Structural variants shape driver combinations and outcomes in pediatric high-grade glioma

Frank P. B. Dubois, Ofer Shapira, Noah F. Greenwald, Travis Zack, Jeremiah Wala, Jessica W. Tsai, Alexander Crane, Audrey Baguette, Djihad Hadjadj, Ashot S. Harutyunyan, Kiran H. Kumar, Mirjam Blattner-Johnson, Jayne Vogelzang, Cecilia Sousa, Kyung Shin Kang, Claire Sinai, Dayle K. Wang, Prasidda Khadka, Kathleen Lewis, Lan Nguyen, Hayley Malkin, Patricia Ho, Ryan O’Rourke, Shu Zhang, Rose Gold, Davy Deng, Jonathan Serrano, Matija Snuderl, Chris Jones, Karen D. Wright, Susan N. Chi, Jacques Grill, Claudia L. Kleinman, Liliana C. Goumnerova, Nada Jabado, David T. W. Jones, Mark W. Kieran, Keith L. Ligon, Rameen Beroukhim, and Pratiti Bandopadhayay

We analyzed the contributions of structural variants (SVs) to gliomagenesis across 179 pediatric high-grade gliomas (pHGGs). The most recurrent SVs targeted MYC isoforms and receptor tyrosine kinases (RTKs), including an SV amplifying a MYC enhancer in 12% of diffuse midline gliomas (DMG), indicating an underappreciated role for MYC in pHGG. SV signature analysis revealed that tumors with simple signatures were TP53 wild type (TP53WT) but showed alterations in TP53 pathway members PPMID and MDM4. Complex signatures were associated with direct aberrations in TP53, CDKN2A and RB1 early in tumor evolution and with later-occurring extrachromosomal amplicons. All pHGGs exhibited at least one simple-SV signature, but complex-SV signatures were primarily restricted to subsets of H3.3K27M DMGs and hemispheric pHGGs. Importantly, DMGs with complex-SV signatures were associated with shorter overall survival independent of histone mutation and TP53 status. These data provide insight into the impact of SVs on gliomagenesis and the mechanisms that shape them.

pHGGs, which encompass DMGs and hemispheric tumors, represent the most common cause of cancer-related deaths in children aged 0–14 years. Targeted sequencing, including exome sequencing, has revealed recurrent single-nucleotide variants (SNVs) in histones including H3.1K27M, H3.3K27M and H3.3G34R, co-occurring with alterations in the TP53 pathway and RTK genes. However, the role of SVs remains underexplored.

SVs represent connections or rearrangements between distant genomic loci. They underlie all somatic copy number (CN) alterations (SCNAs) except whole-chromosome gains and losses, thereby altering more of the genome than any other type of genetic alteration. A single SV can affect dozens to hundreds of genes. In some cases, SVs generate extrachromosomal amplifications (also known as double minutes (DMs)) that can encode hundreds of copies of an oncogene per cell. The effects of SVs on cellular fitness often result from changes in chromatin structure such as disruption of topologically associated domains (TADs) and gene-enhancer interactions. Therefore, in contrast to SNVs, the SVs with the largest effects on selection are often outside the exome and require whole-genome sequencing (WGS) for their characterization. Moreover, these distant effects raise challenges when interpreting the consequences of individual SVs.

Both the frequency with which SVs recur at individual loci and the mechanisms by which they are generated can vary widely across cancers. SV signatures can indicate their formation mechanisms, and recent efforts have begun to characterize these signatures in...
breast cancers and other cancer types. However, whereas SNV signatures have been characterized across tens of thousands of exomes, the relationships between currently described SV signatures across cancer types remain underexplored. For example, high rates of tandem duplications are associated with deficiencies in homologous recombination (HR) only in tumors with very high SV burdens. It remains unclear whether these associations translate to other tumor types, including pHGGs, and which other SV signatures or associated variant-generating processes exist.

The differences across lineages indicate the role of epigenetics in shaping the SVs that are observed in cancer. Mutations in core histones in pHGGs highlight the role of epigenetic dysregulation in these tumors. pHGGs therefore offer a unique perspective on the relationships between patterns of SVs and different alterations in chromatin. Associations between patterns of SVs and other molecular and clinical characteristics of these tumors are also largely unknown.

Historically, the characterization of DMGs lacked pretreatment tissue owing to the risks involved in performing biopsies of midline brain structures. A concern with posttreatment samples is that treatment—often involving ionizing radiation—might alter the SV patterns in these tumors. We leveraged samples from the first multi-institutional North American clinical trials to incorporate biopsies of DMGs and added published pre- and posttreatment samples to assemble the largest pHGG WGS cohort to date. We identified recurrent driver events, stratified pHGGs based upon mechanistically informative SV signatures, and detected genetic events and differences in clinical outcomes associated with these signatures.

Results

Significantly recurrent SVs. We assembled a pHGG WGS cohort including 61 hemisphere tumors and 118 DMGs from 179 children. Of these, 61 were sequenced de novo for this study. The other 118 samples include 18 from Buczkowicz et al., 20 from Taylor et al., 30 from Wu et al. and 50 from Bender et al. (Supplementary Table 1). All sequences were subjected to a single uniform computational pipeline. Among the DMGs, 84 (71%) were from pretreatment biopsies, including 33 obtained from the first multi-institutional North American clinical trial to incorporate diagnostic biopsies. The tumor purity of the pretreatment biopsies was comparable with the observed number of SVs in each bin with a background distribution that considers sequence, epigenetic and other features of each locus (all results in Supplementary Table 4B).

We identified significantly recurring SCNAs using GISTIC and recurrent SNVs using MutSig2CV. The results of both analyses largely agreed with those of prior studies and other known drivers of pHGG (Extended Data Fig. 1b). The most notable finding was a recurrent amplification in 8q24.21, 2 MB telomeric to MYC. This amplicon, which was probably not detected in prior array- and exome-based studies because it lies outside the exome, was present in 28 tumors (16% of the cohort). All but one of these tumors were DMGs, a significant enrichment (P = 0.0016). Most of these amplons excluded MYC itself (Extended Data Fig. 1c). A nonoverlapping peak was also detected that did encompass MYC, owing to two tumors with extra- chromosomal MYC amplons.

We also found that several regions were recurrently amplified together to high levels, including 2p25.1 and the MYCN locus at 2p24.3. This pattern of correlated SCNAs in distinct genomic loci suggests underlying recurrent SVs. We therefore comprehensively cataloged SVs using an assembly-based method with improved ability to detect complex and short SVs compared with standard alignment-based methods. We detected 15,485 SVs (Supplementary Table 3), averaging 87 per tumor, including 1482 (10%) that were 10–300 base pairs (bp) in span; this was a ‘blind spot’ in prior analyses.

To distinguish recurrent SVs, we took two approaches based on methods that we have recently developed. In the first, we conducted a one-dimensional (1D) analysis that identified genomic loci with more SV breakpoints than expected (termed significantly recurrent breakpoints or SRBs, Supplementary Table 4a). This analysis splits the whole genome into 50 kbp bins and compares the observed number of SVs in each bin with a background distribution that reflects the rates at which each bin suffers breakpoints and the genomic distances between them. The bins and background model in this analysis were determined in a prior pan-cancer study; the 2D bin median length was 467 kbp. In both analyses, bins with q values < 0.1 were considered significant.

We identified 10 SRB bins across five TADs (Supplementary Table 4a,b, Fig. 1a and Extended Data Fig. 1d) and two SRJs (Fig. 2a). The most significant SRB was within the MYC TAD, encompassing breakpoints in 28 tumors—more than for any other TAD in the genome. This locus was also a component of an SRJ connecting two adjacent bins at the telomeric end of the MYC

Fig. 1 | SRBs within CCDC26. a, Significance (multiple hypothesis corrected q values, vertical axis) of recurrent breakpoints (genomic positions on the horizontal axis) across the 179 pHGG genomes. b, CN profiles across the MYC TAD for the 15 tumors with recurrent CCDC26–SV, c, H3K27ac ChIP–seq tracks within the TAD containing MYC (green lines) showing an H3K27ac peak at the location of CCDC26–SV in six H3K27M and four H3K27WT pHGGs tumors. Only the top H3K27ac enrichment track originates from a tumor with a CCDC26–SV. Significantly enriched peaks (q value < 0.01) are indicated below each H3K27ac ChIP–seq track. The CCDC26 amplonic boundaries for individual samples are indicated by the paired red arrows at the top. The consensus amplicon is indicated by the red dotted lines and centers on an H3K27ac peak. d, Hi–C heatmap across the MYC–CCDC26 locus from a midline glioma with CCDC26–SV. Increasing interaction frequencies are indicated by brighter shades of red. The black arrowheads indicate significant interaction loops. The track beneath the heatmap indicates RobustTAD left and right boundary scores, which represent the likelihood that TAD boundaries are present. The third row contains a virtual 4C track, in which peaks indicate higher interaction frequencies with an anchor sequence in the MYC promoter, which is highlighted in gray. The fourth row shows H3K27ac ChIP–seq data from the same sample indicating the location of the enhancer peak within the CCDC26–SV consensus amplicon. e, Normalized MYC expression in DMG samples with WT CN profiles at 8q24.21 (n = 92 tumors), CCDC26–SVs (n = 8 tumors) or amplifications of the MYC coding sequence (n = 12 tumors). *P = 0.04 as determined by two-sided Wilcoxon rank sum test. The center line of the boxplot indicates the median, bounds of the box indicate the 25th and 75th percentiles, and whiskers extend from the box to the largest or smallest value no further than 1.5x IQR. f, Schematic illustrating the luciferase reporter used to validate the enhancer in CCDC26, showing the positions of the E1 and E2 sequences with respect to the enhancer within CCDC26. g, Luciferase activity in DIPG13 cells following transduction of the E1, E2 and LUAD enhancer reporters or empty vector controls. Values represent the average of four technical replicates in each of three independent experiments. *P = 1.6 × 10^-5 (E1 versus backbone) and P = 0.89 (LUAD versus backbone), n = 3 independent experiments, nested one-way ANOVA with Tukey’s post–test; boxplot defined as in e.
TAD. The remaining SRBs corresponded to SVs within the TADs of the RTK genes MET ($q = 0.0025$), EGFR ($q = 0.029$) and PDGFR ($q = 0.032$), as well as an SV within the TAD of the transcription factor ID2 (Supplementary Table 4a). This latter SRB was also a component of the second SRJ, which connected ID2 and MYCN.

**Recurrent simple SVs activate MYC by enhancer amplification.** Our 1D and 2D analyses both nominated rearrangements within the MYC TAD. Among the 28 tumors with SVs in this TAD, six contained complex rearrangements that amplified MYC, connecting the amplicon to locations outside the MYC TAD. An additional
15 tumors contained a tandem duplication centering on intron 1 of the long noncoding RNA CCDC26, with a median span of 216 kbp and a minimal common region of amplification (MCR) of only 42 kbp (Fig. 1b). The remaining seven rearrangements exhibited no consistent structure.

The 2Mb region telomeric to MYC has been shown to contain MYC enhancers in lineage-specific locations in several cancer types44. We therefore hypothesized that the CCDC26 amplicon promoted oncogenesis by amplifying an associated neural-lineage enhancer. We analyzed published H3K27ac enhancer tracks generated from H3.3K27M and H3.1pHGGs45 and adult glioblastomas46 and observed H3K27ac enhancer peaks within the MCR of the CCDC26 amplicon (Fig. 1c and Extended Data Fig. 2a). We also confirmed the presence of this enhancer in an independent pHGG assay for transposase-accessible chromatin using sequencing (ATAC-seq) dataset47 and H3K27ac chromatin immunoprecipitation sequencing (ChiP–seq) data from normal neural tissue38 (Extended Data Fig. 3b). The CCDC26–SV amplifies an enhancer that is present across neural lineages but not in enhancer maps from nonneural lineages (hematopoietic and lung tissues) (P = 0.0005; Fisher’s exact test). We conclude that the CCDC26 amplicon centers on a neural-lineage-specific enhancer.

Although this enhancer is present across neural tissues, histone-mutant DMGs exhibit 31% more H3K27ac binding at this locus than H3WT gliomas—the fourth most differential superenhancer between these two groups (q = 0.05; P = 0.0021)48. Intriguingly, the CCDC26 amplicon also occurred almost exclusively in H3.3K27M gliomas (14/97 H3.3K27M versus 1/82 H3WT or H3K27M tumors; q = 0.0018). We conclude that the CCDC26–SV amplifies an enhancer that is present across neural lineages but enriched in H3.3K27M gliomas.

The lineage-specific H3K27ac enhancer at the CCDC26 locus also appears to interact directly with MYC. We evaluated the chromatin topology of an H3.3K27M pHGG that harbored a CCDC26–SV using Hi-C. The CCDC26–SV breakpoints were with the MYC TAD, and the H3K27ac peak within the CCDC26–SV formed an interaction peak with the MYC promoter (Fig. 1d). However, this interaction was not restricted to tumors with CCDC26–SVs. Analysis of Hi-C data generated in a CCDC26–pHGG, two patient-derived H3.3K27M cell lines and induced pluripotent stem cell (iPSC)-derived neural progenitors (Extended Data Fig. 2b) revealed a similar TAD structure and interactions with the MYC promoter. We conclude that the CCDC26–SV amplifies a preexisting MYC enhancer.

The eight pHGGs with the CCDC26–SV also exhibited increased MYC RNA expression compared with tumors without SVs or amplifications in the MYC TAD (8q24.21–WT, n = 92, P = 0.04, Fig. 1e) and similar levels of MYC to those of pHGGs with amplifications of the MYC coding sequence (P = 0.85). Indeed, the absolute CNs of both MYC and the CCDC26–SV were correlated with MYC expression (P = 0.0003 for MYC and P = 0.01 for the CCDC26–SV, Spearman rank correlations; Extended Data Fig. 2c,d). We conclude that CCDC26–SVs and associated enhancer amplifications activate MYC expression.

We next used minimal reporter assays to confirm that the H3K27ac peak in CCDC26 represented a functionally relevant, lineage-specific enhancer. We generated two enhancer reporter systems (E1 and E2), each encompassing slightly more than half of the enhancer, with a small region of overlap (Fig. 1f). We transduced two histone-mutant pHGG cell lines with the reporter constructs and evaluated induction of luciferase expression to mark enhancer activity. In both lines, the E1 enhancer region was sufficient to increase luciferase expression relative to the vector control (P < 0.01 in both cases, n = 3, nested one-way analysis of variance (ANOVA); Tukey’s multiple comparisons; Fig. 1g, Extended Data Fig. 2e). We performed similar experiments in the lung cancer cell line A549 and found no increase in luciferase activity, although previously validated lung adenocarcinoma (LUAD) MYC enhancers49 did induce luciferase activity (P = 0.96 (E1 versus control); P = 0.0071 (LUAD versus control), n = 3, nested one-way ANOVA: Tukey’s multiple comparisons; Extended Data Fig. 2f).

**MYCN activation through enhancer amplification and hijacking.** Somatic enhancer amplification also seems to play a part in the activation of MYCN in pHGG. The SRJ (Extended Data Fig. 3a) connecting ID2 with MYCN represents a set of complex SVs in tumors with high-level amplifications within this region on chromosome 2 (Fig. 2b). ID2 is a transcription factor regulating neural differentiation50, and the ID2 locus is associated with an H3K27ac enhancer that was present across all analyzed pHGG tumor samples (Extended Data Fig. 3b). The MCR of the MYCN–ID2 amplification contains both the ID2-associated enhancer and the coding sequence of MYCN (Extended Data Fig. 3b). The SVs result in juxtaposition of the enhancer in ID2 with MYCN, reducing the distance between the two from the normal 7 Mb to less than 700 kbp (Fig. 2c). These data suggest these SVs hijack the ID2 enhancer to activate MYCN.

We also identified four pHGGs with MYCN amplifications that did not connect to ID2. However, these latter ‘localized’ MYCN amplicons always encompassed more of the immediate neighborhood of MYCN than the complex MYCN–ID2 amplicons. In contrast to MYCN–ID2 amplicons, which only contained a small fraction of the MYCN TAD (23% on average), localized MYCN amplicons contained most of this TAD (60% on average, P = 0.03, t-test), including several enhancers not included in the MYCN–ID2 amplicons (Fig. 2d).

The high-level MYCN amplicons showed typical characteristics of extrachromosomal amplicons, reaching CNs of 50–300 per cell. Other oncopgenes with absolute CNs greater than 10 have been
Fig. 3 | High-level amplicons. **a,** For each genomic locus (horizontal axis), the number of tumors containing a high-level (CN > 10) amplicon is indicated (vertical axis). SRBs are highlighted at the top. **b, c,** Simple and complex SVs exhibit distinct mechanisms to activate PDGFRA. The top and bottom tracks indicate CN and the significance of H3K27ac enrichment (as calculated from eight pHGGs), respectively. SVs are highlighted in red. Selected gene loci are indicated at the bottom. **b,** A simple amplicon of a region with known **PDGFRA** enhancers. **c,** A complex high-level **PDGFRA** amplicon, displayed as in panel (b) with the addition of a track (second from top) indicating the topology of the amplicon. The complex-SV cluster around **PDGFRA** connects several segments on chr4, which are amplified to ten absolute copies. The SV calls support the reconstruction of an extrachromosomal amplicon incorporating **PDGFRA** exons and these segments. **d,** Cancer genes involved in high-level amplicons (>10 copies) within the cohort. Nine of 23 tumors (grouped on right) contained high-level amplicons encompassing loci from two or more chromosomes. These linked loci are marked by an asterisk. The color of each cell represents the number of extra copies due to the amplicon. **e,** Example of a tumor with an extrachromosomal amplicon including two oncogenes from different chromosomes. This tumor shows a cluster of SVs connecting the **EGFR** and **CCNE1** loci. The regions of both oncogenes are amplified to different CNs but in both cases reach several dozen absolute copies. (top) The complexity of the SVs allows for the reconstruction of several possible extrachromosomal amplicons. The CN differences in the bulk profile (middle) could be explained by either a mix of different circles or by more complex circles incorporating some segments repeatedly. The SV calls also reveal that a small fraction of the **EGFR** amplicons in this patient already show the **EGFRvIII** variant. The bottom two tracks show the genes of interest at the location and q value H3K27ac track calculated from eight pHGG tumors.
shown to reside on extrachromosomal amplicons in various cancer types. Although projections on the linear reference genome resulted in typical complex patterns, it was possible to construct circular amplicons containing MYCN and, in the MYCN–ID2 cases, ID2. Indeed, these circular amplicons represent an optimal solution to explain the joint CN and SV profiles in this region. We therefore sought to validate this using metaphase fluorescence in situ hybridization (FISH) on a pHGG cell line derived from a tumor with a MYCN–ID2 rearrangement. We found abundant extrachromosomal amplicons containing both the MYCN and ID2 loci (Fig. 2f). These appeared to reflect multiple subclonal amplifications, including some containing additional oncogenes such as MDM4 (Fig. 2g). The CN of MYCN was consistently higher than that of ID2, raising the possibility that ID2 was incorporated into a subset of preexisting MYCN amplicons during tumor development. Alternatively, MYCN might be further amplified within cells that already harbor MYCN–ID2 amplicons.

These data suggest that MYCN–ID2 rearrangements are an example of enhancer hijacking, bringing a stronger enhancer in ID2 next to MYCN on amplicons without the endogenous elements of the MYCN TAD, whereas amplifications of MYCN without ID2 amplify enhancers within the MYCN TAD.

Recurrent SV around RTKs suggest extrachromosomal DNA. The remaining SRBs all involved RTK genes that are known to be amplified and oncogenic in pHGG: PDGFRA, EGFR and MET. These loci, along with MYC and MYCN, were also the only regions with high-level amplifications that recurred in at least three patients (Fig. 3a). The RTK SRBs also comprised both simple SVs that presumably amplify local enhancers (Fig. 3b and Extended Data Fig. 4c) and SVs that appeared to reflect complex extrachromosomal amplicons that integrate distant sites and reach as many as 200 copies (Fig. 3c).

Overall, 35 of 179 tumors showed at least one >50 kbp amplicon with an absolute CN greater than ten. Among 34 of these 35 tumors, the high-level amplicons contained at least one well-known oncogene, and, apart from the coding sequence of the oncogene, they recurrently incorporated the same genomic loci around the oncogene (Extended Data Fig. 4a,b for PDGFRA and EGFR). In several pHGGs, we detected SVs that allowed for the reconstruction of circular extrachromosomal amplicons containing multiple oncogenes from different chromosomes (Fig. 3d,e). We again validated this by performing FISH on a tissue slide of a tumor with a high-level amplification and SVs connecting segments on chr8 (including GATA4) and chr10 (including FGR2). This showed massively increased numbers of foci for both the GATA4 and FGR2 probes. In many cases, the signal of both probes overlapped, indicating colocalization (Extended Data Fig. 4e). The number of copies per cell was highly variable, and the amplicons were distributed in heterogeneous clusters throughout the nucleus. All these features have been associated with extrachromosomal amplicons. These data suggest that only a subset of pHGGs develop high-level amplicons, which recurrently contain the same (presumably regulatory) sequences in addition to the target oncogene and can contain segments originating from different chromosomes.

To further understand the structure of these amplicons, we first focused on high-level amplicons containing PDGFRA. These amplicons span more than 2.5 Mbps and are superimposed on low-level amplicons of the surrounding region, often starting from the centromere. The amplicons included KIT in 80% of cases and KDR in 60%. All but one (14 of 15, 93%) of the amplicons in the PDGFRA TAD amplified PDGFRA itself, often to the highest CNs reached in the region (Extended Data Fig. 4a). The sole exception was the amplification of a short sequence centromeric to PDGFRA containing H3K27ac peaks (Fig. 3b) that have been shown to interact with the PDGFRA promoter in pHGGs and adult GBMs, suggesting use of enhancer amplification to activate PDGFRA. Indeed, this region was included in the PDGFRA amplicon in nearly all tumors (14 of 15, 93%) (Extended Data Fig. 4a). These data suggest that SVs in pHGG recurrently incorporate an upstream enhancer-rich region into high-level PDGFRA amplicons.

The high-level EGFR and MET amplicons also extended beyond the RTK coding sequence to recurrently involve associated enhancers (Extended Data Fig. 4b–d). The EGFR amplicons showed a skew towards the enhancers in SEC61G, which drive EGFR expression in extrachromosomal amplicons in adult GBM. Both EGFR and MET amplicons showed subclonal SVs within the coding sequence, potentially allowing for expression of alternate transcripts. In two tumors, these resulted in the EGFRvIII variant. EGFR amplicons also showed complex subclonal structures with incorporation of distant oncogenes (Fig. 3e). MET amplicons skewed towards a region including enhancers in CAPZA2 (Extended Data Fig. 4d).

We conclude that the RTK gene amplicons are shaped by the epigenetic machinery necessary to drive their expression.

SV signatures relate to genetic and epigenetic tumor states. The discovery of these two distinct classes of recurrent SVs, simple enhancer amplifications and complex extrachromosomal DNA (ecDNA)-based amplicons, raises the question of whether these alterations are part of distinct variant patterns in different pHGGs. Unsupervised identification of SV signatures can reveal tumor subgroups with distinct SV-generating mechanisms. Recent work has identified SV signatures that are present across cancer types, albeit in tissue-specific conformations and with tissue-specific variant associations. We performed a manual review of individual SVs to assess their probable mechanisms of formation as described above; however, automated classifiers have recently been developed for genome-wide analyses. Using methods developed by the Pan-Cancer Analysis of Whole Genomes group, we detected 10,385 complex and/or clustered SVs. Among these, automated methods further classified 21% by their probable formation mechanism. We therefore combined both approaches to identify SV signatures, using the more precise formation mechanisms when available and denoting the remaining complex events as complex-NOS (not otherwise specified).

We thus obtained nine SV signatures (Fig. 4a). Six of these signatures were complex (breakage fusion bridge cycle (BFB), chromothripsis, chromoplexy, complex-NOS, DMs and tyfonas) and three were simple (deletion, duplication, translocation/inversion). The DM signature comprised only DMs, of which 6% were complex DMs according to the categorization of Hadi et al.1. Templated insertion chains contributed 6% of the translocation/inversions, 1% of the deletions and 1% of the complex-NOS SV signatures. Inversions contributed 35% of the translocation/inversion signature and 13% of the duplication signature. Four signatures were composed entirely of a single type of feature: chromothripsis, BFB, chromoplexy and tyfonas (Fig. 4b).

We next looked for possible causes and consequences of the pHGG SV signatures by testing for associations between the activity of each signature and the presence of recurrent and known oncogenic variants (Supplementary Table 3). As expected, the DM signature was associated with ecDNA amplifications of MYCN and ID2. The complex-NOS SV signature was closely associated with focal TP53 disruption and loss of 17p (encompassing TP53) and ant correlating with oncogenic mutations in PPM1D, ACRV1 and HIST1H3B (Extended Data Fig. 5a). Notably, we observed complex-SV signatures in pHGGs with high SV counts (q = 0.68, P < 2.2 × 10^-15, Spearman’s). Unlike several high-SV-count adult cancers with disrupted DNA damage response (DDR) and HR/BRCA, where tandem duplication signatures were dominant, pHGGs with the simple tandem duplication signature tended to have few SVs. None of the genes previously implicated in tandem...
duplication signatures or loss of HR in adults\(^2\) reached a significant level of association with any signature in pHGG.

We next asked whether pHGGs separated into subsets with different DNA damage and damage response characteristics based on patterns of both SVs and SNVs. We detected an anticorrelation between the complex-SV signatures and the three simple-SV signatures (Fig. 4c). Evaluation of SNV mutation patterns revealed 14 SNV signatures, including signatures similar to known aging, APOBEC (COSMIC signature SBS13 (ref. 47)), HR deficiency (SBS3 (ref. 25)) and hypermutation SNV signatures\(^2\) (Extended Data Fig. 5b–d).

**Fig. 4 | SV signatures in pHGG.** a, Contribution of different SV features to each of the identified SV signatures. The horizontal axis indicates SV features: deletion (del), duplication (dup), inversion (inv), BFB, complex DM (cpxdm), DM (dm), template insertion chains (tic) and translocation (tra). The vertical axis indicates the fraction of SVs with each of these features within the nine identified SV signatures. b, Heatmap indicating the contribution of each SV type to each signature. Deeper red color indicates a larger fraction of all SVs with each type contributing to each signature. The size of the squares indicates the fraction of all SVs contributing to a signature that belong to this SV type. c, Heatmap indicating the correlation between the SV signatures as determined by Pearson analysis. Shading in red indicates positive correlation coefficient and blue indicates anticorrelation. d, Normalized SV and SNV signature activities within the complex-SV and SNV-dominant clusters. Each column represents an individual tumor and rows indicate signature activities in the individual samples. Contribution of complex-SV signatures within the complex-SV group are represented by the colors shown (\(n=179\) tumors). e, Genetic variants (SV and SNVs) significantly enriched in the complex-SV and SNV-dominant clusters (\(n=179\) tumors), with correlations reaching \(q<0.1\). \(q\) values were calculated using Fisher’s exact tests.
We also performed signature analyses using alternative methods; these generated similar results (Extended Data Figs. 6, 7).

The entire pHGG cohort separated into two groups reflecting different amplitudes of the 9 SV and 14 SNV signatures (Fig. 4d). One cluster (complex-SV) was dominated by complex-SV signatures (q values ranging from 0.02 for DM to 1.4 × 10−30 for complex-NOS; Extended Data Fig. 8a). SBS3 and SBS13 were also enriched in this cluster, following their close correlation with the complex-NOS SV signature (Extended Data Fig. 8b). The complex-SV cluster was enriched for TP53 inactivation (q < 0.1, Fig. 4c), SVs surrounding and amplification of PDGFRα, EGFR and MET (q < 0.1). By contrast, the other cluster (SNV-dominant) was dominated by simple-SV signatures (q < 8 × 10−3), lacked TP53 disruption (q < 0.1) and was enriched for PPM1D mutations. This cluster seemed to be driven instead by a combination of SNVs including ACVR1, PPM1D, H3.1K27M and PIK3CA mutations (q < 0.1). Both clusters included hemispheric and midline gliomas. H3.3K27M showed no enrichment in either cluster (q = 0.46). These data suggest that pHGG genomes are shaped by at least two distinct variant-generating processes, which are associated with distinct driver combinations.

Signatures indicate two groups, complex-SV and SNV-dominant.

We next evaluated whether SV signatures could inform pHGG subtypes. Currently, pHGGS are classified according to their location and histone mutations; different histone mutations are known to be associated with distinct recurrent SNVs and SCNAS. We confirmed these known relationships and detected two additional associations with SVs (Fig. 5a,b and Extended Data Fig. 8c). The SV in CCDC26 resulting in MYC enhancer amplification was enriched in H3.3K27M gliomas (q = 0.008), and H3.1K27M pHGGs were enriched for a focal deletion of CDKN2C with breakpoints in the adjacent gene FAF1 (q = 0.04).

We also found that inclusion of SV signatures could identify two pHGG subtypes, both derived from H3.3K27M pHGGs. Most H3.3K27M pHGGs exhibited high complex-SV signature activity, but 42% did not. The combinations of genetic alterations in known cancer-related genes differed significantly between the H3.3K27M complex-SV and H3.3K27M SNV-dominant groups (q = 6.0 × 10−4; Fig. 5c). Indeed, the H3.3K27M SNV-dominant pHGGs were as different from the H3.3K27M complex-SV pHGGs as were the H3.1K27M pHGGs (Extended Data Fig. 9a). The H3.3K27M SNV-dominant pHGGs showed a variant pattern resembling H3.1K27M DMGs (Fig. 5d); mutations in PPM1D, ACVR1 and PIK3CA, gains of 1q encompassing MDM4, and amplifications in CCDC26 encompassing MYC enhancers were enriched in these tumors relative to H3.3K27M complex-SV pHGGs (all q < 0.09 except for CCDC26 amplifications, where q = 0.11). By contrast, loss of TP53 (by SNV or SCNA) and amplifications of PDGFRα and MYC were depleted (all q < 0.05) in the H3.3K27M SNV-dominant tumor group. These data suggest that the propensity of pHGGs to develop complex SVs influences the combination of driver alterations they accrue, even within groups defined by their histone mutations.

Across the DMGs with more than 20% complex-SV signature activity (denoted H3.3K27M complex-SV, including H3.3K27M (n = 43) and H3.1K27M (n = 3) DMGs), the TP53 pathway was inactivated almost universally through direct disruption of TP53 (44 of 46 cases, 96%; Fig. 5a). By contrast, the majority of DMGs with less complex signature activity (denoted H3.3K27M SNV-dominant, H3.3K27M (n = 30) and H3.1K27M (n = 21)) lacked direct TP53 disruption (37 of 51, 73% TP53WT; q = 2.3 × 10−5) but appeared to suppress the TP53 pathway through other mechanisms. Mutations in PPM1D were more prevalent in this group, although they were still a minority (7 of 30 H3.3K27M, 2 of 21 H3.1K27M, 20% in total; versus 1 of 46 H3.3K27M-complex tumors; q = 0.008). It is possible that gains of 1q, encompassing MDM4, also served to suppress the TP53 pathway in these tumors. Although 1q spans approximately 2,580 genes, we observed two sources of evidence that their prevalence in SNV-dominant DMGs was related to MDM4 and TP53 pathway suppression. First, MDM4 was significantly overexpressed in 1q-amplified pHGGs of all types in our cohort (q = 0.004; Extended Data Fig. 9b). Second, 1q gains were the only arm-level SCNAS that were anticolored with disruption of TP53 (q = 2.8 × 10−4 in H3.3K27M-DMGs; q = 0.0003 across all pHGGs; Extended Data Fig. 9c,d) apart from gains of chromosome 2 (q < 0.0025). By contrast, seven of the other 13 significantly recurrent arm-level SCNAS were positively correlated with TP53 disruption in H3.3K27M DMGs (all q < 0.018), presumably owing to the role of TP53 in generating aneuploidies. Indeed, across TP53−/− H3.3K27M DMGs, gains of 1q were among the most common genetic events, observed in 85% of tumors (16 of 20 H3.3K27M, 17 of 19 H3.1K27M), compared with 31% of TP53-disrupted H3.3K27M DMGs (P = 1.4 × 10−6). No pHGGs in our cohort exhibited focal high-level amplifications of MDM2. These data suggest that direct disruption of TP53 contributes to a different pattern of SVs compared with other mechanisms of TP53 pathway inactivation, primarily including alterations of PPM1D and MDM4.

Focusing on the tumors that harbored significantly recurrent SVs, we observed two groups. One group contained tumors with high-level amplions of PDGFRα, EGFR, MET, MYC and MYCN (oncogene-amp). By contrast, the second group showed amplification of presumed enhancer elements within the TADS of these oncogenes without amplification of their coding sequences (enhancer-amp). The oncogene-amp pHGGs exhibited significantly higher activity of complex-SV signatures (P = 4.0 × 10−7; Fig. 6a,b). The two groups also harbored inactivating alterations in different DDR genes (Fig. 6a). Oncogene-amp pHGGs were enriched for TP53 SNVs (69% of oncogene-amp versus 18% of enhancer-amp pHGGs, q = 0.01) and RB1 deletions (23% of oncogene-amp versus 0% of enhancer-amp pHGGs, q = 0.16). By contrast, enhancer-amp pHGGs were enriched with PPM1D SNVs (29% of enhancer-amp versus 0% of oncogene-amp pHGGs, q = 0.03) and gains of 1q.
encompassing MDM4 (71% of enhancer-amp versus 34% of oncogene-amp pHGGs, \( q = 0.16 \)). In summary, alterations in TP53 and RB1 are associated with complex-SV signatures and high-level amplifications of oncogenes, whereas PPM1D SNVs and 1q gains more frequently occur with simple-SV signatures and amplifications of enhancer elements near oncogenes. These data raise the possibility that alterations in the DDR shape not only the processes that generate SVs but also the types of driver alterations they exhibit in MYC, MYCN and RTK genes.

Temporal evolution of genetic variants. The correlation between the presence of a variant and the activity of a signature by itself cannot tell us anything about the direction of the link between the two. This is most obvious for the inactivation of tumor suppressors through CN loss or the amplification of oncogenes and their associations with complex-SV signatures. These events could be direct consequences of the activity of this signature. On the other hand, these genetic variants could drive survival after catastrophic SVs, increase genomic instability and thereby drive...
the activity of the signature after their initial random occurrence in tumor evolution.

Specifically, we considered two hypotheses regarding tumor evolution. First, disruption of DDR could be an early event that activates the complex-SV-generating process and culminates in the development of a specific class of genetic events, including the high-level amplicons described earlier. Alternatively, both disruption of the DDR and high-level oncogene amplification could...

**Fig. 6 | Context of significantly recurrent SVs.**

a, Comut plot for enhancer-amp and oncogene-amp significantly recurrent SVs. Enhancer-amp significantly recurrent SVs generate focal amplifications in the TAD of an oncogene without amplifying the protein-coding sequence; oncogene-amp significantly recurrent SVs generate high-level (CN > 3.4) amplifications or fusions of the coding sequence. The top two rows show associated metadata. The next seven rows indicate the genes affected by the significantly recurrent SVs. The bottom seven rows show genes in DDR pathways. Significant associations with the two groups are illustrated with pie charts below the plot, based on Fisher’s exact test.

b, Enhancer-amp pHGGs show significantly lower combined complex-SV signature activity than oncogene-amp pHGGs ($P = 4 \times 10^{-7}$; two-sided Wilcoxon test; $n = 52$ tumors; center line of the boxplot indicates the median, bounds of the box indicate the 25th and 75th percentiles, and whiskers extend from the box to the largest or smallest value no further than $1.5 \times$ IQR).

c, Timing analysis of somatic variant acquisition in enhancer-amp (c) and oncogene-amp (d) pHGGs based on a Bradley–Terry model. The horizontal axis shows the log odds of the variant being an early event. The distributions indicate the results of 100 random subsamples of the data ($n = 17$ pHGGs (c), $n = 35$ pHGGs (d)). The center line of the boxplot indicates the median, bounds of the box indicate the 25th and 75th percentiles, and whiskers extend from the box to the largest or smallest value no further than $1.5 \times$ IQR. Only variants in the SRSV-affected genes and pathways (growth factors and MYC signaling) and in DDR genes altered in more than two samples are shown.
happen later in tumor development as a consequence of complex SVs involving these genes.

Notably, we observed no effects of therapy on SV patterns, suggesting that the SVs occurred during gliomagenesis. Although radiation treatment has been shown to induce DNA breaks\(^5\), we found no differences in the number of SVs per sample (median 35 versus 42; \(q = 0.6\)) or in the activity of the complex-SV signatures (median 24\% versus 28\%; \(q = 0.7\)) between pretreatment biopsy and autopsy samples (Extended Data Fig. 9e,f).

We performed a timing analysis reflecting the relative ordering of mutations and SCNAs during gliomagenesis\(^5\). Focusing on the subset of pHGGs with simple enhancer amplifications (enhancer-amp pHGGs), we found that the focal amplification of the MYC enhancer in \(CCDC26\) was one of the earliest variants in these samples (Fig. 6c), occurring earlier than alterations in \(PPM1D\) and \(1q/MDM4\) gain. By contrast, amplification of the MYC isoform and RTK genes in the oncogene-amp samples happened after the loss of the tumor suppressors \(TP53\), \(RB1\) and \(CDKN2A/B\) (Fig. 6d).

These data suggest that simple tandem duplications can arise in tumors without major disruptions of DDR, potentially contributing to tumor initiation, whereas the creation of complex high-level oncogene amplicons requires prior direct genetic disruption of \(TP53\), \(RB1\) or \(CDKN2A/B\).
H3.1K27M and H3.3K27M gliomas, respectively (Extended Data Fig. 10). This large WGS cohort also allowed us to time focal SCNAs based on the ratio of SNVs acquired before and after each change in each CN14. We found that losses of TP53, CDKN2A/B and RB1 preceded RTK gene amplifications across the H3.3K27M, H3.3CNR and H3.3WT subgroups of pHGG. In H3.3K27M DMGs, simple amplifications of the MYC enhancer in CCDC26 were early events, whereas complex amplifications of MYC itself occurred later in tumor development.

**Complex-SV DMGs are associated with shorter survival.** We suspected that the differences in SV-generating processes across DMGs could be associated with clinical phenotypes including survival. First, we confirmed the known association47,48 between H3.1K27M and longer overall survival (OS) compared with H3.3K27M (9.3 versus 16.1 months; P = 0.0004; Extended Data Fig. 9e,top) and the lack of association between TP53 mutation and OS within H3.3K27M pHGGs (P = 0.72; Extended Data Fig. 9e, bottom). To address SV signatures specifically, we also investigated the correlations between the numeric values of the combined complex-SV signature and OS. Across all DMGs, this complex-SV signature was significantly anti-correlated with OS (Fig. 7a; P = 0.001).

The combined complex-SV signature was also significantly associated with shorter survival in a multivariate Cox regression analysis of DMGs that controlled for the known predictors of survival47 (histone SNV and age) and for TP53 status (Fig. 7b; P = 0.038). This analysis confirmed a significantly increased hazard ratio for H3.3K27M compared with both H3.1K27M and H3WT DMGs and a lack of significant associations between TP53 disruptions and OS in multivariate analyses as previously described49. However, associations with age did not reach significance, probably owing to our low representation of the under-three and over-ten age groups. Although all patients with DMGs in our study died from their disease, the combined effects of these factors caused survival differences of several months. For example, children with DMGs with at least 20% complex-SV activity survived a median of 9.6 months, about 3 months less than the 12.3-month survival of children with less than 20% complex-SV activity (Fig. 7c).

**Discussion**

These analyses found recurrent SVs, including a tandem duplication in 12% of all DMGs encompassing a MYC enhancer; revealed distinct SV signatures; and indicated two classes of DMG, whose driver alterations were either largely complex SVs or dominated by SNVs.

The MYC enhancer amplifications highlight an underrecognized role for MYC in pHGGs. MYC is the most frequently amplified gene across all cancers, with focal amplifications observed in 15% of tumors15. By contrast, MYC amplifications only occur in 5% of pHGGs. The observation of MYC enhancer amplification in pHGGs, without amplification of MYC itself, enables us to start to address this discrepancy. Although tissue-specific amplifications of MYC enhancers occur in other cancers15, CCDC26 duplication is a pHGG-specific occurrence apparently driven by differences in enhancers across cell types. Altogether, when including high-level MYCN amplifications, 14% of pHGGs harbored SVs predicted to activate MYC pathways. Given this high rate, the role of MYC in pHGG formation requires further study.

Although both SNV-dominant and complex-SV pHGGs activate MYC signaling pathways, they do so in strikingly different ways. Whereas SNV-dominant pHGGs amplify only the MYC enhancer in CCDC26, pHGGs with complex-SV signatures contain high-level amplifications of both the MYC coding sequence and segments of CCDC26, PVT1 or other distant regions. Amplifications of MYC-PVT1 have been reported in DMGs and other cancers57,58. These additional segments could contain independent oncogenic activity, as has been proposed for PVT1, or they could represent regulatory elements that have been hijacked to drive MYC expression. The complex MYC amplicons are often extrachromosomal, as indicated by their circular topology and high CN. In this respect, MYC serves as an example for other oncogenes, including MYCN, PDGFRα, EGFR and MET. Extrachromosomal amplicons (also known as DMs) containing recurrent oncogene–enhancer combinations occur in several cancers12,20,27. However, their regulatory elements differ from those of pHGG and appear to reflect the tissue specificity of regulatory loci19.

DMs have been shown to originate as byproducts of chromothripsis46. Our data suggest that in pHGG they often contain multiple oncogenes from different chromosomes. These DMs would therefore either require simultaneous chromothripsis of two chromosomes or need to develop sequentially by a less-clear mechanism. Our data also suggest multiple variants of DMs within individual pHGGs. These could be correlated with different descendants of the initial DM, as suggested by recent mechanistic and long-read sequencing studies29,47. In cases where both oncogenes are integrated into a DM that is subsequently amplified, the number of copies of each oncogene should be identical. However, pHGGs often exhibit different amplification levels of these oncogenes, suggesting sequential incorporation into the amplicon. The exact mechanism for this remains elusive; the possibilities range from sequential chromothripsis events47 to reversible DM integration in proximity to oncogenes15, or deletions within the DMs. It is tempting to speculate that the evolution and optimization of DMs46 could contribute to the rapid, lethal growth of pHGGs and their poor response to available therapies. RTK inhibition is still a promising goal in pHGG, but our study highlights that understanding how these DMs evolve might provide insight into resistance mechanisms.

We also observed an association between H3.3K27M, complex-SV signatures and TP53 loss. Although TP53 disruption is known to be associated with higher SV burden46, the reason for its association with H3.3K27M instead of H3.1K27M is unclear. H3.1K27M and H3K27M pHGGs also included both complex-SV and SNV-dominant tumors, although H3.1K27M DMGs were enriched in the SNV-dominant subgroup. The split into complex-SV and SNV-dominant types observed in H3.3K27M pHGGs could also occur in H3.1K27M and H3WT tumors—indeed, these distinctions may exist in other tumor types—but our cohort was insufficient to address this possibility.

We found TP53 disruption to be an early event in tumors with complex SVs. TP53 disruption also precedes and might facilitate survival after chromothripsis in medulloblastoma48. Notably, although almost all DMGs with complex-SV signatures were TP53 disrupted, not all TP53-disrupted DMGs showed complex-SV signatures. In addition, hemispheric pHGGs with complex-SV signatures were frequently TP53WT but often harbored early loss of CDKN2A/B. This indicates that although TP53 loss and H3.3K27M are correlated with complex-SV signatures, they are neither necessary nor sufficient, either alone or in combination, for the generation of complex-SV signatures in pHGG.

Finally, we found variants in known cancer genes in 98.3% (176 of 179) of pHGGs, substantially expanding the share of patients with identified potential drivers compared with those reported by prior exome-sequencing-based studies. Many of our observed alterations were in noncoding regions of the genome, targeting regulatory elements such as enhancers. WGS also allowed us to determine which patients had complex-SV or SNV-dominant signatures, which were associated with survival, controlling for histone and TP53 status. The association between the complex-SV signatures and survival might be causative and indicate potential therapeutic targets or it could represent a quantifiable biomarker for underlying factors such as genome instability. In any case, these findings indicate that both research and clinical sequencing of these tumors should encompass the whole genome.
Sample acquisition. This study included published ^23 data available under EGAS0000001000575, EGAS000000101139, EGAS000000100572 and EGAS000000100192. Novel data were generated from samples obtained from the DPEG-BATs clinical trial (NCT01182350) and the Dana-Farber Tissue Bank or collaborating institutions, under protocols approved by the institutional review board of the Dana-Farber/Harvard Cancer Center with informed consent (DFCI protocols 10417, 10201 and DFCI 1929), without participant compensation. DNA and RNA were extracted from single DMG cores, pHGGi biopsies and autopsy samples using Quantel-mxPrep DNA/RNA extraction kits. Previously published pHG Gi WGS data^22 and paired RNA sequencing (RNA-seq) data were acquired from public repositories.

WGS. Library preparation for paired-end WGS was performed^16. Genomic DNA from public repositories. Sample acquisition.

SNV and SCNA analyses. SNVs were detected using Mutect2 and filtered for common sequencing artifacts, gnomad single-nucleotide polymorphisms (SNPs) and SNVs present in a panel of WGS normal samples. Significance of recurrent SNVs in nonhypermutable samples (SNV counts <100,000 per sample) was determined with MutSigCV^45. SCNAS were called using the GATK4 CNV pipeline with normalization against a panel of blood normals from 184 samples (174 from this cohort, ten from TCGA). Purity and ploidy were determined using ABSOLUTE^46. All SCNAS call were purity- and ploidy-adjusted. Significantly recurrent SCNAS were identified using GISTIC2.0 (ref. ^37) with the following parameters: arm_thresh=0.1; del_thresh=0.7; arm_pval=0; broad_length_cutoff=0.5; cap=3.6; conf_level=0.99; max_sample_segs=3500; and qv_thresh=0.1.

Structural variant detection and significance analysis. SvABA^33 was used to call SVs in paired tumor normal mode with default parameters. In addition to filtering germline SVs against the paired normal, telomeres and centromeres were blacklisted. Significance of recurrence analysis was performed separately for breakpoints (1D) and juxtapositions (2D) adapted from pan-cancer analyses^34.

The analysis of recurrent breakpoints (1D)^35 binned the genome into 50 kbp bins with 300 bp overlap. The bins were annotated by the overlapping TAD with TAD names derived from COSMIC Cancer Gene Census^36 gene presence. Germinal zone TAD boundaries from GSE77565 (ref. ^27) were used as the closest available normal neural progenitor. Eligible territory was defined by masking low-complexity genomic loci based on https://github.com/lh3/sgdp-fermi/releases/download/v1.0.35/sl.d350.bed.gz. Only one SV per sample per bin was counted. Pushbook^38 was used to calculate a background model for the likelihood of a breakpoint in each bin, based on six covariates: replication timing (from http://mskliblab/flashHook/hg19/RT_NHEK_Keratinocytes_Int92817591_hg19.rds), GC content (from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/gcBase/hg19.gcBase.txt.gz), presence of SINE elements (http://www.repeatmasker.org/fishHook/hg19/RT_NHEK_Keratinocytes_Int92817591_hg19.rds), bins with 500 bp overlap. The bins were annotated by the overlapping TAD using a background model determined from 2658 cancers across several types^21. The analysis of recurrent breakpoints (1D) was performed using the gGnome::plots function. Events were called on the gGraph output from JabBa using the gGnome::events function. Events were assessed for enrichment in a background model determined with MutSigCV. SVs were stratified according to the span between the two breakpoints (0–30 kbp, 0.03–1 Mbp, >1 Mbp, interchromosomal); read orientation (deletion, duplication, inversion and interchromosomal); and whether they were clustered, as determined by clusterSV^47. This was analyzed with Bayesian NMF using SignatureAnalyzer^32,24. Jblab was used to generate genome graphics^35. SV events were called on the gGraph output from Jblab using the gGnome::events function. Events were mapped to individual SVs with gGnomeDesignations when available. SVs without annotations from Jblab/gGnome were classified as ‘complex-NO’ if they had cluster size >2 according to the clusterSV method or involved an inversion, translocation, deletion or duplication based on the orientation of their supporting reads. The count matrix was analyzed with Bayesian NMF using SignatureAnalyzer^48.

RNA-seq analysis. RNA-seq data were available for 112 of 179 tumors (57 sequenced de novo and 55 previously published^47). For de novo samples, cDNA libraries were prepared using the Tru-Seq Strand Specific Large-Insert Kit and sequenced to a depth of 50 million paired ends using Illumina HiSeq. All reads were aligned to the hg19 reference genome using STAR and quantified with RNA-SeqC following the GTEX analysis pipeline^47. Counts were normalized using the VST transform as implemented in DESeq2 (ref. ^38) and batch-corrected with COMBAT^49 as implemented in sva^50.

ChIP. Active Motif was used to perform ChIP–seq. Cells were fixed with 1% formaldehyde (15 min) and quenched with 0.125 M glycine. Chromatin was isolated by adding lysing buffer then disrupted with a Dounce homogenizer. Lystases were sonicated and DNA was sheared to an average length of 300–500 bp with Agilent’s EpiSpin probe (cat. no. 53081). DNA was then prepared by treatment with RNase, protein K and heat for de-crosslinking, followed by clean-up with SPRI beads (Beckman Coulter) and quantitation by Claristor (BMG Labtech).

An aliquot of chromatin (30 μg) was precleared with protein A agarose beads (Invitrogen). Genomic DNA regions of interest were isolated using 4 μg of antibody against H3K27ac. Complexes were washed, eluted from the beads with sodium dodecyl sulfate buffer and treated with RNase and protein K. Crosslinks were reversed by incubation overnight at 65 °C. ChIP DNA was purified by phenol–chloroform extraction and ethanol precipitation.

ChIP–seq. Illumina sequencing libraries were prepared from the ChIP and input DNAs by the standard consecutive enzymatic steps. Sequencing was performed on an automated system (Apollo 342, Wafergen Biosystems/Takara). After PCR amplification, the DNA libraries were sequenced on Illumina's NextSeq 500 (75-nucleotide (nt) reads, single end). Reads were aligned to hg19 using the BWA algorithm (default settings). Duplicate reads were removed, and only those with mapping quality ≥25 were used for further analysis. Alignments were extended in silico at their 3′-ends to a length of 200 bp and assigned to 32-nm bins along the genome. Published H3K27ac ChIP–seq sequencing data from primary DMGs were downloaded from GSE128745 (ref. ^47). Peaks were called using MACS2 (ref. ^51) peakcall with -B –SPMR to square the fragment pileup per million reads track. The bigwig files were used to calculate fold enrichment and a value track with MACS2 bdgcmp, transformed into bigwig files with rtracklayer and visualization with CN and SV calls in gtrack. Additional bigwig files for adult GBM H3K27ac, pHG Gi ATAC-seq and noncancerous or nonbrain tissues were downloaded from GSE54792 (ref. ^52), GSE126319 (ref. ^53) and the Encode project^54, respectively.

Hi-C. Library generation and sequencing. In situ Hi-C libraries were generated from 5 million cultured H3.3^55 or H3.3^56 glioblastoma cell lines (HSI-019 and HSI-031) and H3.3^61 primary tumors (HSI-031 and 039) following published protocols^68 with minor modifications. Briefly, the steps were as follows: (1) crosslinking cells with formaldehyde; (2) digesting the DNA using a 4-cutter restriction enzyme (for example, DpnII) within intact permeabilized nuclei; (3) filling in, blunting the resulting 5′-overhangs and ligation the blunt ends; (4) shearing the DNA; (5) pulling down the biotinylated ligation junctions with streptavidin beads; (6) library amplification and (7) analyzing these fragments using paired-end sequencing.

Quality control for efficient sonication was performed through a combination of agarose DNA gel electrophoresis with appropriate size selection using an Agilent Bioanalyzer 2100. library amplification and (7) analyzing these fragments using paired-end sequencing.

Hi-C. Library generation and sequencing. In situ Hi-C libraries were generated from 5 million cultured H3.3^55 or H3.3^56 glioblastoma cell lines (HSI-019 and HSI-031) and H3.3^61 primary tumors (HSI-031 and 039) following published protocols^68 with minor modifications. Briefly, the steps were as follows: (1) crosslinking cells with formaldehyde; (2) digesting the DNA using a 4-cutter restriction enzyme (for example, DpnII) within intact permeabilized nuclei; (3) filling in, blunting the resulting 5′-overhangs and ligation the blunt ends; (4) shearing the DNA; (5) pulling down the biotinylated ligation junctions with streptavidin beads; (6) library amplification and (7) analyzing these fragments using paired-end sequencing.

Quality control for efficient sonication was performed through a combination of agarose DNA gel electrophoresis with appropriate size selection using an Agilent Bioanalyzer 2100.

Data processing. Additional Hi-C files for neural progenitor cells were downloaded from a previous publication71. Additional Hi-C files for neural progenitor cells were downloaded from a previous publication71. Additional Hi-C files for neural progenitor cells were downloaded from a previous publication71. Additional Hi-C files for neural progenitor cells were downloaded from a previous publication71. Additional Hi-C files for neural progenitor cells were downloaded from a previous publication71. Additional Hi-C files for neural progenitor cells were downloaded from a previous publication71.
Luciferase reporter. Cell lines. The pHGG cell line DIPG13 (ref. 75) was a gift from the Michelle Monje laboratory. BT245 was obtained from the Keith Ligon laboratory. Cell lines were grown in ULA flasks in a 1:1 ratio of neurobasal A (Gibco) and Dulbecco's modified Eagle medium (DMEM)/F-12 (Gibco) and 1% of each of HEPS buffer solution (1 M), sodium pyruvate solution (100 mM). DMEM, NEAA solution (100x), Glutamine (100x), streptocin (10x). The culture medium was supplemented with epidermal and fibroblast growth factor (H-EGF and H-GF), StemCell Tech., Inc.) at 20 ng ml$^{-1}$, platelet-derived growth factors (H-PDGF-AA and H-PDGF-BB, StemCell Tech., Inc.) at 10 ng ml$^{-1}$, heparin solution (0.2%; StemCell Tech., Inc.) at 2 μg ml$^{-1}$ and 300 B-27 Minus Vitamin A (Invitrogen). Cells were passaged every 2–4 days and were dissociated into single cells at the time of passage using Accutase (StemCell Tech., Inc.).

Cell line authenticity and mycoplasma surveillance. Cell line authenticity was confirmed using short tandem repeat (STR) profiling. All cell lines were monitored for the presence of mycoplasma using the MycoAlert Mycoplasma Detection Kit (Lonza) following the manufacturer's protocol.

Luciferase reporter construction. A lentiviral firefly luciferase reporter system was constructed from pGL4.26 (Promega) and the pKLO.1 backbone via Gibson Assembly. The pKLO.1 backbone was digested with FastDigest KfiI and EcoRI. The minimal promoter firefly luciferase reporter cassette was PCR-amplified from pGL4.26 using the lucminP primer set (Supplementary Table 6) using NEB Q5 polymerase. These two fragments were assembled into the lentiviral firefly luciferase reporter using the NEBuilder HiFi DNA Assembly Cloning Kit according to the manufacturer's instructions. The DNA sequence in the H3K27ac peak in the consensus CCDC26-SV amplicon was split into two fragments (E1 and E2) and PCR-amplified from DIPG13 genomic DNA with the primers listed in Supplementary Table 6 using NEB Q5 polymerase. The resulting E1 and E2 fragments were cloned into the vector using the KPN1 and NEHI restriction sites. The lentiviral constitutively active pLX313-Renilla construct was obtained from Addgene (plasmid no. 118016) to serve as an intrinsic control.

Viral production. HEK-293T (ATCC CRL-3216) cells were cultured in T75 tissue culture treated flasks in DMEM (Gibco) supplemented with 10% fetal bovine serum (FBS; Gemini Bio). Lipofectamine 3000 (Invitrogen) was used to transfect with the plasmid of interest, in addition to packaging plasmids VSV-G and psPAX2, according to the manufacturer's protocol. Media were replaced transfect with the plasmid of interest, in addition to packaging plasmids VSV-G and psPAX2, according to the manufacturer's protocol. Media were replaced

Luciferase reporter readout. The Dual-Glo Luciferase Assay System (Promega) was used following the manufacturer's protocols for all measurements 4 days post spinfection (2 days post puro selection).

Visualization and reconstruction of complex MYCN and RTK amplicons. JdbA$^{15}$ was used to generate cancer genome graphs using svABA SV, GATK CNV, and absolute purity and plodly as inputs. Tracks were visualized in gnomes/ gTrack, which was also used to calculate distances between loci in the cancer genome. Extrachromosomal amplicons were inferred by using a subset of only circular path segments with CN > 20 and reconstructed with the gGene walk() function.

MYCN amplicons recurrently incorporated into the amplicons were determined based on the distribution of amplicons around the oncogene. TADs adjacent to the amplified oncogene were divided into 10 kbp windows. The average CN per 10 kbp window was calculated for all tumors with an amplicon of CN > 5 anywhere in the TAD of the oncogene (using the germline zone TAD boundaries from GSE72556 (ref. 78)). Among tumors with an amplification of CN > 5 anywhere within the TAD of the oncogene, the fraction of tumors with an amplification in each 10 kbp window was determined. The location of probable enhancer elements, which are necessary to drive expression of the amplified oncogene, were inferred from the direction of the skew of the observed distribution compared with the expected symmetric normal distribution.

GATA4/3 FGFR2 interphase enumeration FISH. Probe specifics. GATA4/3 FGFR2 enumeration was analyzed with FISH. Human bacterial artificial chromosomes (BACs) covering the GATA4 gene region were identified using the University of California Santa Cruz (UCSC) August 2021 Assembly hg38. The GATA4 clones (RP11-2141B23, RP11-2351S5 and RP11-737E8) were labeled by nick translation with Spectrum Orange (Abbott Molecular), and the 3 FGFR2 clones (RP11-8781D10, RP11-69817, CT2-2291K12 and CT2-3237E5) were labeled by nick translation with Spectrum Green. Labeled clones were combined to create an enumeration probe set.

Slide processing for paraffin-embedded tissue samples. Slides were placed in a 90°C oven for 15 min. Slides were then deparaffinized with xylene (two times, 15 min each) at room temperature (RT), dehydrated in 100% ethanol for 5 min at RT, and placed in 10 mM citric acid (pH 6.0) and microwaved for 10 min. Following this, the slides were immersed in 2x standard saline citrate (SSC) for 5 min at 37°C, followed by digestion in 0.2% pepsin working solution (1.2 g pepsin per 600 ml, 0.9% NaCl pH 1.5) at 37°C for 12 min. Immediately after digestion, the slides were dehydrated using an ethanol series (70, 85, 100%) for 2 min each at RT. A working solution of GATA4/3 FGFR2 (2 μg per 20 μl) was made by mixing 2 μl of concentrated 3 FGFR2 probe and 1 μl of concentrated GATA4 probe with 7 μl of LSI/WCP hybridization buffer (Abbott Laboratories). The working solution was applied to the target areas, coverslipped, co-denatured with a Thermobrite at 83°C for 5 min and hybridized overnight in a 37°C humidified oven. Following hybridization, slides were soaked in RT 2xSSC/0.1% NP-40 to remove coverslips, placed in 2xSSC/0.1% NP-40 at 74°C for 2 min and then placed in RT 2xSSC/0.1% NP-40 for 2 min. The slides were stained with 4',6-diamidino-2-phenylindole (DAPI) (Vector Laboratories) and coverslipped.

ID2/MYCN metaphase FISH. Probe specifics. ID2/MYCN enumeration was analyzed with FISH. BACs covering the ID2 gene region were identified using the UCSC August 2021 Assembly hg38. The ID2 clone (CTD-2131H8) was labeled by nick translation with Spectrum Green (Abbott Molecular), and the MYCN probe was commercially available from Abbott Molecular. The ID2 probe and MYCN probe were combined to create a enumeration probe set.

Slide processing for metaphase samples. Slides were air-dried at RT overnight. Following this, the slides were immersed in 2xSSC for 30 min at 37°C. The slides were denatured using an ethanolic series (70, 85, 100%) for 2 min each at RT. A working solution of ID2/MYCN was made by mixing 2 μl of concentrated ID2 probe and 1 μl of concentrated MYCN probe with 7 μl of LSI/WCP hybridization buffer (Abbott Laboratories). The working solution was applied to the target areas, coverslipped, co-denatured with a Thermobrite at 73°C for 5 min and hybridized overnight in a 37°C humidified oven. Following hybridization, slides were soaked in RT 2xSSC/0.1% NP-40 to remove coverslips, placed in 2xSSC/0.1% NP-40 at 74°C for 2 min and then placed into RT 2xSSC/0.1% NP-40 for 2 min. The slides were stained with 10% DAPI (Vector Laboratories) and coverslipped.

SNV signature analysis. De novo SNV signature extraction was performed using Bayesian NMF in SignatureAnalyzer$^{24}$. The resulting SNV signatures were compared with the COSMICv3 SBS signatures using cosine similarity to annotate known etiologies and signature names. DeconstructSig$^{25}$ was used to extract SBS signatures with the highest degree of similarity to the de novo signatures, including a designation of 'unknown'.

Signature integration and definition of signature clusters with similar variant-generating processes. To better understand the information contained in each of the nine SV and 14 SNV signatures, consensus clustering was applied to the tumor × signature proportion matrix, comprising the 23 values representing the proportions of each of the SV and SNV signatures of all SV and SNV signatures in each tumor. The proportions for each signature were median-centered across all tumors before consensus clustering with the ConsensusClusterPlus R package. The following parameters were set: pItem = 0.01, pGene = 0.005, pItem = 0.99, pGene = 0.1, clusterAlg = 'hc', distance = ' spearman'. The resulting most stable and informative clusters were named ‘complex-SV’ and ‘SNV-dominant’, after the signatures with the highest enrichment in the cluster.

Chromothripsis and extrachromosomal amplicons. To define regions of chromothripsis, we used Shatterbox. Regions in the genome with CN > 10 extending for more than 50 kbp were defined as probably extrachromosomal or derived from an extrachromosomal stage.
Comut plots and variant combination matrix. SNVs were annotated using Oncotator. SVs were annotated and linked to a gene based on whether the SV breakpoints were in exons of the gene (named 'coding SV'), intronic ('intron SV') or in the TAD of the gene ('TAD SV'). Absolute purity and ploidy-adjusted CN was determined for each gene using the width-weighted mean CN from all segments overlapping the gene.

To create the variant combination matrix, we used a subset of only Cancer Gene Census genes and SNVs and genes that showed significantly recurrent variants in this cohort. For SNVs, the variant classification was simplified to truncating, 'snvs = (’Nonsense_Mutation’, ’Frame_Shift_Del’, ’Frame_Shift_Ins’, ’Splice_Site’, ’Start_Codon_Snp’, ’Start_Codon_Snp’, ’Translation_Start_Site’) and missense.snvs = (’Missense_Mutation’, ’In_Frame_Del’, ’Stop_Codon_Del’, ’DE_Novo_Start_In_Frame’, ’DE_Novo_Start_Out_Frame’, ’Nonstop_Mutation’, ’In_Frame_Ins’, ’START_Codon_Ins’). SCNs of genes with a ploidy- and/or ploidy-adjusted CN of C < 0.8 were annotated as 'homdel', CN > 5.4 as amp, CN > 10 as ExCn amp and amplifications covering only parts of a gene with a CN > 3.1 as ’part.amp’ based on the CN histogram across all tumors defining recurrent CN states. A genetic variant had to recur in at least three samples (excluding the hypermutant samples for SNVs) to be kept in the matrix. For SV in the TAD of a gene ('TAD SV'), this threshold was increased to at least ten. GISTIC peaks in each sample were used to incorporate the SCNs of lower amplitude.

clonolist oncoprinter was used to visualize variants. Column order represents the samples within each subgroup determined by hierarchical clustering (HC). HC with one minus Spearman rank correlation metric was applied (Extended Data Fig. 6c) and HC with the common similarity metric on the respective subsets of the variant combination matrix (Fig. 6a and Extended Data Fig. 6c) with average linkage in all cases. Genes of interest were manually selected based on the variants with the highest enrichment in the subgroups.

Distances between sample groups in variant space. We calculated Jaccard distances between genetic profiles across 369 variants in genes from the Cancer Gene Census for each pair of tumors. Tumor groups were determined by mutations in H3.1 or H3.3, and the complex-SV signature contributed more or less than 20% to all SV signature activity.

Variant timing analysis. The palmpeset R package was used to determine single-patient timelines of SNVs. SNVs were classified into clonal versus subclonal based on their cancer cell fraction (variant allele frequency adjusted by local CN and/or ploidy) and/or SNVs overlapping with CNAs could further be timed into early or late depending on whether they occurred before or after the SCNA. MutationTimer was used to determine the timing of SCNAS in individual patients, including clone clusters as input. Mobster was used to define clone clusters based on the distribution of the absolute cancer cell fractions. The resulting molecular time for the SCNA segments was assigned to the GISTIC peaks present in the respective samples with a width-weighted mean and categorized based on the timing quartiles. For each subgroup, in addition to the timing GISTIC peaks, single-patient timed SNVs in consensus cancer genes were tallied into winning tables reflecting the frequency of this variant being an early event using published code. The BradleyTerryScalable R package was used to estimate the winning probability and the subgroup fraction for each variant, which is a measure of the probability of this variant being an early event in the tested subgroup. The Bayesian maximum a posteriori probability estimate was used to fit the model as previously described. To control for outlier samples, the analysis was performed on 100 random samples of 70% of each subgroup. The resulting distributions for the strength parameters (on a log scale) were plotted for variants recurring at least three times in the tested subgroup.

Survival analysis. Univariate correlations for differences in survival were analyzed using the Kaplan–Meier method, and significance was determined by log-rank test. Spearman rank correlation tests were used to determine correlations between OS and complex-SV signature activity. This was possible because all children with DIPG died within the observed period, resulting in an absence of censored data. Variables included in the multivariate analysis (Cox model) were histone SNV, age and TP53 status combined with complex-SV signature activity.

Statistics and reproducibility. No statistical method was used to predetermine sample size. Three previously published samples were excluded because their fastq files could not be successfully realigned using our pipelines. Exclusion criteria were pre-established. The experiments were not randomized. The investigators were not blinded to allocation during experiments or outcome assessment. All statistical analyses were performed in R 3.6.3. Unless otherwise indicated, statistical comparisons were performed using Fisher's exact tests or Wilcoxon tests, as appropriate. The data met the assumptions of the statistical tests used. Unless otherwise specified, data were assumed to be not normally distributed, but this was not formally tested. P values less than 0.05 were considered significant. Multiple testing was accounted for by using false discovery rate q values unless otherwise indicated. In all box plots, the boxes represent the interquartile range (25th and 75th percentiles), and the central line indicates the median. Statistical comparisons for the luciferase reporter were performed in Prism 9 using nested one-way ANOVA and Tukey's multiple comparison test.
22. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* 534, 47–54 (2016).
23. Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. *Nature* 578, 190–201 (2020).
24. Kim, J. et al. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.* 48, 600–606 (2016).
25. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* 578, 94–101 (2020).
26. Degasperi, A. et al. A practical framework and online tool for mutational signature analyses: show intragenome variation and driver dependencies. *Nat. Cancer* 1, 249–263 (2020).
27. Angus, L. et al. The genomic landscape of metastatic breast cancer highlights changes in mutation and signature frequencies. *Nat. Genet.* 51, 1450–1458 (2019).
28. Pagut, S. et al. Biological in a series of 130 pediatric diffuse intrinsic Pontine gliomas. *Childs Nerv. Syst.* 31, 1773–1780 (2015).
29. Roujeau, T. et al. Stereotactic biopsy of diffuse pontine lesions in children. *J. Neurosurg.* 107, 1–4 (2007).
30. Cage, T. A. et al. Feasibility, safety, and indications for surgical biopsy of intrinsic brainstem tumors in children. *Childs Nerv. Syst.* 29, 1313–1319 (2013).
31. Gupta, N. et al. Prospective feasibility and safety assessment of surgical biopsy for patients with newly diagnosed diffuse intrinsic pontine glioma. *Neuro Oncol.* 20, 1547–1555 (2018).
32. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal genomic copy-number alteration in human cancers. *Genome Biol.* 12, R41 (2011).
33. Wala, J. A. et al. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* 28, 581–591 (2018).
34. Zhang, X. et al. Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. *Nat. Genet.* 48, 176–182 (2015).
35. Krug, B. et al. Pervasive H3K27 acetylation leads to ERV expression. *Nature* 575, 699–703 (2019).
36. Nikbakht, H. et al. Spatial and temporal homogeneity of driver mutations in diffuse intrinsic pontine glioma. *Nat. Commun.* 7, 11185 (2016).
R.B. has received grant funding from Deerfield Therapeutics for an unrelated project. PB consults for QED Therapeutics, R.B. consults for and owns equity in Scorpio Therapeutics, and reports consulting or advisory roles for Novartis, Merck (I), Gilead Sciences (I) and ViVo Healthcare (I); research funding from Novartis; patents, royalties other intellectual property—Prognostic Marker for Endometrial Carcinoma (US patent application 13/91456, filed June 6, 2013), SF3B1 Suppression as a Therapy for Tumors Harboring SF3B1 Copy Loss (international application No. WO/2017/177191, PCT/US2017/026693, filed July 4, 2017), Compositions and Methods for Screening Pediatric Gliomas and Methods of Treatment Thereof (international application No. WO/2017/132574, PCT/US2017/015448, filed 1/27/2017). K.L.L. reports grants and personal fees from BMS, grants from Amgen, personal fees and other from Travera LLC, personal fees from Rarecyte, grants from Tragara, grants from Lilly, grants from Deciphera, grants from X4, grants from Charles River Labs and grants from Specicare, outside the submitted work, and has a patent US20160032359A1 pending. M.W.K. is now an employee of Day One Biopharmaceuticals. L.C.G. is the Co-Founder of Thromboprotea Inc. All the remaining authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s43018-022-00403-z.
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43018-022-00403-z.

Correspondence and requests for materials should be addressed to Keith L. Ligon, Rameen Beroukhim or Pratiti Bandopadhayay.

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

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2022

Author contributions
F.D., M.W.K., K.L.L., P.B. and R.B. conceived the project. F.D., O.S., N.G., T.Z., J.W., J.T., D.H., A.C., A.H., K.K., M.B.J., K.S.K., C.S., D.W., P.K., H.M., P.H., R.R., R.G., D.D., J.S., M.S., K.W., S.C., J.G., C.K., J.G., N.J., D.J., M.W.K., K.L.L., R.B. and PB acquired samples and generated and analyzed the data. F.D., R.B. and PB wrote the manuscript. All authors edited the manuscript. K.L.L., P.B. and R.B. supervised the overall study.

Competing interests
R.B. and P.B. receive grant funding from the Novartis Institute of Biomedical Research for an unrelated project, R.B. receives grant funding from Merck and
Extended Data Fig. 1 | See next page for caption.
Extended Data Fig. 1 | Sample characteristics and significantly recurrent variants. (a) Purity of pretreatment biopsy and autopsy samples were not significantly different (p = 0.5, two-sided Wilcoxon, n = 174 tumors, center line of the boxplot indicates the median, bounds of the box the 25th and 75th percentiles and whiskers extend from the box to the largest or smallest value no further than 1.5x IQR). (b) Significantly recurrent SNVs in nonhypermutant tumors (n = 179 tumors). (c) Significantly recurrent SCNAs (n = 179 tumors). All of these SCNAs have been noted except for a non-protein-coding locus in 8q.24.21, near MYC—which is also within a separate recurrently amplified locus. (d) Q-Q plot for the analysis of significantly recurrent SV breakpoints. The most SRBs are within the long noncoding RNA CCDC26, within the TAD encompassing MYC (based on n = 179 tumors). (e) Representative examples of the enhancer amplification through simple tandem-duplications within the long noncoding RNA encoding CCDC26.
Extended Data Fig. 2 | See next page for caption.
Extended Data Fig. 2 | Lineage specificity of the enhancer peak in CCDC26. (a) ATAC-seq (top) and H3K27ac ChIP-seq (bottom) enrichment (vertical axis) of samples from different lineages (indicated on right; ‘27 y’ indicates the sample was obtained from a 27-year-old person) across the TAD encompassing MYC (horizontal axis). The location of the MYC coding sequence is highlighted in red. The CCDC26 amplicon boundaries for the 15 samples with the amplicon are indicated by the paired red arrows at the top. The consensus amplicon is indicated by the green dotted lines and centers on an H3K27ac peak present only in glial samples. (b) Hi-C heatmaps depicting DNA interaction profiles (5 kb resolution) from a midline glioma (top), iPSC-derived neural progenitor cells (2nd from top) and two cell lines harboring H3.3K27M mutations (bottom). Red and white indicate high and low interaction frequencies, respectively. MYC interacts more frequently with the H3K27ac peak within CCDC26 (black oval) relative to neighboring loci. The minimal common region of the CCDC26 amplicon is indicated at the bottom of the heatmaps (SV MCR; blue rectangle). (c, d) Correlation between MYC expression and genomic CN of (C) its enhancer amplified in the CCDC26-SV (p = 0.01, two-sided Spearman rank correlation test; samples with MYC CN > 2.5 excluded, n = 94 tumors), or (d) the MYC coding sequence (p = 0.0003, two-sided Spearman rank correlation test; n = 114 tumors). (e, f) Lineage specificity of E1 enhancer activity in (E) neural lineage/BT245 p-value(E1 vs. Backbone) = 0.0056; p-value(LUAD vs. Backbone) = 0.99 and (f) lung epithelial lineage/A549, p-value(E1 vs. Backbone) = 0.96; p-value(LUAD vs. Backbone) = 0.0071, n = 3 independent experiments, Nested One-Way Anova: Tukey’s Multiple Comparisons. Analogous to Fig. 1g; center line of the boxplot indicates the median, bounds of the box the 25th and 75th percentiles and whiskers extend from the box to the largest or smallest value no further than 1.5x IQR.
Extended Data Fig. 3 | Significantly recurrent juxtaposition between MYCN and ID2. (a) (top) Count matrix showing all possible juxtapositions between pairs of genomic loci. (bottom) Illustration of the principle behind the analysis of recurrent juxtapositions, as exemplified by the MYCN-ID2 loci. First, we count the number of SVs connecting each pair of genomic loci. Using a background model for the probability of juxtapositions generated from an analysis of 2658 cancers, we then determine the probability of observing this number of SVs due to chance alone, corrected for multiple hypothesis testing. This analysis revealed the MYCN-ID2 juxtaposition as the only significantly recurrent juxtaposition in the window shown. (b) Overlay of amplification frequencies on ChIP-seq data in the ID2 and MYCN loci. The top two tracks show, among tumors with MYCN-ID2 rearrangements (top track, n = 4 tumors) or MYCN amplifications without ID2 involvement (second track, n = 4 tumors), the percentage of tumors with amplifications (y-axis) at each genomic locus (x-axis). The bottom eight tracks indicate H3K27ac ChIP-seq profiles across these loci for four H3K27M and four H3WT pHGG tumors. Coding sequences of ID2 and MYCN are highlighted with yellow and red lines respectively. Significantly enriched H3K27ac peaks (q-value < 0.01) are indicated below each ChIP-seq track. The small region at the ID2 locus that is amplified in all MYCN-ID2 pHGGs shows an H3K27ac peak in the ChIP tracks from all six pHGG tumor samples. Tumors that amplify MYCN without ID2 take in a much larger region of the MYCN TAD into the amplicon. (c) G-track plots indicating copy-number profiles and genome topology after consideration of local SVs, for two examples of pHGGs with focal MYCN amplicons without incorporation of ID2. For both tumors the CN and SV profiles support several possible reconstructions of extrachromosomal circular amplicons. All are limited to the neighborhood of MYCN, presumably incorporating endogenous enhancers from the MYCN TAD.
Extended Data Fig. 4 | Structures of recurrent RTK amplicons. (a) Average amplicon profile for all pHGGs with amplifications in the PDGFRA TAD reaching a CN of at least four (n = 15 tumors). The top track shows the percentage of those tumors with amplifications (vertical axis) at each location (horizontal axis). The track below shows the average CN across all tumors with amplifications in the PDGFRA TAD. The segments included in the PDGFRA amplicon in 80% of tumors are highlighted in the red box. Most amplicons range over several Mb, often including KIT. (b) Average amplicon profile for all pHGGs with amplifications in the EGFR TAD reaching at least four copies (n = 7 tumors) displayed as in (A). The segments included in all the EGFR amplicons are highlighted in the red box. The pHGG EGFR amplicons always include upstream enhancers elements around SEC61G. (c) Structure of a simple EGFR-TAD amplicon that encompasses enhancers that are also amplified in all tumors with EGFR amplifications. (d) SVs, CN tracks and reconstructions for all pHGGs with high-level MET amplifications. The observed high-level MET amplicons are a few 100kbp in size. Three out of four MET-amplified pHGGs incorporate a downstream region including an enhancer (bottom H3K27ac track) into the amplicon. For all four MET-amplified pHGGs possible reconstructions of the extrachromosomal amplicon are shown above the CN and SV track. (A-D) From bottom to top the tracks show: the genes of interest at the location, a q-value H3K27ac track calculated from eight pHGG tumor samples, the CN and SV for the indicated tumor at the location and reconstructions of possible extrachromosomal amplicons if applicable. (e) FISH with probes for the GATA4 locus on chr8 (red) and the FGFR2 locus on chr10 (green) in tissue from a pHGG with SVs within high-level amplicons that connect these two loci. Scale bar indicates 2 µm. Representative image from n = 200 nuclei.
Extended Data Fig. 5 | See next page for caption.
**Extended Data Fig. 5 | Associations between SV signatures, genetic variants and SNV signatures.** (a) The statistical significance (as determined by Wilcoxon Tests) of positive (enriched, shown in red) and negative (depleted, shown in blue) associations between each SV signature and of all recurrently altered somatic genetic alterations that are documented in the Cancer Gene Census\(^\text{44}\). Shading within each box indicates level of significance as determined by the q value. (b) De-novo extracted SNV signatures (based on \(n=179\) tumors). (c) Cosine similarity between de novo extracted SNV signatures and the COSMICv3 SBS-signatures. (d) SNV signature activity in every tumor. The hypermutant tumors on the left show signatures associated with hypermutation in COSMICv3. Signature 3, which is similar to the SBS3 homologous recombination deficiency signature, features prominently in many nonhypermutant tumors.
Extended Data Fig. 6 | See next page for caption.
Extended Data Fig. 6 | SV signatures in pHGG based on size, SV type and complexity.  (a) The horizontal axis indicates the size and type of SVs. Del stands for deletion, dup for duplication, inv for inversion, and int for interchromosomal rearrangement. The vertical axis indicates the fraction of SVs within each signature that are contributed by each SV type (based on n = 179 tumors). (b) The statistical significance of positive (enriched) and negative (depleted) associations (Wilcoxon Tests) between each SV signature and all recurrently altered somatic genetic alterations that are documented in the Cancer Gene Census44. (c) Consensus clustering of the normalized SNV and SV signature activities in each tumor sample (columns). Rows indicate signature activities (top) and potentially oncogenic variants (bottom). (D) Correlations between SV and SNV signatures. Signature labels from this analysis are indicated on the left; the nearest COSMICv3 signatures are indicated on the right, with their proposed mechanisms in parentheses. Complex-SV signatures show a close correlation with APOBEC and homologous recombination deficiency SNV signatures (SBS3). q-values are based on Spearman rank correlations. (e) Enrichment analysis for signature activities in each cluster from panel B. FDR q-values are based on Wilcoxon tests. (f) Significance of signature cluster associations for all variants with correlations reaching q < 0.1; q-values are based on Fisher's exact tests. Tumors in the complex-SV clusters are enriched for copy-number changes in cancer genes and SNVs in TP53, whereas simple-SV pHGGs tend to exhibit SNVs in different cancer genes. (g) Number of SVs per tumor in each cluster (n = 179 tumors). All differences are significant to q < 0.003 by Wilcoxon tests, center line of the boxplot indicates the median, bounds of the box the 25th and 75th percentiles and whiskers extend from the box to the largest or smallest value no further than 1.5x IQR.
Extended Data Fig. 7 | See next page for caption.
Extended Data Fig. 7 | Chromothripsis and homology length at the breakpoints in size/SV-type/complexity SV signatures. (a, b) SV signature activities in samples that contain or lack (a) chromothriptic regions or (b) extrachromosomal amplifications (n = 179 tumors, center line indicates the median). (c) SV signature analysis including homology length channels reveals five signatures. The horizontal axis indicates the size and type of SVs. Del stands for deletion, dup for duplication, inv for inversion, and int for interchromosomal rearrangement. The vertical axis indicates the amount of SVs within each signature that are contributed by each SV type. (d) SV signature activity in every tumor of the homology SVsigs. Tumors with higher SV counts on the left show complex-SV signature activity whereas tumors with lower SV counts on the right show a mix of simple-SV signatures mimicking the SVsig distribution without homology information (n = 179 tumors). (E) By sample correlation between SVsigs with and without homology.
Extended Data Fig. 8 | See next page for caption.
Extended Data Fig. 8 | Associations between SV-defined groups and extended comut plot split by histone groups. (a) Correlations between SVs and SNV signatures in included in the COSMICv3 signatures database. SNV signatures with well-established links to mechanisms are indicated in parentheses. These include mismatch repair (MMR), homologous recombination (HR) deficiency, APOBEC and aging. q-values were calculated using Spearman rank correlations. (b) Enrichment analysis for signature activities present in Complex-SV and SNV-dominant clusters. q-values were calculated using Wilcoxon tests. (c) Comut plot of the 176/179 (98.3%) tumors with somatic variants in at least one well-known oncogene. Columns represent tumors, ordered within histone mutation-defined subgroups by HC of all potential driver variants. The top two rows show signature metadata.
Extended Data Fig. 9 | See next page for caption.
Extended Data Fig. 9 | Associations between SV + histone defined groups. (a) Jaccard distances between tumor pairs (vertical axis), calculated from the combination of variants in each tumor, across subgroups of H3K27M DMGs. Tumor groups were determined by H3.1 or H3.3 mutations and the combined complex-SV signatures exceeding 20% of all SV signature activity. Tumors were paired within or between these groups, as indicated on the horizontal axis. All differences were significant with q < 0.005 (FDR-corrected two-sided Wilcoxon test) unless indicated otherwise. n = 165 DMGs) (b) Association between MDM4 expression (vertical axis) and CN (horizontal axis). MDM4 gains universally represent arm-level gains of 1q. ** indicates adjusted p = 0.004, ANOVA with two-sided Tukey post-test, n = 114 tumors. (c, d) Volcano plot indicating the significance (vertical axis; FDR-corrected Fisher’s exact tests) of associations between genetic variants and pHGG subgroups (horizontal axis). (C) Arm-level SCNAs in TP53-disrupted (n = 97 tumors) vs. TP53WT pHGGs (n = 77 tumors). TP53 disruption represented SNVs (n = 88 tumors, often with copy loss) or copy loss alone (n = 9 tumors). (d) Arm-level SCNAs in TP53-disrupted (n = 56 tumors) vs. TP53WT H3K27M mutant DMGs (n = 39 tumors). TP53 disruption represented SNVs (n = 53 tumors often with copy loss) or copy loss alone in (n = 3 tumors). Only significantly recurrent arm-level SCNAs are shown. (e, f) Comparison between pretreatment biopsy and autopsy samples. These groups exhibit no significant differences in (e) the number of SVs per sample (q = 0.6, Wilcoxon, n = 174 tumors) or (f) the activity of the combined complex-SV signatures (q = 0.7, Wilcoxon, n = 174 tumors). Center line of the boxplots indicates the median, bounds of the box the 25th and 75th percentiles and whiskers extend from the box to the largest or smallest value no further than 1.5x IQR. (g) Kaplan–Meier plot indicating OS for (top) H3.1K27M and H3.3K27M DMGs and (bottom) H3.3K27M DMG with and without TP53 SNVs. p-values are from log-rank tests. Error bands show the 95% confidence interval.
Extended Data Fig. 10 | See next page for caption.
Extended Data Fig. 10 | Timing analysis of somatic variant acquisition in histone mutation-defined pHGG subgroups. For each subgroup the individual (per-sample) timing of recurrent variants is fed into a Bradley-Terry model. This results in a strength parameter for each variant which is indicated on the x-axis in log scale and can be interpreted as the relative log odds of the variant being an early event in this subgroup. Each distribution indicates the results of 100 random subsamples of the respective subgroup. Only potential driver variants recurrent in more than two samples are shown. Subgroups: (a) H3.1K27M, n = 24 tumors (b) H3.3K27M, n = 73 tumors (c) H3.3G34R, n = 14 tumors (d) H3WT, n = 63 tumors (e) hypermutant pHGGs, n = 5 tumors. Center line of the boxplot indicates the median, bounds of the box the 25th and 75th percentiles and whiskers extend from the box to the largest or smallest value no further than 1.5x IQR.
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.

- 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
No software used for data collection

Data analysis
All software used is described in the methods section including links to repositories. Publicly available software included: R (3.6.3), Prism (9), BWA (0.7.1), SAMTools (1.10, https://github.com/samtools/samtools), Picard (2.21.6, https://github.com/broadinstitute/picard), Mutect2 (4.1.4.1, https://github.com/broadinstitute/gatk), GATK4 (4.1.4.1, https://github.com/broadinstitute/gatk), MutSigCV (1.3.01, https://www.genepattern.org/modules/docs/MutSigCV), ABSOLUTE (1.0.6, https://software.broadinstitute.org/cancer/gsa/absolute_download), GISTIC (2.0, https://software.broadinstitute.org/cancer/gsa/gistic), SvABA (1.1.0, https://github.com/walah/svaba), clusterSV (1.0.0, https://github.com/cancerct/ClusterSV), SignatureAnalyzer (0.0.8, https://github.com/getlab/signatureanalyzer), STAR (2.5.2b, https://github.com/alexdobin/STAR), RNA-SeqQC (2.3.5, https://github.com/getlab/maeseq), DESeq2 (3.10, https://bioconductor.org/packages/release/bioc/html/DESeq2.html), ComBat (3.0, https://www.genepattern.org/modules/docs/ComBat), MACH2 (2.2.7.1, https://pypli.org/project/MACH2), Juicer (1.6, https://github.com/aidenlab/juicer), Jalba (v1.0, https://github.com/msklib/jalba), gTrack (1.0.0, https://github.com/msklib/gTrack), Palimpsest (https://github.com/FunGeST/Palimpsest), MutationTimeR (1.0.0, https://github.com/gerstung-lab/MutationTimeR), Mobster (1.0.0, https://github.com/yhehir/mobster), BradleyTerryScalable (v0.1.0, https://github.com/EllaKaye/BradleyTerryScalable), Fishhook (1.0.0, https://github.com/msklib/fishhook), ChromHMM (1.2.0, https://compbio.mit.edu/ChromHMM), Java (13.0.1), gGnome (v1.0, https://github.com/msklib/gGnome), deconstructSigs (v1.8.0, https://github.com/raerose/deconstructSigs), ConsensusClusterPlus (3.10, https://bioconductor.org/packages/release/bioc/html/ConsensusClusterPlus.html), Shatterseek (v0.4, https://github.com/parklab/ShatterSeek), Oncotator (v1.9.9.0, https://github.com/broadinstitute/oncotator/releases), cbiportal (v4.1.4, https://www.cbiportal.org/oncoprinter)

Main custom analysis code is available at: https://github.com/FrankDubois/phHG_SVs. All custom code to connect and reformat the outputs of the publicly available software as well as code to generate the figures is available upon request.

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

De novo generated sequencing data from this study are accessible under dbGaP Accession Number phs002380.v1.p1
Previously published sequencing data4-7 that were re-analysed here are available under accession code EGAS00001000575, EGAS00001001139, EGAS00001000572, EGAS00001000192, GSE128745, GSE4792, GSE126319 and the Encode project. TAD boundaries are from GSE77565. Cosmic signatures and cancer genes are available at: https://cancer.sanger.ac.uk/cosmic/download. Source data have been provided as Source Data files. All other data supporting the findings of this study are available from the corresponding author on reasonable request.

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 [X] Behavioural & social sciences [ ] Ecological, evolutionary & environmental sciences

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

Life sciences study design

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

Sample size
No statistical test was used to determine the sample size. We collected all available published pHGG whole genome sequencing data and sequenced samples available through clinical trials at our institutions. The resulting sample size was 179 gliomas, representing all available data to us, which was sufficient to detect molecular and genetic associations.

Data exclusions
Three previously published samples were excluded because their fastq files could not be successfully realigned using our pipelines. Exclusion criteria were pre-established.

Replication
Our experiments with the in vitro luciferase assays were replicated in three independent experiments. All replicates were used in the analysis.

Randomization
No randomization was performed in this study because the data was analyzed as a cohort and did not require randomization between groups. For associations within the cohorts, we adjusted for covariates by performing Cox-Regression, when applicable. Overall, this study was designed as a descriptive cohort study and was not intended to compare known groups which would require randomization.

Blinding
Investigators were not blinded because the analysis were done in an automated fashion without risk of operator bias.

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 |

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)
SU-DIG-I3 (DIPG13) cells, derived from a patient biopsy was provided by the creator of the line, Dr. Michelle Monje at
Human research participants

Policy information about studies involving human research participants

Population characteristics
Population characteristics are described in supplementary table 1.

Recruitment
This study includes published data available under EGAS00001000575, EGAS00001001139, EGAS00001000572 and EGAS00001000192. Novel data was generated from samples obtained from the DIPG-BATs clinical trial (NCT01182350), the Dana-Farber Tissue Bank or collaborating institutions, under protocols approved by the institutional review board of the Dana-Farber/Harvard Cancer Center with informed consent.

Ethics oversight
Protocols were approved by the institutional review board of the Dana-Farber/Harvard Cancer Center (10-321 and 10-417).

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

Clinical data

Policy information about clinical studies
All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration
NCT01182350

Study protocol
https://clinicaltrials.gov/ct2/show/NCT01182350

Data collection
This study includes published data available under EGAS00001000575, EGAS00001001139, EGAS00001000572 and EGAS00001000192. Novel data was generated from samples obtained from the DIPG-BATs clinical trial (NCT01182350), the Dana-Farber Tissue Bank or collaborating institutions, under protocols approved by the institutional review board of the Dana-Farber/ Harvard Cancer Center with informed consent (DFCI protocols 10417, 10201 and DFCI 19293). Data was collected from 2011 until 2017.

Outcomes
Overall survival was obtained from the DFCI protocols 10417, 10201, 19293 as well as published data. This manuscript does not report pre-defined primary or secondary outcomes of any clinical trial.