**;7(5) doi 10.1101/cshperspect.a020610.
56. Suva ML, Tirosh I. The Glioma Stem Cell Model in the Era of Single-Cell Genomics. Cancer Cell **2020**;37(5):630-6 doi 10.1016/j.ccell.2020.04.001.
57. Liu C, Sage JC, Miller MR, Verhaak RG, Hippenmeyer S, Vogel H, *et al.* Mosaic analysis with double markers reveals tumor cell of origin in glioma. Cell **2011**;146(2):209-21 doi 10.1016/j.cell.2011.06.014.
58. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. Nature **2020**;581(7809):434-43 doi 10.1038/s41586-020-2308-7.
59. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, *et al.* A method and server for predicting damaging missense mutations. Nat Methods **2010**;7(4):248-9 doi 10.1038/nmeth0410-248.
60. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods **2010**;7(8):575-6 doi 10.1038/nmeth0810-575.
61. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res **2011**;39(17):e118 doi 10.1093/nar/gkr407.
62. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics **2015**;31(5):761-3 doi 10.1093/bioinformatics/btu703.
63. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res **2019**;47(D1):D886-D94 doi 10.1093/nar/gky1016.
64. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res **2018**;28(11):1747-56 doi 10.1101/gr.239244.118.
65. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res **2019**;47(D1):D941-D7 doi 10.1093/nar/gky1015.
66. Lee J, Lee AJ, Lee JK, Park J, Kwon Y, Park S, *et al.* Mutalisk: a web-based somatic MUTation AnaLyIS toolKit for genomic, transcriptional and epigenomic signatures. Nucleic Acids Res **2018**;46(W1):W102-W8 doi 10.1093/nar/gky406.

Figure legends

Figure 1. Experimental strategy. General scheme of our methodological pipeline using targeted sequencing for variant discovery and in silico analyses of large-independent cohorts spaning from fetal to older ages (BrainVar and GTEx databases). DNA obtained from 418 samples derived from normal brain and non-brain tissue were analyzed using MIPs, capturing genes associated with brain tumors, pan-cancer, and focal cortical dysplasia. Samples were deep-sequenced, and called variants were validated using ion torrent ultra-deep-sequencing. Brain samples from BrainVar (n=166) and GTEx (n=1640) databases were analyzed to discover oncogenic variants and to evaluate copy number variants, mutational signatures and aging correlations.

Figure 2. Non-diseased brains harbor low allele frequency cancer reported mutations. (**A**) Correlation between MIPS and Ion torrent VAFs of 12 unique variants detected in the normal brain. (**B**) List of validated brain-specific somatic variants showing general information such as variant allele frequency, pathogenicity prediction, and presence in cancer databases. Additionally, we describe detected germline variants with functional impact for each individual. N/R= Not reported, NSYND3=highest score in predicted pathogenicity, UNCERTAIN= Uncertain pathogenicity of variants with clinical significance, CLINSIG. (**C**) Distribution frequency of genes affected by the detected oncogenic variants found in the brain. (**D**) Distribution frequency of the most affected driver genes found in LGG with pathogenicity relevance (Intogen database), black arrows indicate overlap with our discovered genes. (**E**) Number of oncogenic mutations found in CXW (circles, two-tailed Fisher exact test, p=0.025) as a function of age (years). (**F**) Comparison of the number of pathogenic mutations found in CX (n=53), CXG (n=92) and CXW (n=94) and HC (n=69) (Fisher exact test, CXW vs. CXG p=0.028, HC vs. CXG p=0.182, CXW vs. HC p=0.4).

Figure 3. Oncogenic mutations are enriched in the white matter and glial cells. (**A**) Schematic illustrating the discovery of the *IDH1* R132H mutation in the normal PFC of a 17y/o individual. The mutation was identified in two adjacent WM samples and not present elsewhere, including GM from the same section or GM/WM from the following brain section. White matter is mainly composed of neuronal axons, astrocytes, oligodendrocytes, and OPCs, while GM is a combination of neurons with glial cells. (**B**) Illustration of the two focal and distant pathogenic mutations found within the same brain. (**C**) Schematic of nuclear sorting protocol to isolate neuronal (NEUN+) and non-neuronal cells (NEUN-). Nuclei were evaluated using single-cell RNAseq. TSNE plot of 3700 NEUN+ nuclei, showing an exclusive presence of excitatory and inhibitory neurons but not glia (upper panel). Evaluation of 1800 NEUN- nuclei showing the presence of glial cells but not neurons (lower panel). (**D**) Fold change gene expression of NEUN- vs. NEUN+ nuclei subdivided by different brain cell-types. (**E**) Genotyping of the *IDH1* R132H mutation by ddPCR. Graph shows the ratio of mutant/wild type droplets analyzed in different sorted populations (each data point corresponds to 300 sorted nuclei), showing a nominal enrichment in the NEUN- glial fraction. Genomic DNA without the *IDH1* R132H mutation was used as a control for the ddPCR reaction (CTRL).

Figure 4. Somatic mutations detectable by RNAseq do not accumulate with age in the normal brain. Evaluation of somatic mutations using RNA-MuTect in 1,640 GTEx and 166 BrainVar non-diseased brain samples. Suspected RNA editing bases A>G, and T>C were

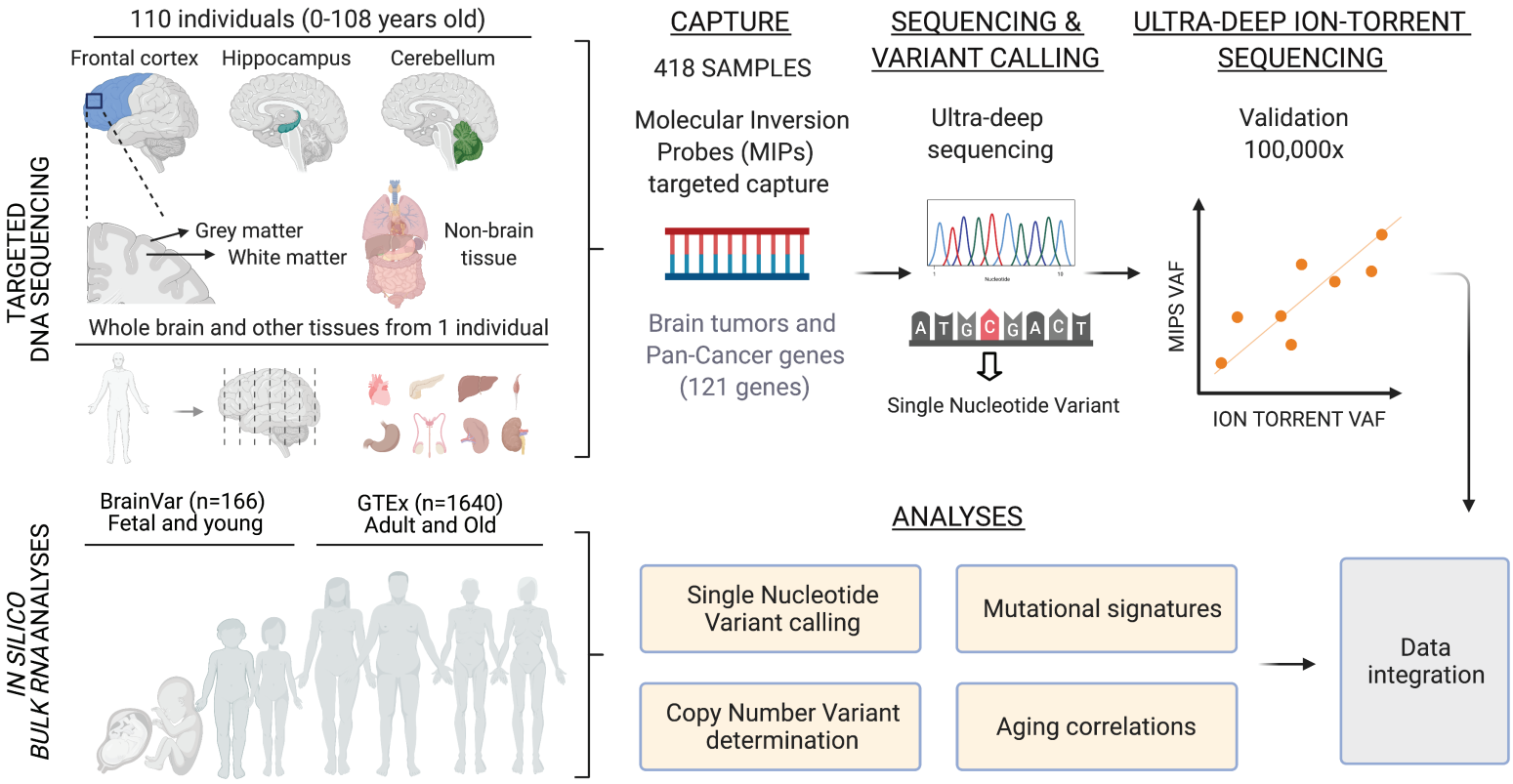
removed to reduce false positives as well variants with VAF>40%. Dotplots showing the proportion of samples with at least one mutation across age and forest plots of the aging incidence rate ratio for all mutations (**A**), predicted pathogenic and non-pathogenic mutations (**B**), and disruptive (nonsense, splice site) and non-disruptive mutations (3' UTR, 5' UTR, 5' flank, or nonstop) (**C**). Error bars are the Clopper-Pearson 95% confidence interval of the sample proportion. Forest plots also include standardized RNA integrity score, and standardized total mappable read count, with horizontal lines indicating 95% confidence intervals. Incidence rate ratio was estimated using mixed-effects negative binomial model with donor ID as a random-effect.

Figure 5. Copy number variants are found in the normal brain. Visualization of detected somatic copy number variants (sCNVs) in the GTEx and BrainVar databases (**A**). Color codes indicate different age ranges, brain regions and types of alteration (gain, loss or loss of heterozygosity (LOH)). Upper bar plot summarizes the counts of the different alteration types and the bar plot on the left side also summarizes the alteration types but sorted by chromosome. Labels along the left y-axis refer to chromosome arms and the percentages displayed in the right y-axis represent the frequency of events in each chromosome arm. (**B**) Distribution of sCNVs events across different ages from prenatal to elder. The number of individuals under each age range is described under the age label. (**C**) Graphic representation of two representative sCNVs and the genes located in that region, one involving the whole chromosome 19 found at 19% clonality, and the second involving the q-arm of chromosome 22 found at 24% clonality.

Figure 6. Mutations found in normal brain exhibit signatures present in brain tumors. Mutational signature analysis of normal brain and skin (VAF<15%). Number of mutations evaluated is described next to the tissue label and graphs show bases substitution, signatures and spectrum obtained for each tissue. Colored circles next to each signature represent that the signature was observed in cancer (green=brain cancer, purple=skin cancer).

Author Manuscript Published OnlineFirst on August 13, 2021; DOI: 10.1158/2159-8290.CD-21-0245
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

Figure 1



Author Manuscript Published OnlineFirst on August 13, 2021; DOI: 10.1158/2159-8290.CD-21-0245
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

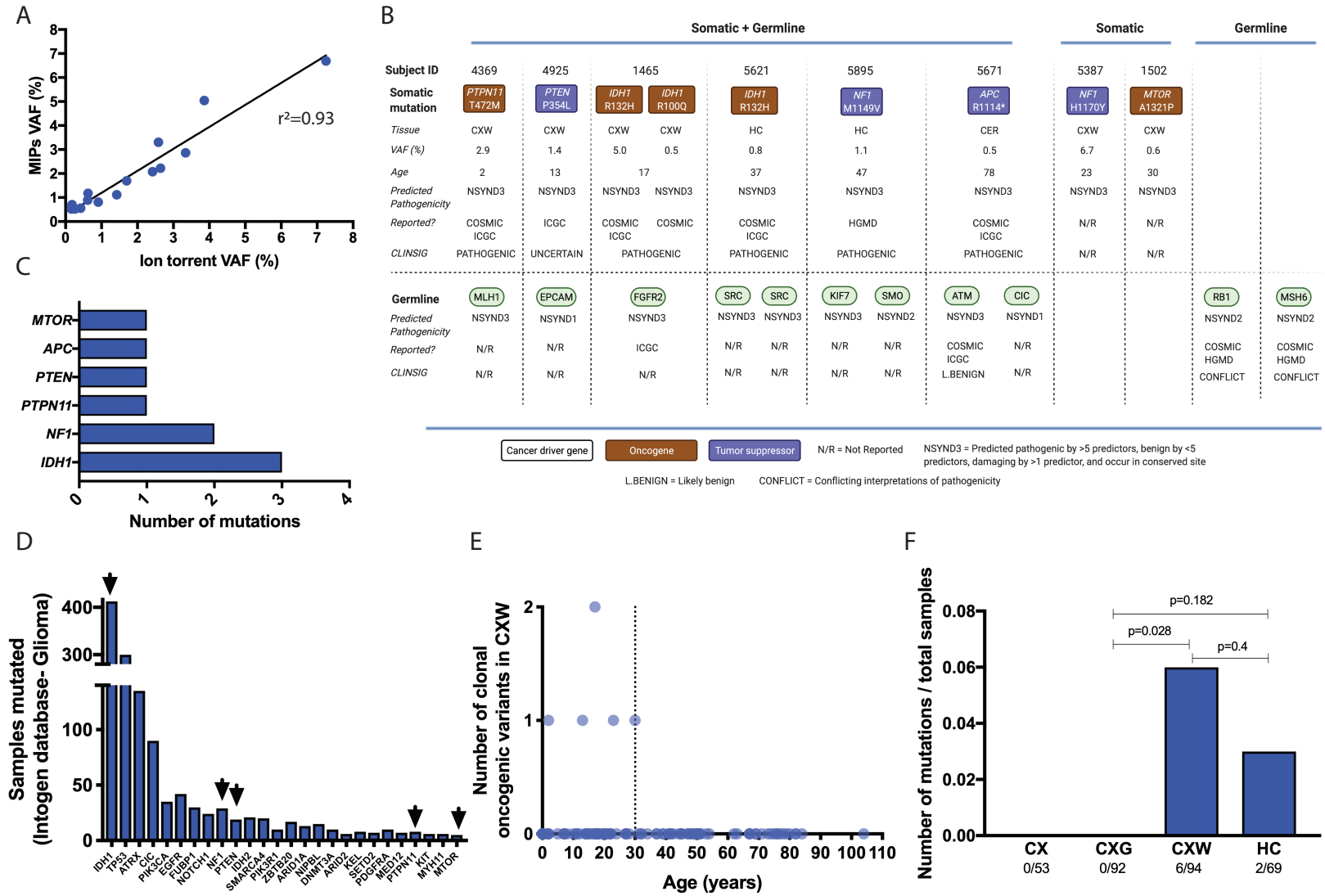
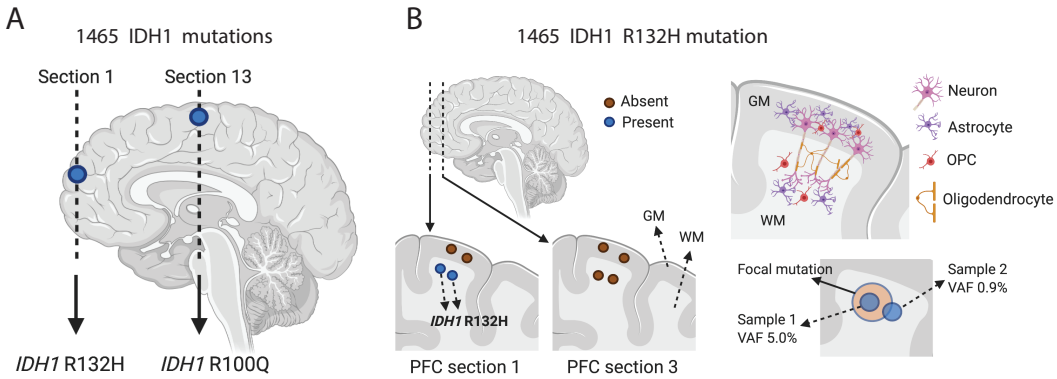
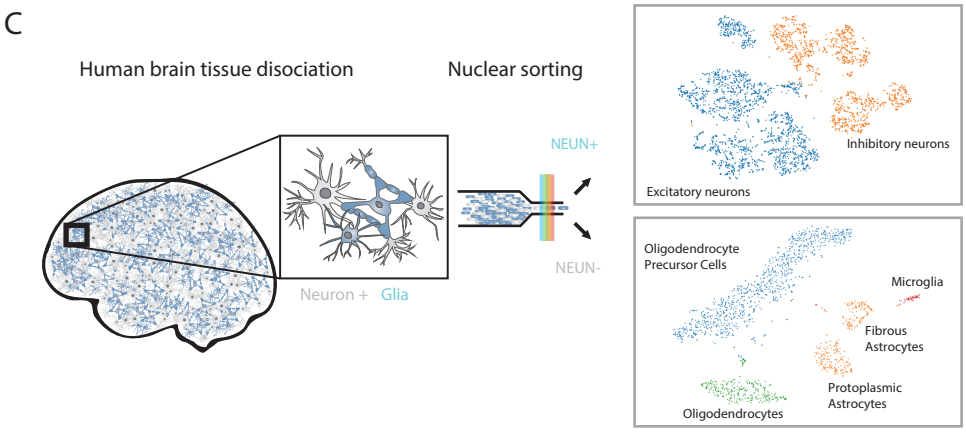
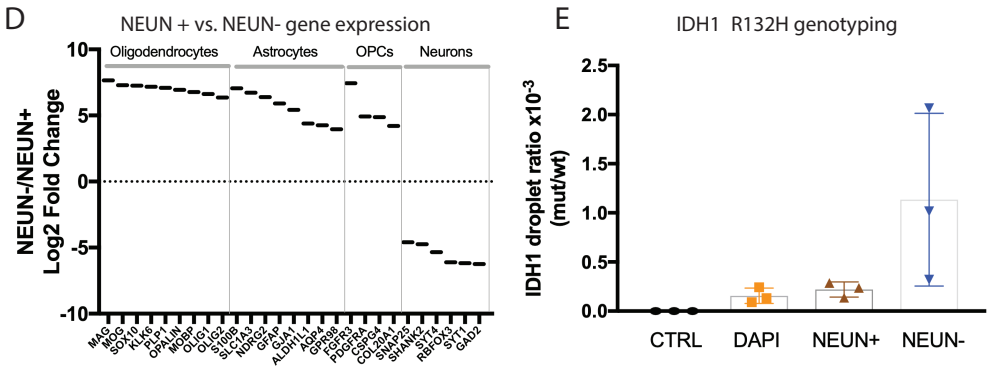


Figure 3

Author Manuscript Published OnlineFirst on August 13, 2021; DOI: 10.1158/2159-8290.CD-21-0245
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

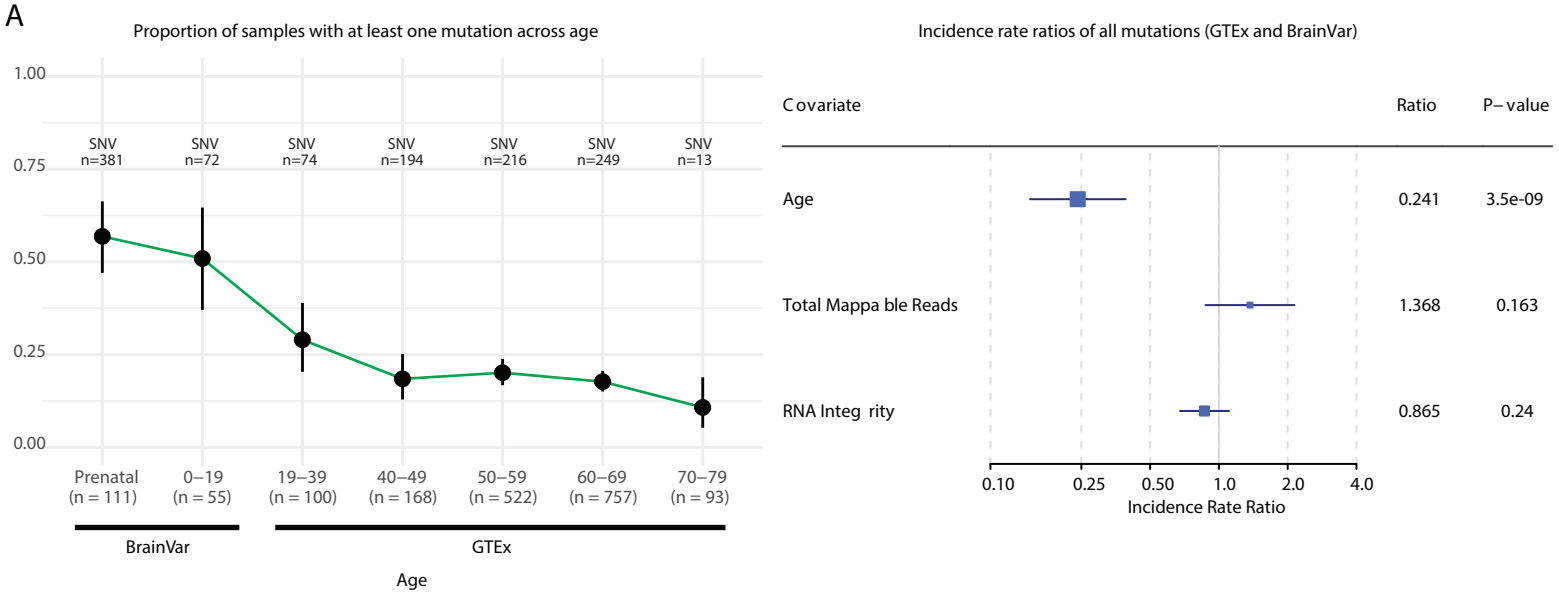




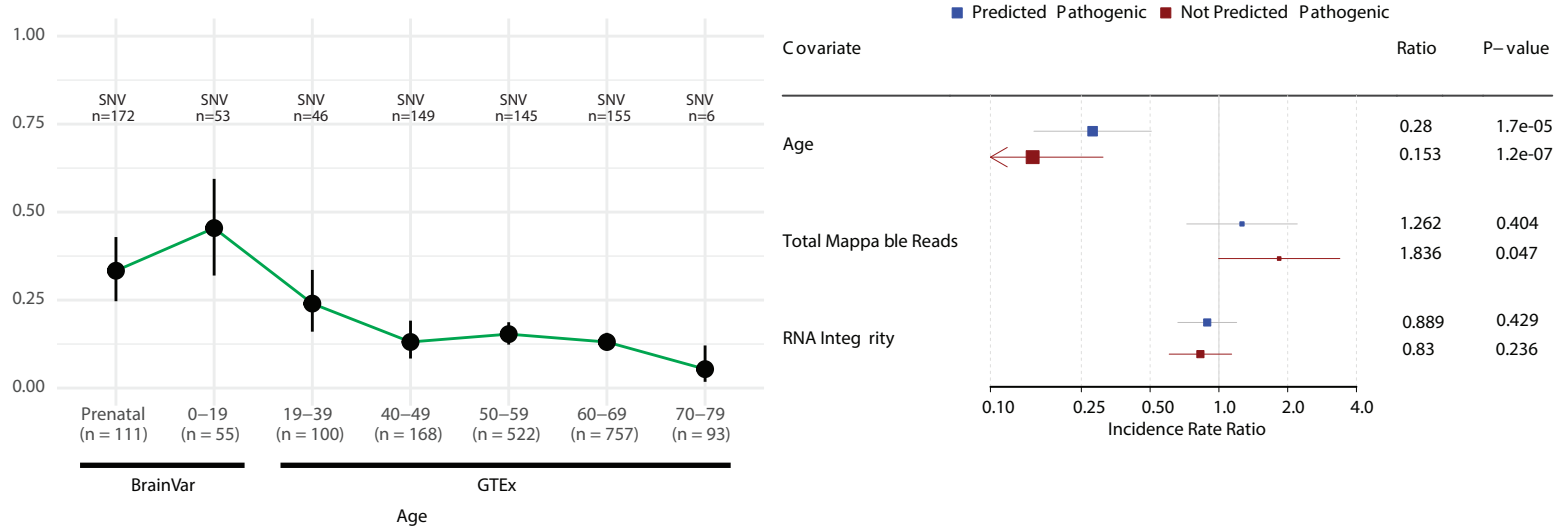


Author Manuscript Published OnlineFirst on August 13, 2021; DOI: 10.1158/2159-8290.CD-21-0245
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

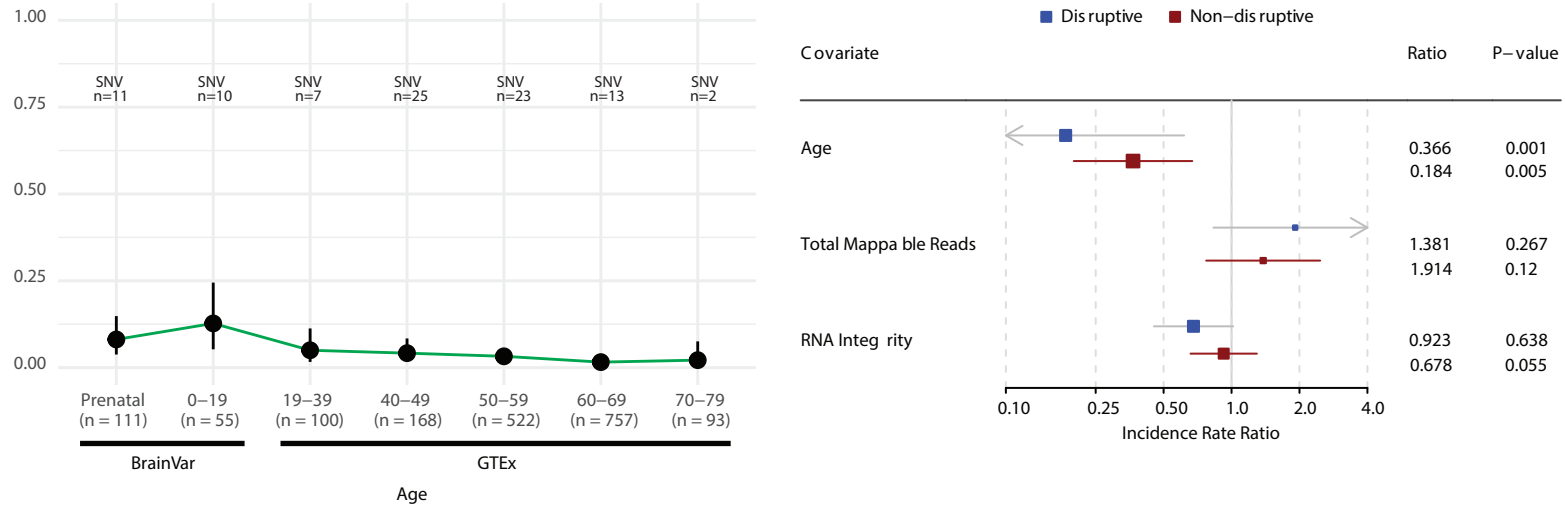
A



B



C



Author Manuscript Published OnlineFirst on August 13, 2021; DOI: 10.1158/2159-8290.CD-21-0245
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

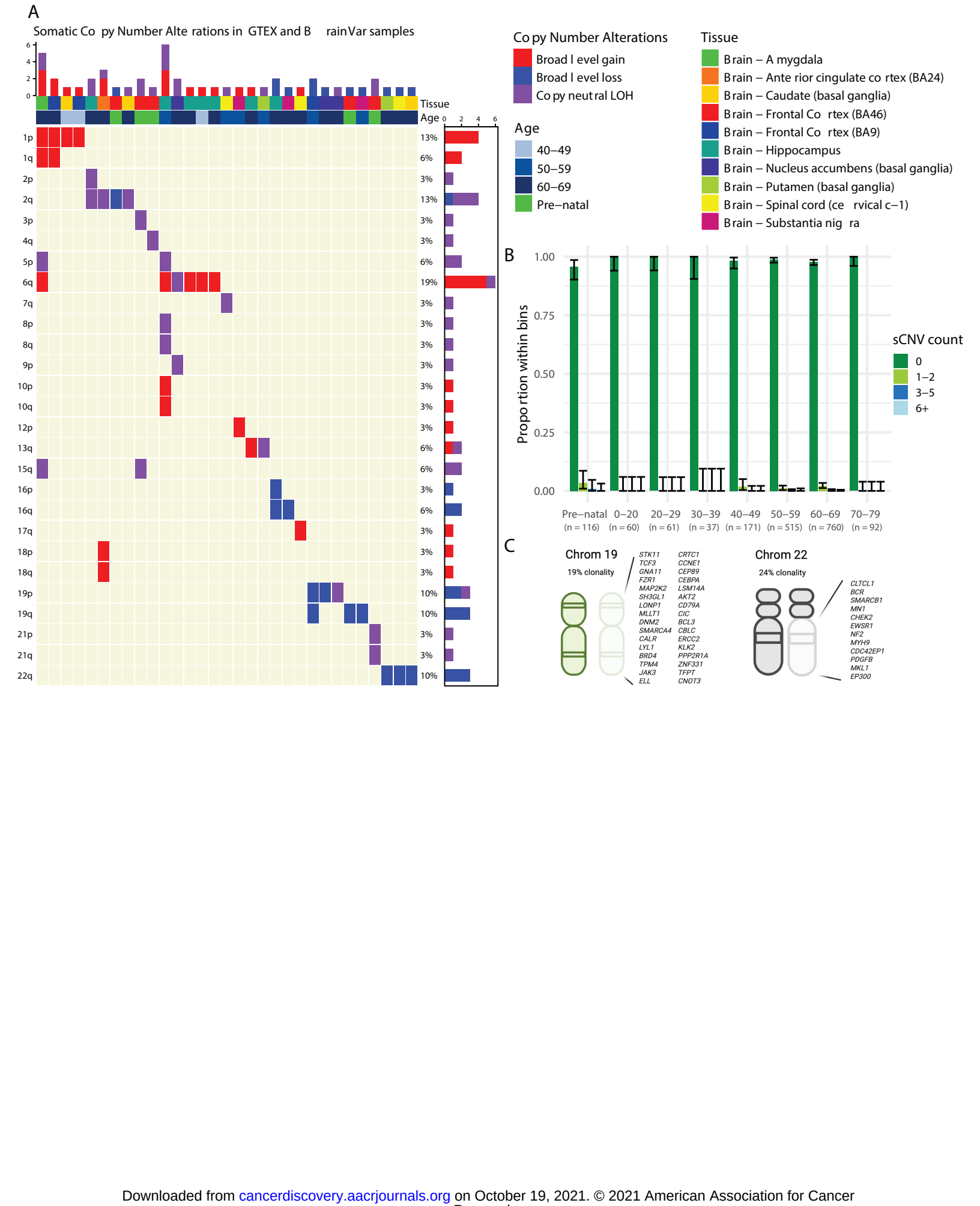
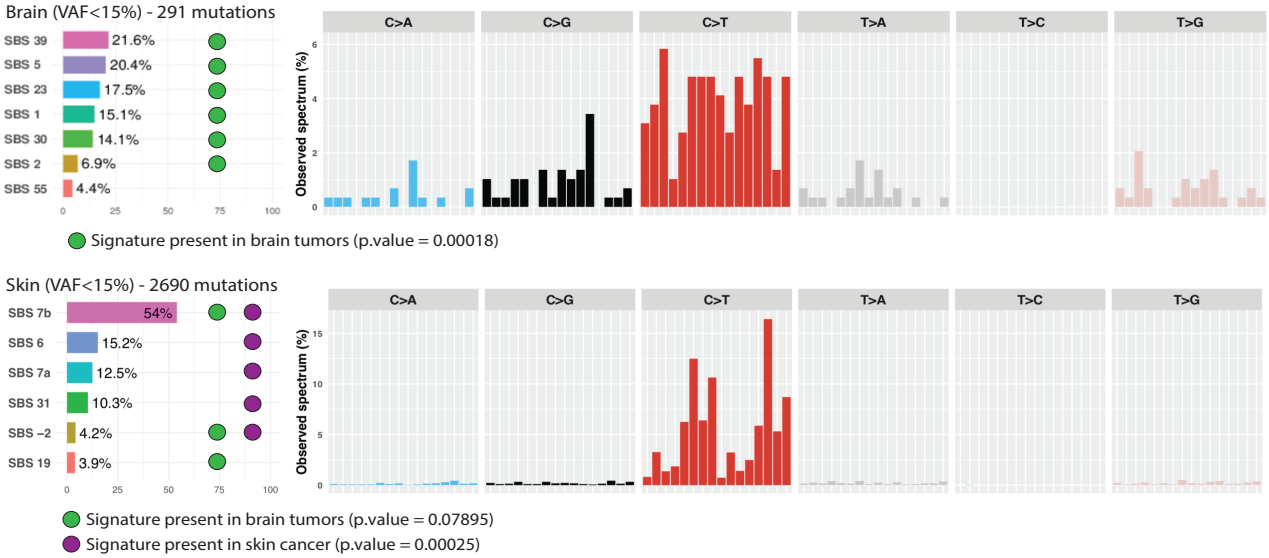


Figure 6

Author Manuscript Published OnlineFirst on August 13, 2021; DOI: 10.1158/2159-8290.CD-21-0245



AACR American Association for Cancer Research

CANCER DISCOVERY

Rates and patterns of clonal oncogenic mutations in the normal human brain

Javier Ganz, Eduardo A Maury, Basheer Becerra, et al.

Cancer Discov Published OnlineFirst August 13, 2021.