SUPPLEMENTARY METHODS

Table of Contents

1. Participants and glioma samples 3

1.1 Sample selection 3

1.2 Glioma tumour classification 3

1.3 Sequencing of Genomics England samples 4

2. Variant calling and filtering 5

2.1 Removing alignment bias introduced by soft clipping of semi-aligned reads 6

3. Mutational signature analysis 6

4. Copy number alteration profiling 6

4.1 Stage 1: Initial profiling of copy number alterations 7

4.2 Stage 2: Evaluation of profile concordance with variant allele frequency distributions 7

4.3 Stage 3: Quality assessment 9

4.4 Stage 4: Re-profiling of copy number alterations 10

5. Structural variants 10

5.1 Structural variants calling 10

5.2 Structural variant classification 11

5.3 Hotspots of simple structural variation 12

5.4 Genome regions enriched for chromothripsis 15

5.5 Structural variation signatures 16

6. Neoantigen and immune escape prediction 17

6.1 HLA allele detection 17

6.2 Neoantigen prediction 17

6.3 Detection of immune evasion mechanisms 17

7. Detection of recurrent broad and focal copy number alterations 18

7.1 Preparing input copy number segmentation file 18

7.2 Running GISTIC 19

7.3 Prioritising likely gene targets of focal amplifications and deletions 19

7.4 Assignment of segments to tumours overlapping recurrent CNAs 20

8. Telomere analysis methods *21*

9. Extrachromosomal DNA detection *21*

10. Annotation of consequence of structural variants *22*

11. Driver analysis methods *22*

11.1 Mutation annotation **22**

11.2 Protein-coding driver identification **23**

11.3 Non-coding driver identification **26**

11.5 OncoKB annotation of driver mutations **28**

11.6 Lollipop plots of driver gene mutations **29**

11.7 Identification of clinically actionable driver alterations **29**

12. Timing driver mutations *30*

12.1 Preparing MutationTimeR input files **30**

12.2 Running MutationTimeR **31**

13. Mutational signature attribution *31*

13.1 Annotation of doublet base substitution (DBS) mutations **31**

14. Druggability and essentiality analysis of driver genes *31*

15. Clinical data collection and outcome measures *32*

16. Statistical analysis *32*

17. Correlating genomic and clinical features *33*

1. Participants and glioma samples

Whole Genome Sequencing (WGS) for glioma was performed through the Genomes project: the 100,000 Genomes project was approved by the East of England – Cambridge South Research Ethics Committee (REC reference: 14/EE/1112). Only individuals from the 100,000 Genomes project for whom WGS data were available and who consented for their data to be used for research purposes were included in the analyses. All data were de-identified before use. The research was conducted within the Research Environment (a secure data platform complying with the Department of Health Information Governance policy and meeting the data security standards of the Information Governance Toolkit of the Health and Social Care Information Centre) using LabKey (Main Programme Version 8 release, 28/11/2019).

1.1 Sample selection

Tumour sample purity and sequencing data quality affect the sensitivity and precision of variant calling[1] and we therefore excluded samples meeting either of the following criteria:

1. High sample cross-contamination. Tumours were excluded if cross-contamination of the tumour sample or matched germline sample was >1%, as estimated by VerifyBamID[2].
2. Low tumour sample purity. Estimating tumour sample purity is difficult when purity is low[3] and we therefore used distributions of single-nucleotide variant (SNV) allele frequencies (VAFs) to identify low-purity samples, as low average SNV VAF can be indicative of low purity[4]. Tumour samples with median SNV VAF <0.15 were therefore excluded. SNV VAF distributions were also reviewed manually to check that no samples were excluded due to high intra-tumour heterogeneity, which can increase the proportion of SNVs with low VAFs.

1.2 Glioma tumour classification

Tumour sequence information was integrated into glioma tumour classification according to the CNS WHO v5 criteria[5]. The following molecular features were defined:

- IDH mutation – any missense mutation altering codon 132 of IDH1 or codon 172 of IDH2
- *TERT* promoter mutation – either chr5:1295135 G>A or chr5:1295113_G >A

- Chr7 gain – at least half of chr7 has a total copy number higher than the called tumour ploidy (>2 for diploid, >4 for tetraploid)
- Chr10 loss – at least half of chr10 has a copy number call consistent with loss of heterozygosity (e.g. 1/0, 2/0)
- Chr1p/chr19q co-deletion – at least half of chr1p and at least half of chr19q has a copy number call consistent with loss of heterozygosity (e.g. 1/0, 2/0)
- *CDKN2A* bi-allelic loss – copy number at the transcription start site (TSS) of *CDKN2A* is 0/0
- *EGFR* amplification – copy number at the TSS of *EGFR* is >5

Specifically:

- GBM tumours with IDH mutation are classified in CNS WHO v5 as Astrocytoma, grade 4.
- Diffuse IDH wildtype astrocytomas are classified in CNS WHO v5 as GBM, grade 4 if fulfilling one or more of the following criteria:
    - Combined loss of chr10 and gain of chr7
    - *EGFR* amplification
    - *TERT* promoter mutation
- Astrocytomas with 1p/19q co-deletion are classified in CNS WHO v5 as Oligodendrogliomas

1.3 Sequencing of Genomics England samples

Samples were prepared using an Illumina TruSeq DNA library preparation kit and sequenced on a HiSeq X, generating 150 base pair (bp) paired-end reads. Tumour and constitutional DNA samples were sequenced to average depths of 100x and 33x respectively. Poor sequencing quality outliers were identified using principal component analysis and removed based on the following quality metrics: percentage of mapped reads; percentage of chimeric DNA fragments; average insert size; AT/CG dropout; and unevenness of local coverage. Illumina's North Star pipeline (v2.6.53.23) was used for the primary whole genome sequencing analysis. Sequence reads were aligned to the *Homo sapiens* GRCh38Decoy assembly using Isaac (v03.16.02.19)[6]. After the outlined quality control steps, tumour and germline sequencing data for 403 fresh-

frozen adult glioma samples were obtained from the 100kGP main programme version 8 release and taken into our analysis (Supplementary Data 33).

2. Variant calling and filtering

Single nucleotide variant (SNV) and small insertion and deletion (indel) calling was performed using Strelka (v2.4.7). In addition to the default Strelka filters, we applied the following filters:

- Variants with a germline allele frequency (AF) > 1% in a Genomics England dataset.
- Variants with a population germline AF > 1% in GnomAD[7].
- Recurrent somatic variants with frequency > 5% in a Genomics England dataset. A 5% cut-off was chosen based on the frequency of recurrent non-synonymous variants in Cancer Gene Census genes[8].
- Variants overlapping simple repeats as defined by Tandem Repeats Finder[9].
- Indels in regions with high levels of sequencing noise where ≥ 10% of the base calls in a window extending 50bp either side of the indel have been filtered out by Strelka due to the poor quality.
- Indels within 10bp of 100kGP or GnomAD v3 germline indel with AF > 1%.
- Variants in regions of poor mappability where the majority of overlapping 150bp reads do not map uniquely to the variant position.
- SNVs resulting from systematic mapping and calling artefacts present in both tumour and normal 100kGP sample sets. We tested whether the ratio of tumour allele depths at each somatic SNV site was significantly different to the ratio of allele depths at this site in a panel of normal samples (PoN) using Fisher's exact test. The PoN was composed of a cohort of 7,000 non-tumour genomes from the Genomics England dataset. At each genomic site only individuals not carrying the relevant alternate allele were included in the count of allele depths. The mpileup function in bcftools (v1.9) was used to count allele depths in the PoN. To replicate Strelka filters duplicate reads were removed and quality thresholds set at mapping quality ≥5 and base quality ≥5. All somatic SNVs with Fisher's exact test phred score <80 were filtered, with the threshold determined by optimising precision and recall calculated from a TRACERx truth set[10].

2.1 Removing alignment bias introduced by soft clipping of semi-aligned reads

The Isaac --clip-semialigned parameter invokes the soft clipping of read ends until five consecutive bases are matched with the reference genome[6]. This soft clipping therefore results in the loss of support for alternate alleles occurring within five bases of each read end, leading to artefactually low VAFs. To address allelic bias introduced by this clipping, we introduced FixVAF to soft clips all reads by 5bp at each end, regardless of whether any of the bases are variant sites or whether the reads support reference or alternate alleles[11]. Reads containing small insertions and deletions at variant positions were ignored.

3. Mutational signature analysis

Single base substitution (SBS), doublet base substitutions (DBS) and small insertions and deletions (ID) signatures were inferred from filtered Strelka input VCFs using SigProfilerMatrixGenerator v1.1.30[12] and SigProfilerExtractor v1.1.4[13]. Signatures were extracted *de novo* and decomposed to known COSMIC signatures (v3.2). The following parameters were used: "minimum_signatures= 1, maximum_signatures=25, nmf_replicates=500, nmf_init='random', matrix_normalization='gmm', min_nmf_iterations=10000, max_nmf_iterations=1000000, nmf_test_conv=10000", otherwise default parameters were used.

The optimal number of *de novo* signatures extracted was determined by seeking to maximise the mean sample cosine distance while maintaining the average stability >0.9, all while ensuring that the decomposed COSMIC signatures were plausible in gliomas based on a reference pan-cancer comparison[14]. For SBS96 the chosen signature solution was 9 signatures, for ID83 this was 2 signatures, and for DBS78 this was 2 signatures (Supplementary Fig. 12).

4. Copy number alteration profiling

Clonal and sub-clonal somatic copy number alterations (CNAs) were detected using an iterative procedure incorporating Battenberg v2.2.8[15] (Supplementary Fig. 29). This procedure comprises four stages: (stage 1) initial CNA profiling; (stage 2) evaluation of CNA profile concordance with

variant allele frequency (VAF) distributions; (stage 3) quality assessment; and for those samples that fail quality assessment, (stage 4) CNA re-profiling with alternative purity and ploidy estimates and repeat of stages 2 and 3.

4.1 Stage 1: Initial profiling of copy number alterations

Battenberg was used to detect clonal and sub-clonal CNAs and to estimate sample purity and tumour ploidy[15]. Briefly, numbers of reads supporting single-nucleotide polymorphism (SNP) reference and alternate alleles were counted in tumour and normal samples using alleleCount-FixVAF, which removes reference bias introduced by ISAAC[16]. Heterozygous SNPs were phased using SHAPEIT2 v2.r904[17] and A and B alleles assigned. These data were then segmented using piece-wise constant fitting[18] and sub-clonal copy-number segments identified using t-tests. Sample purity and tumour ploidy were estimated using the method described by Van Loo, et al. [19]. As sequencing data were aligned to hg38, SNP positions were converted to hg37 before phasing, and output segments converted back to hg38.

4.2 Stage 2: Evaluation of profile concordance with variant allele frequency distributions

Expected variant allele frequency is dependent of the fraction of tumour cells containing the variant, the tumour copy number profile, the number of chromosome copies with the variant (its multiplicity) and the sample purity[4]. Given the tumour copy number profile and an estimated sample purity, we can therefore expect to observe enrichment of variants with allele frequencies approximating particular values (representing variants present in all tumour cells)[19]. Failure to observe such enrichment would suggest that either the copy number profile or sample purity is incorrect, and we therefore assessed Battenberg output validity using SNV VAF distributions (Supplementary Fig. 30).

When evaluating SNV VAF distributions, only autosomal genome segments with copy number states of 1:1, 1:0, 2:2, 2:1, 2:0 and no evidence of sub-clonal copy number changes were considered. Each of these five copy number states was evaluated separately, as the possible variant multiplicities and expected clonal SNV VAFs differ between them[19] (Supplementary Fig.

30a). Copy number states corresponding to genome regions containing <5% of all SNVs were not considered. Expected locations of VAF distribution peaks were computed as:

(1)

$$\frac{\rho_{\text{Battenberg}} m}{2\left(1-\rho_{\text{Battenberg}}\right)+\rho_{\text{Battenberg}}\psi_v}$$

Where $\rho_{\text{Battenberg}}$ is the sample purity estimated by Battenberg, ψ_v the ploidy of the tumour at the variant site, and *m* the variant multiplicity, which can equal 1 or 2 in copy number states of 2:2, 2:1 and 2:0 and only 1 in states of 1:1 and 1:0 (Supplementary Fig. 30b). VAF distribution peaks were called using kernel density estimation, implemented in peakPick v0.11[20], with peaks corresponding to densities <0.3 being excluded (Supplementary Fig. 30c). For each copy number state, the expected peak location corresponding to the highest considered variant multiplicity was matched to the observed VAF distribution peak with the greatest VAF, whilst any other expected peak location was matched to the observed peak with the most similar VAF (Supplementary Fig. 30d). Tumour heterogeneity can prevent VAF peak detection, and therefore for samples where >1 expected peak locations were considered, the expected peak furthest from the respective matched observed peak (in terms of VAF) was discarded (Supplementary Fig. 30e). Sample purity (ρ_i) was re-estimated for each remaining expected peak location using the matched observed peak VAF (Supplementary Fig. 30f):

(2)

$$\rho_i = \frac{2a}{m+\omega(2-\psi_s)}$$

Where w is the VAF of the matched observed peak and ψ_s the ploidy of the respective copy number state. Greater variant numbers improve our ability to call peaks, and therefore a single new purity estimate (ρ_{new}) was computed as the weighted average of the peak-wise purity estimates (and used when re-profiling samples) (**Supplementary Fig. 2g**):

(3)

$$\rho_{\text{new}} = \sum_i \frac{n_i \rho_i}{N q_i}$$

Where n_i is the number of SNVs in genome regions of the respective copy number state, N is the number of SNVs in genome regions of all considered copy number states, and q_i is the number of considered variant multiplicities for the respective copy number state. When assessing CNA profile quality, the weighted average of the difference between the purity estimated by Battenberg and the peak-wise purity estimates was used:

(4)

$$\eta = \sum_i \frac{n_i |\rho_i - \rho_{\text{Battenberg}}|}{N q_i}$$

4.3 Stage 3: Quality assessment

Multiple criteria were used to assess Battenberg output validity:

- VAF distribution peaks observed at the correct locations (defined as h<5%).
- DPClust[15] identified a clonal mutation cluster (defined as a cluster containing ≥5% of all SNVs with a CCF of between 0.9 and 1.1).
- DPClust identified no "super-clonal" mutation clusters (defined as clusters containing ≥5% of all SNVs with CCFs >1.1).
- Where Battenberg identifies most of the genome to be tetraploid (2:2), a peak in the SNV VAF distribution in 2:2 regions corresponding to a variant multiplicity of 1 is observed.
- No single homozygous deletion >10Mb is called.

Samples satisfying all criteria were deemed to pass and their CNA profiles and purity estimates were used in subsequent analyses. Samples not satisfying at least one criterion were deemed to fail and were re-profiled (*i.e.* proceeded to stage 4).

4.4 Stage 4: Re-profiling of copy number alterations

Samples failing quality assessment were re-profiled a maximum of three times using Battenberg with new purity and ploidy estimates. Samples failing quality assessment after three re-profiling attempts were not considered in subsequent analyses. The new purity (ρ_{new}) was estimated in stage 2, whilst the new ploidy (ψ_{new}) was estimated using the method described by Van Loo, et al. [19]:

(5)

$$\psi_{\text{new}} = \frac{\rho_{\text{Battenberg}} \left(\psi_{\text{Battenberg}} - 2\right) + 2\rho_{\text{new}}}{\rho_{\text{new}}}$$

5. Structural variants

5.1 Structural variants calling

We identified structural variants (SVs; also referred to as rearrangements) using a graph-based consensus approach including Delly[21], Lumpy[22] and Manta[23], and support from CNAs. Delly was run with post-filtering of somatic SVs using all normal samples, and Lumpy and Manta were run with default parameters. Rearrangements from the three SV callers were excluded if <2% of tumour reads at the rearrangement breakpoint site supported the rearrangement, if any reads in the matched normal supported the rearrangement, or if either rearrangement breakpoint was in a centromeric or telomeric region, or on a non-standard reference contig (not chromosomes 1-22, X or Y). Remaining rearrangements were merged using a modified version of PCAWG Merge SV, which is a graph-based approach that identifies and merges rearrangements from multiple callers, allowing 400bp slop for ambiguity in rearrangement breakpoint position[24]. Rearrangements were included in the final call set if they were identified by at least two SV callers, or by one SV caller but with a rearrangement breakpoint <3kb from a CNA segment boundary.

xTea was used to identify somatically acquired long interspersed nuclear element (LINE-1) retrotransposition events[25]. Other retrotransposition categories, including Alu elements, SINE-VNTR-Alu elements and processed pseudogene, were not considered as they collectively

comprise ≤3% of retrotransposition events across human cancers[26]. Retrotransposition and other SV-generating events are mechanistically distinct[26] and we therefore excluded retrotransposition events from our SV analyses. SVs identified using the graph-based consensus approach were classified as potential retrotransposition events and excluded if: (1) xTea identified a transduced region within 10kb of either rearrangement break point in ≥1% of glioma samples, or (2) xTea identified a transduced region within 10kb of either rearrangement break point in the same sample. A threshold of 10kb was chosen as most somatically acquired transductions span regions <10kb from a canonical LINE-1 element[27].

5.2 Structural variant classification

ClusterSV was used to group rearrangements identified by the graph-based consensus approach into footprints and clusters based on their proximity within the genome[24]. Rearrangement footprints represent rearrangement breakpoint sets that are positionally associated, whilst rearrangement clusters represent mechanistically associated rearrangements sets. The string approach proposed by Li, et al. [24] was used to describe rearrangement footprints. Rearrangement clusters were classified as simple events (deletions, tandem duplications, balanced inversions, balanced translocations, unbalanced translocations, or simple unclassified) or complex events (chromoplexy, chromothripsis, or complex unclassified) if they comprised ≤2 or ≥3 individual rearrangements respectively.

Established criteria were used to identify chromothripsis events[28,29]. Rearrangement clusters were defined as chromothripsis events if it met all the following criteria:

- At least six interleaved intra-chromosomal rearrangements.
- A series of four adjacent genome segments oscillating between two copy number states, or five adjacent genome segments oscillating between three copy number states.
- No evidence (FDR >0.2) that the distribution of intra-chromosomal fragment join orientations diverged from a multinomial distribution with equal probabilities for the four orientation categories (deletion-like, duplication-like, head-to-head inversion, and tail-to-tail inversion).

Rearrangement clusters were defined as chromoplexy if it met all the following criteria:

- Consisted of between three and thirty rearrangements.
- Comprised a rearrangement chain spanning at least three chromosomes[24]. SV chains were identified using a graph-based approach, in which vertices represent breakpoints, connected by an edge if they fell within 1Mb of each other and were not involved in the same rearrangement.
- At least 50% of the rearrangement footprints in the cluster represented balanced translocations, either with a deletion bridge between the break ends, or no observed copy number change.

5.3 Hotspots of simple structural variation

Structural variation rates differ across the genome and are influenced by local genomic features[24]. We therefore identified genome regions enriched for simple SVs using a permutation-based approach as per Glodzik, et al. [30], considering genomic features associated with structural variation occurrence. Each simple SV class was considered separately (deletions, tandem duplications, balanced inversions, balanced inter-chromosomal translocations, unclassified simple SVs). Individual rearrangements that form parts of complex SVs were not considered in this analysis.

Part 1: Evaluating associations between SV rates and local genomic features

Negative binomial regression was used to test associations between genomic features and the number of SVs from each simple class[30]. The models comprised the following features: average total copy number across the bin in the glioma sample set, the presence of genes highly or lowly expressed in glioma, GC content, ALU repeats, segmental duplications, other genomic repeats, fragile sites, replication timing, and DNase, H3K36me3 and H3K9me3 peaks[30]. Highly and lowly expressed genes were defined as those with mean RSEM value in the top 25% and bottom 75% of protein-coding genes in glioma TCGA samples with available RNA-Seq data[31]. Genomic repeats were obtained from the UCSC Genome Browser[32]. Segmental duplications were obtained from the Segmental Duplication Database in GRCh38 (https://humanparalogy.gs.washington.edu)[33].

Fragile sites were taken from Bignell, et al. [34]. Replication timing data from neural progenitor cells were obtained from ReplicationDomain (https://www2.replicationdomain.com)[35]. DNAse-seq data (ENCFF417MBH) and ChIP-seq data for histones H3K36me3 (ENCFF129ASR) and H3K9me3 (ENCFF927KUC) were obtained for neural progenitor cells from ENCODE[36].

Part 2: SV simulation

Simulation of SVs was used to test whether the number of SVs observed in a genome region was greater than expected by chance given local features[30]. SVs were simulated for each simple SV class, preserving the number and length of SVs observed in the glioma sample set. SVs were simulated by dividing the genome into non-overlapping 1Mb bins, in which genomic features (listed above) were summarized. To aid comparison, all genomic features were normalized to a mean of zero and standard deviation of one. Numbers of break ends expected in each bin were estimated using the effect estimates from negative binomial regression models. For each observed SV, an SV was then simulated by sampling a bin with probabilities proportional to the expected numbers of break ends in each bin. For inter-chromosomal SVs, a partner break end was simulated by selecting a bin under probabilities proportional to the expected numbers of break ends in each bin, excluding bins on the same chromosome, and then choosing a position from the bin at random. For intra-chromosomal SVs, a partner break end was simulated by selecting a position either upstream or downstream equal in distance to the distance between the two break ends in the observed SV. If either break end fell within a telomere or centromere then the SV was re-simulated. SVs were simulated 1,000 times to compute the expected number of SVs in each 1Mb bin.

Part 3: SV hotspot identification

Genome regions containing greater numbers of SV break ends than expected were identified using piece-wise constant fitting (PCF)[30]. SV break ends were first positionally sorted and the distance between adjacent break ends calculated. PCF was then applied to the $\log_{10}$ of the intermutational distances (IMD). The observed (d_i^{obs}) and expected (d_i^{exp}) number of break ends per base pair for each PCF segment (i) is then computed as:

(6)

$$d_i^{obs} = \frac{a_i}{s_i}$$

(7)

$$d_i^{exp} = \frac{\sum_{j=1}^{n} b_j}{ns^{bin}}$$

Where a_i is the number of break ends in the segment, s_i is the length of the segment in base pairs, b_j is the expected number of SVs in bin j, n is the number of bins overlapping the segment, and s^{bin} is the bin size (1Mb). An enrichment factor (β_i^{simple}) is then computed for each PCF segment as:

(8)

$$\beta_i^{simple} = \frac{d_i^{obs}}{d_i^{exp}}$$

PCF requires parameters k_{min} (the minimum number of mutations in a segment) and γ (that controls the smoothness of the segmentation). False discovery rates (FDRs) at each β^{simple} value were estimated by applying PCF to both the observed and simulated SV sets and dividing the mean number of segments with a β^{simple} value at least as great in the simulated SV sets by the number of segments with a β^{simple} value at least as great in the observed SV set. FDR values equal to zero were changed to the lowest non-zero FDR value observed, whilst FDR values greater than one were set to one. Optimal γ and k_{min} values were chosen by repeating this process for γ values between one and twenty, and k_{min} values between two and twenty, and selecting values that maximized the number of hotspots identified, whilst minimizing FDR. In the final analysis γ=10 and k_{min}=4 was used. SV hotspots were considered potential artefacts and excluded if the SVs at the hotspot were not supported by CNAs and the SV hotspot had a Rearrangement Dispersion Score less than 1[37]. Overlapping SV hotspots were merged.

Part 4: Fragile site classification

SV hotspots were classified as fragile sites if they satisfied at least three of the following six criteria:

- Were late replicating[38]. Replication timing data from neural progenitor were obtained from ReplicationDomain (https://www2.replicationdomain.com). Late replicating regions were defined as those with mean Repli-Seq values ≤0.
- Had low gene-density[39]. A threshold of five genes per Mb was set.
- Overlapped a gene greater than 300kb in size. This threshold was chosen as fragile sites generally occur in geome regions containing genes at least 300kb in size[40].
- The overlapping gene of greatest size was the focus of the SV enrichment. This was assessed by computing the ratio between SV break point densities in the overlapping gene of greatest size and intergenic regions flanking 1Mb upstream and downstream. A ratio threshold of five was set.
- Overlapped a fragile site reported by Bignell, et al. [34]. These fragile sites were mapped from NCRI36 to GRCh38 co-ordinates using LiftOver[32].
- Overlapped a fragile site reported by Wala, et al. [37]. These fragile sites were mapped from GRCh37 to GRCh38 co-ordinates using LiftOver[32].

SV hotspots were not classified as fragile sites if they contained an identified or previously established driver gene[41].

5.4 Genome regions enriched for chromothripsis

A permutation-based approach was used to identify genome regions enriched for chromothripsis. Chromoplexy and unclassified complex SVs were not similarly considered due to low frequency. First, the genome was split into non-overlapping 1Mb bins and the observed number of tumour samples with chromothripsis footprints (g_i^{obs}) overlapping each bin (*j)* counted. Chromothripsis footprint positions were then permuted 100,000 times by randomly sampling regions of the genome equal in size to the footprints. The expected number of tumour samples with chromothripsis footprints (g_i^{exp}) overlapping each 100kb bin was next estimated as the mean number of tumour samples with chromothripsis footprints overlapping the bin

across all permutations. A chromothripsis enrichment factor (β_j^{chromo}) was calculated for bin (j) as:

(9)

$$\beta_j^{chromo} = \frac{g_j^{obs}}{g_j^{exp}}$$

FDRs at each β^{chromo} value were estimated by computing β^{chromo} for each bin in both the observed and permuted chromothripsis sets and dividing the mean number of bins with a β^{chromo} value at least as great in the permuted chromothripsis sets by the number of bins with a β^{chromo} value at least as great in the observed chromothripsis set. FDR values equal to 0 were changed to the lowest non-0 FDR value observed, a maximum FDR of 1 applied. Regions with an FDR < 5% were considered enriched for chromothripsis events.

5.5 Structural variation signatures

Due to low numbers of complex SVs, SV signatures were extracted using only simple SVs (deletions, tandem duplications, balanced and unbalanced inversions, and balanced and unbalanced inter-chromosomal translocations). 133/402 tumours with at least ten simple SVs were considered when extracting SV signatures. Deletion and tandem duplication size distributions are multi-modal and these variants were therefore classified as <10kb, 10kb to 1Mb, or >1Mb[24]. Deletions and tandem duplications were additionally classified as late, mid or early replicating, using neural progenitor cell Repli-Seq data from ReplicationDomain (https://www2.replicationdomain.com) and thresholds of <-0.5, -0.5 to 0.5, and >0.5. Mechanisms of fragile site instability differ from other SVs and we therefore separately considered deletions and tandem duplications at fragile sites[24]. SV signatures were extracted using a hierarchical Dirichlet process (HDP) as implemented in the hdp R-package (v0.1.5)[24]. HDP structure was initialized with one common grandparent node, a parent node for each of the astrocytoma IDH mutation, glioblastoma and oligodendroglioma histologies, and a child node for each of the 133 tumour samples. Four separate Markov chain Monte Carlo posterior sampling

chains were run with 10,000 burn-in iterations, extracting five SV signatures, which we refer to as SV1-5.

6. Neoantigen and immune escape prediction

6.1 HLA allele detection

HLA alleles were detected from germline bam files using HLA-LA v1.0.1[42] (https://github.com/DiltheyLab/HLA-LA).

6.2 Neoantigen prediction

Input file preparation

Input somatic and phased germline-somatic VCFs were prepared for neoantigen detection as described (https://pvactools.readthedocs.io/en/latest/pvacseq/input_file_prep/proximal_vcf.html) in order to correct for potential germline effects on candidate peptides/neoantigens[43]. Briefly, a VCF of combined germline and somatic variants was created using GATK's CombineVariants and then phased with GATK's ReadBackedPhasing[44]. This was then annotated using VEP with the plugins "Downstream" and "Wildtype".

pVACseq

pVACseq v1.5.4 (https://pvactools.readthedocs.io/en/latest/pvacseq.html[45]) was used to identify candidate neoantigens. Default parameters were used in filtering and identification of candidate neoantigens (*i.e.* binding < 500nM).

6.3 Detection of immune evasion mechanisms

Detection of allelic imbalance and loss of heterozygosity at HLA-A, HLA-B and HLA-C alleles was carried out using LOHHLA[46] (https://bitbucket.org/mcgranahanlab/lohhla/src/master/). The top-ranked HLA alleles called by HLA-LA were compared with genomic fasta files from the IPD-IMG HLA database v3.5[47] (https://github.com/ANHIG/IMGTHLA/). HLA alleles called by HLA-LA not

found in the IPD-IMG HLA database were not considered for LOHHLA. Tumours were designated as exhibiting LOH at HLA-A, HLA-B, or HLA-C alleles using criteria as per McGranahan et al., (i.e. at least one HLA allele has an allelic imbalance P-value < 0.01 and copy number < 0.5). For tumors where ≥ 1 HLA allele could not be assessed by LOHHLA and the remaining alleles showed no evidence of LOH the status was designated as "Undetermined".

Tumours were designated as "HLA mutated" if containing at least one nonsynonymous mutation in HLA-A, HLA-B or HLA-C.

Nonsynonymous mutations and LOH copy number status were assessed in a panel of "immune escape genes" (IEGs): B2M, CALR, CANX, CIITA, ERAP1, ERAP2, HSPBP1, IRF1, PDIA3, PSMA7, PSME1, PSME2, PSME3, TAP1, TAP2, TAPBP as per Cornish, et al. [48].

7. Detection of recurrent broad and focal copy number alterations

7.1 Preparing input copy number segmentation file

For every tumour passing copy number quality control measures, a copy number segmentation file was generated as input for GISTIC from Battenberg per-tumour segmentation output. From each copy number segment identified by Battenberg, the chromosomal coordinates, major (nMaj) and minor (nMin) copy number calls were obtained. In the case of subclonal copy number segments, nMaj and nMin were taken from the subclone with the largest tumour fraction.

Per-segment normalised copy number was calculated differently depending on whether the tumour had been identified as having undergone whole genome doubling (WGD) where the ploidy was assumed to be 4 (tetraploid), otherwise the ploidy was assumed to be 2 (diploid).

For diploid tumours per-segment normalised copy number was calculated by:

(10)

$$SegCN = (nMaj + nMin) - 2$$

For diploid tumours with a germline male karyotype (XY), copy number at chrX was normalised by:

(11)

$$SegCN = (nMaj + nMin) - 1$$

For tetraploid tumours per-segment normalised copy number was calculated by:

(12)

$$SegCN = \frac{(nMaj + nMin) - 4}{2}$$

For tetraploid tumours with a germline male karyotype (XY), copy number at chrX was normalised by:

(13)

$$SegCN = (nMaj + nMin) - 2$$

$SegCN$ was thresholded to a maximum of 2.0 and minimum of -2.0.

7.2 Running GISTIC

GISTIC v2.0.2.3[49] was run to identify recurrent arm-level copy number events, as well as focal amplifications and deletions (https://github.com/broadinstitute/gistic2). The following parameters were used:

"-conf 0.99 -broad 1 -qvt 0.25 -genegistic 1 -gcm extreme -brlen 0.5 -rx 0 -twoside 1 -scent median -armpeel 1 -arb 1 -refgene hg38.UCSC.add_miR.160920.refgene.mat"

7.3 Prioritising likely gene targets of focal amplifications and deletions

Candidate target genes at focal amplifications and deletions were annotated according to the following criteria:

1. Overlap with genes at focal amplifications and deletions previously reported in a pan-cancer TCGA study as found in Supplementary Table 3 of Zack, et al. [50] Comparisons were made both with the overall pan-cancer GISTIC analysis, as well as GISTIC analysis restricted to the given tumour type. Special consideration was given to genes specifically highlighted in the Zack et al., 2013 analysis as being prioritised candidates
2. Overlap with Cosmic Cancer Gene Census genes and whether their annotated role (Oncogene (OG)/tumour suppressor gene (TSG)/ambiguous) is consistent with the copy number change (OG with amplifications, TSG with deletions)
3. Overlap with driver genes identified as being significantly mutated in this study, and whether the driver's likely role (Oncogene (OG)/tumour suppressor gene (TSG)/ambiguous) is consistent with the copy number change (OG with amplifications, TSG with deletions)

On the basis of the above criteria, "consensus" driver genes were manually assigned to peaks. To maximise detection of overlapping genes, comparisons were made with all potential gene synonyms as available from the HUGO gene nomenclature name committee[51] (https://www.genenames.org/).

7.4 Assignment of segments to tumours overlapping recurrent CNAs

Alterations from the broad analysis with $Q<0.05$ were taken to indicate significantly recurrent arm-level events. Copy number segments comprising greater than half of the total chromosome arm size were taken to indicate arm-level events.

In the case of focal events identified by GISTIC, the "wide region" was used to compare potential extent of overlap with Battenberg copy number segments. Segments were defined as overlapping focal events if either the segment interval comprised greater than half of the focal region, or vice versa, using pybedtools and bedtools v2.3.0[52,53] (https://github.com/arq5x/bedtools2).

Segment copy numbers were assigned to the following categories on the basis of comparison of major and minor copy number calls and tumour ploidy (tetraploid if called as whole genome doubled (WGD), otherwise diploid):

0 – homozygous deletion (0/0)

1 – loss of heterozygosity (diploid: 1/0, 2/0; tetraploid: 1/0, 2/0, 3/0, 4/0)

2 – other loss (diploid: none; tetraploid: 1/1, 2/1)

3 – no change (diploid: 1/1; tetraploid: 2/2, 3/1)

4 – gain (diploid: 2 > total copy number ≤ 5; tetraploid: 4 > total copy number ≤ 10)

5 – big gain (diploid: total copy number > 5; tetraploid: total copy number > 10)

As above, in the case of subclonal copy number changes, major and minor copy numbers were taken from the subclone with the largest tumour fraction.

For the purposes of generating a matrix indicating a 1/0 presence/absence of arm-level and focal recurrent CNAs in each tumour, focal/arm losses were assigned as "1" on the basis of an overlapping segment assigned as 0 or 1 as per above, otherwise "0", and focal/arm gains were assigned as "1" on the basis of an overlapping segment assigned as 4 or 5 as per above, otherwise "0".

8. Telomere analysis methods

Telomere content was estimated from tumour and germline bam files by the methods Telomere Hunter v1.1.0[54] (https://pypi.org/project/telomerehunter/) and Telomerecat v3.3.0[55] (https://github.com/cancerit/telomerecat) using default parameters. Telomere content was normalized by (14) $log2(\frac{Tumour}{Germline})$. Telomere content estimates were compared between methods as well as across age groups, tumour purity and WGD status (Supplementary Figs. 31-35).

9. Extrachromosomal DNA detection

Potential extrachromosomal DNA (ecDNA) molecules were detected from tumour bam files using AmpliconArchitect v1.2[56] (https://github.com/virajbdeshpande/AmpliconArchitect). Briefly, per-tumour "seed" regions were prepared from Battenberg copy-number segmentation output if a segment was >100kb and the total copy number was > 5. AmpliconArchitect was then run using these "seed" regions to extract overlapping sequence reads from the tumour bamfile and construct candidate amplicons.

Candidate amplicons were classified using AmpliconClassifier v0.4.6 (https://github.com/jluebeck/AmpliconClassifier) into the following categories: 1) Cyclic (truly circularised ecDNA) 2) Complex non-cyclic 3) Linear amplification 4) No amp/invalid. Amplicons were highlighted if containing a known highly amplified oncogene (*MDM2*, *MYC*, *EGFR*, *CDK4, ERBB2, SOX2, TERT, CCND1, E2F3, CCNE1, CDK6, MDM4, NEDD9, MCL1, AKT3, BCL2L1, ZNF217, KRAS, PDGFRA, AKT1, MYCL, NKX2-1, IGF1R, PAX8*; as per Kim et al., 2020).

10. Annotation of consequence of structural variants

The final set of high-quality structural variant (SV) calls were annotated by SNPeff v5.1[58] (http://pcingola.github.io/SnpEff/) according to GRCh38 ensembl v101 reference to identify their likely effect on gene transcript and/or protein function. Annotations were restricted to canonical protein-coding transcripts in ensembl v101. These annotated SV calls were further prioritised using "simple sv annotation" (https://github.com/AstraZeneca-NGS/simple_sv_annotation) to highlight exon-level deletion events in candidate driver genes.

11. Driver analysis methods

11.1 Mutation annotation

Somatic mutations were annotated to GRCh38 ensembl v101 using the variant effect predictor[59] (VEP). The following parameters were used: "vep -i <input_vcf> --assembly GRCh38 –no_stats –cache –offline –symbol –protein -o <output> --vcf –canonical –dir <ref_dir> --hgvs –hgvsg –fasta <GRCh38_fasta> --plugin CADD,<CADD_score_file> --plugin UTRannotator,<GRCh38_uORF_reference>". The <CADD_score_file> was obtained using CADD

v1.6[60-62] (https://cadd.gs.washington.edu/) with scores obtained for all SNV and indel mutations using the CADD software (https://github.com/kircherlab/CADD-scripts/), before being utilised by the VEP CADD plugin. The plugin "UTRannotator"[63] (https://github.com/ImperialCardioGenetics/UTRannotator) was used to annotate the potential impact of five prime untranslated region (5' UTR) mutations.

11.2 Protein-coding driver identification

Protein-coding driver genes were identified using the IntOGen pipeline[64] (https://bitbucket.org/intogen/intogen-plus/src/master/; downloaded February 2021; https://intogen.readthedocs.io/en/latest/).

Pre-processing of input mutations

Somatic mutations passing the filtering criteria described previously across a tumour cohort were subject to initial sample and mutation pre-processing. In the case of multiple tumours from the same patient, the primary tumour was used (in the case of primary/recurrence pair), alternatively the tumour with the highest purity was used. Hypermutated tumours were flagged for exclusion from downstream driver gene identification if containing > 10,000 mutations and having an outlier mutation count (count > 1.5* interquartile range (IQR) + upper quartile (UQ)). Mutations found to be present in a Hartwig Panel of Normals were further excluded. Unless otherwise specified, mutations were mapped to canonical protein-coding transcripts from Ensembl v101.

Running driver identification methods

Seven driver gene identification methods were run:

1. dNdSCV[65] (https://github.com/im3sanger/dndscv) is designed to detect genes under positive selection that show an excess of nonsynonymous (missense, nonsense, essential splice) mutations after correction for local trinucleotide context.
2. OncodriveFML[66] (https://bitbucket.org/bbglab/oncodrivefml/src/master/) aims to detect driver genes that show an enrichment of mutations with high functional impact. CADD

v1.6 scores were used as measure of functional impact[60-62] (https://cadd.gs.washington.edu/).

3. OncodriveCLUSTL[67] (https://bitbucket.org/bbglab/oncodriveclustl/src/master/) is a method designed to detect driver genes that are enriched for linear mutation clusters.
4. cBaSE[68] (http://genetics.bwh.harvard.edu/wiki/sunyaevlab/cbase) aims to detect driver genes under positive selection that exhibit a significant mutation count bias after correction by tri-nucleotide context.
5. MutPanning[69] (https://github.com/vanallenlab/MutPanningV2) is designed to detect driver genes that exhibit an enrichment of mutations with unusual nucleotide contexts compared to a background model.
6. HotMaps3D[70] (https://github.com/KarchinLab/HotMAPS) detects driver genes containing missense mutations that are spatially clustered together in the 3d structure of the protein. Protein structures obtained from the protein data bank (PDB) were downloaded in March 2020.
7. smRegions[71,72] (https://bitbucket.org/bbglab/smregions/src/master/) is designed to detect genes containing an enrichment of nonsynonymous mutations in regions of interest, such as protein domains, after correcting for trinucleotide context. This analysis utilised information from protein family (pfam) domains which were mapped to ensembl v101 canonical transcripts.

Combination of driver identification methods

The driver combination procedure considered the top-40 ranked genes and their association *P* and *q*-values in each of the seven driver identification methods. Briefly, genes assigned as Tier 1 or Tier 2 somatically mutated genes in the COSMIC Cancer Gene Census[73] (https://cancer.sanger.ac.uk/census; v92 downloaded February 2021) were designated as "CGC" genes and represented a "truth" set of known drivers. Through comparison of the relative enrichment of known drivers in the top ranked gene lists a per-method weighting was obtained. Per-method ranked lists were combined using Schulze's voting method to generate a consensus ranking, with combined *P*-values estimated using a weighted Stouffer Z-score method.

Driver candidates were then classified into the following tiers:

Tier 1 – Candidates where the consensus ranking is higher than the ranking of the first gene with stouffer $q > 0.05$. These represent high confidence drivers.

Tier 2 – Candidates not meeting the criteria for Tier 1 but which are CGC genes, and show a combined stouffer $q_{CGC}<0.25$. Representing a "rescue" of known cancer drivers.

Tier 3 – Candidates not meeting the criteria for Tier 1 or Tier 2 but which have stouffer $q < 0.05$. These represent lower confidence drivers.

Tier 4 – Candidates not meeting criteria for Tier 1 or Tier 2 and stouffer $q > 0.05$. These represent candidates that are not likely to be drivers.

Post-processing of candidate drivers

Candidate driver genes were filtered on the basis of the following annotations:

1. "AUTOMATIC FAIL" – a candidate driver gene would be excluded from further consideration if annotated by at least one of the following:
    a. "TIER4" – categorised into Tier 4 by the combination procedure
    b. "1_METHOD" – only significant (Q<0.1) in 1/7 methods (non-CGC genes)
    c. "EXPRESSION" – gene has very low or absent expression in the relevant TCGA tumour type
    d. "OLFACTORY_RECEPTOR" – gene in list of olfactory receptor genes
    e. "KNOWN_ARTIFACT" – gene is in a known list of artifacts or long genes (e.g. TTN)
2. "MANUAL REVIEW" – if a gene is not excluded on the basis of any "AUTOMATIC FAIL" filters, it is retained as a candidate driver
    a. "GERMLINE" – non-Tier 1-CGC gene has ≥1 mutations per sample and oe_syn/ms/lof > 1.5 based on GnomAD v2.1 constraint metric estimates (https://gnomad.broadinstitute.org/downloads#v2-constraint)
    b. "SAMPLE_3_MUTS" – non-CGC gene where there are ≥3 mutations in ≥1 tumour

c. "LITERATURE" – non-CGC gene where there are no literature annotations according to CancerMine[74] (http://bionlp.bcgsc.ca/cancermine/; downloaded February 2021)

3. "AUTOMATIC PASS" – is not flagged by any "AUTOMATIC FAIL" or "MANUAL REVIEW" filters

Candidate driver roles were assigned on the basis of the dN/dS ratios for missense (*wmis*) and nonsense (*wnon*) mutations for the given gene derived from dNdSCV (https://bitbucket.org/intogen/intogen-plus/src/master/core/intogen_core/postprocess/drivers/role.py):

- A "distance" metric was calculated by (15) $distance = \frac{(wmis - wnon)}{\sqrt{2}}$
- Candidate drivers where $distance > 0.1$ represent those with an excess of missense to nonsense mutations and are assigned as "Oncogenes"
- Candidate drivers where $distance < 0.1$ represent those with an excess of nonsense to missense mutations and are assigned as "Tumour suppressor genes (TSGs)"
- Otherwise, the role of the candidate driver is unclear and is assigned as "Ambiguous"

In the case of multiple cohorts being run representing subsets of a given tumour type, a "consensus" role was designated comparing between each subtype role ("Oncogene" if ≥1 cohort and no other cohorts assigned as "TSG", "TSG" if in ≥1 cohort and no other cohorts assigned as "Oncogene", otherwise "ambiguous").

Gene candidates were annotated by their overlap with any IntOGen cohorts from a previous 2020.02.01 pan-cancer analysis (https://www.intogen.org/download?file=IntOGen-Cohorts-20200201.zip) as well as from a pan-cancer TCGA analysis[75].

11.3 Non-coding driver identification

Defining sets of non-coding regions

Regions from candidate non-coding elements overlapping coding sequence (CDS) or exon regions from canonical protein-coding transcripts were removed using bedops v2.4.39[76] (https://github.com/bedops/bedops).

The following sets of non-coding regions were defined:

1. Core promoters (n=19,283). Defined based on transcription start site (TSS) of canonical protein-coding transcripts: 200bp < TSS < 50bp. Coding sequence (CDS) regions were removed.
2. Distal promoters (n=19,296). Defined based on TSS of canonical protein-coding transcripts: 2kb < TSS. CDS regions were removed.
3. Five-prime untranslated regions (5'-UTRs; n=18,613). Defined based on canonical protein-coding transcripts. CDS regions were removed.
4. Three-prime untranslated regions (3'-UTRs; n=18,806). Defined based on canonical protein-coding transcripts. CDS regions were removed.
5. LincRNAs (n=16,510). Based on exon regions from transcripts annotated as lincRNAs in the ensembl v101 GTF. Exon regions from canonical protein-coding transcripts were removed.
6. miRNAs (n=1,793). Based on regions from transcripts annotated as miRNAs in the ensembl v101 GTF. Exon regions from canonical protein-coding transcripts were removed.
7. Non-canonical splice regions (n=18,163). Defined from regions extending 30bp into the intron from essential splice donor or acceptor sites in canonical protein-coding transcripts. Exon regions from canonical protein-coding transcripts were removed.
8. Enhancers (n=130,996). Defined from ensembl v101 regulatory elements annotated as "Enhancer". Exon regions from canonical protein-coding transcripts were removed.
9. Open chromatin regions (n=95,344). Defined from ensembl v101 regulatory elements annotated as "Open chromatin". Exon regions from canonical protein-coding transcripts were removed.

10. CTCF sites (n=173,711). Defined from ensembl v101 regulatory elements annotated as "CTCF sites". Exon regions from canonical protein-coding transcripts were removed.
11. TF binding sites (n=29,259). Defined from ensembl v101 regulatory elements annotated as "TF binding sites". Exon regions from canonical protein-coding transcripts were removed.
12. GeneHancer Brain elements (n=155,352). Defined from elements annotated as active in brain tissue, brain cell types and glioma cell types in the GeneHancer v5 database[77] (date downloaded: 2021). Exon regions from canonical protein-coding transcripts were removed.

Detecting non-coding drivers

OncodriveFML[66] was run on sets of non-coding regions according to the following amended parameters from the protein-coding analysis: "indel-max" – indels are treated as a set of substitutions, with the functional impact of the indel mutation being the maximum of all the substitutions, and the background simulated as substitutions. A q<0.01 threshold was taken to indicate regions of significance.

11.4 SNV mutations exhibiting extreme strand bias

SNV mutations that otherwise pass filtering criteria as previously detailed were further scrutinised by whether they exhibited excessive strand bias (Strelka INFO field "SNVSB=" > 10). This highlighted a large number of missense mutations causing a recurrent missense change in *CACNA1E (p.Ile95Leu*) that exhibited excessive strand bias and therefore were likely to be false-positive mutation calls. Therefore, it is likely that *CACNA1E* itself is likely a false-positive driver candidate.

11.5 OncoKB annotation of driver mutations

Nonsynonymous mutations in the 682 gene transcripts considered by OncoKB v3.3 were annotated using the OncoKB API (https://www.oncokb.org/)[78]. In the first instance, the HGVSg

identifier was used, however in rare instances this failed a combination of gene symbol, consequence and HGVSp were used to map mutations to OncoKB annotations.

Annotation of oncogenic mutations

Nonsynonymous mutations in candidate driver genes were annotated as "Oncogenic" if either of the following criteria were met:

1. The mutation is annotated by OncoKB as "Oncogenic", "Likely Oncogenic" or "Predicted Oncogenic"
2. The driver role is "Oncogene", consequence is "missense" and mutation is recurrent (seen in 3+ tumours in cohort)
3. The driver role is "TSG" or "ambiguous" and either
    a. Consequence is protein-truncating ("splice acceptor", "splice donor", "frameshift", "stop lost", "stop gained", "start lost")
    b. Consequence is "missense" and mutation is recurrent (seen in 3+ tumours in cohort)

Nonsynonymous mutations not meeting these criteria were considered as a variant of uncertain significance (VUS).

11.6 Lollipop plots of driver gene mutations

Lollipop plots of driver gene mutations were generated using the R package trackViewer[79] (https://github.com/jianhong/trackViewer). Pfam protein domains mapping to the ensembl v101 canonical transcripts were plotted. The protein position was taken from the first position in the HGVSp annotation, other than for splice donor and acceptor mutations where the codon nearest to the HGVSc transcript position was assigned as the protein position.

11.7 Identification of clinically actionable driver alterations

OncoKB[78] v3.17 was used to annotate clinically actionable genomic alterations. The following features were considered for annotation:

1. Nonsynonymous mutations in 696 genes considered by OncoKB
2. Exon-deletion events in EGFR:
    a. Exon 2-7 deletion – EGFRvIII
3. Copy number at TSS of 696 genes considered by OncoKB
    a. Deletion – copy number LOH and total copy number lower than ploidy
    b. Amplification – total copy number at least 5 for diploid tumours, at least 10 for tetrapoloid tumours
4. Tumour mutational burden – coding mutations per Mb > 10
5. Microsatellite instability (MSI) – high levels of one or more SBS signatures related to mismatch repair (MMR) deficiency – SBS6, SBS14, SBS15, SBS21, SBS26, SBS44

12. Timing driver mutations

The relative evolutionary timing of candidate driver mutations was obtained using MutationTimeR[80] (https://github.com/gerstung-lab/MutationTimeR).

12.1 Preparing MutationTimeR input files

Copy number input for MutationTimeR was prepared from Battenberg segmentation files, with the clonal frequency of each segment taken as the tumour purity. In the case of subclonal calls, the clonal frequency was calculated by multiplying the tumour purity by the clonal fraction.

The clusters input for MutationTimeR was prepared from DPClust cluster estimates. The VAF proportion was calculated by multiplying the estimated cluster CCF by the tumour purity. Superclonal clusters (CCF>1.1) were removed.

VCF input for MutationTimeR was obtained from the small somatic SNV/indel variant VCFs which had been filtered as previously described. For SNVs, alt and ref depths were obtained using FixVAF (https://github.com/danchubb/FixVAF). For indels, ref and alt depths were obtained from Tier2 Strelka TAR and TIR fields respectively. Only mutations within Battenberg copy-number segments were retained (note: for male XY tumours with only 1 copy of the X chromosome copy

number information is restricted to the pseudoautosomal region (PAR) and battenberg was not run on the Y chromosome).

12.2 Running MutationTimeR

MutationTimeR was run with 1,000 bootstraps. For tumours previously defined as having undergone whole genome doubling (WGD), the parameter "isWgd" was set to TRUE. Mutations were then classified into estimated simple clonal states (as per Fig.1a of Gerstung, et al. [80]):

- "Clonal [early]" – Mutation on ≥ 2 copies per cell
- "Clonal [late]" – Mutation on 1 copy per cell, no retained allele
- "Clonal [NA]" – Mutation on 1 copy per cell, either on amplified or retained allele
- "Subclonal" – Mutation on < 1 copy per cell

13. Mutational signature attribution

SeqInfo VCFs produced as part of SigProfilerMatrixGenerator (https://github.com/AlexandrovLab/SigProfilerMatrixGenerator) were used to map somatic mutations from input VCFs to their SBS96, DBS78 or ID83 contexts and then to the final SigProfilerExtractor COSMIC v3.2 decomposed signature probabilities.

Dinucleotide mutations called originally by Strelka as adjacent SNVs were re-annotated by VEP. When deriving timing estimates from MutationTimeR the timing from one of the two adjacent SNVs was taken for a given dinucleotide mutation.

13.1 Annotation of doublet base substitution (DBS) mutations

Per-tumour VCFs containing DBS mutations, either directly called originally by Strelka, or originally called by Strelka as two adjacent SNVs and reconstructed as DBS mutations, were created and mutation consequences re-calculated using VEP as above.

14. Druggability and essentiality analysis of driver genes

Proteins of identified driver genes were annotated with pharmacological and druggability data using canSAR's Cancer Protein Annotation Tool (CPAT)[81]. Essential and selective genes were ascertained from the ShinyDepMap[82] analysis, using efficacy and selectivity score thresholds of < -0.56 and > 0 respectively.

15. Clinical data collection and outcome measures

Participants from 20 neurosurgical units in the United Kingdom were included. The clinical governance committee of each participating hospital approved the 100,000 genomes project. Ethical approval for inclusion of clinical data was gained as part of this project through NHS East of England. Individual units were approached to supply additional clinical and treatment information of participants. The following demographic and operative parameters were captured: age, sex, date of surgery, tumour type, laterality tumour location, surgery type, post-operative complications (within 30 days), requirement for CSF diversion, WHO diagnosis/grade, histopathological report, molecular features (ATRX/IDH/1p19q/TERT/MGMT status (%)/p53), treatment (chemotherapy and radiotherapy (number of cycles, type), immunotherapy, trial participation, repeat or previous surgery), radiological extent of resection (comparing the pre- and postoperative magnetic resonance imaging (MRI) findings (≤72 hours postoperatively) using the extent of the fluid-attenuated inversion recovery signal in patients with low grade glioma and the extent of contrast enhancement on T1-weighted images for patients with high grade glioma) as previously classified[83] by (gross total resection (GTR), no residual tumor; subtotal resection (SR), residual volume ≤10 cm^3; partial resection (PR), >10 cm^3 residue), treatment response (progression vs pseudoprogression vs stable disease[84]), survival (last seen alive vs date of death), Eastern Cooperative Oncology Group[85] performance status, any other comments.

16. Statistical analysis

Descriptive statistics were used to characterise the patient population. Statistical analysis was performed using GraphPad Prism v9 and The R Foundation for Statistical Computing Platform v4 statistical package. Chi-square test was used to assess the statistical significance of observed differences between different tumour cohorts. Kaplan-Meier plots were used to assess median

overall survival and progression free survival for tumour subgroups. Cox proportional hazards regression was used to predict survival based on EGFR extrachromosomal DNA amplification (adjusted for age, sex and MGMT methylation status). False discovery rate (FDR) q-values were calculated using the Benjamini-Hochberg method unless otherwise stated.

17. Correlating genomic and clinical features

Key genomic and clinical features were extracted for pairwise statistical association testing. Each feature was categorized as either “Binary”, “Categorical”, “Survival” (in the case of overall survival and progression free survival) or “Continuous” (Supplementary Data 31).

Pairwise comparisons between binary and categorical features were carried out using Fisher’s exact test (python 3.9.19; FisherExact v1.4.2). Pairwise comparisons between binary/categorical and continuous features were carried out using the Kruskal-Wallis test (python 3.9.19; SciPy v1.13.1). Pairwise comparisons between continuous features were carried out using Spearman’s rank correlation test (python 3.9.19; SciPy v1.13.1). Comparisons between genomic and clinical features with OS and PFS were carried out using cox’s proportional hazards test (python 3.9.19; lifelines v0.29.0) with a step-size of 0.1, including the following covariates: patient age at diagnosis, patient sex.

18. Germline analysis

Germline-related analyses made use of the Aggregated Variant Calls (AggV2) resource provided by Genomics England. This briefly comprised an aggregated analysis of 78,195 germline genomes. Further details are available at Genomics England (https://re-docs.genomicsengland.co.uk/aggv2/). Of the 400 participants with diffuse glioma used in our landscape analysis, germline variant calls were available for 393.

18.1 Ancestry prediction

Population ancestry predictions were provided by Genomics England. Briefly, they were obtained for aggV2 samples by projecting calculated principal components (PCs) onto five broad super-

populations from 1000 genomes project phase 3 (African, Admixed American, East Asian, European, South Asian). Further details are provided by Genomics England (https://re-docs.genomicsengland.co.uk/ancestry_inference/). Ancestry predictions were obtained for 393/400 participants in this study.

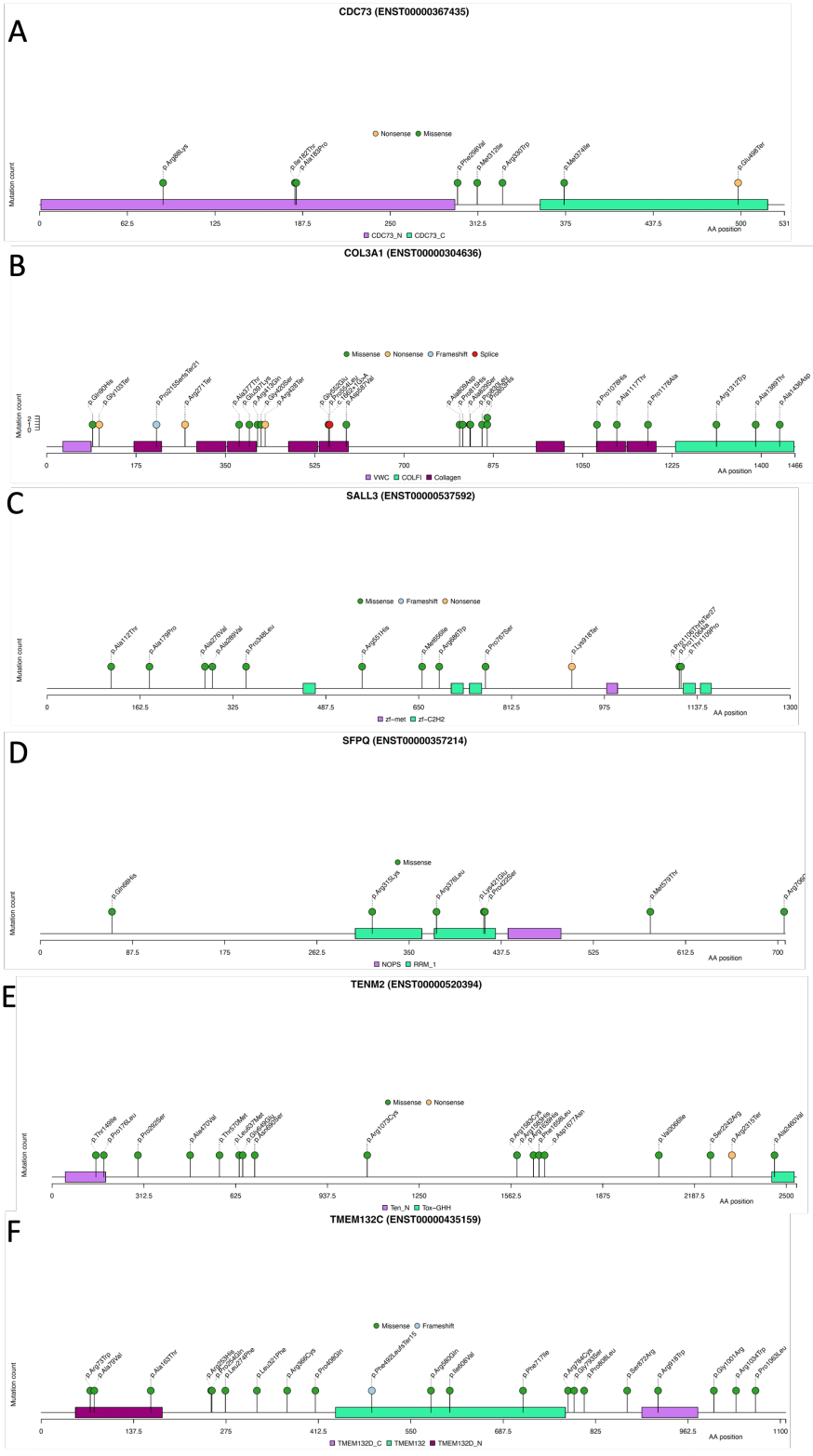
18.2 Germline variants

Germline variant genotypes and annotations were obtained from aggV2 VCFs for query common single nucleotide polymorphisms (SNPs) as well as pathogenic mutations in query genes for 393/400 participants in our landscape analysis. Query common SNPs comprised glioma risk SNPs identified through genome-wide association studies[86] as well as two MGMT promoter SNPs[87,88]. Variant genotypes were retained if exhibiting GQ >= 30 and DP >= 5 (Supplementary Data 34). Potential pathogenic variants were restricted to those in a panel of genes (*ATM, BRCA1, BRCA2, MBD4, MGMT, MLH1, MSH2, MSH6, NTHL1, OGG1, POLE, POT1, TP53*). Candidate pathogenic variants were retained under the following criteria:

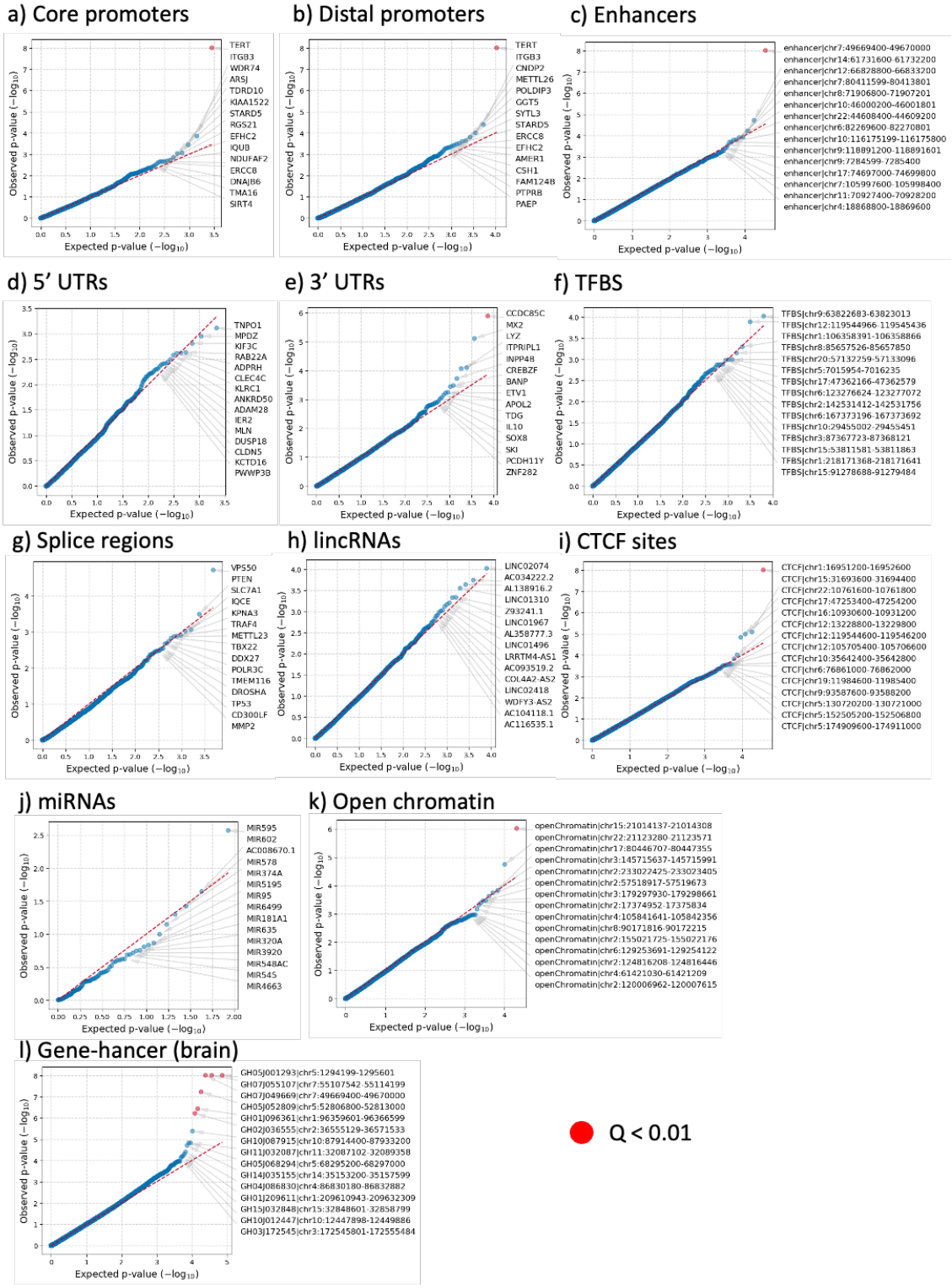
- Passing AggV2 variant filtering
- GnomADg_v3 AF < 0.005, topmedg_AF < 0.005 and GEL_AF < 0.005
- Ensembl VEP v98 consequence mapping to canonical transcript is nonsynonymous
- ClinVar (release 20190219) clinical significance is “Pathogenic” or “Pathogenic/Likely pathogenic”, or for POLE variants the variant alters codon Leu424

SUPPLEMENTARY FIGURES

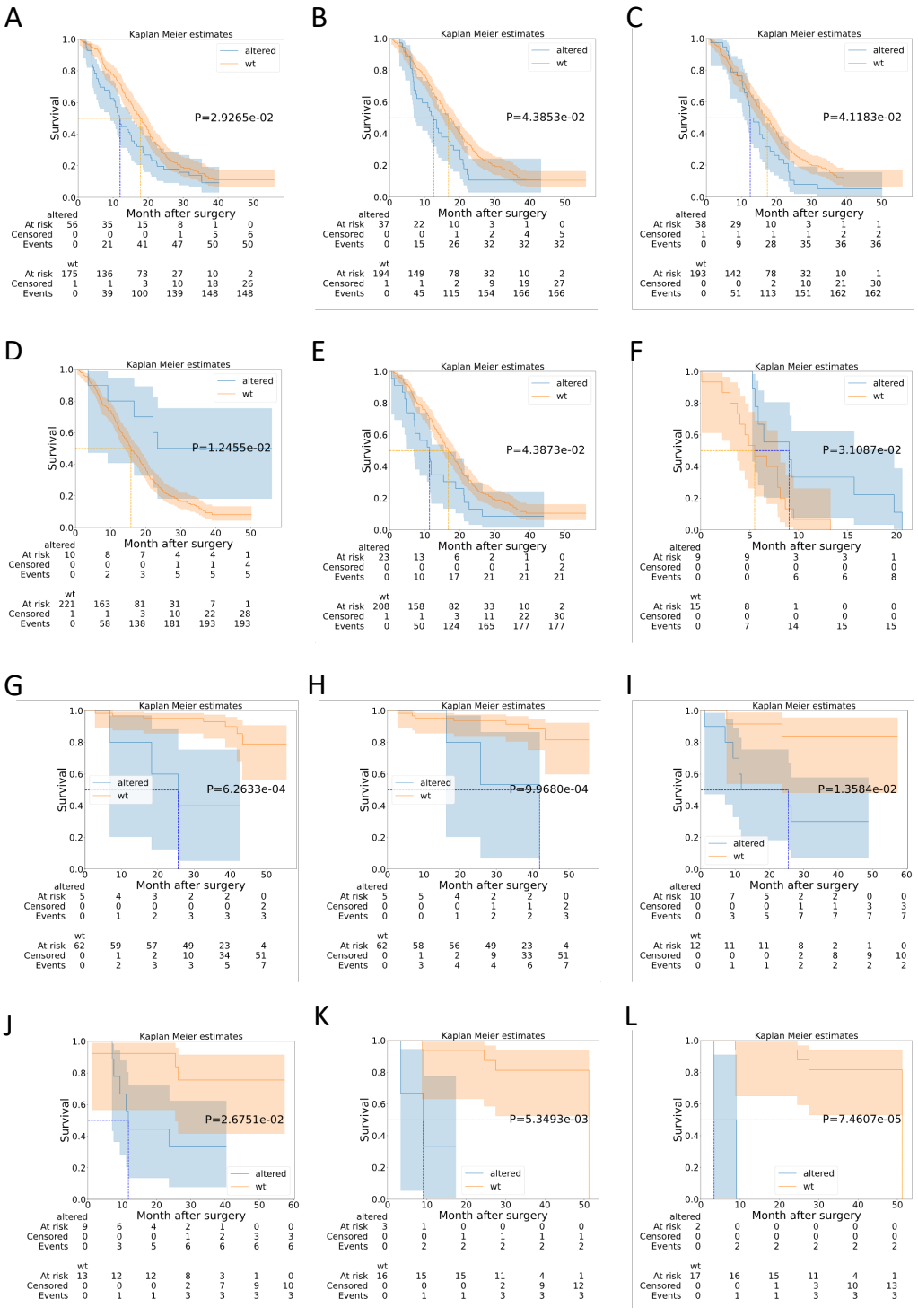
Supplementary Figure 1: Lollipop plots of candidate glioma driver genes. (A) *CDC73*; (B) *COL3A1*; (C) *SALL3*; (D) *SFPQ*; (E) *TENM2*; (F) *TMEM132C*.



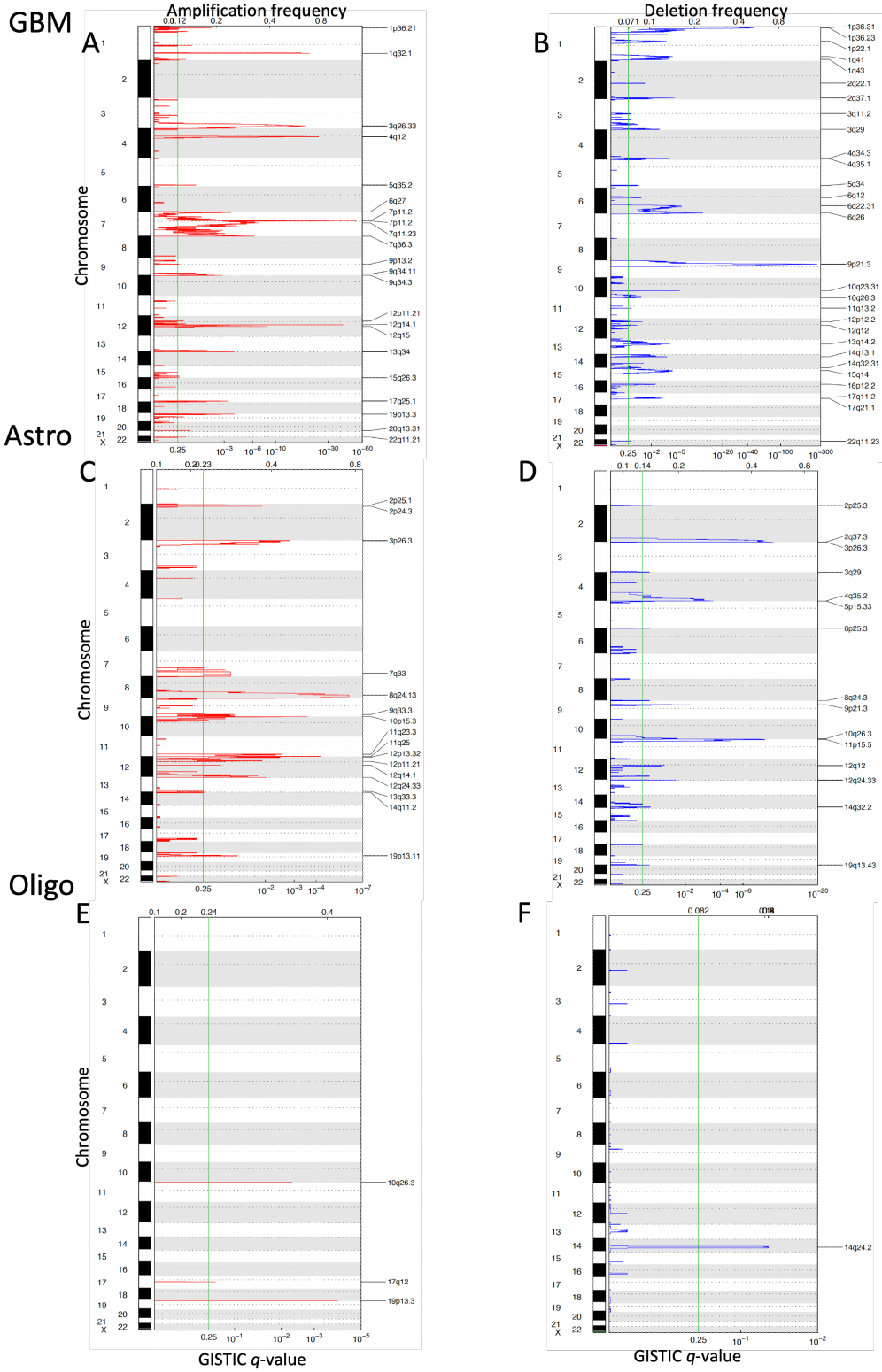
Supplementary Figure 2: Quantile-quantile plots of OncodriveFML analysis of non-coding elements. Red circles indicate non-coding elements with *Q*<0.01. (a) Core promoters; (b) Distal promoters;(c) Enhancers; (d) 5' untranslated regions (UTRs); (e) 3' UTRs; (f) Transcription factor binding sites (TFBS); (g) Splice regions; (h) lincRNAs; (i) CTCF sites; (j) miRNAs; (k) Open chromatin; (l) Gene-hancer elements (brain).



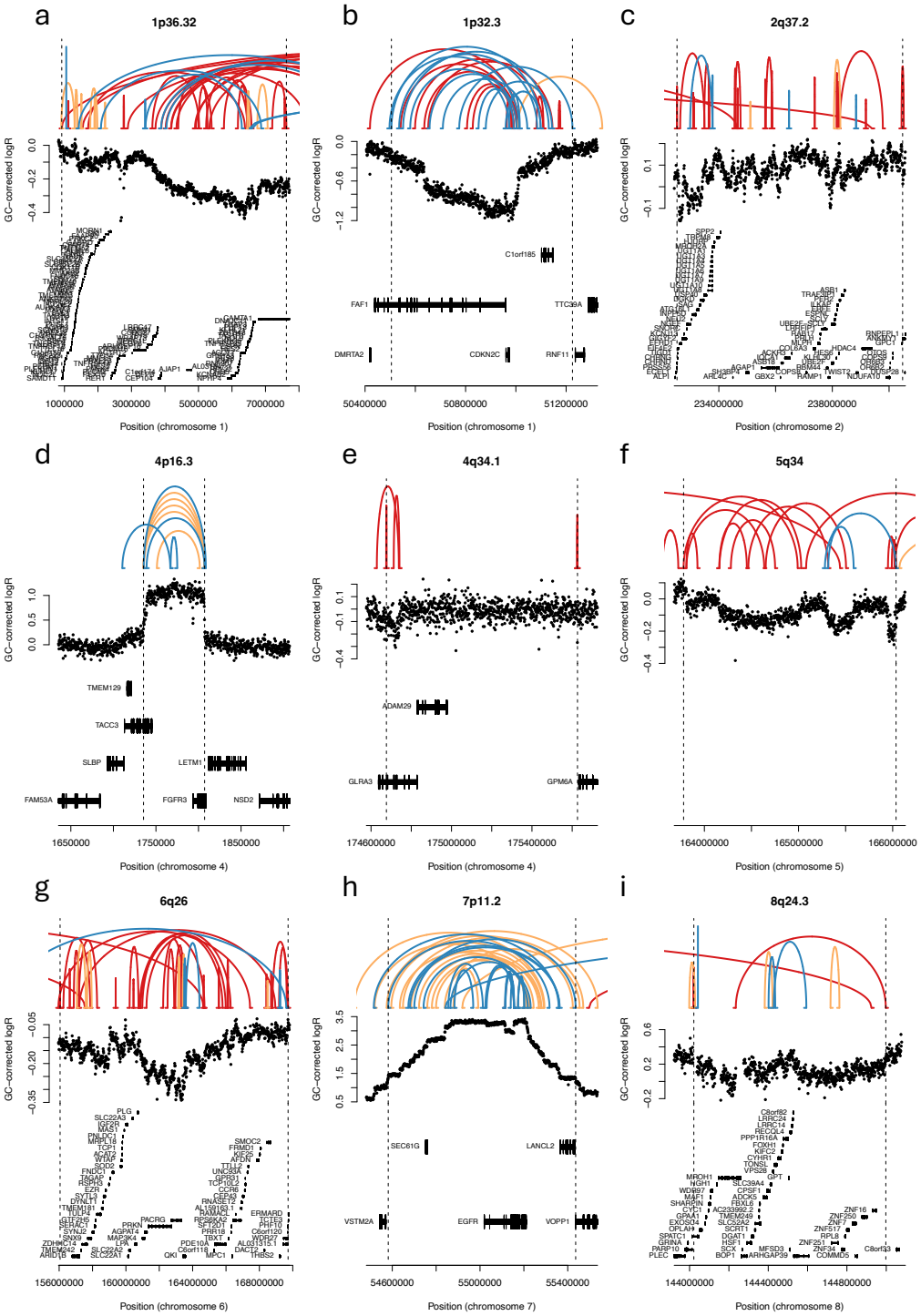
Supplementary Figure 3: Kaplan-meier survival curves. (A-E) Primary GBM *TP53* mutations, *CDKN2A* LOH, 2q37.2 recurrent SV, 8q24.3 recurrent SV and 10q23.2 recurrent SV; (F) Recurrent GBM 1p36.32 recurrent SV; (G-H) Primary astrocytoma *PIK3CA* mutations and rs7837822 variant; (I-J) Recurrent astrocytoma SBS40 activity and IEG LOH; (K-L) Recurrent oligodendroglioma *BCOR* mutation and non-coding chr7:55107542-55114199 mutation. Unadjusted two-sided P-value from logrank association test. Total n=403.

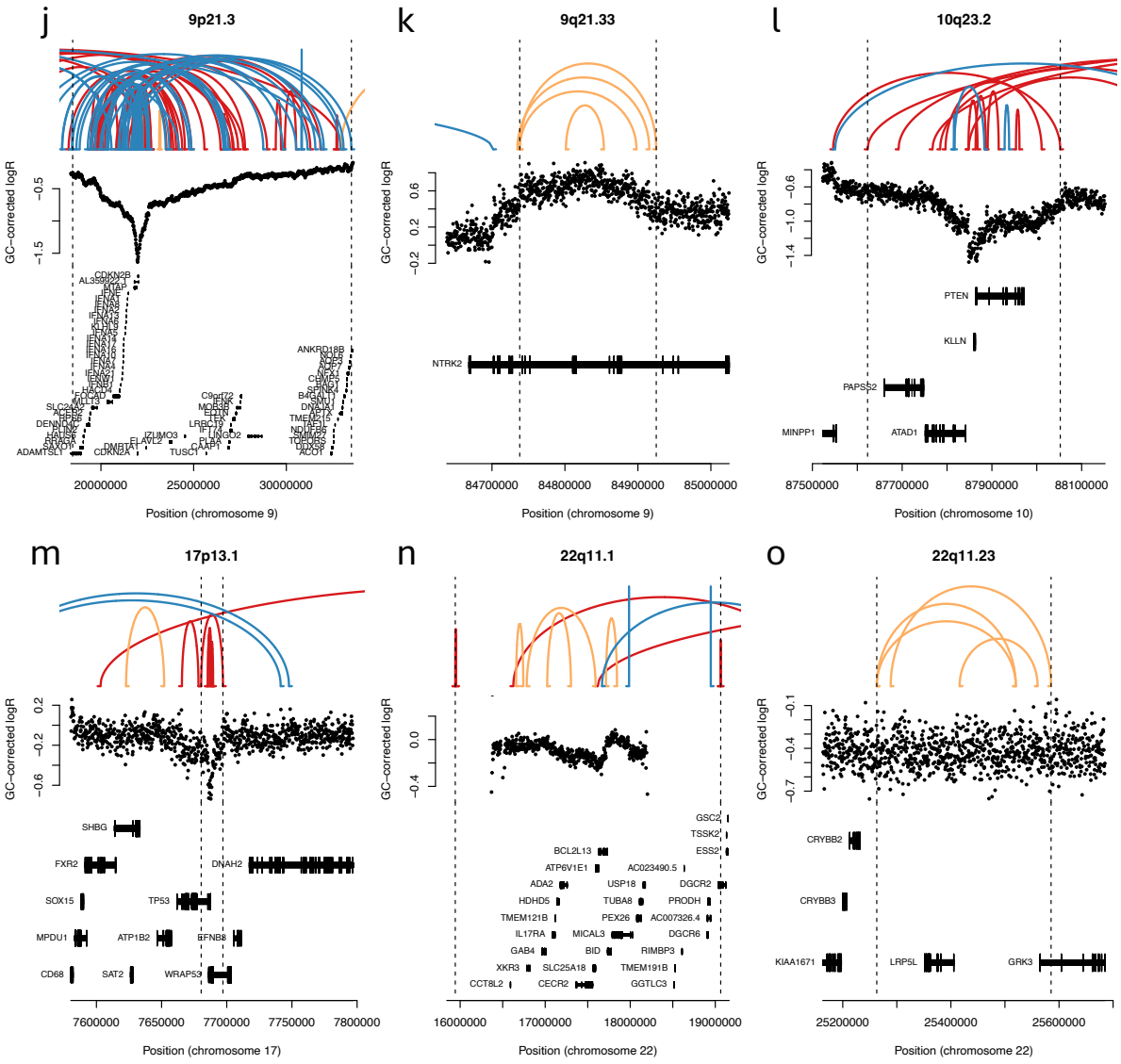


Supplementary Figure 4: Focal amplifications and deletions identified by GISTIC. (A-B) GBM amps and dels; (C-D) Astrocytoma amps and dels; (E-F) Oligodendroglioma amps and dels.

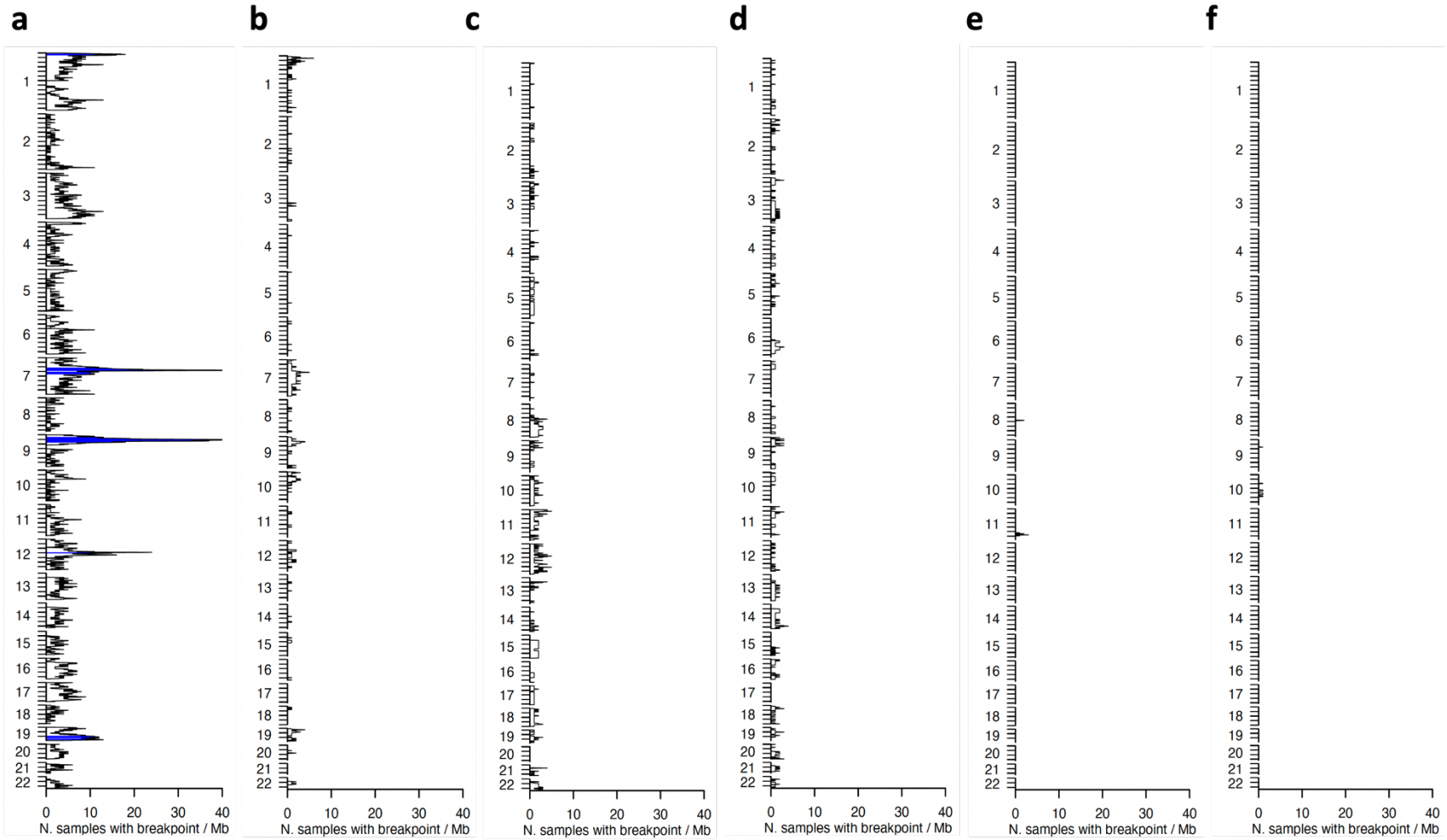


Supplementary Figure 5: Genomic plots of recurrent structural variant hotspots. Shown are genomic locations of breakpoint boundaries for all identified recurrent structural variant (SV) hotspots not in fragile sites (Supplementary Table 16). SV classes are coloured accordingly: red, deletion; blue, unclassified; yellow, tandem duplication. (a) 1p36.22; (b) 1p32.3; (c) 2q37.2; (d) 4p16.3; (e) 4q34.1; (f) 5q34; (g) 6q26; (h) 7p11.2; (i) 8q24.33; (j) 9p21.3; (k) 9q21.33; (l) 10q23.2; (m) 17p13.1; (n) 22q11.1; (o) 22q11.23.

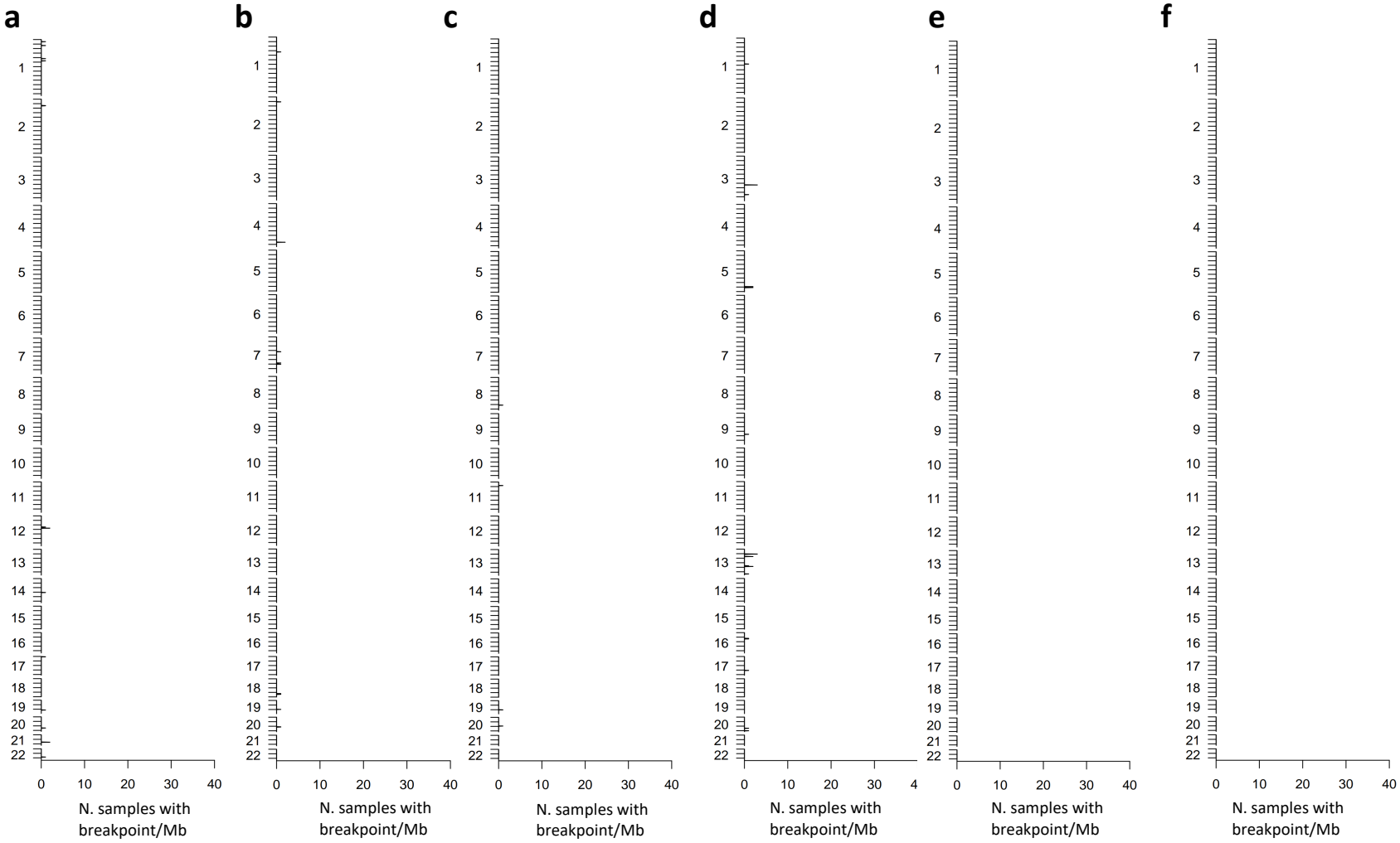



j
9p21.3
GC−corrected logR
Position (chromosome 9)
k
9q21.33
NTRK2
Position (chromosome 9)
l
10q23.2
PTEN
KLLN
PAPSS2
MINPP1
ATAD1
Position (chromosome 10)
m
17p13.1
SHBG
FXR2
DNAH2
SOX15
TP53
MPDU1
ATP1B2
EFNB3
CD68
SAT2
WRAP53
Position (chromosome 17)
n
22q11.1
GSC2
TSSK2
ESS2
BCL2L13
ATP6V1E1
AC023490.5
ADA2
USP18
DGCR2
HDHD5
TUBA8
PRODH
TMEM121B
PEX26
AC007326.4
IL17RA
MICAL3
DGCR6
GAB4
BID
RIMBP3
XKR3
SLC25A18
TMEM191B
CCT8L2
CECR2
GGTLC3
Position (chromosome 22)
o
22q11.23
CRYBB2
CRYBB3
KIAA1671
LRP5L
GRK3
Position (chromosome 22)

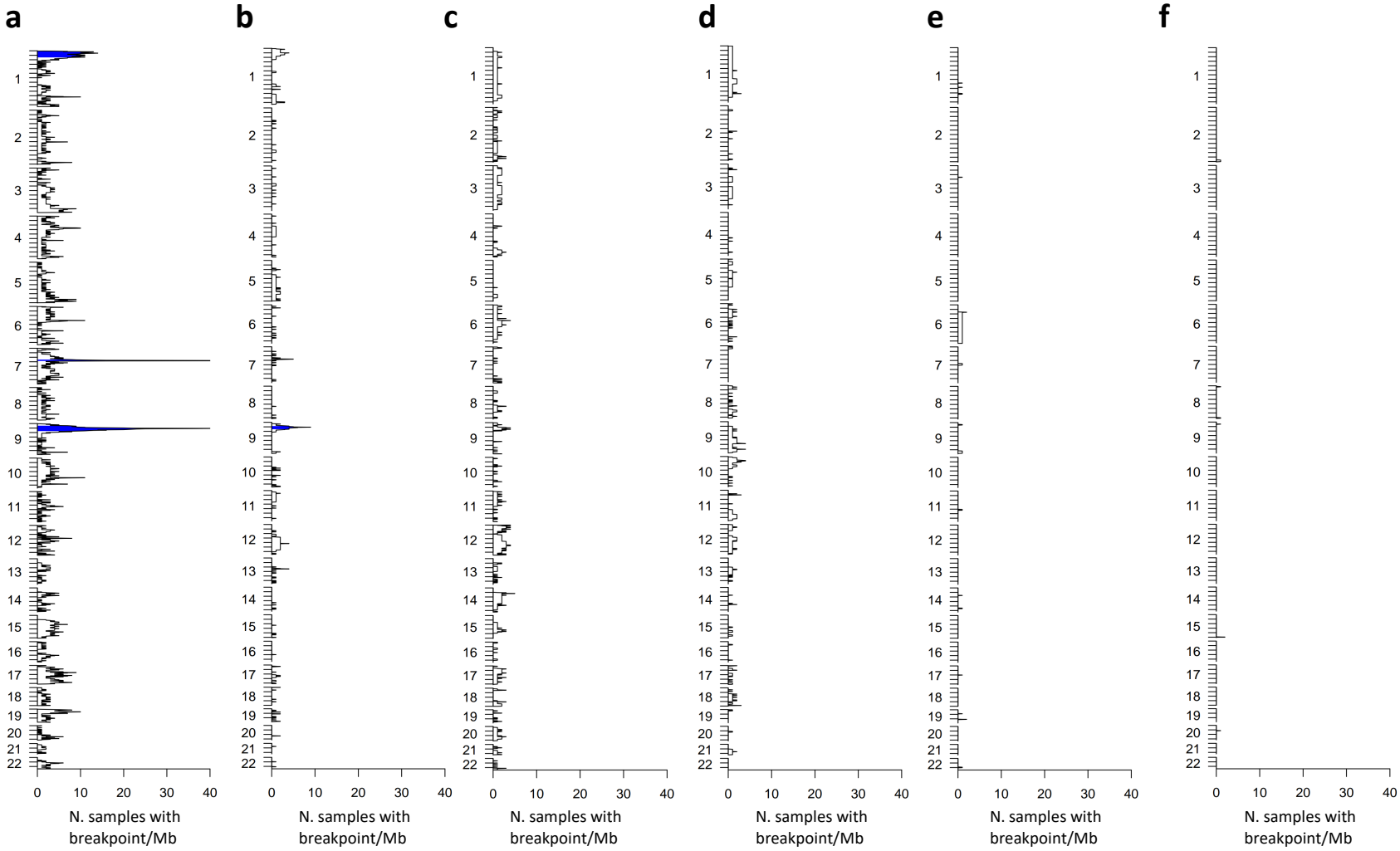
Supplementary Figure 6: Frequency of chromothripsis events across glioma subtypes. Regions enriched for chromothripsis at a 5% FDR greater than 5Mb are coloured blue. (a) primary GBM; (b) recurrent GBM; (c) primary astrocytoma; (d) recurrent astrocytoma; (e) primary oligodendroglioma; (f) recurrent oligodendroglioma.



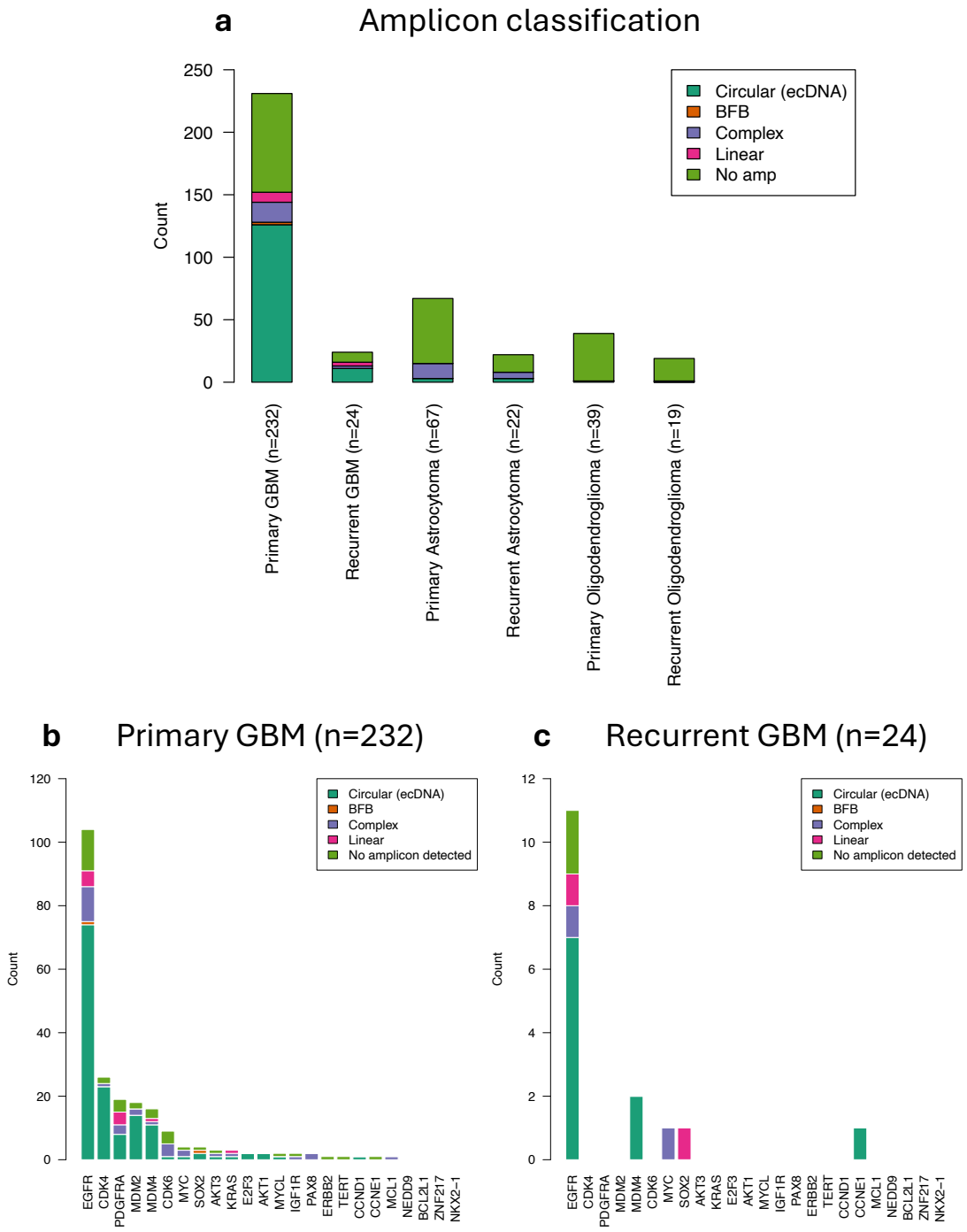
Supplementary Figure 7: Frequency of chromoplexy events across glioma subtypes. Regions enriched for chromoplexy at a 5% FDR greater than 5Mb are coloured blue. (a) primary GBM; (b) recurrent GBM; (c) primary astrocytoma; (d) recurrent astrocytoma; (e) primary oligodendroglioma; (f) recurrent oligodendroglioma.



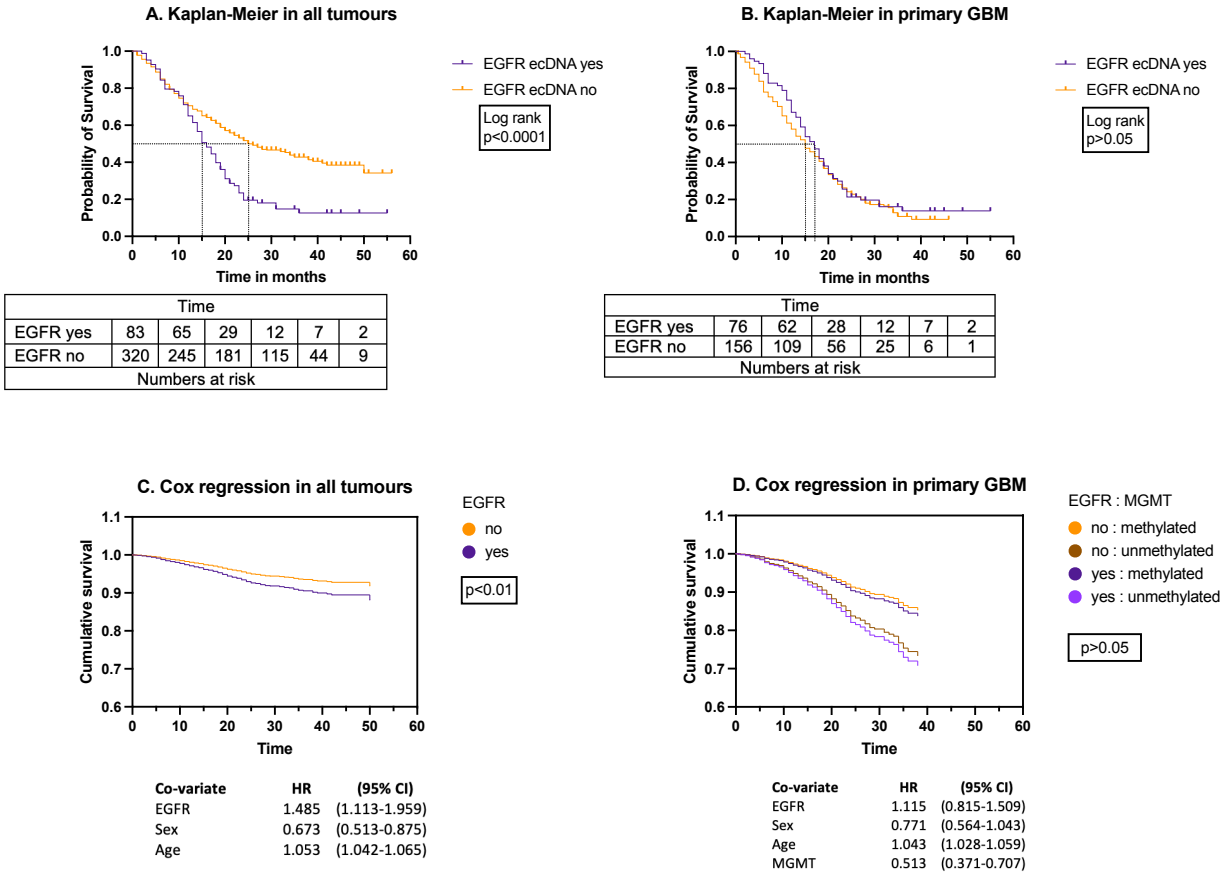
Supplementary Figure 8: Frequency of complex unclassified structural variant events across glioma subtypes. Regions enriched for events at a 5% FDR greater than 5Mb are coloured blue. (a) primary GBM; (b) recurrent GBM; (c) primary astrocytoma; (d) recurrent astrocytoma; (e) primary oligodendroglioma; (f) recurrent oligodendroglioma.



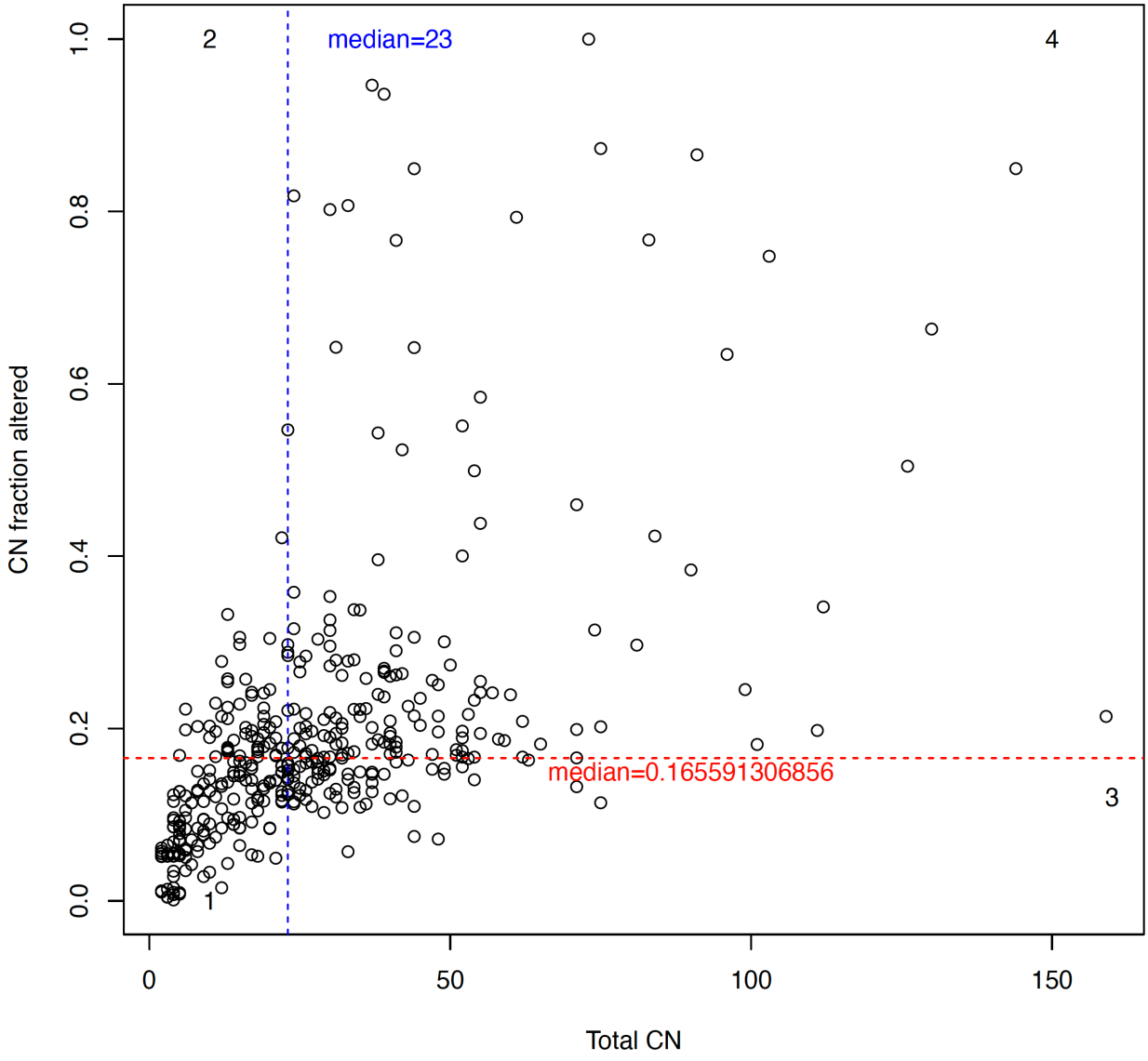
Supplementary Figure 9: Detection of extrachromosomal DNA (ecDNA) in glioma subtypes. (a) Amplicon classification of tumours carrying ≥ 1 amplicon across tumour subtypes (*e.g.* tumour counted as "Circular (ecDNA)" if ≥ 1 circularised amplicon detected, otherwise "BFB" if ≥ 1 BFB amplicon detected until "No amp" where no valid amplicon detected). (b-c) Classification of amplicons containing commonly amplified oncogenes in primary and recurrent GBM tumours respectively. Classification restricted to gene amplifications with a total copy number ≥ 5 in diploid tumours and ≥ 10 in tetraploid tumours (*i.e.* "big gains"). Counts based on data from Supplementary Data 20. Total n=403.



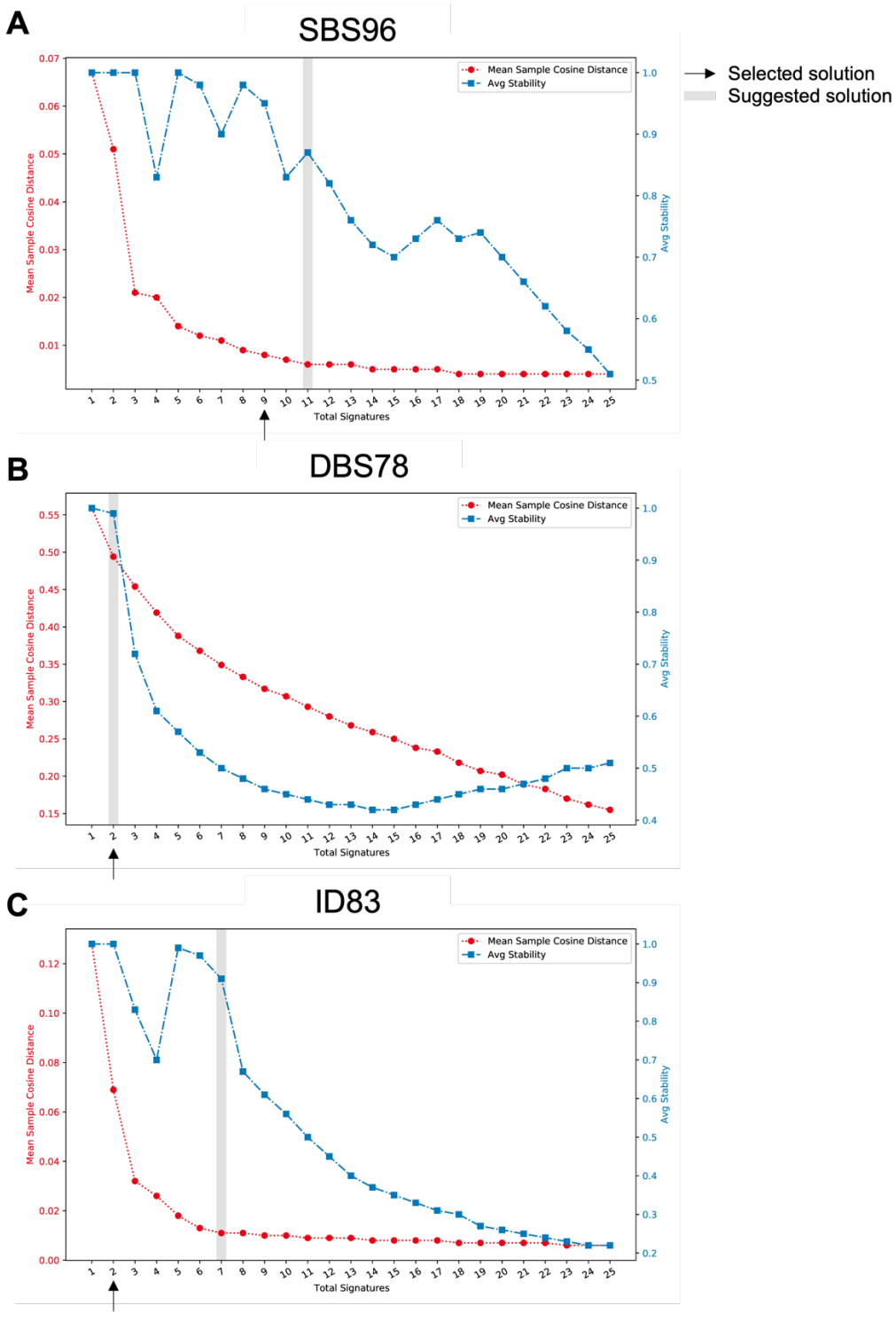
Supplementary Figure 10: Association between survival and extrachromosomal DNA (ecDNA)-mediated EGFR amplification in diffuse glioma. (A-B) Kaplan-Meier plots of overall survival (OS) in all tumours (A) and primary GBM (B). The table below depicts the number of patients at risk at each individual time point. (A) There is a statistically significant difference in mOS between tumours where ecDNA-mediated EGFR amplification is present (purple) compared to tumours where it is absent (orange) in all tumours (log rank test (LRT) *P*<0.0001). (B) There is no statistically significant difference in OS (LRT *P=0.3932*) when comparing the presence of ecDNA-mediated EGFR amplification in primary GBM; (C-D) Cox logistic regression (adjusted for age and gender in all tumours, and additionally MGMT methylation status in GBM) analysis comparing predicted survival in patients with ecDNA-mediated EGFR amplification in all tumours (C) and in only primary GBM (D). The table below each graph details the hazards ratio (HR) and associated confidence intervals (CIs) for each co-variate. (C) There was a statistically significant difference in predicted survival in all tumours comparing ecDNA-mediated EGFR amplification (purple) to no EGFR amplification (*P=0.0061*). (D) Cox logistic regression performed on primary glioblastoma shows no significant difference (*P=0.4891*). There is a strong association of predicted survival with MGMT methylation status (independent of ecDNA-mediated EGFR amplification, HR 0.51, CI 0.37-0.71). *P*-values are unadjusted and two-sided. Total n=403.



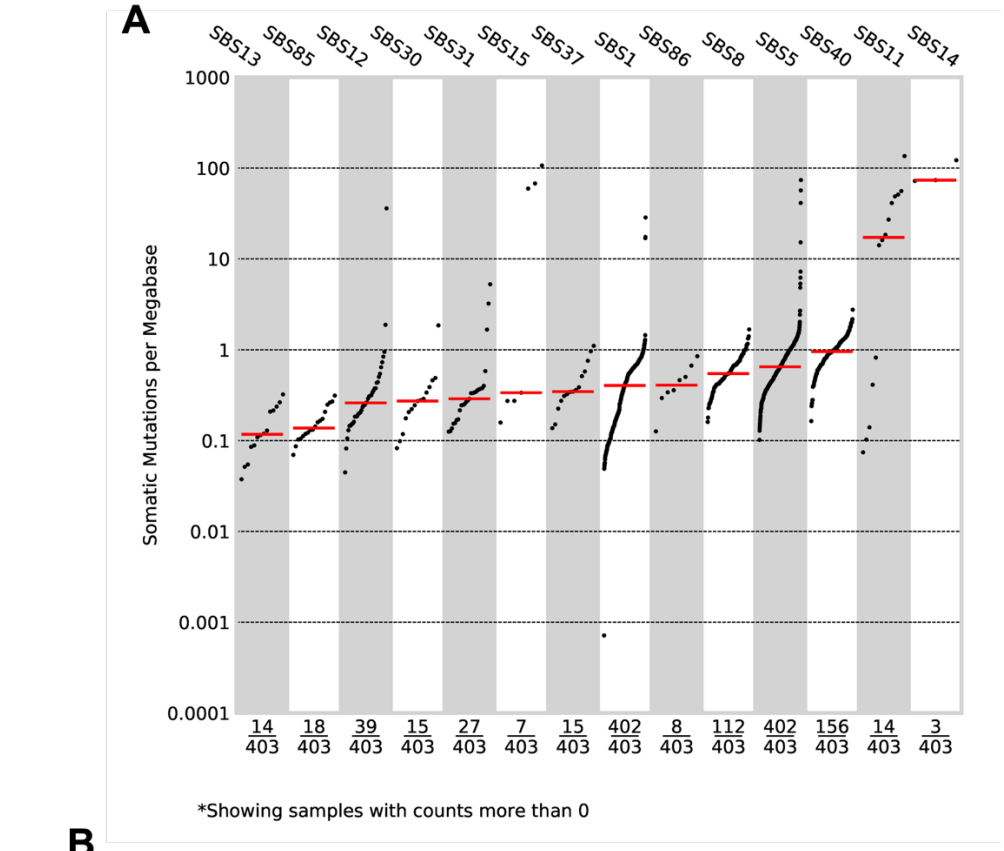
Supplementary Figure 11: Categorizing glioma tumours based on copy number profiles. Distributions of the genomic fraction affected by a copy number alteration (CNA) were compared against the total number of CNAs. Four groups were assigned: 1 – genomic fraction altered by CNA and total CNA number less than the cohort median – representing "Copy number quiet" tumours; 2 – genomic fraction altered by CNA greater than cohort median and total CNA number less than cohort median – representing "Broad copy number" tumours; 3 – genomic fraction altered by CNA less than cohort median and total CNA number greater than cohort median – representing "Focal copy number" tumours; 4 - genomic fraction altered by CNA and total CNA number greater than the cohort median - representing "Copy number disrupted" tumours. Total n=403.

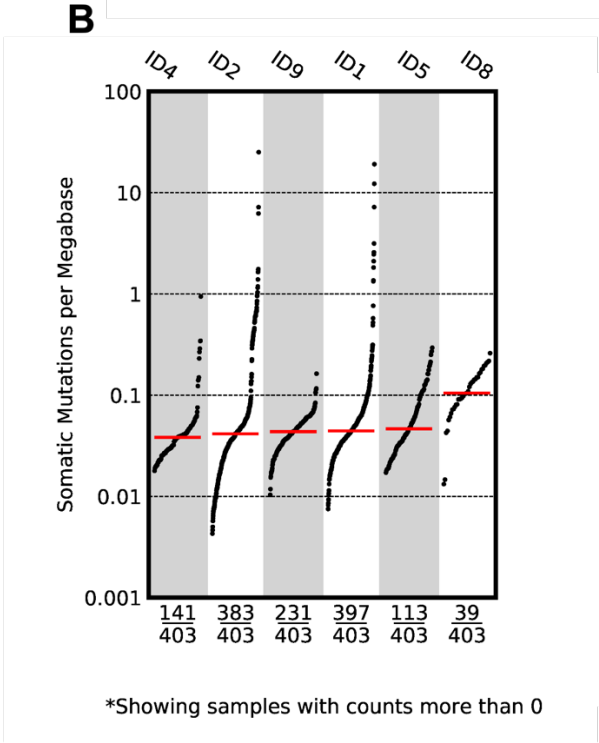


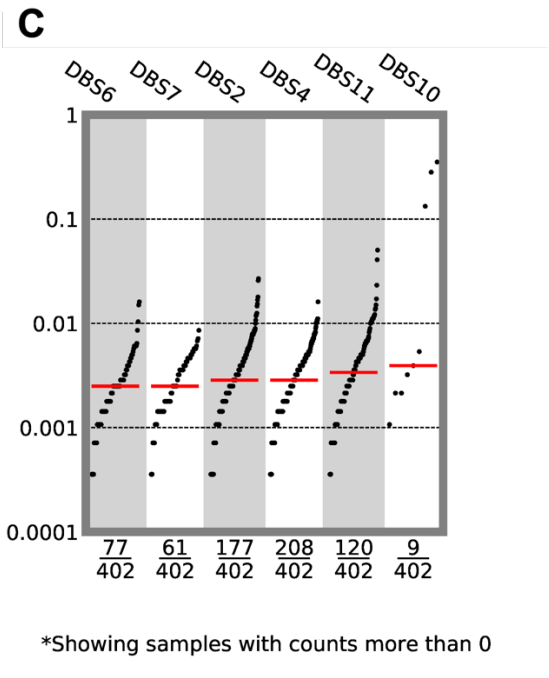
Supplementary Figure 12: Selection of SBS96, DBS78 and ID83 *de novo* signature solutions. (A) SBS96; (B) DBS78; (C) ID83.



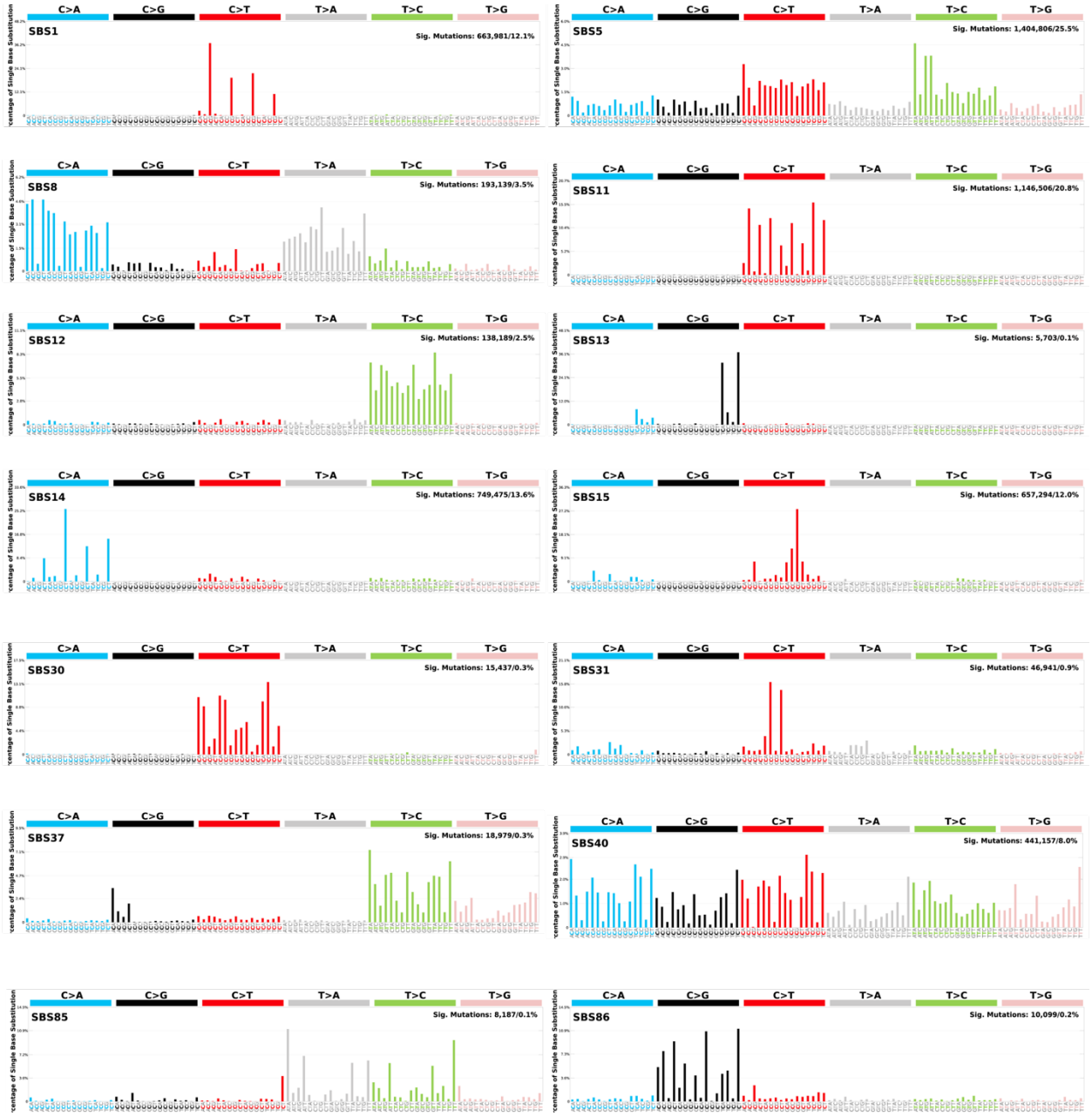
Supplementary Figure 13: SBS96, ID83 and DBS78 signatures by tumour mutational burden (TMB). (A) SBS96; (B) ID83; (C) DBS78.



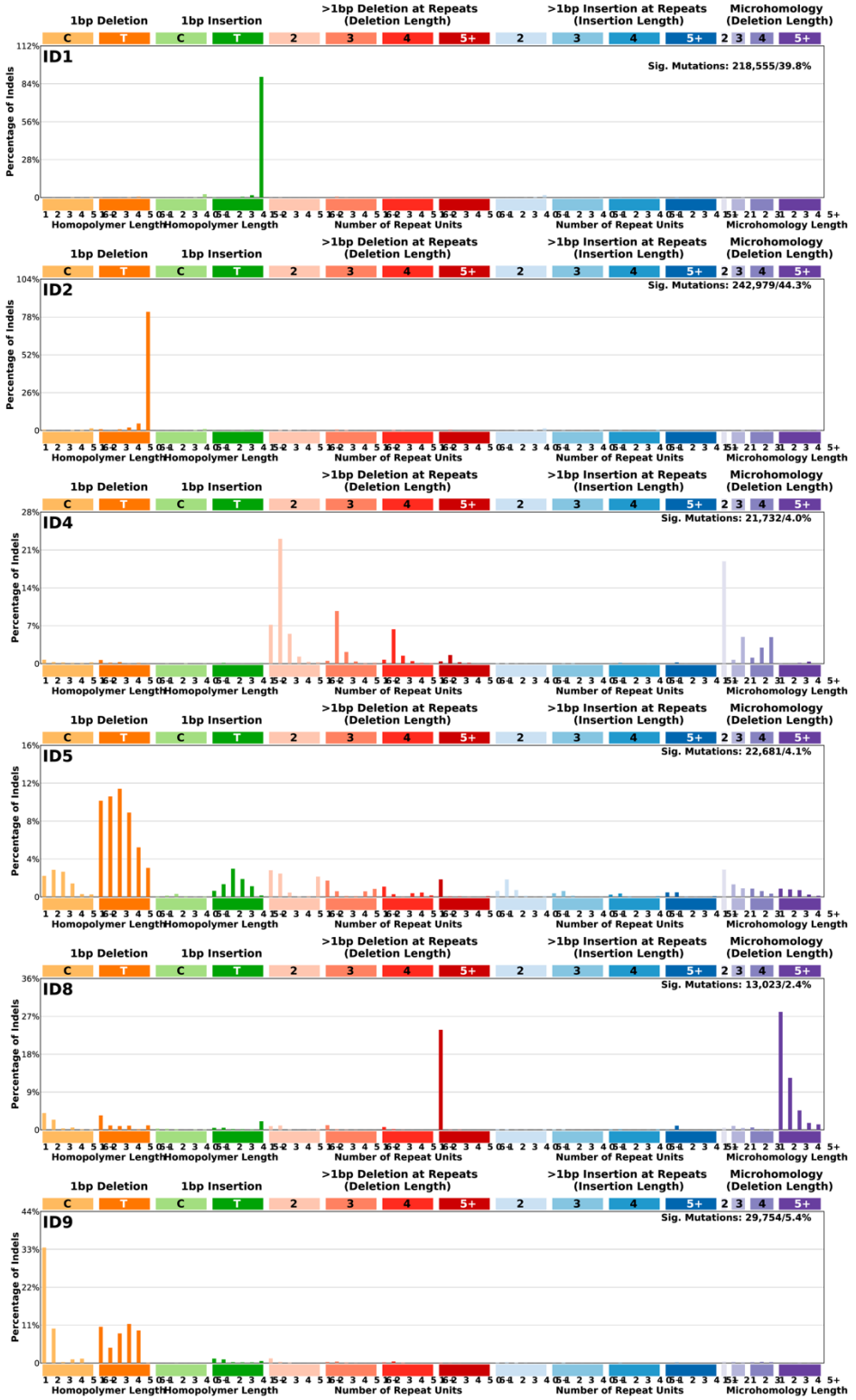




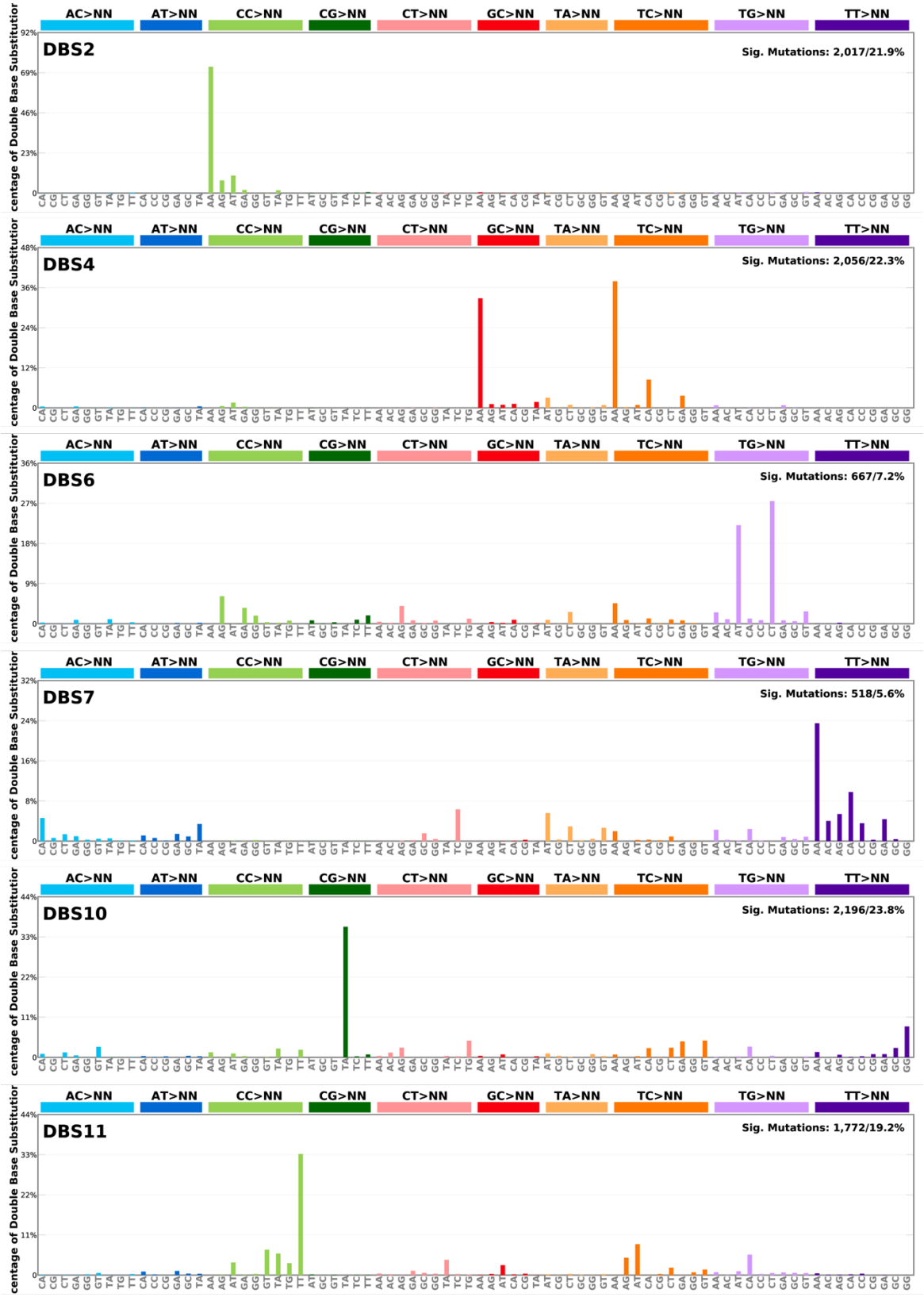
Supplementary Figure 14: SBS96 signature probabilities by mutational context.



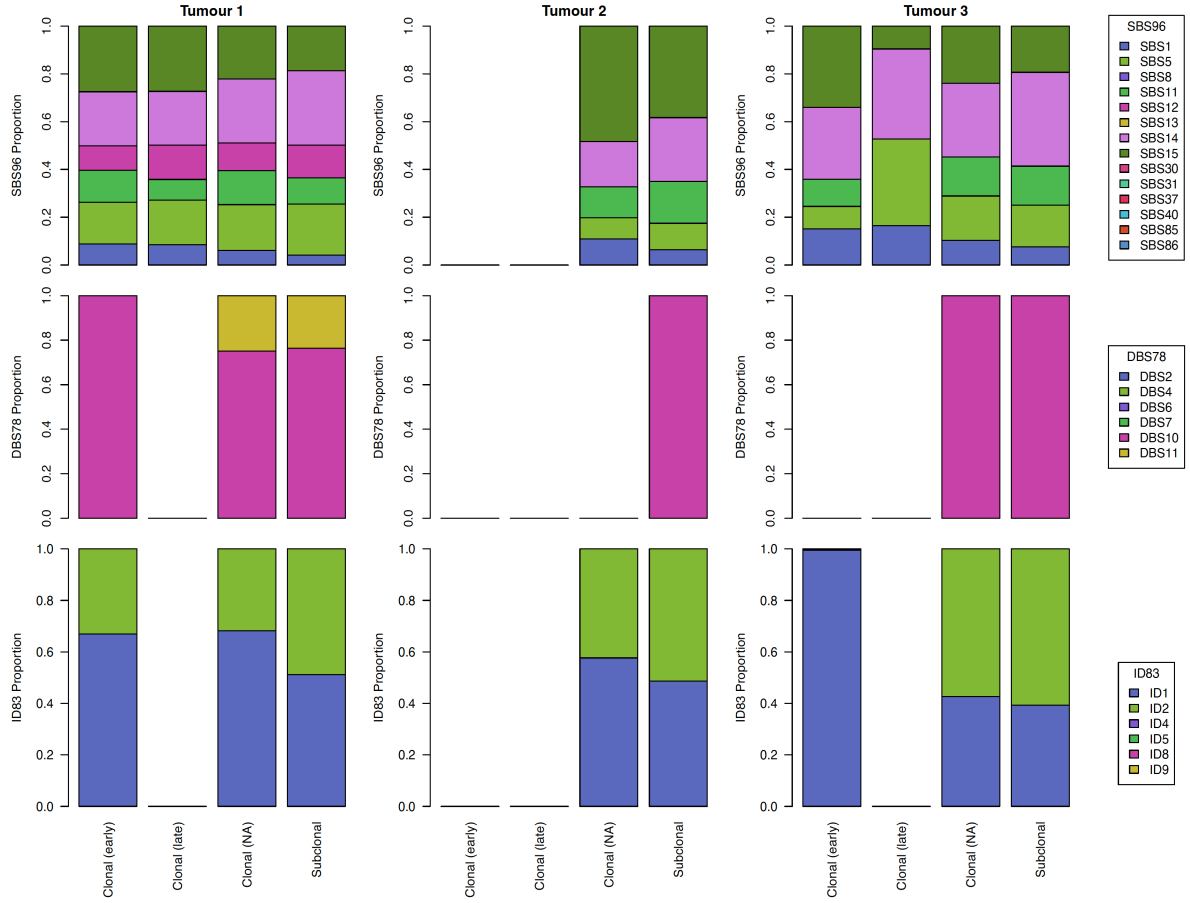
Supplementary Figure 15: ID83 signature probabilities by mutational context.



Supplementary Figure 16: DBS78 signature probabilities by mutational context.

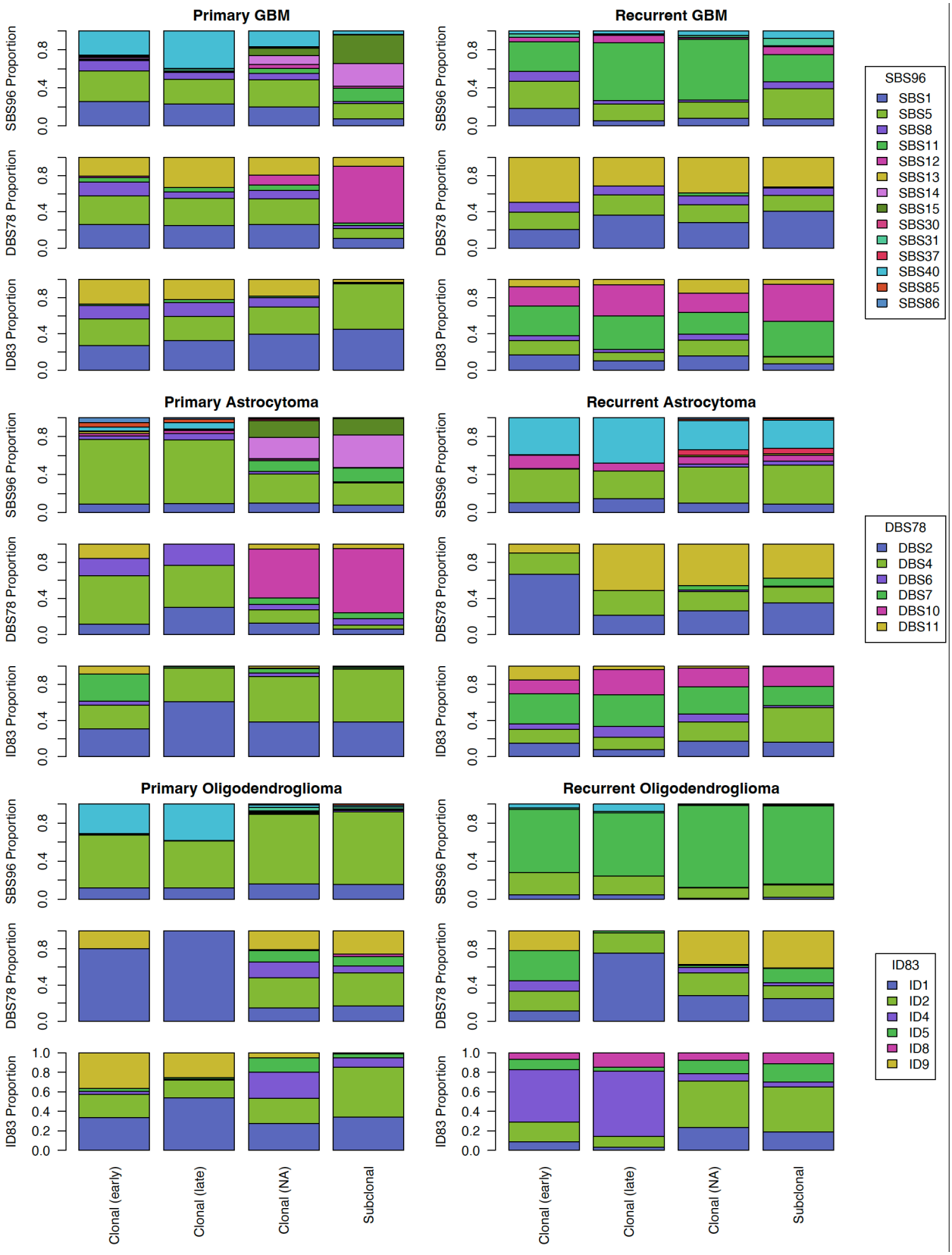


Supplementary Figure 17: Timing of mutational signatures across three hypermutated glioma tumours.

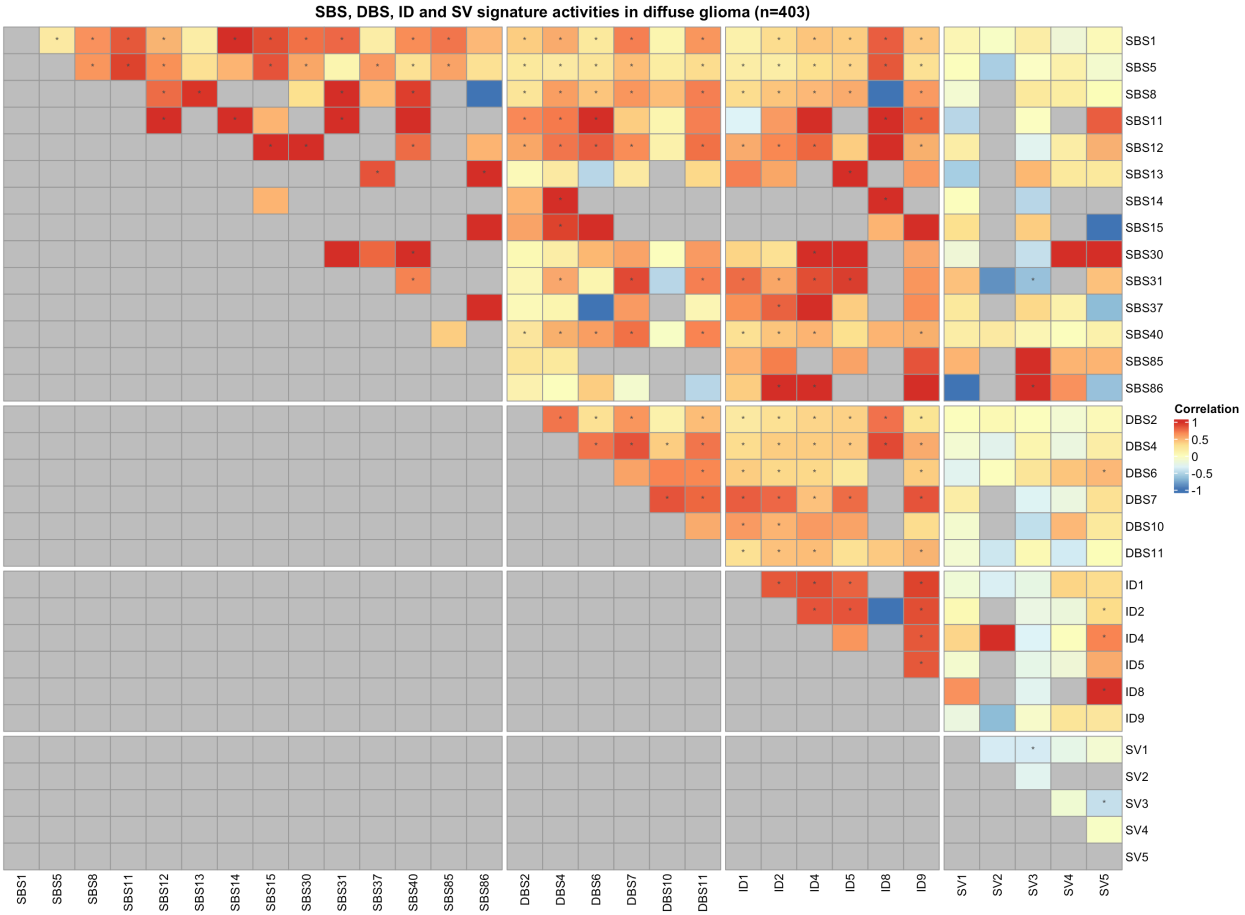


	Tumour 1				Tumour 2				Tumour 3			
	Clonal (early)	Clonal (late)	Clonal (NA)	Sub clonal	Clonal (early)	Clonal (late)	Clonal (NA)	Sub clonal	Clonal (early)	Clonal (late)	Clonal (NA)	Sub clonal
SBS96	3489	23	591939	191874	0	0	389	788313	28	3	295810	684124
DBS78	1	0	376	132	0	0	0	778	0	0	333	651
ID83	142	0	38017	13497	1	0	39	39157	1	0	35527	88275

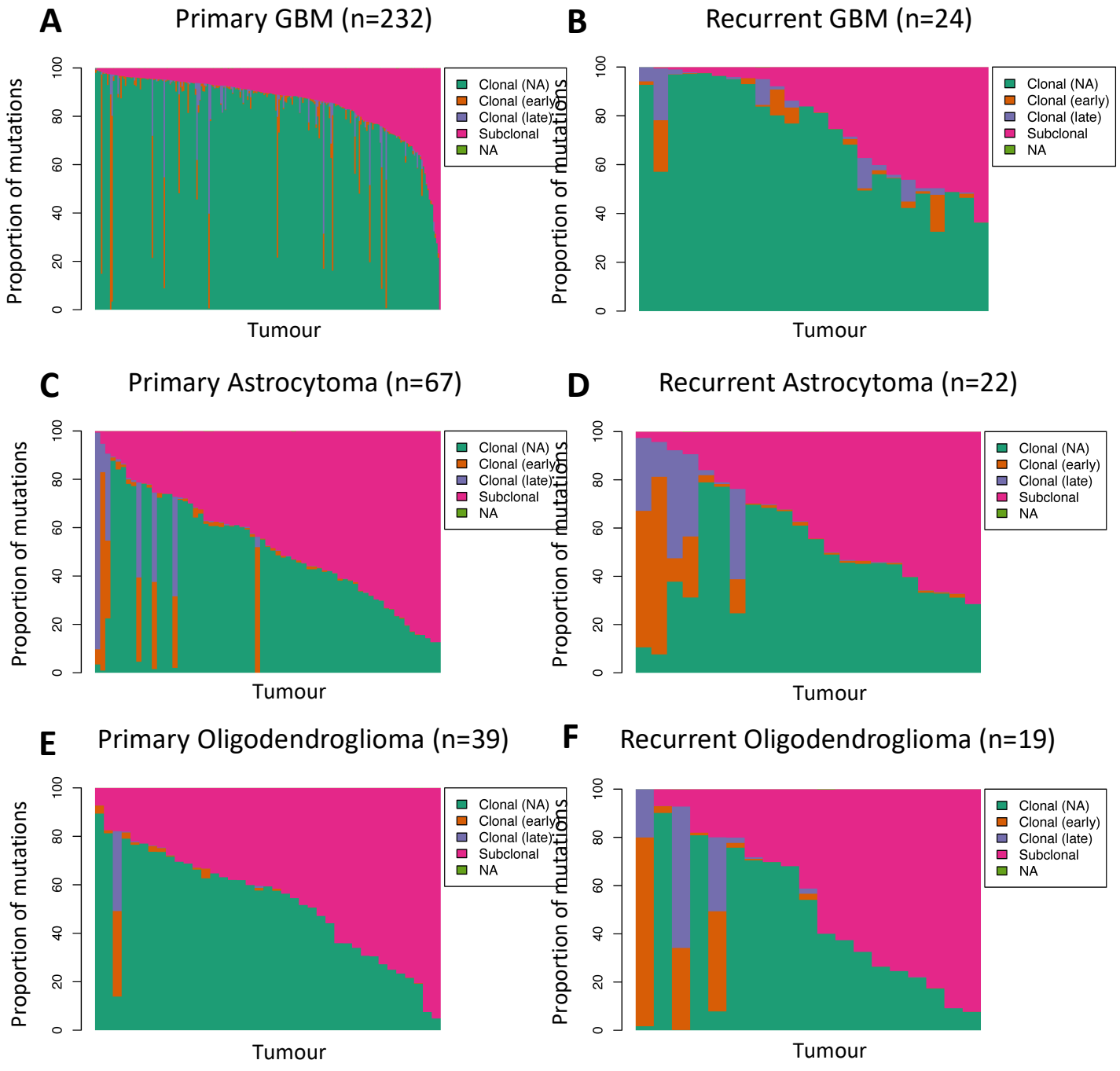
Supplementary Figure 18: Timing of mutational signatures across glioma subtypes.



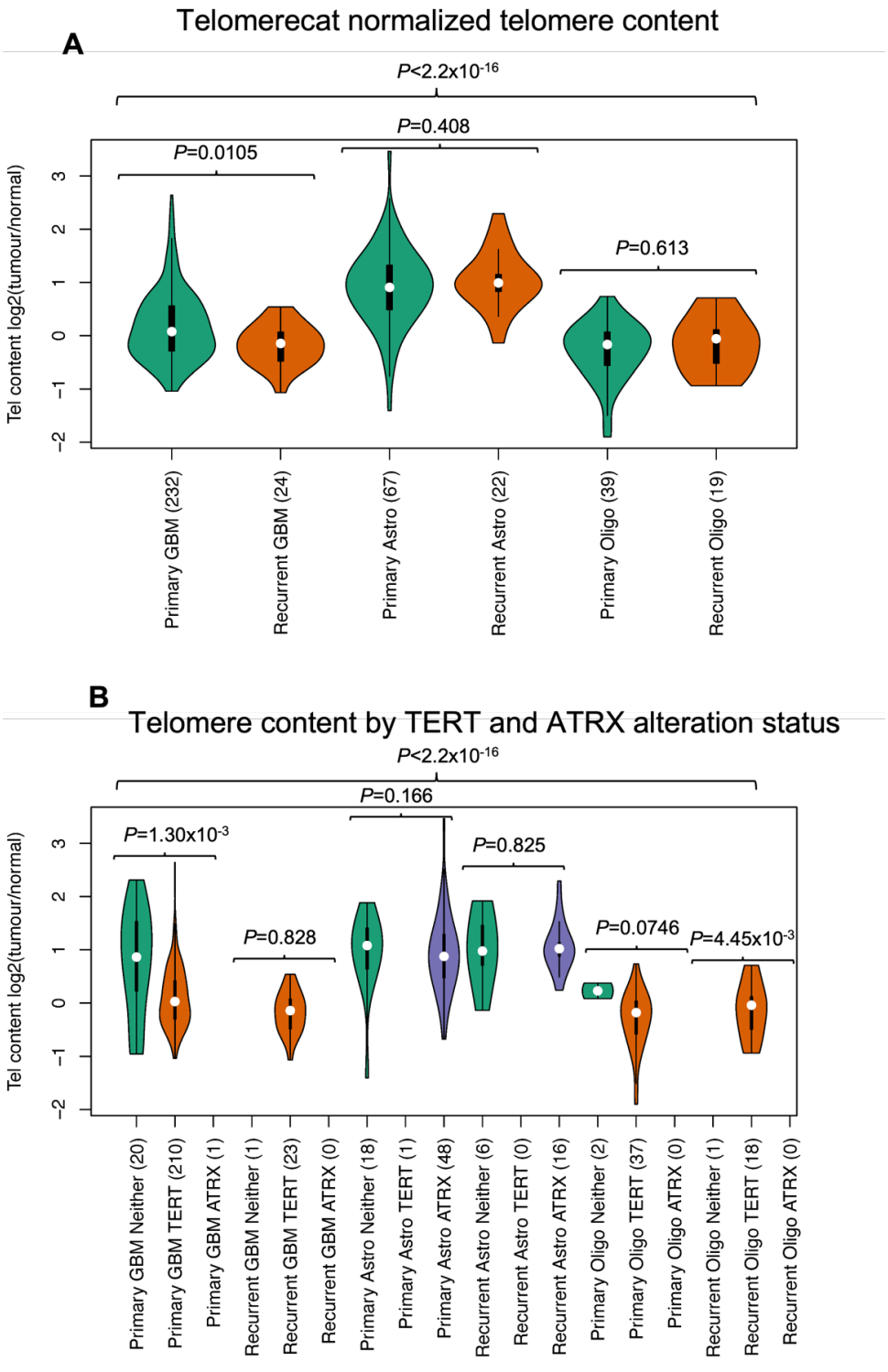
Supplementary Figure 19: SBS, DBS, ID and SV signature activity correlations in diffuse glioma. Pairwise correlations in signature activities assessed using spearman's test (Supplementary Data 24). * indicates correlations with $P < 0.05$. *P*-values are unadjusted and two-sided. Total n=403.



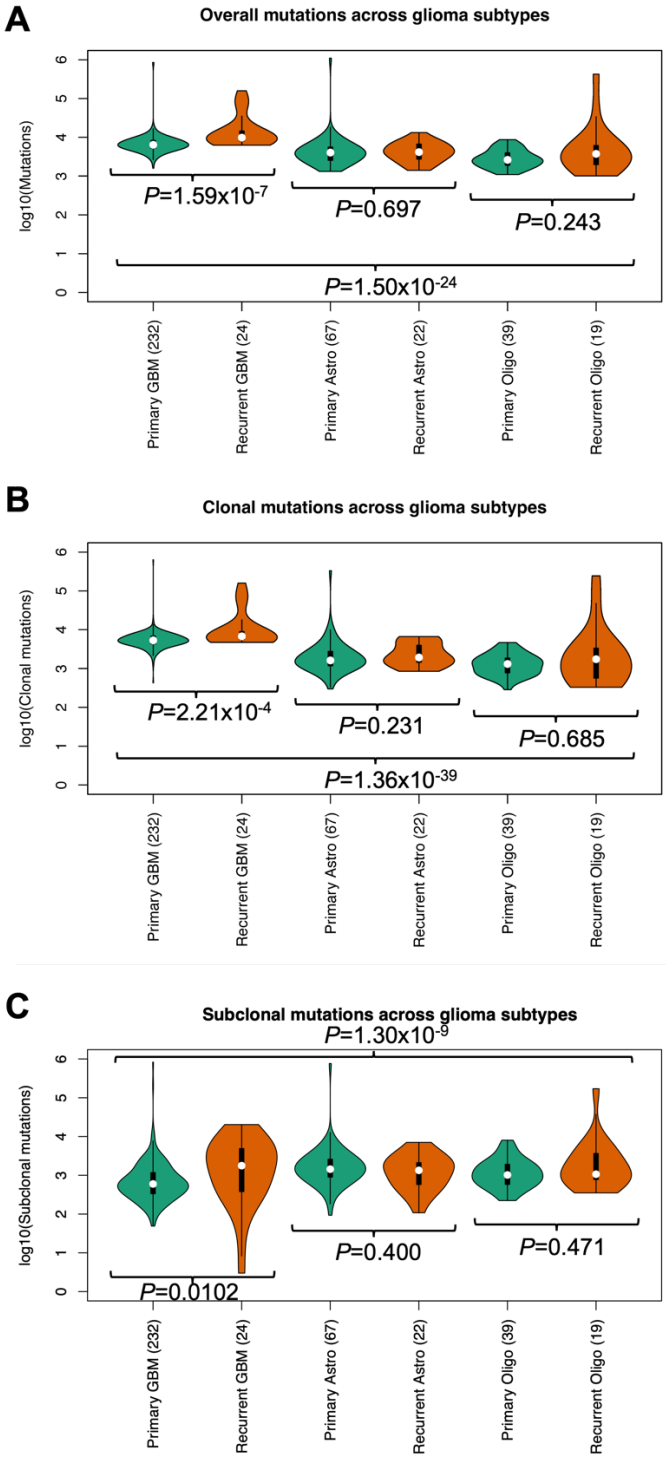
Supplementary Figure 20: SNV and indel mutation clonality across glioma subtypes. (A) Primary GBM; (B) recurrent GBM; (C) primary astrocytoma; (D) recurrent astrocytoma; (E) primary oligodendroglioma; (F) recurrent oligodendroglioma. Total n=403.



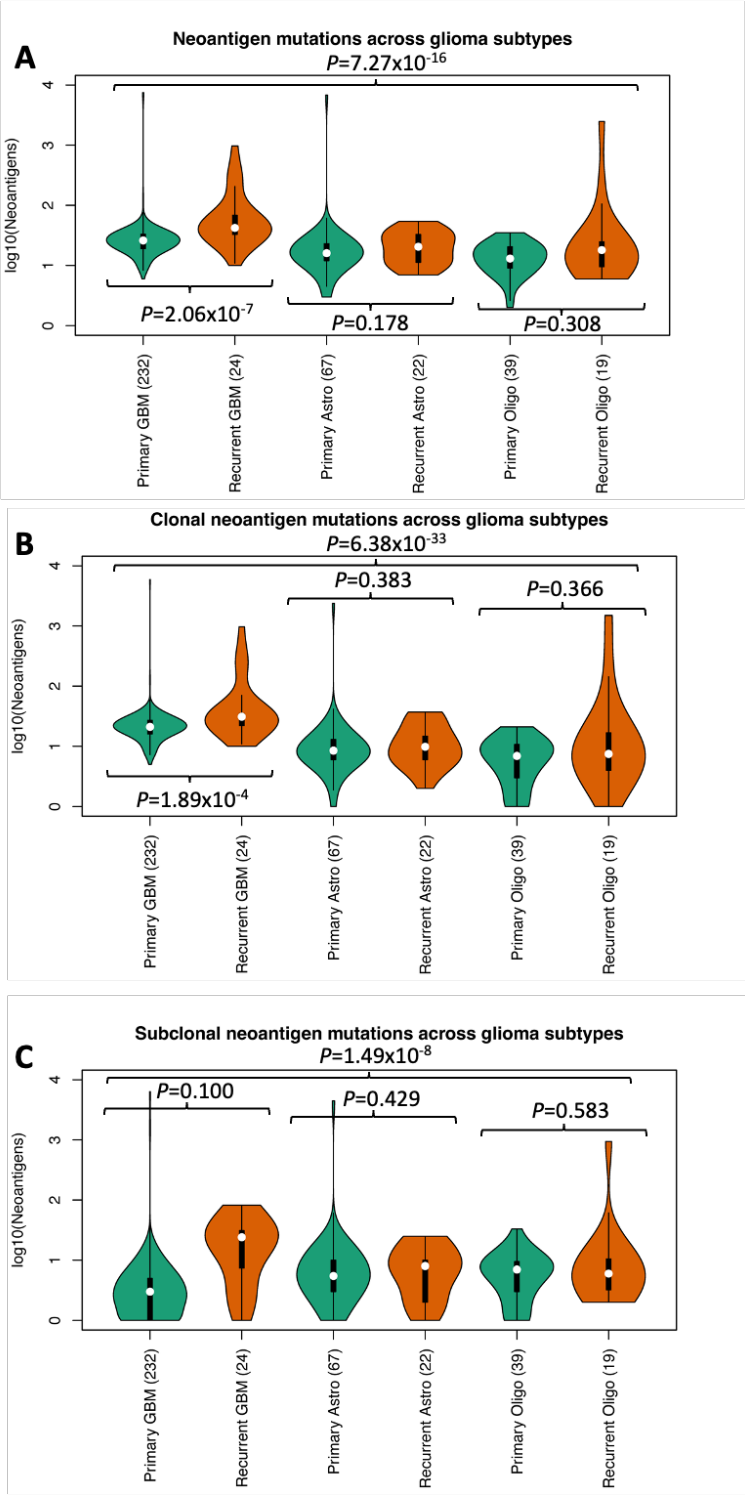
Supplementary Figure 21: Telomere content by glioma subtype and telomere maintenance mechanism. *P*-values are unadjusted and two sided and estimated using the Kruskal-Wallis test. In the box plots, the center line represents the median, and the box bounds represents the inter-quartile range. (A) Normalized telomere content estimated by telomerecat; (B) Telomere content by TERT and ATRX alteration status. Total n=403.



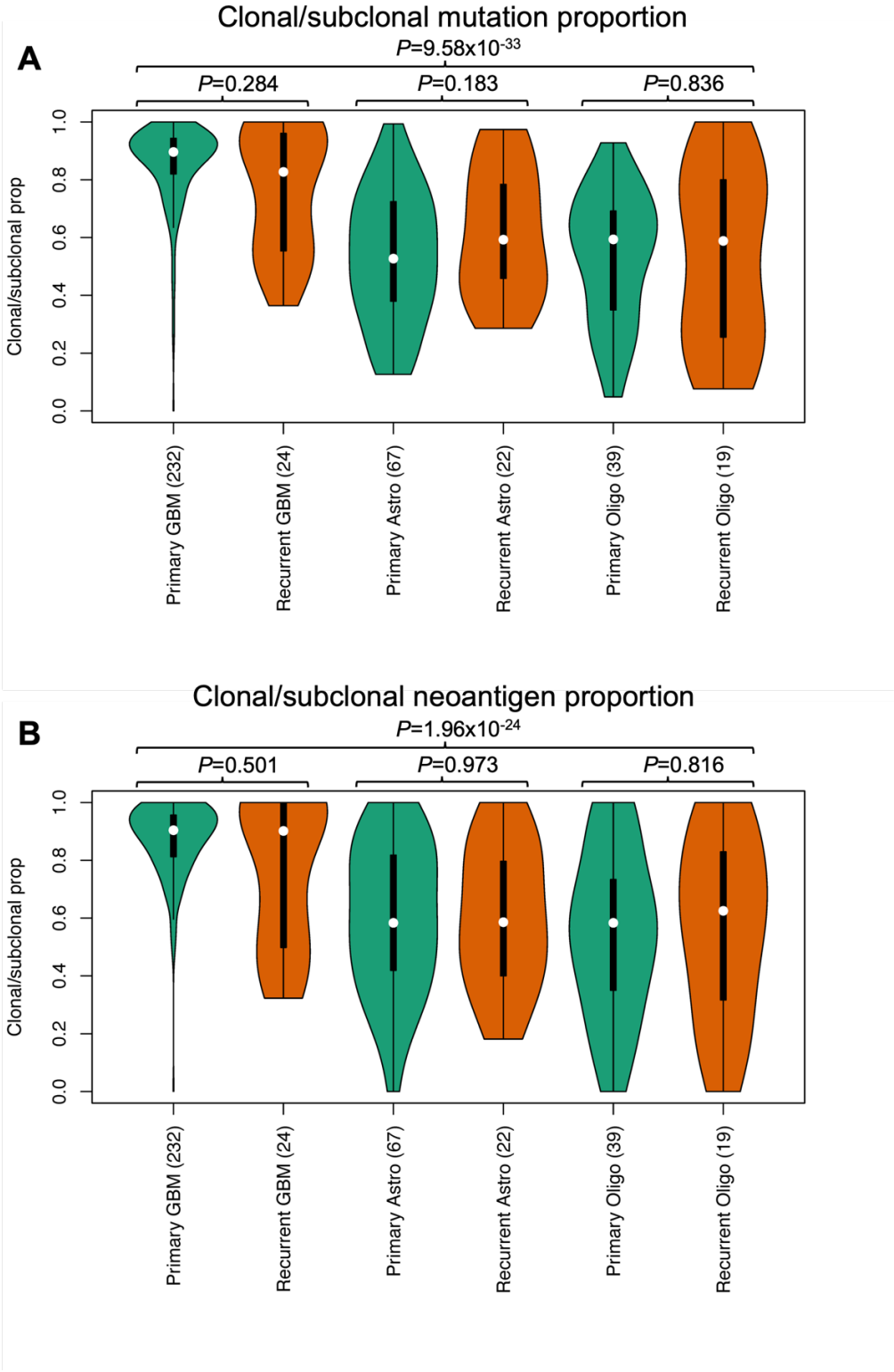
Supplementary Figure 22: Clonal and subclonal mutation counts by glioma subtype. *P*-values are unadjusted and two-sided and estimated using the Kruskal-Wallis test between mutation count estimates. In the box plots, the center line represents the median, and the box bounds represents the inter-quartile range. (A) Total mutations; (B) Clonal mutations; (C) Subclonal mutations. Total n=403.



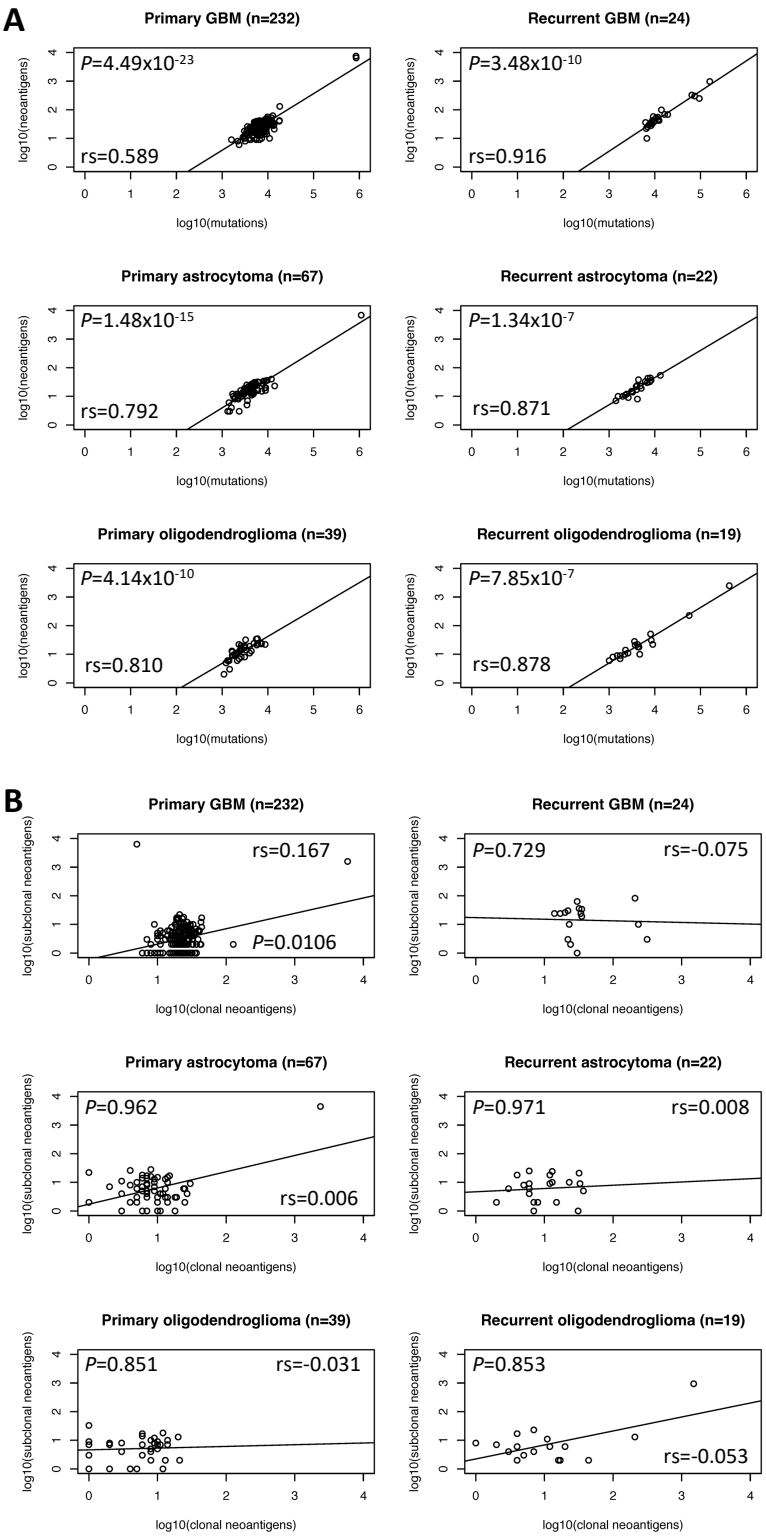
Supplementary Figure 23: Clonal and subclonal neoantigen counts by glioma subtype. *P*-values are unadjusted and two-sided and estimated using the Kruskal-Wallis test between neoantigen count estimates. In the box plots, the center line represents the median, and the box bounds represents the inter-quartile range. (A) Total neoantigens; (B) Clonal neoantigens; (C) Subclonal neoantigens. Total n=403.



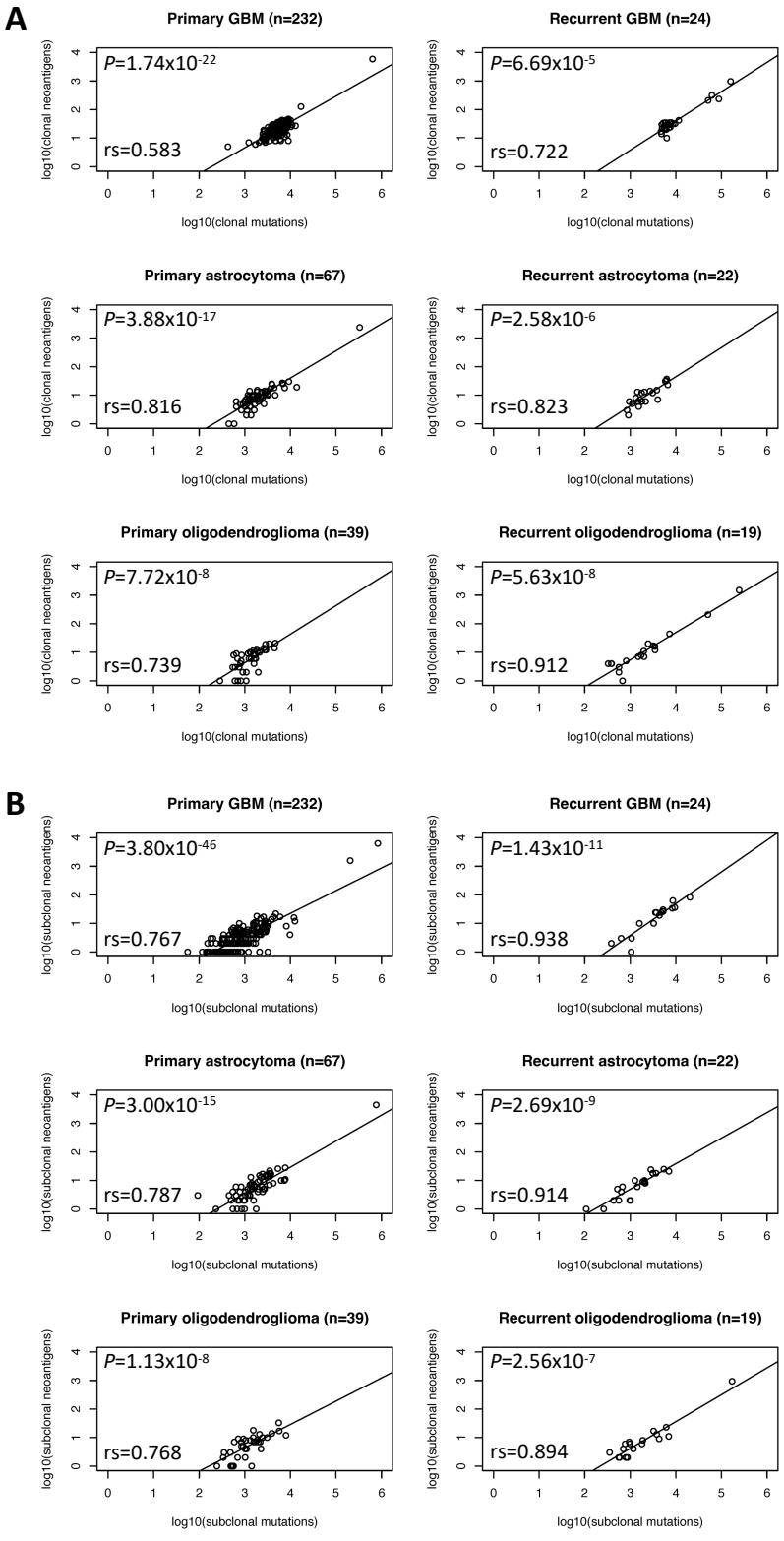
Supplementary Figure 24: Proportions of clonal/subclonal mutations and neoantigens by glioma subtype. *P*-values are unadjusted and two-sided and estimated using the Kruskal-Wallis test. In the box plots, the center line represents the median, and the box bounds represents the inter-quartile range. (A) Clonal/subclonal mutation proportion; (B) Clonal/subclonal neoantigen proportion. Total n=403.



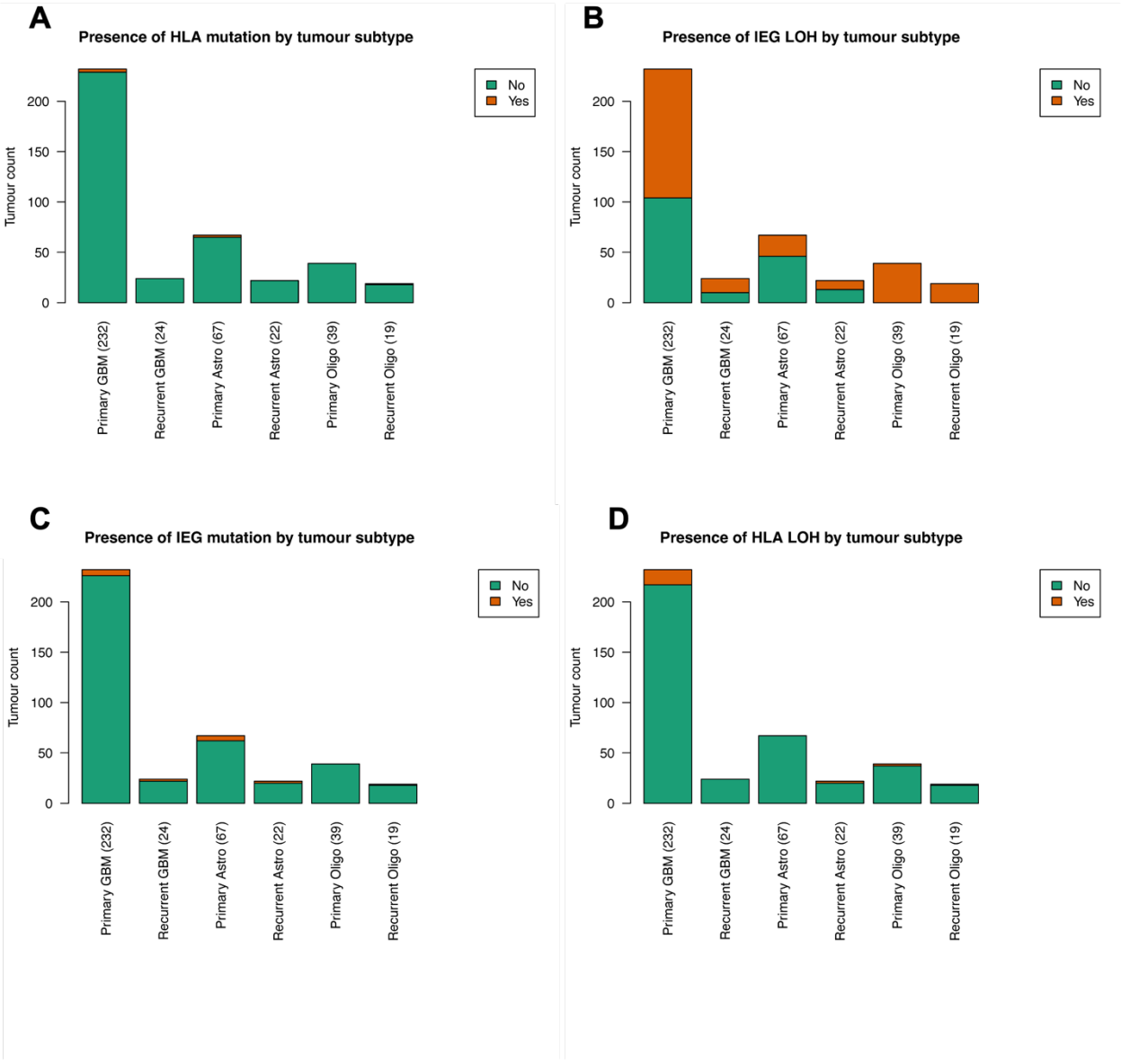
Supplementary Figure 25: Counts of mutations and neoantigens across glioma subtypes. (A) Total mutations and neoantigens; (B) Clonal and subclonal neoantigens. Reported *P*-values and Spearman's rho values (rs) are from Spearman's correlation test. *P*-values are unadjusted and two-sided. Total n=403.



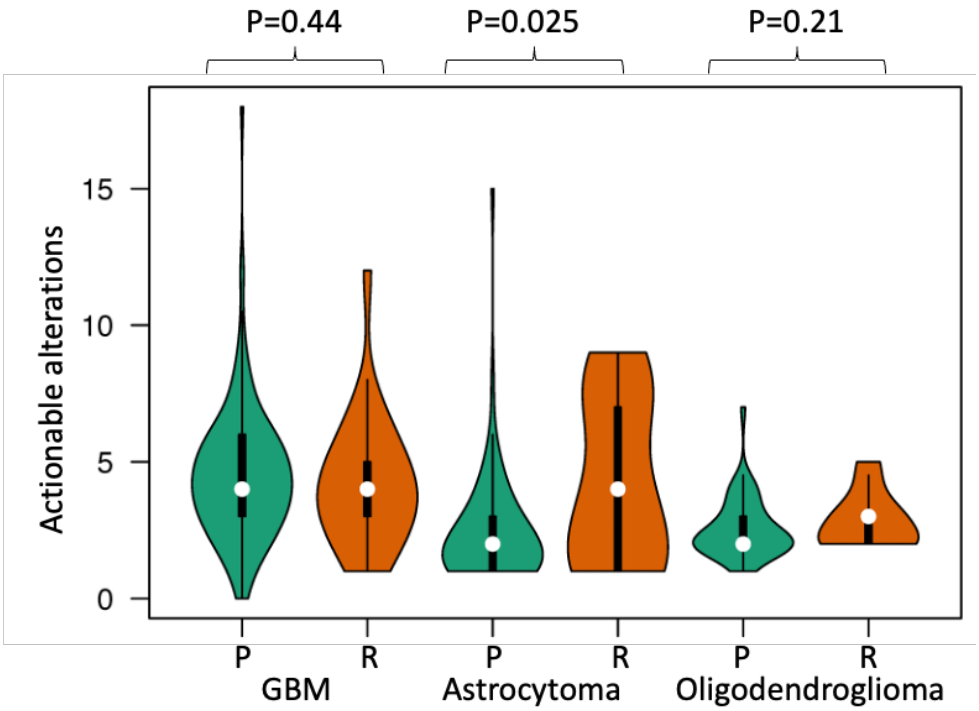
Supplementary Figure 26: Counts of total mutations and neoantigens across glioma subtypes. (A) Clonal mutations and neoantigens; (B) Subclonal mutations and neoantigens. Reported *P*-values and Spearman's rho values (rs) are from Spearman's correlation test. *P*-values are unadjusted and two-sided. Total n=403.



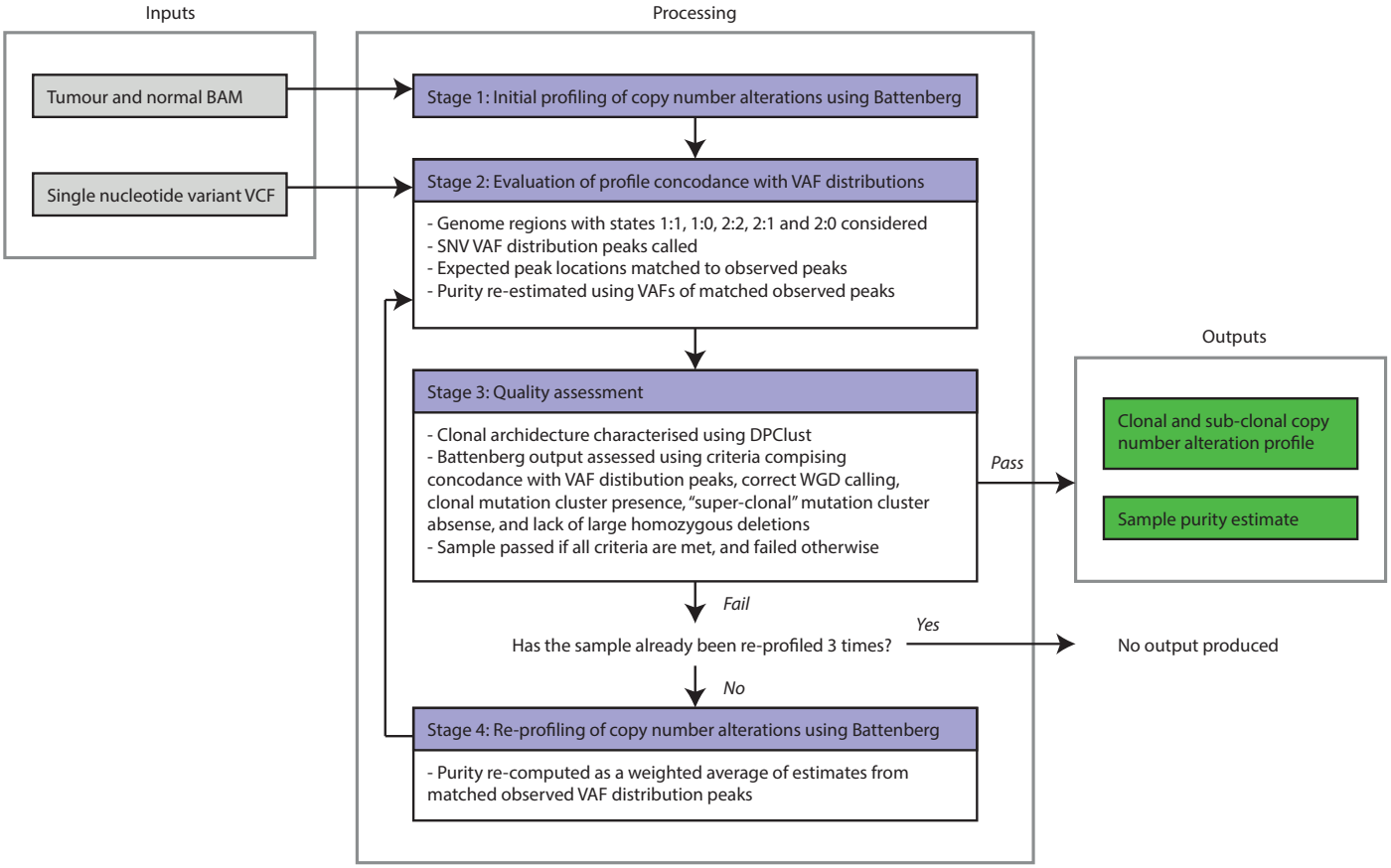
Supplementary Figure 27: Immune escape mechanisms by glioma subtype. IEG; immune escape gene. (A) HLA mutation; (B) IEG LOH; (C) IEG mutation; (D) HLA LOH. Total n=403.



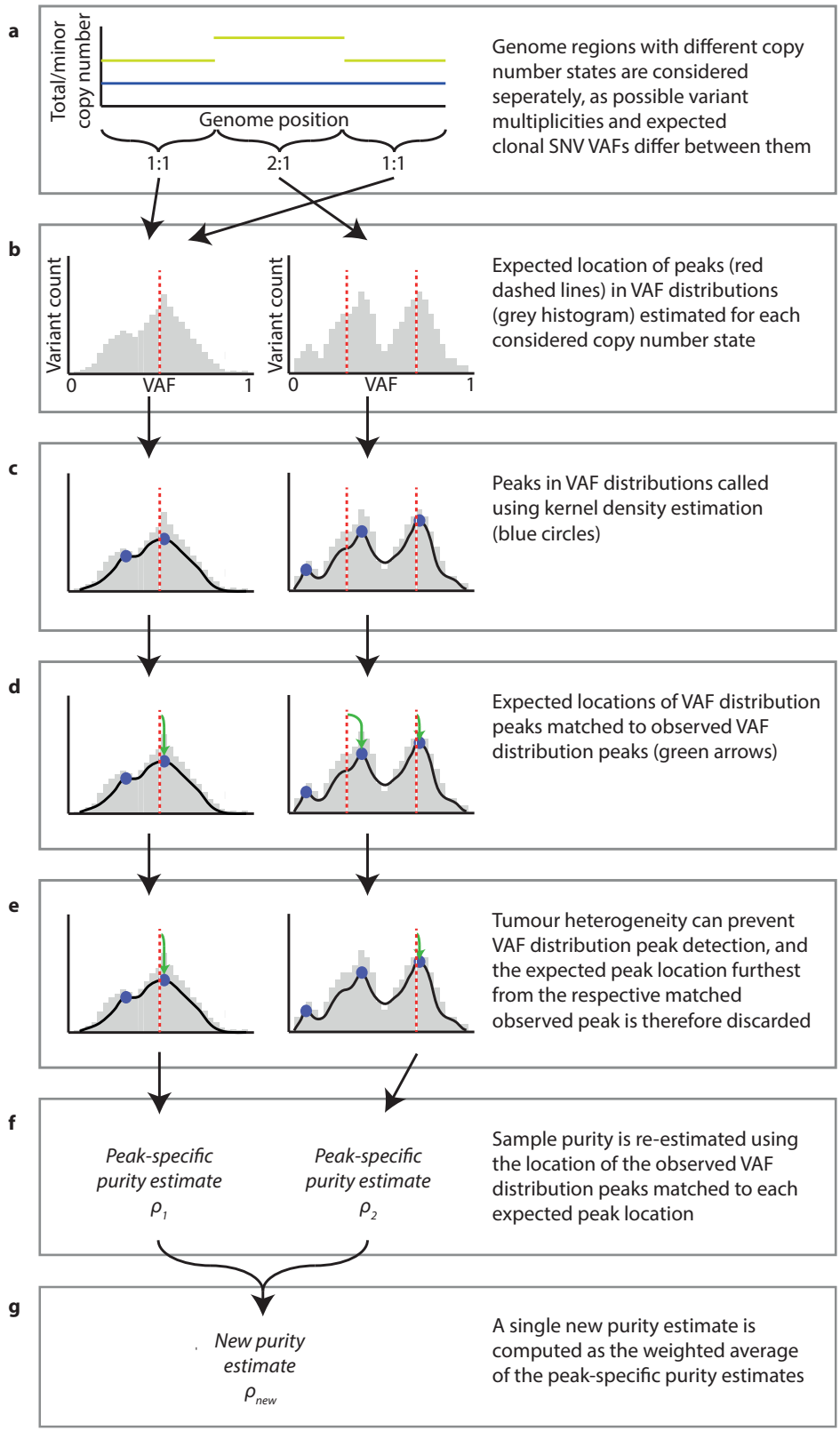
Supplementary Figure 28: Per-tumour actionable alterations across glioma subtypes. *P*-values are unadjusted and two-sided and estimated using the Kruskal-Wallis test. P, primary; R, recurrence. In the box plots, the center line represents the median, and the box bounds represents the inter-quartile range. Total n=403.



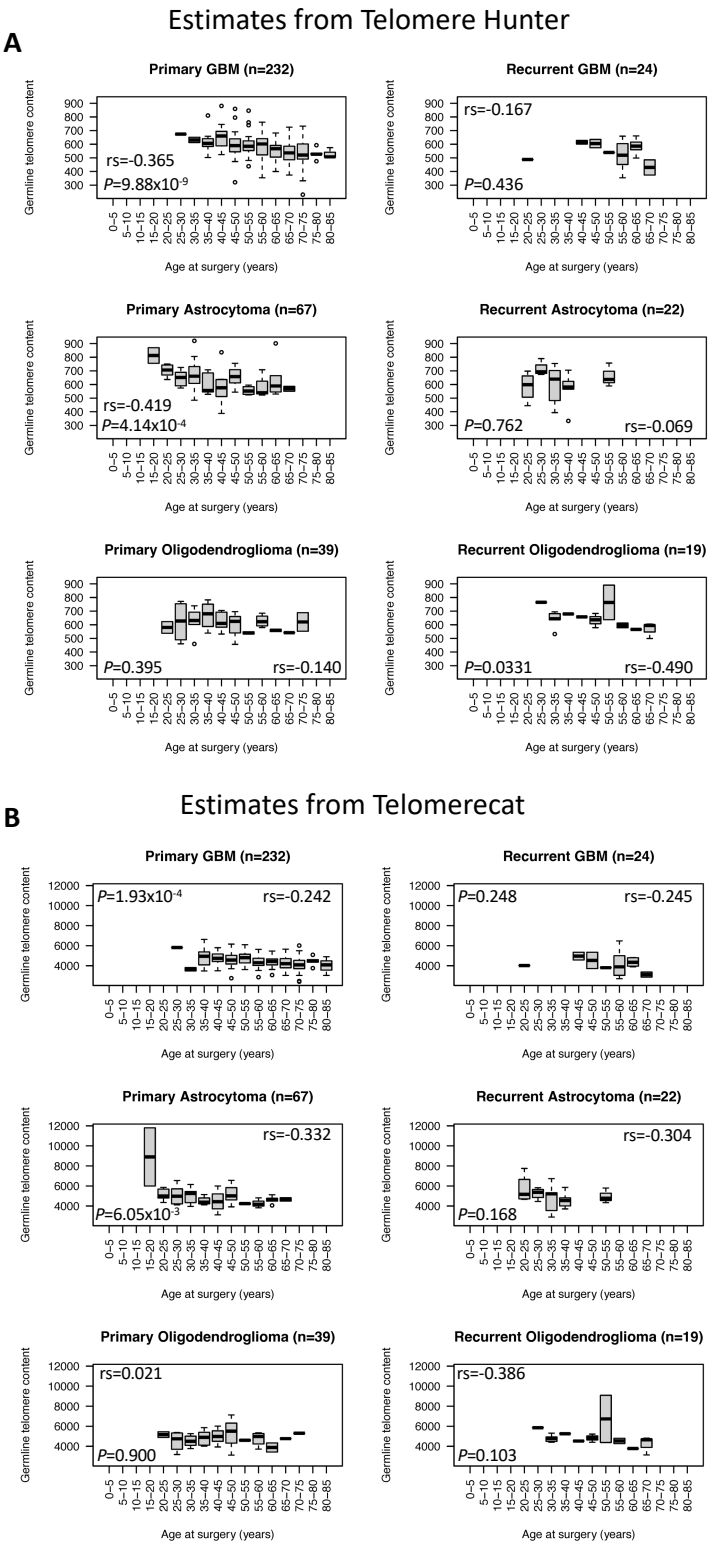
Supplementary Figure 29: CNV calling pipeline overview. BAM, binary sequence alignment map; VCF, variance call format; SNV, single nucleotide variant; VAF, variance allele frequency; WGD, whole genome duplication.



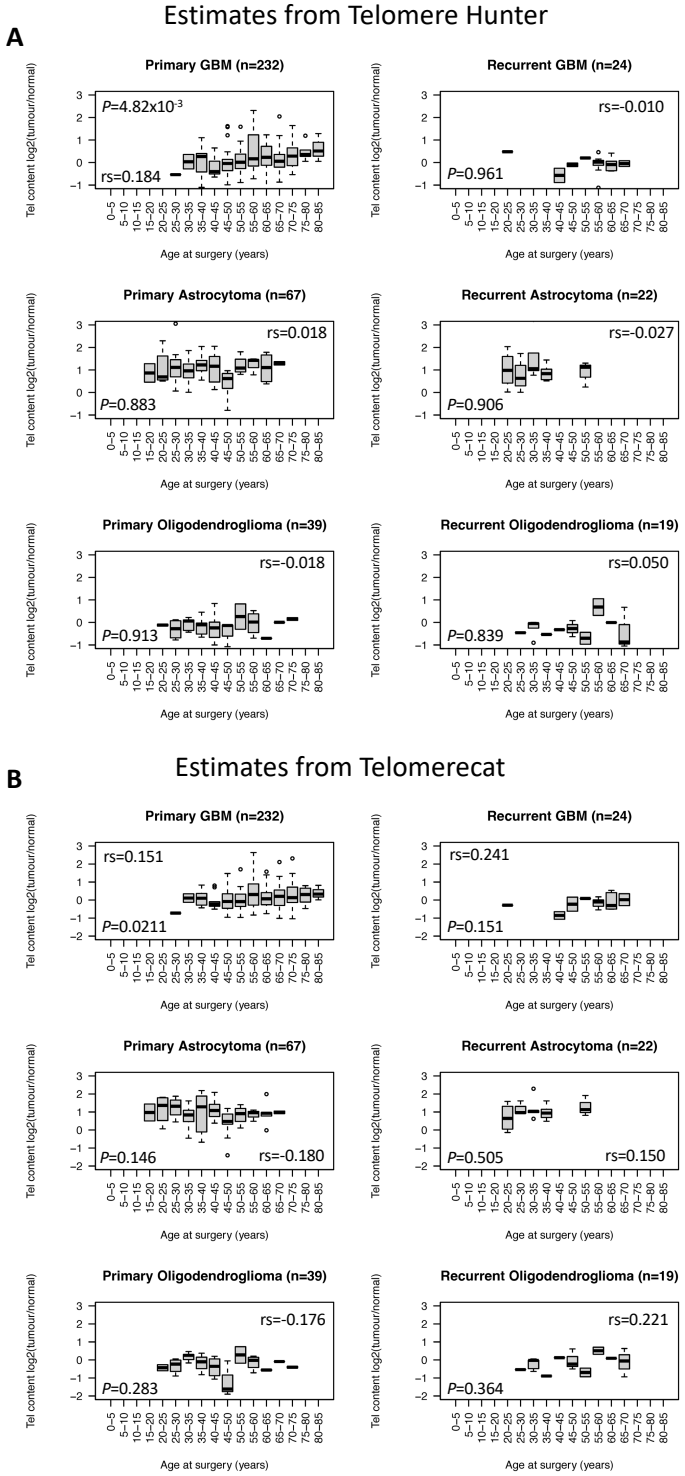
Supplementary Figure 30: Overview of stage 2 of the CNV calling pipeline. SNV, single nucleotide variant; VAF, variant allele frequency.



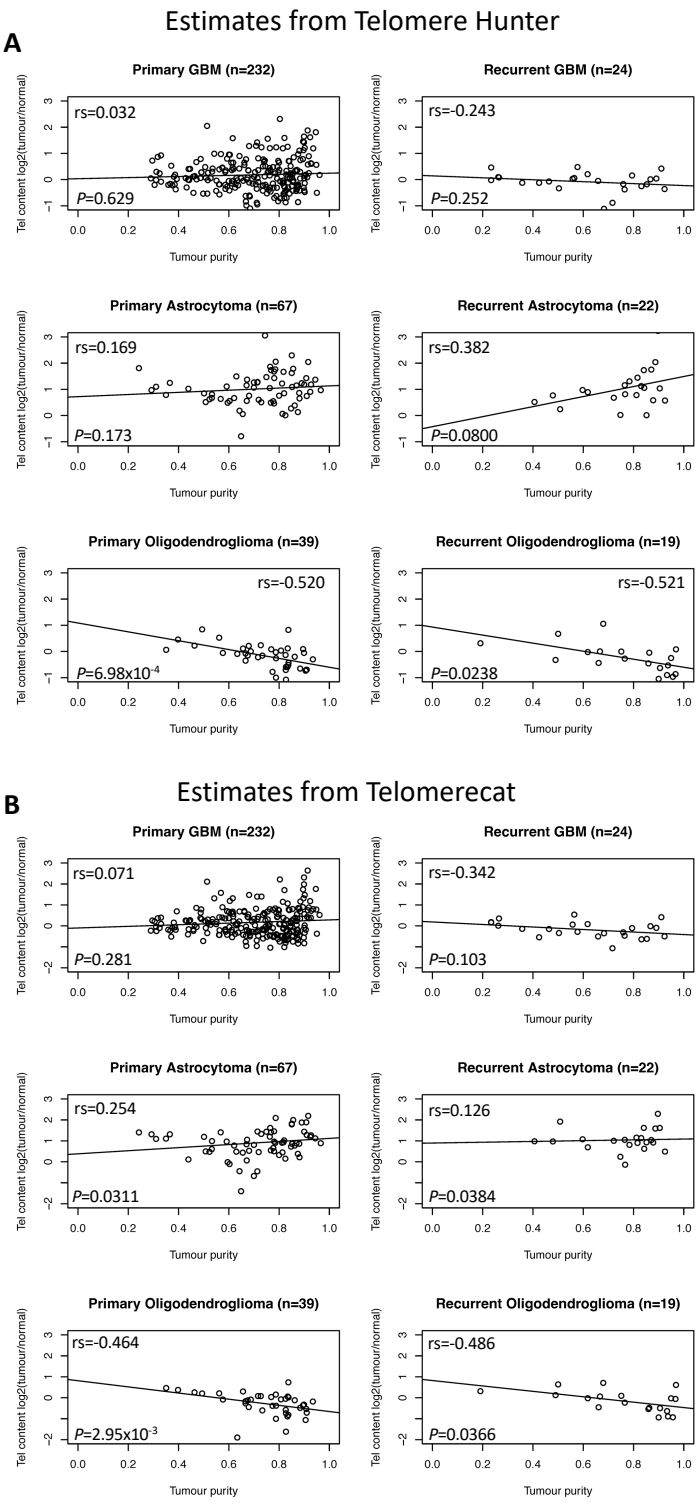
Supplementary Figure 31: Germline telomere content by age across glioma subtypes. Reported *P*-values and Spearman's rho values (rs) are from Spearman's correlation test. *P*-values are unadjusted and two-sided. (A) Telomere Hunter; (B) Telomerecat. Total n=403.



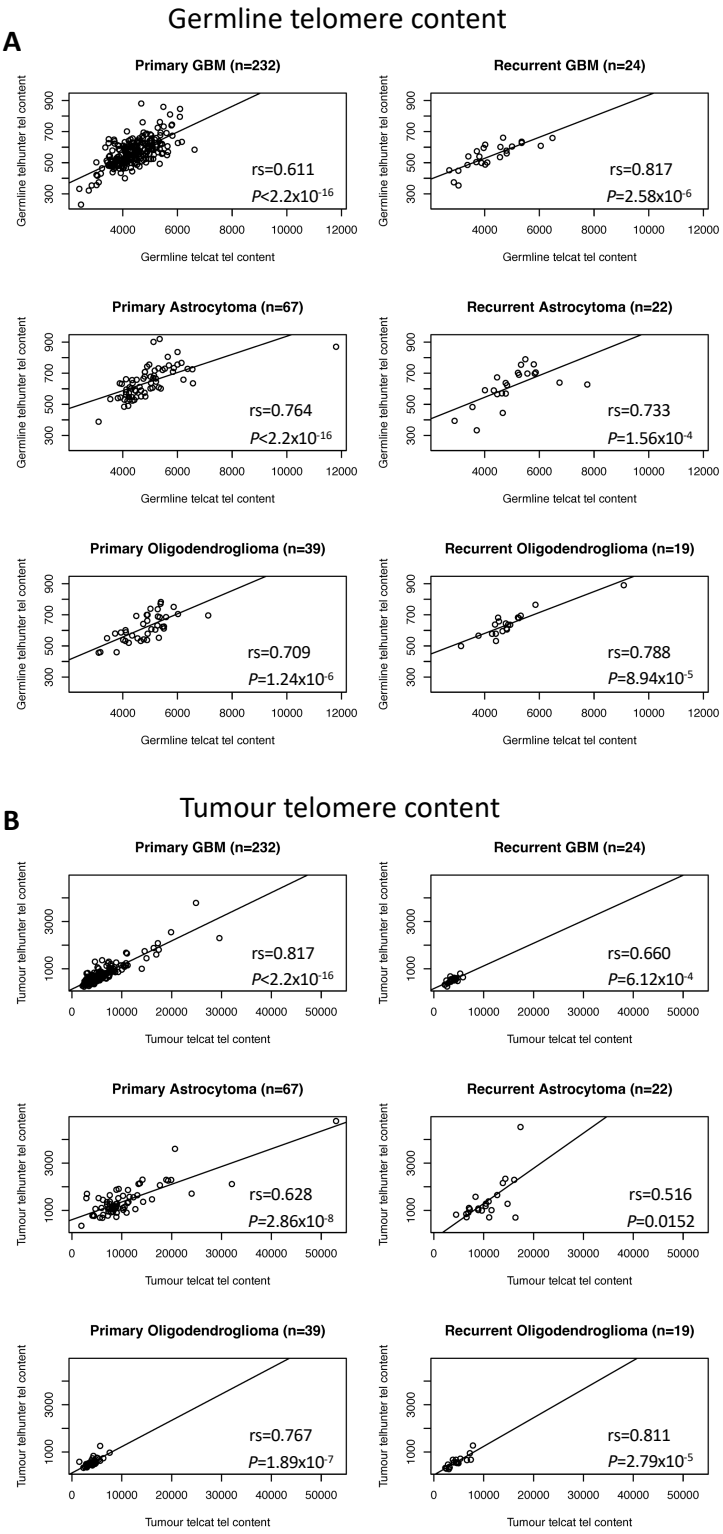
Supplementary Figure 32: Normalised telomere content by age across glioma subtypes. Telomere content normalized by log2(tumour/normal). Reported *P*-values and Spearman's rho values (rs) are from Spearman's correlation test. *P*-values are unadjusted and two-sided. (A) Telomere Hunter; (B) Telomerecat. Total n=403.



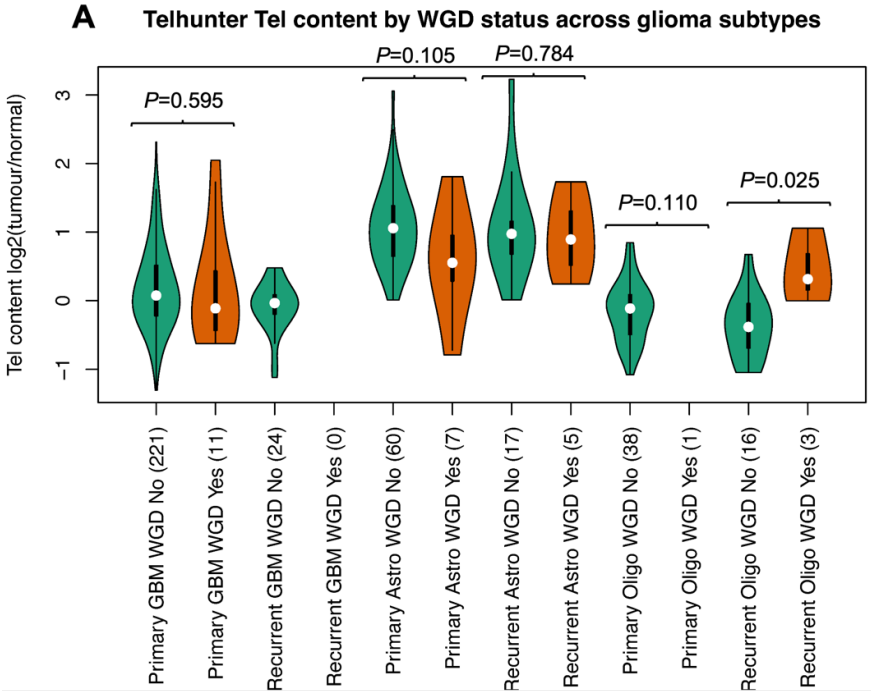
Supplementary Figure 33: Normalised telomere content by tumour purity across glioma subtypes. Telomere content normalized by log2(tumour/normal). Reported *P*-values and Spearman's rho values (rs) are from Spearman's correlation test. *P*-values are unadjusted and two-sided. (A) Telomere Hunter; (B) Telomerecat. Total n=403.

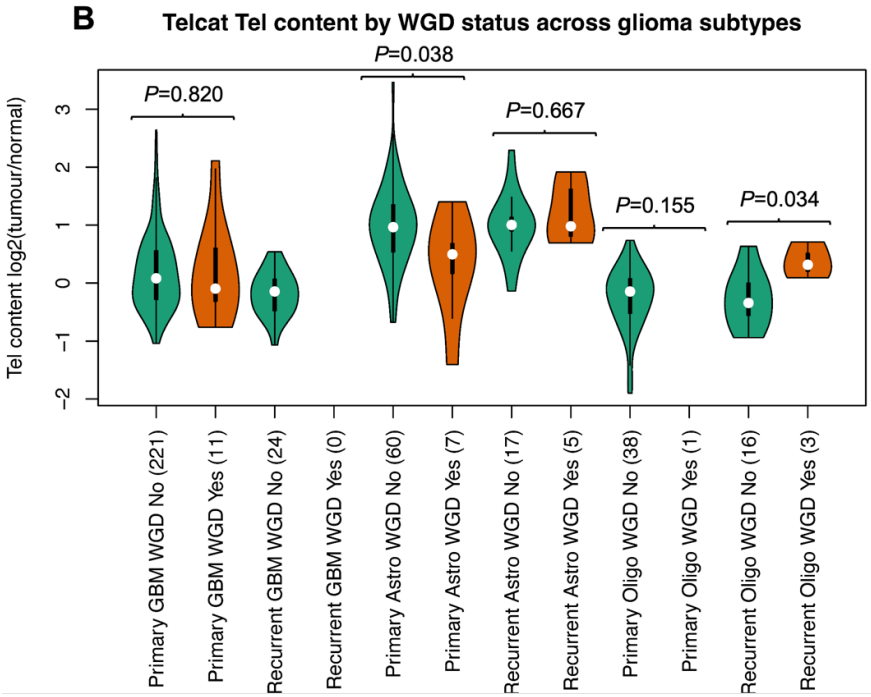


Supplementary Figure 34: Germline and tumour telomere content estimates by Telomere Hunter and Telomerecat. Reported *P*-values and Spearman's rho values (rs) are from Spearman's 0correlation test. *P*-values are unadjusted and two-sided. (A) Germline telomere content; (B) Tumour telomere content. Total n=403.



Supplementary Figure 35: Normalised telomere content by WGD status across glioma subtypes. Unadjusted and two-sided *P*-values estimated using the Kruskal-Wallis test. (A) Telomere content estimated using Telomere Hunter; (B) Telomere content estimated using Telomerecat. Total n=403.





SUPPLEMENTARY REFERENCES

1. Saunders, C.T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811-7 (2012).
2. Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet* **91**, 839-48 (2012).
3. Dentro, S.C. *et al.* Portraits of genetic intra-tumour heterogeneity and subclonal selection across cancer types. *bioRxiv* (2018).
4. Dentro, S.C., Wedge, D.C. & Van Loo, P. Principles of Reconstructing the Subclonal Architecture of Cancers. *Cold Spring Harb Perspect Med* **7**(2017).
5. Louis, D.N. *et al.* The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro Oncol* **23**, 1231-1251 (2021).
6. Raczy, C. *et al.* Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics* **29**, 2041-3 (2013).
7. Karczewski, K.J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020).
8. Tate, J.G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47**, D941-D947 (2019).
9. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-80 (1999).
10. Jamal-Hanjani, M. *et al.* Tracking genomic cancer evolution for precision medicine: the lung TRACERx study. *PLoS Biol* **12**, e1001906 (2014).
11. Cornish, A.J. *et al.* Reference bias in the Illumina Isaac aligner. *Bioinformatics* **36**, 4671-4672 (2020).
12. Bergstrom, E.N. *et al.* SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genomics* **20**, 685 (2019).
13. Islam, S.M.A. *et al.* Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *bioRxiv*, 2020.12.13.422570 (2022).
14. Alexandrov, L.B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101 (2020).
15. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994-1007 (2012).
16. Cornish, A.J., Chubb, D., Frangou, A., Hoang, P.H. & Kaiser, M. Correcting reference bias from the Illumina Isaac aligner enables analysis of cancer genomes. *bioRxiv* (2019).
17. Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-81 (2011).
18. Nilsen, G. *et al.* Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* **13**, 591 (2012).
19. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* **107**, 16910-5 (2010).
20. Weber, C.M., Ramachandran, S. & Henikoff, S. Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. *Mol Cell* **53**, 819-30 (2014).
21. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333-i339 (2012).
22. Layer, R.M., Chiang, C., Quinlan, A.R. & Hall, I.M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15**, R84 (2014).
23. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220-2 (2016).
24. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112-121 (2020).
25. Chu, C. *et al.* Comprehensive identification of transposable element insertions using multiple sequencing technologies. *Nat Commun* **12**, 3836 (2021).

26. Rodriguez-Martin, B. *et al.* Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat Genet* **52**, 306-319 (2020).
27. Tubio, J.M.C. *et al.* Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**, 1251343 (2014).
28. Cortes-Ciriano, I. *et al.* Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat Genet* **52**, 331-341 (2020).
29. Korbel, J.O. & Campbell, P.J. Criteria for inference of chromothripsis in cancer genomes. *Cell* **152**, 1226-36 (2013).
30. Glodzik, D. *et al.* A somatic-mutational process recurrently duplicates germline susceptibility loci and tissue-specific super-enhancers in breast cancers. *Nat Genet* **49**, 341-348 (2017).
31. Hoadley, K.A. *et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929-944 (2014).
32. Haeussler, M. *et al.* The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res* **47**, D853-d858 (2019).
33. She, X. *et al.* Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**, 927-30 (2004).
34. Bignell, G.R. *et al.* Signatures of mutation and selection in the cancer genome. *Nature* **463**, 893-8 (2010).
35. Weddington, N. *et al.* ReplicationDomain: a visualization tool and comparative database for genome-wide replication timing data. *BMC Bioinformatics* **9**, 530 (2008).
36. The ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699-710 (2020).
37. Wala, J.A. *et al.* Selective and mechanistic sources of recurrent rearrangements across the cancer genome. *bioRxiv*, 187609 (2017).
38. Barlow, J.H. *et al.* Identification of early replicating fragile sites that contribute to genome instability. *Cell* **152**, 620-32 (2013).
39. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899-905 (2010).
40. Le Tallec, B. *et al.* Common fragile site profiling in epithelial and erythroid cells reveals that most recurrent cancer deletions lie in fragile sites hosting large genes. *Cell Rep* **4**, 420-8 (2013).
41. Brennan, C.W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155**, 462-77 (2013).
42. Dilthey, A.T. *et al.* HLA*LA-HLA typing from linearly projected graph alignments. *Bioinformatics* **35**, 4394-4396 (2019).
43. Hundal, J. *et al.* Accounting for proximal variants improves neoantigen prediction. *Nat Genet* **51**, 175-179 (2019).
44. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-8 (2011).
45. Hundal, J. *et al.* pVACtools: A Computational Toolkit to Identify and Visualize Cancer Neoantigens. *Cancer Immunol Res* **8**, 409-420 (2020).
46. McGranahan, N. *et al.* Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell* **171**, 1259-1271 e11 (2017).
47. Robinson, J. *et al.* IPD-IMGT/HLA Database. *Nucleic Acids Res* **48**, D948-D955 (2020).
48. Cornish, A.J. *et al.* The genomic landscape of 2,023 colorectal cancers. *Nature* **633**, 127-136 (2024).
49. Mermel, C.H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41 (2011).
50. Zack, T.I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat Genet* **45**, 1134-40 (2013).
51. Seal, R.L. *et al.* Genenames.org: the HGNC resources in 2023. *Nucleic Acids Res* **51**, D1003-D1009 (2023).

52. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-2 (2010).
53. Dale, R.K., Pedersen, B.S. & Quinlan, A.R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**, 3423-4 (2011).
54. Feuerbach, L. *et al.* TelomereHunter - in silico estimation of telomere content and composition from cancer genomes. *BMC Bioinformatics* **20**, 272 (2019).
55. Farmery, J.H.R., Smith, M.L., Diseases, N.B.-R. & Lynch, A.G. Telomerecat: A ploidy-agnostic method for estimating telomere length from whole genome sequencing data. *Sci Rep* **8**, 1300 (2018).
56. Deshpande, V. *et al.* Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. *Nat Commun* **10**, 392 (2019).
57. Kim, H. *et al.* Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. *Nat Genet* **52**, 891-897 (2020).
58. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80-92 (2012).
59. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
60. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).
61. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**, D886-D894 (2019).
62. Rentzsch, P., Schubach, M., Shendure, J. & Kircher, M. CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med* **13**, 31 (2021).
63. Zhang, X., Wakeling, M., Ware, J. & Whiffin, N. Annotating high-impact 5'untranslated region variants with the UTRannotator. *Bioinformatics* **37**, 1171-1173 (2021).
64. Martinez-Jimenez, F. *et al.* A compendium of mutational cancer driver genes. *Nat Rev Cancer* **20**, 555-572 (2020).
65. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041 e21 (2017).
66. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol* **17**, 128 (2016).
67. Arnedo-Pac, C., Mularoni, L., Muinos, F., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. *Bioinformatics* **35**, 4788-4790 (2019).
68. Weghorn, D. & Sunyaev, S. Bayesian inference of negative and positive selection in human cancers. *Nat Genet* **49**, 1785-1788 (2017).
69. Dietlein, F. *et al.* Identification of cancer driver genes based on nucleotide context. *Nat Genet* **52**, 208-218 (2020).
70. Tokheim, C. *et al.* Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. *Cancer Res* **76**, 3719-31 (2016).
71. Porta-Pardo, E. & Godzik, A. e-Driver: a novel method to identify protein regions driving cancer. *Bioinformatics* **30**, 3109-14 (2014).
72. Martínez-Jiménez, F., Muiños, F., Lopez-Arribillaga, E., Lopez-Bigas, N. & Gonzalez-Perez, A. Disruption of ubiquitin mediated proteolysis is a widespread mechanism of tumorigenesis. *bioRxiv*, 507764 (2018).
73. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer* **18**, 696-705 (2018).

74. Lever, J., Zhao, E.Y., Grewal, J., Jones, M.R. & Jones, S.J.M. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. *Nat Methods* **16**, 505-507 (2019).
75. Bailey, M.H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385 e18 (2018).
76. Neph, S. *et al.* BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**, 1919-20 (2012).
77. Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)* **2017**(2017).
78. Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol* **2017**(2017).
79. Ou, J. & Zhu, L.J. trackViewer: a Bioconductor package for interactive and integrative visualization of multi-omics data. *Nat Methods* **16**, 453-454 (2019).
80. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578**, 122-128 (2020).
81. di Micco, P. *et al.* canSAR: update to the cancer translational research and drug discovery knowledgebase. *Nucleic Acids Res* **51**, D1212-D1219 (2023).
82. Shimada, K., Bachman, J.A., Muhlich, J.L. & Mitchison, T.J. shinyDepMap, a tool to identify targetable cancer genes and their functional connections from Cancer Dependency Map data. *Elife* **10**(2021).
83. Berger, M.S., Deliganis, A.V., Dobbins, J. & Keles, G.E. The effect of extent of resection on recurrence in patients with low grade cerebral hemisphere gliomas. *Cancer* **74**, 1784-91 (1994).
84. Wen, P.Y. *et al.* Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. *J Clin Oncol* **28**, 1963-72 (2010).
85. Oken, M.M. *et al.* Toxicity and response criteria of the Eastern Cooperative Oncology Group. *Am J Clin Oncol* **5**, 649-55 (1982).
86. Melin, B.S. *et al.* Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumors. *Nat Genet* **49**, 789-794 (2017).
87. Rapkins, R.W. *et al.* The MGMT promoter SNP rs16906252 is a risk factor for MGMT methylation in glioblastoma and is predictive of response to temozolomide. *Neuro Oncol* **17**, 1589-98 (2015).
88. Brown, C.C. *et al.* A genome-wide association analysis of temozolomide response using lymphoblastoid cell lines shows a clinically relevant association with MGMT. *Pharmacogenet Genomics* **22**, 796-802 (2012).