SUPPLEMENTARY INFORMATION FOR

Novel mutations target distinct subgroups of medulloblastoma.

Giles Robinson[1,2,3*], Matthew Parker[1,4*], Tanya A. Kranenburg[1,2*], Charles Lu[1,5], Xiang Chen[1,4], Li Ding[1,5,6], Timothy N. Phoenix[1,2], Erin Hedlund[1,4], Lei Wei[1,4,7], Xiaoyan Zhu[1,2], Nader Chalhoub[1,2], Suzanne J. Baker[1,2], Robert Huether[1,4,8], Richard Kriwacki[1,8], Natasha Curley[1,2], Radhika Thiruvenkatam[1,2], Jianmin Wang[1,9], Gang Wu[1,4], Michael Rusch[1,4], Xin Hong[1,5], Jared Beckford[1,9], Pankaj Gupta[1,9], Jing Ma[1,7], John Easton[1,4], Bhavin Vadodaria[1,4], Arzu Onar-Thomas[1,10], Tong Lin[1,10], Shaoyi Li[1,10], Stanley Pounds[1,10], Steven Paugh[1,11], David Zhao[1,9], Daisuke Kawauchi[1,12], Martine F. Roussel[1,12], David Finkelstein[1,4], David W. Ellison[1,7], Ching C. Lau[1,13], Eric Bouffet[1,14], Tim Hassall[1,15], Sridharan Gururangan[1,16], Richard Cohn[1,17], Robert S. Fulton[1,5,6], Lucinda L. Fulton[1,5,6], David J. Dooling[1,5,6], Kerri Ochoa[1,5,6], Amar Gajjar[1,3], Elaine R. Mardis[1,5,6,18], Richard K. Wilson[1,5,6,19], James R. Downing[1,7], Jinghui Zhang[1,4], Richard J. Gilbertson[1,2,3].

[1]St Jude Children's Research Hospital - Washington University Pediatric Cancer Genome Project and Departments of [2]Developmental Neurobiology, [3]Oncology, [4]Computational Biology and Bioinformatics, [7]Pathology, [8]Structural Biology, [9]Information Sciences, [10]Biostatistics, [11]Pharmaceutical Sciences, [12]Tumour Biology and Genetics, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, USA.
The Genome Institute[5], Departments of Genetics[6], Medicine[19], and Siteman Cancer Center[18], Washington University School of Medicine in St. Louis, St. Louis, Missouri 63108, USA.
[13]Texas Children's Cancer and Hematology Centers, 6701 Fannin St., Ste. 1420, Houston, TX 77030.
[14]The Hospital for Sick Children, 555 University Avenue, Toronto, Ontario, Canada, M5G 1X8.
[15]The Royal Children's Hospital, 50 Flemington Road, Parkville Victoria 3052 Australia.
[16]Duke University Medical Center, 102382, Durham, NC 27710
[17]The School of Women's and Children's Health, University of New South Wales, Kensington, NSW, Australia.

*These authors contributed equally to the work.

TABLE OF CONTENTS

SUPPLEMENTARY INFORMATION FOR 1

Novel mutations target distinct subgroups of medulloblastoma. 1

TABLE OF CONTENTS 2

SUPPLEMENTARY METHODS 5

- Patient samples 5
- Identification of medulloblastoma subgroups 5
- Immunohistochemistry 6
- Mouse studies 6
  - Genetic mouse models 6
  - Electroporation of shRNAs or DDX3X constructs into LRL and cell tracking 7
  - shRNA Lentiviral Production 7
  - Lower Rhombic Lip (LRL) cell cultures 8
- Next Generation Sequencing 8
  - Illumina Library Construction 8
  - Analysis of whole-genome sequencing (WGS) data 9
  - Experimental validation of genetic alterations 9
  - Telomere Length Analysis 10
  - Background Mutation Rate Calculation 10
- Recurrence screening for somatic sequence variations 10
- Mapping of human assembly hg18 coordinates to hg19 10
- Significance of Mutated Genes 11
- Significance of mutational difference among subgroups 11
- Pathway analysis of mutant genes 12
- SNP 6.0 Analysis 12
- Fluorescence in situ Hybridization (FISH) 13
- Structural modeling 13

SUPPLEMENTARY RESULTS 14

- Cluster Analysis of Expression Profiles 14
- Structural modeling of Mutations in Medulloblastoma 14
  - *DDX3X* 14
  - *KDM* 16

Additional analysis of WGS data 17
Complex copy number variations of Chromosome 12 in SJMB004 17
Analysis of the complex genomic profile for SJMB008 17
Chromothripsis in Medulloblastoma 18
Telomere Analysis 18
Comparison to Parsons et al., Science 2011: “The Genetic Landscape of Childhood Medulloblastoma” (Ref. 45 main manuscript). 19
SUPPLEMENTARY FIGURES 21
Supplementary Figure 1. Unsupervised Hierarchical Clustering Dendrogram. 22
Supplementary Figure 2. Bootstrap co-assignment probability estimates for hierarchical cluster analysis. 23
Supplementary Figure 3. Bootstrap subgroup assignment probability estimates. 24
Supplementary Figure 4. Genome coverage. 25
Supplementary Figure 5. Coverage of coding exons of the 136 genes selected for recurrence screening 26
Supplementary Figure 7. Copy number heatmaps generated by for WGS and SNP cases 28
Supplementary Figure 8. The mutation spectrum of Medulloblastoma 29
Supplementary Figure 9. Copy number alteration, mutant allele fraction and LOH in SJMB008 (G3-16). 32
Supplementary Figure 10. Relationship of copy number alteration and mutant allele fraction in SJMB008 (G3-16). 33
Supplementary Figure 11. Comparison of mutant allele frequencies across the samples with the highest number of total mutations 34
Supplementary Figure 12. OTX2 copy number variations 35
Supplementary Figure 13. Chromothripsis in SJMB008 and 38. 36
Supplementary Figure 14. *DDX31* copy number variations 39
Supplementary Figure 15. Genomically validated in-frame fusions proteins detected by CREST 40
Supplementary Figure 16. Structural analysis of KDM mutations. 42
Supplementary Figure 17. DDX3X additional mutational analysis 43
Supplementary Figure 18. T275, G302 and G325 of DDX3X stabilize bound RNA. 45
Supplementary Figure 19. DDX3X mutation M370R. 46
Supplementary Figure 20. DDX3X mutation T275M 47
Supplementary Figure 21. DDX3X mutation G302V. 48
Supplementary Figure 22. DDX3X mutation G325E. 49

Supplementary Figure 23. Additional protein plots for recurrently mutated genes in Figure 1.50
Supplementary Figure 24. Complex copy number variation in SHH-09 (SJMB004) 51
Supplementary Figure 25. Estimation of total telomere content using WGS data 53
Supplementary Figure 26. Comparison of ages of patients sequenced. 54
SUPPLEMENTARY REFERENCES 55

SUPPLEMENTARY METHODS

Patient samples

93 primary medulloblastoma and paired matched peripheral blood samples were obtained with informed consent through an institutional review board approved protocol at St Jude Children's Research Hospital. All tumors had a histopathological diagnosis of medulloblastoma by central pathology review (D.W.E.) and were obtained at time of diagnostic surgery. Snap frozen and formalin fixed paraffin imbedded materials were available for analysis. RNA was extracted from the snap frozen medulloblastomas using STAT-60. DNA was extracted from the snap frozen medulloblastomas and patient-matched peripheral white blood cells using the DNeasy kit (Qiagen).

Criteria for whole genome sequencing were a minimum of 5µg of tumor DNA with matching minimum 5µg of peripheral white blood cell DNA. Quant-iT PicoGreen (Invitrogen) assay was used to quantify double stranded genomic DNA for sequencing. 37 patient samples met these criteria and were used for the discovery cohort. 56 samples were selected for the validation screen. Validation tumor and germline DNA was whole genome amplified (WGA) using REPLI-g (Qiagen), per manufacturer's instructions. The WGA material was subsequently used for solution capture of the target regions identified in the discovery cohort. Clinical data for all patients providing tumor samples is summarized in Supplementary Table 1.

Identification of medulloblastoma subgroups

Medulloblastoma samples were segregated into the previously described subgroups (WNT, SHH, subgroup-3, and subgroup 4) by mRNA expression profiling and immunohistochemistry[51]. Briefly, mRNA expression profiles were generated using total RNA isolated from human tumors and the U133 Plus 2.0 microarray (Affymetrix, Santa Clara, CA). Gene expression data were normalized using the MAS 5.0 algorithm. The data were then transformed and variance stabilized by addition of small factor of 20 that shrinks the effects of small numbers and then taking the natural logarithm. The median absolute difference (MAD) of these transformed signals was calculated for each probe set across all samples on each array separately within species. The data was then imported into Spotfire Decision Site (Palo Alto CA, USA) and for each probe set and subject z-scores were calculated by computing the mean and standard deviation across subjects within each probeset (see Supplementary Figure 1).

As a separate method to assess the robustness of our subgroup assignments, the global statistical test of clustering[52] was applied to determine whether the multivariate distribution of expression profiles deviated significantly from that of a single multivariate normal distribution. A significant result of this test indicates that the data depart from a single multivariate normal distribution and thus suggests that two or more distinct clusters may exist in the data set. The median absolute difference (MAD) of these transformed signals was calculated for each probe set across all samples on each array separately within species. Log-expression data for the 2,750 probe-sets with the greatest MAD score was used to perform hierarchical cluster analysis by unweighted paired-group method with average linkage

(UPGMA) on Euclidean distances. Bootstrap-aggregation (also known as bagging) analysis[53] was performed to evaluate the robustness of hierarchical clustering results (Supplementary Figure 2). This analysis repeated probe-set selection by MAD score and hierarchical clustering by UPGMA on Euclidean distances for each of 1,000 bootstrap data sets obtained by selecting subjects at random with replacement. The bootstrap results were used to compute estimates of the probability of co-assignment of each pair of subjects to the same subgroup and each subject's probability of assignment to each subgroup (Supplementary Figure 3). The pairwise co-assignment probability estimates were defined as the proportion of bootstraps that assign the pair to the same cluster among the set of bootstraps that included both subjects of the pair. Each subject's subgroup assignment probability estimate was defined as the proportion of bootstraps that assign the subject to the subgroup among bootstraps that include the subject. All gene expression data are available through GEO (accession number: GSE37418)

Immunohistochemistry

Immunohistochemistry was performed to provide an additional assessment of tumor subgroup as previously described and to subgroup cases for which expression profiles were not available[54]. The four antibodies applied were β-catenin (BD#610154; 1:800; antigen retrieval, citrate buffer 20 min bond), GAB1 (Abcam #ab27439; 1:50; antigen retrieval, citrate buffer 20 min Bond), filamin A (Fitzgerald #10R-F113A; 1:100; antigen retrieval, TRIS buffer 30 min BenchmarkXT), and YAP1 (Santa Cruz#sc-101199; 1:50; antigen retrieval, citrate buffer 20 min Bond). WNT medulloblastomas were identified by strong widespread or focal nuclear β-catenin immunoreactivity. SHH medulloblastomas were identified by YAP1, GAB1, and Filamin A immunoreactivity in absence of strong widespread or focal nuclear β-catenin (Supplementary Table 1).

Mouse studies

Genetic mouse models

All animal experiments were performed under approved ACUC protocols. CD1 mice were used for electroporation studies of the lower rhombic lip (LRL) and to harvest cells for *in vitro* neurosphere assays. To generate a cre-inducible *Pik3ca*^{E545K} allele, we used homologous recombination to introduce a lox-puro-STOP-lox cassette immediately upstream of the exon containing the initiation codon, and replaced exon 9 with an exon containing the E545K mutation along with a Neo cassette flanked by Frt sites in intron 9. ES cells carrying the targeted allele were identified by Southern blot, injected into blastocysts to create chimeric mice, and then bred to Flpase-transgenic mice to excise the Neo cassette and create mice with a germline *Pik3ca*^{E545K} allele in which the mutation is expressed from the endogenous locus following cre-mediated excision of the stop cassette. *Blbp-Cre, Ctnnb1*$^{lox(ex3)/lox(ex3)}$ and *Tp53*$^{flx/flx}$ mice were bred with *Pik3ca*^{E545K} mice and monitored for signs of tumor development. Mouse tumors comprised at least 85% tumor cells. The *Pik3ca*^{E545K} mice have a knock-in of the E545K mutation in exon 9 of Pik3ca and a lox-stop-lox cassette inserted upstream of the first exon, so the mutant allele is silent in the absence of cre activity, and it is expressed from the endogenous promoter following cre excision of the stop cassette.

Immunohistochemistry using anti trimethyl H3 (lys 27) antibody (Millipore ABE44) was performed on paraffin embedded tissue sections of three murine medulloblastomas from *Blbp-Cre, Ctnnb1*$^{lox(ex3)/lox(ex3)}$ *Tp53*$^{flx/flx}$ mouse; *Cdkn2c*$^{-/-}$, *Ptch1*$^{+/-}$ mouse; and *Myc* subtype mouse medulloblastoma.

Electroporation of shRNAs or DDX3X constructs into LRL and cell tracking

In utero electroporation and cell tracking was performed as previously described (main manuscript ref. 5). Briefly, pregnant mice bearing E12.5-E13.5 embryos were deeply anaesthetized and the uterus externalized under sterile conditions. For shRNA conditions, a mixture of 3 shRNAs targeting each gene (FUGW-H1-RFPturbo) were mixed with a control eGFP plasmid (FUGW-H1-eGFP): all plasmids 1μg/μl. For overexpression studies, the following plasmids were used: EBcat (stable Ctnnb1 plasmid; Nusse Lab; Addgene #24312). Ddx3x wild-type or mutant cDNAs were cloned into pcDNA vector and mixed with control eGFP plasmid. 1μl of plasmid solution containing 0.01% Fast Green (Sigma) was injected into the fourth ventricle. Embryos were harvested at noted time points, fixed overnight in 4% paraformaldehyde, cryoprotected in 30% sucrose, and sectioned at 30 microns. Sections were imaged on a Zeiss D1 or 510 confocal microscope with Axiovision/Zen software and the line measurement tool used to determine the distance cells migrated from the dorsal brainstem to the pontine grey nucleus (PGN). Several steps were taken to ensure uniformity of measurement between embryos targeted with the same or different DNA constructs. Measurements were taken from multiple sagittal sections ($\geq$5) across of the width of the P1 brainstem. Equivalent sections were selected from each embryo and a 'common line of reference' drawn from the dorsal brainstem to the ventral surface of the PGN. Cell migration distances were recorded as a proportion of this line. To control for equivalence of electroporation between embryos and constructs, all embryos were co-electroporated with eGFP. >98% of labeled cells in embryos electroporated with shRNA or control FUGW-H1-RFPturbo vectors contained >98% GFP$^+$/RFP$^+$ cells and equivalent migration statistics were obtained by recording measures of GFP$^+$, RFP$^+$ or GFP$^+$/RFP$^+$ cells. Cell measurements were recorded from GFP$^+$/RFP$^+$ cells in shRNA electroporated embryos relative to GFP$^+$/RFP$^+$ control embryos. Cell measurements were recorded from GFP$^+$ cells in DDX3X cDNA electroporated embryos relative to GFP$^+$ control embryos.

shRNA Lentiviral Production

shRNAs were cloned into the pFUGWH1-RFP^{Turbo} construct (gift from Temple Lab). Briefly, oligonucleotides were synthesized containing a XbaI recognition sequence, followed by sequence-specific 19 nucleotide stretch designed to target either the open reading frame (ORF) or 3'UTR, the loop sequence (TTCAAGAGA), the reverse complement of the targeting sequence, and finally an EcoRI recognition sequence. Three shRNAs were produced to target each gene of interest, two in the ORF and one in the 3'UTR.

Lentivirus encoding each shRNA was produced by cotransfecting the shRNA-expressing lentiviral plasmid with plasmids pVSV-G and pCMVd8.9 into 293FT cells. Viral containing media were collected, filtered, and concentrated using PEG-*it* Virus Precipitation Solution (System Bioscience). Viral titers were measured by serial dilution on 293FT cells followed by flow cytometric analysis after 72 hr.

Knockdown of the gene of interest by the shRNAs was confirmed in RFP^+ transduced cells 48 hr post-transfection. RNA was isolated using 5 Prime PerfectPure RNA Cultured Cell Kit with on column DNase digestion and reverse transcribed to cDNA using Applied Biosytems High Capacity cDNA Reverse Transcription Kit. qPCR for the gene of interest was performed on an Applied Biosystems 7900HT Fast Real Time PCR System using Fast *SYBR* Green Master Mix (Applied Biosystems). Knockdown was assessed using the $2^{-\Delta\Delta CT}$ method[55].

Gene Target	Targets	19mer
Cdh1	ORF	AGAAGGAGGTGGAGAAGAA
	ORF	CCAACTGGCTGGAGATTAA
	3'UTR	GGAGAGAAAAGGAGAGAAA
Ddx3x	ORF	GCAAATACCTGGTGTTAGA
	ORF	GTACAGGCCGTGTGGGAAA
	3'UTR	TGAAATAGGTTTAGGAGAA
Mll2	ORF	CTGCAGAAGTGAAGAGTTT
	ORF	GCAAATGGAATGTGAAATT
	3'UTR	CGATAGTCCTGTAGAATTT
Gabrg1	ORF	CGATAAAGCAGATGATGAA
	ORF	CCTAAGTACTGGAGATTAT
	3'UTR	GTTCATACAAGCAGAGAAA
Kdm6a	ORF	CTTAAAAGCTGAAGGGAAA
	ORF	GCACACAATTAATGGAGAA
	3'UTR	CAGCAACGTCACAAAGATA

Lower Rhombic Lip (LRL) cell cultures

Embyonic E 14.5 CD1 wildtype mice were obtained from euthanized pregnant females. The LRL was pared away from the brainstem as previously described (see Reference 5 in main manuscript). LRL isolates were pooled together and dissociated in neurobasal media containing 10 units/ml papain (Worthington) activated with NAC. After dissociation cells were cultured in Ultra Low Adherent $10cm^2$ dishes (Corning) in Neurobasal medium (Invitrogen) containing 2 mM L-glutamine, N2 supplement (Invitrogen), B27 supplement (Invitrogen), 20 ng/ml hrEGF (Invitrogen), 20 ng/ml hrbFGF (Invitrogen) and 50 µg/ml BSA in 5% CO_2. Primary spheres were broken to a single cell suspension and 5 x 10^5 cells infected with pooled shRNAs against the gene of interest (2 – 3 particles of each per cell, 6 – 9 particles per cell total). Upon formation of secondary spheres, single cells were plated at 5,000 cells/mL in triplicate on Ultra Low Adherent 6-well plates (Corning) and neurospheres were counted 7 days after plating.

Next Generation Sequencing

Illumina Library Construction

All methods in the library construction and whole genome DNA sequencing have been described previously[56,57]. Detailed information regarding runs and lanes generated for the 37 normal/tumor pairs is included in Supplementary Tables 12 and 13. All WGS and SNP mapping data are available through dbGaP (accession number: phs000409.v1.p1).

Analysis of whole-genome sequencing (WGS) data

WGS mapping, coverage and quality assessment, SNV/indel detection, tier annotation for sequence mutations, prediction of deleterious effects of missense mutations, and identification of loss-of-heterozygosity were described previously (Zhang et al. Nature *In Press*). Structural variations were analyzed using CREST (see ref.9 main manuscript) and annotated as previously described (Zhang et al. Nature *In Press*). The reference human genome assembly hg18 was used for mapping 14 samples (SJMB001, SJMB002, SJMB003, SJMB004, SJMB006, SJMB008, SJMB010, SJMB011, SJMB012, SJMB013, SJMB014, SJMB015, SJMB016, SJMB017) while hg19 was the reference genome for mapping the remaining samples. The WGS analytical result for the 14 "hg18 samples" were "lift-over" to hg19.

Single nucleotide variations (SNVs) were classified into the following four tiers. Tier 1: Coding synonymous, nonsynonymous, splice site, and non-coding RNA variants; Tier 2: Conserved variants (cutoff: conservation score greater than or equal to 500 based on either the phastConsElements28way table or the phastConsElements17way table from the UCSC genome browser, and variants in regulatory regions annotated by UCSC annotation (Regulatory annotations included are targetScanS, ORegAnno, tfbsConsSites, vistaEnhancers,eponine, firstEF, L1 TAF1 Valid, Poly(A), switchDbTss, encodeUViennaRnaz, laminB1, cpgIslandExt); Tier 3: Variants in non-repeat masked regions and Tier4: the remaining variants.

CNVs were identified by evaluating the difference of read depth for each tumour and its matching normal using the novel algorithm CONSERTING (COpy Number SEgmentation by Regression Tree In Next-Gen sequencing, manuscript in preparation, Zhang et al. Nature *In Press*). Three samples were re-analyzed after reviewing the initial computational results. Both SJMB015 and SJMB019 have a large number of CNV segments. For SJMB015, a larger threshold (0.3 instead of 0.125) in the segmental merging step was applied to account for the increased noise observed. The over-segmentation in SJMB019 was caused by suboptimal quality of matching germline sample. Therefore, the final CNV analysis only included the tumor sample with no matching germline. For SJMB008, the reference for 2x copy number was selected manually to account for a discrepancy between LOH and CNV analysis. Confidence for a CNV segment boundary was assessed using a series of criteria, including length of the flanking segments, difference of CNV between neighboring segments, presence of sequence gaps on the reference genome, presence of SV breakpoints and CNV ascertained from the matching germline sample. Manual curation of CNV calls for several samples (SJMB001, SJMB004, SJMB008, SJMB011, SJMB019, SJMB031) was also carried out using coverage data to refine CNV segments in these cases.

Experimental validation of genetic alterations

For 14 WGS samples (SJMB001-4, SJMB006, SJMB008, SJMB010-017), all predicted tiers 1-3 SNVs, Indels in coding regions and structural variations were validated using Nimblegen/Roche solid-phase custom capture followed by Illumina sequencing as described previously (Zhang et al, Nature, in press). For the remaining cases, all SNVs and indels in protein coding regions were validated by either 454 or Sanger sequencing. All SVs were validated by PCR followed by Sanger sequencing. Primers for SV

validation were designed using Primer3[58] to the 1000bp flanking the predicted structural variation breakpoint.

Telomere Length Analysis

The total number of telomeric reads were assessed by searching the next generation sequencing .bam file (mapped by BWA) for reads containing the repetitive telomeric motif $(TTAGGG)_4$ [59] The total numbers of reads were then normalized to the average genomic coverage.

Background Mutation Rate Calculation

The background mutation rate was calculated using validated and high quality tier 3 mutations (i.e. mutations in non-coding, non-regulatory and non-repetitive regions) normalized against all tier3 regions with effective coverage (i.e. covered by >10x in both tumor and matching normal).

Recurrence screening for somatic sequence variations

We performed recurrence screening of a cohort of 56 meduloblastoma tumor samples consisting of 6 WNT, 8 SHH, 11 Group 3, 19 Group 4, and 12 unclassified cases. 136 genes (listed Supplementary Table 14) were chosen on the basis of recurrence in the discovery cohort, previously described genes of importance in MB and genes in critical pathways associated with MB. Custom capture of coding exons were performed by Becknam Coulter Genomics followed by Illumina sequencing with a library insert size of 300bp. Detection of putative SNV and indel follows the same protocol as the WGS data analysis. Since only tumor samples were sequenced, known germline variations in dbSNP (excluding validated mutations in COSMIS, OMIMSNP and ClinicalVar), NHLBI Exome Sequencing Project (http://evs.gs.washington.edu/EVS/ downloaded on 11.21.2011) and germline variations identified by PCGP were removed. Non-silent, putative coding variations including both SNVs and indels were validated by Sanger sequencing in both tumor and normal samples when available.

Mapping of human assembly hg18 coordinates to hg19

LiftOver of “hg18 samples” (n=14) to human genome assembly hg19 was performed using the UCSC LiftOver program (http://genome.ucsc.edu/cgi-bin/hgLiftOver). Quantification of the confidence level in each LiftOver process was determined by recursively adjusting “Minimum ratio of bases that must remap” (-minMatch), a built-in parameter in LiftOver. To achieve this goal, a strict LiftOver was carried out with -minMatch=1.00. All successful entries are classified as “complete”. For any entries that did not meet this criteria, we re-ran LiftOver using a lower threshold (-minMatch=0.50). All entries that fail under these lower stringency conditions are classified as “failed”. Any remaining entries is subject to a recursive binary search with minMatch between 0.50-1.00 until reaching the highest -minMatch that can still be remapped to hg19. This entry will be classified as “partial” and the highest minMatch value recorded.

Significance of Mutated Genes

In order to assess the significance of the non-silent mutations in SJMB dataset, we used the Significantly Mutated Gene (SMG) test, part of the MuSiC package developed at the Genome Institute (Dees et al., submitted, http://gmt.genome.wustl.edu). The SMG test is used to identify genes that have significantly higher mutation rates (MR) than the background mutation rate (BMR) when multiple mutational mechanisms (coding indels and single nucleotide substitutions, splice site mutations, truncation mutations, etc.) are considered. BMR is typically reported as number of mutated bases per total bases. Our method defines total bases as the number of bases with adequate coverage in the alignment data. Therefore, we first count the number of bases with sufficient aligned read-depth based upon user-defined coverage limits set independently for the tumor and normal BAM files for each sample in the cohort (in this case, 37 SJMB samples with WGS BAMs). For the purposes of our algorithm, these counts are divided into three reference sequence-based categories, 1 category each for A and T bases, CpG bases, and C and G bases not connected as CpG. We also divide the mutations into seven categories currently: A and T transitions, A and T transversions, CpG transitions, CpG transversions, C and G (non CpG) transitions, C and G (non CpG) transversions, and lastly, an "indel" category, for which we use the entirety of the covered space in the sample-set when comparing indel-affected bases versus available bases. In order to determine the BMR for each of the mutation categories, we then divide the number of mutations found in each category by the total number of bases available across the cohort in which such a call can be made. For the additional samples that were used for recurrence screening and that did not undergo whole-genome sequencing, we estimated the number of bases covered in each category by taking the average of the appropriate values obtained from the 37 WGS cases. We restricted this coverage estimation to the 136 genes used in the recurrent screening.

For each gene, P-values are generated for each mutation category assessing the significance of the MR versus the BMR for that category. We used a convolution test to combine the per-category P-values to generate a combined P-value per gene to identify significantly mutated genes. In addition, false discovery rates are also calculated for the combined P-value. We typically evaluate our SMG test results by establishing a P-value or FDR threshold (typically 0.2 or less for FDR), and filtering the results based on these thresholds. Genes with convolution test P-values < 0.05 and FDR < 0.2 are considered "significantly mutated genes".

Significance of mutational difference among subgroups

The significance of mutational difference among subgroups was calculated by Kruskal Wallis test for the four features shown in Figure 1b of the main manuscript: BMR (Background mutation rate), the number of nonsilent mutations, and the number of SVs and the total length of CNV amplifications. P values were calculated for the unadjusted raw values as well as age and gender adjusted values. To adjust for age and gender, medulloblastoma subgroup, age and gender were used as predictors in the generalized linear models (GLMs). Poisson regression was used for all models except for BMR where

ordinary linear regression was used because BMR is not an integer. The final models were selected by AIC using the *step* function in R.

Pathway analysis of mutant genes

All non-silent SNVs and indels (HQ Putative and Valid) were used to determine significantly affected pathways utilizing a method which improves upon standard fisher test-based gene-set enrichment analyses. This method uses a hypergeometric distribution to model the number of pathway genes hit on each chromosome. The hypergeometric distribution used for each chromosome can be thought of as randomly selecting genes without replacement from the collection of genes located on the chromosome and each gene is labeled as on the pathway or not on the pathway (Pathways included; KEGG, BioCarta, and several pathways curated from careful analysis of the literature). The number of genes drawn from the chromosome is equal to the total number of genes (pathway or not) that were hit on the chromosome. We use this sampling process as a model of the distribution of the number of pathway genes hit on a chromosome by "chance."

This model forms the basis for determining the distribution of other measures of how frequently pathway genes are hit by "chance." Distributions for the total number of pathway genes hit for one subject, the number of subjects in a cohort with at least one pathway gene hit, and the total number of pathway hits observed for all subjects in a cohort are derived by convolution of the hypergeometric distributions assumed for each chromosome of each subject.

P-values are computed by comparing the observed values of these statistics to their distribution under the model of chance. After correcting for multiple hypothesis testing (FDR) a threshold of 0.05 was used to determine the most significantly affected pathways.

SNP 6.0 Analysis

SNP microarray profiles were generated in collaboration with the Hartwell Center for Bioinformatics and Biotechnology at St Jude Children's Research Hospital using the Affymetrix GeneChip Human Mapping 6.0 assay. Purity and integrity of DNA samples was confirmed by UV spectrophotometry and by agarose gel electrophoresis. Processing of DNA samples was performed according to the Affymetrix SNP protocol. For each probe set and each array, the raw signal is extracted from the CEL file. It is defined as the average log-transformed intensity of probes belonging to the probe set. For each probe set, the raw difference is computed by subtracting the raw signal of the control array from the raw signal of the tumor array. An iterative normalization-segmentation procedure was applied to the raw differences. Probe sets were categorized as copy number probe set, genotype probe set with heterozygous call in control, genotype probe set with homozygous call in control, or genotype with no call in control. For each chromosome, shift and scale transformations were applied to the raw differences of each probe set category so that each category had the same mean and standard deviation. The results are called the adjusted differences. A circular binary segmentation algorithm was applied to the adjusted differences, if any change-points are detected, steps were repeated within

each detected segment. The adjusted differences and genotype calls are summarized for each segment. The summary statistics involving loss-of-balance (proportion of markers called heterozygous in control that are not called heterozygous in tumor) and mean adjusted difference are used to select a set of reference segments. The aligned differences are defined as the adjusted differences minus the mean adjusted difference of the reference segments.

Fluorescence in situ Hybridization (FISH)

Dual color interphase FISH was performed on 6-8μm formalin fixed paraffin embedded tissue sections. Probes were derived from bacterial artificial chromosomes (BAC) clones and were labelled with FITC or rhodamine fluorochromes and were used to assess copy number. The probes used included the following: Chrom 6p22 DCDC2 RP11-72O5 with Chromosome 6q23 SGK1 RP11-692B5; 17p13 HIC1 RP11-357O7/RP11-806J5 with 17q control RP11-368A16/RP11-661H23; MYC CTD-3056O22/2267H22 with 8p control RP11-1077A8/RP11-867P1; MYCN RP11-355H10/RP11-348M12 with 2q control RP11-296A19/RP11-384O8 (Ref.54).

Structural modeling

Structures of the DEAD-box helicase domain of Human DDX3X (PDB ID: 2I4I) and Drosophila Vasa (PDB ID: 2DB3) were obtained from the Protein Databank (PDB) (www.pdb.org) (July 2011 release). The structure for DDX3X represents an open, apo conformation of the DEAD-box helicase domain and that for Vasa represents a closed, RNA-bound conformation. There is high sequence and structural similarity between the two domains; for example, the two sequences exhibited 50% identity (Supplementary Figure 17b) and the two structures, 2I4I (residues 167-404) and 2DB3 (residues 222-459), exhibited an RMS deviation of 1.072 Å over 195 Cα atoms. Sequence alignments were performed using the program Clustalw using default settings and were displayed using Espript. Structural alignments were calculated using the “align” command in the molecular display program Pymol. The mutated amino acids of Human DDX3X in MB are conserved in the DEAD-box domain of Drosophila Vasa (Supplementary Figure 17b). Modeling of the mutations was performed using both the closed form of the Vasa DEAD-box domain (teal) and Human DDX3X (green). Graphics were generated using pymol, including the visual depiction of steric clashes of mutated sites with surrounding residues (illustrated as colored disks indicating overlap of Van der Waals radii). Orthologs of Human DDX3X were identified manually through analysis of the SwissProt and TrEMBL (UniProt release 2011_08).

SUPPLEMENTARY RESULTS

Cluster Analysis of Expression Profiles

Hierarchical cluster analysis of our expression profiles produced the dendrogram shown in Supplementary Figure 1. The global statistical test of clustering was very significant ($p < 0.0001$), indicating that the multivariate distribution of expression profiles is not accurately modeled by a single multivariate normal distribution. This suggests that the data may contain several distinct subgroups.

The bagging analysis indicated that four distinct subgroups can be robustly recapitulated across 1,000 bootstrap samples of the data (Supplementary Figure 2). Assignment of subjects to three subgroups shows a tendency for further subdivision of the largest subgroup (Supplementary Figure 2B). Assignment of subjects to four subgroups shows a similar tendency (Supplementary Figure 2C) but assignment of subjects did not give a co-assignment probability greater than 0.75 for any non-identical pair (Supplementary Figure 2D).

Assignments of subjects to four subgroups were very reproducible across the bootstrap data sets (Supplementary Figure 3). Each subject's most frequent bootstrap subgroup assignment occurred with probability estimate greater than 50%; the probability of only two subjects' most frequent bootstrap assignment occurred with estimated probability less than 75%. The bagging assignments were very consistent with the assignments obtained for the original data. The most frequent bagging assignment differed from the observed assignment for only one subject.

Immunohistochemistry was performed to provide an additional assessment of tumor subgroup as previously described[54] and to subgroup cases for which expression profiles were not available (n=4; SJMB002, SJMB111, SJMB118, SJMB123).

Tumors that did not have expression profiles or immunohistochemistry, which clearly defined them as a member of one of the four subgroups were called "unclassified" and assigned to the group "U".

Structural modeling of Mutations in Medulloblastoma

DDX3X

The DEAD-Box RNA helicase protein DDX3X is involved in several cellular processes, including mRNA export, transcriptional regulation and translational control (see main text for references). DDX3X helicase domain is comprised of two RecA-like domains [termed DEAD (residues 211-403) and helicase (residues 414-575)] connected by a non-canonical linker of 11 amino acids[60]. The DEAD domain binds RNA and is the catalytic ATPase. Four missense mutations were identified from multiple cases of MB and found to cluster within the DEAD domain at positions Thr 275 (mutated to Met; T275M), Gly 302 (mutated to Val; G302V), Gly 325 (mutated to Glu; G325E), and Met 370 (mutated to Arg; M370R). Structural modeling was employed to assess the influence of these mutations on the structure and function of the DEAD domain using the structure of the closed, RNA-bound form of the Human DDX3X homolog Drosophila Vasa (Supplementary Figure 17). Three of the four positions (Thr 275, Gly 302, and Gly 325) clustered on the RNA binding surface (Supplementary Figure 18), and the fourth position is present within the core of the DEAD domain (Supplementary Figure 19).

Threonine 275 is located at the amino terminus of a surface exposed helix with its backbone carbonyl participating in typical α-helical hydrogen bonds, while its side chain hydroxyl creates a hydrogen bond with the backbone amine of Leu 278 (Supplementary Figure 18). The backbone amide of Thr 275 potentially interacts with the RNA via a helix dipole (Supplementary Figure 18). This position is conserved as a Thr or Ser in 32 different human DExD-box helicases, which indicates the importance of a side chain hydroxyl at this position[60]. Mutation of this position to a Met (T275M) would eliminate the interactions of the hydroxyl as well as cause a large steric clash with a bound RNA molecule (Supplementary Figure 20). This position appears to have evolved to stabilize and cap the N-terminus of this helix. Mutation of residues surrounding this position has previously been shown to cause defects in RNA binding and ATPase activity. These results establish that disruptive amino acid substitutions within this region of the protein have detrimental effects on RNA binding and potentially ATP hydrolysis.

The Glycine residue at position 302 is located in a structured turn between the central β-sheet and a surface exposed α-helix. The position is conserved in known DDX3X proteins (Supplementary Figure 17) and partially conserved in 37 Human DEAD-Box proteins (86%)[61]. In Human DDX3X the backbone amide group of Gly 302 hydrogen bonds with the side chain hydroxyl of Thr 323 on an adjacent helix, while in RNA bound Vasa, the amide group interacts with a phosphate group of the bound nucleic acid. The amino acid Gly, lacking a side chain, allows the bound RNA to pack against its Cα atom. Introduction of a Val side chain at position 302 would introduce a steric clash with the RNA molecule, eliminating the protein:RNA hydrogen bond and inhibiting RNA binding (Supplementary Figure 21). In summary, this mutation appears to sterically hinder RNA binding, which would reduce the affinity of DDX3X for its RNA substrates.

The third site of a missense mutation present on the RNA binding surface is Glycine 325. This residue is absolutely conserved in known DDX3X proteins and highly conserved in 37 Human DEAD-Box proteins (95%)[61]. Its carbonyl group participates in a typical α-helical hydrogen bond with the amide group of Asp 329, while the amide is positioned to interact with the bound RNA (Supplementary Figure 18)[62]. Mutation of position 325 to a larger glutamic acid (G325E) would occlude the backbone amide and create a repulsive electrostatic clash with a negatively charged RNA molecule (Supplementary Figure 22). The Glu side chain would be forced to project into the interface because the surrounding residues (Phe 357 in particular) would prevent other rotomers of Glu from being sampled through steric hindrance. Similar to what was observed with Gly 302 and Thr 275, mutation of Gly 325 to a larger residue (Glu in an MB sample) would likely adversely affect interactions between the DDX3X protein and RNA.

The fourth missense mutation occurred at Methionine 370, which is located in a loop near the C-terminus of the DEAD domain. This conserved side chain protrudes into a cavity comprised of hydrophobic residues Ile 364, Val 365, Leu 338, Cys 341, Leu 270, and Leu 344 and one polar residue Arg 376 (Supplementary Figure 19). Mutation of this position to an Arg would likely destabilize the hydrophobic core of the protein in this region. First, the mutant Arg side chain is longer than that of the natural Met, which would introduce steric clashes, and second, the polar guanidinium group of the Arg side chain would be incompatible with this hydrophobic pocket. Overall, the M370R mutation may destabilize the folded structure of the DEAD domain and indirectly affect the function of DDX3X.

The four novel, MB-associated missense mutations (T275M, G302V, G325E, and M370R) within DDX3X are localized to the DEAD domain. Structural modeling indicates that the mutations either disrupt RNA/Protein interactions or destabilize the protein. Previous studies have identified missense mutations within DDX3X in breast cancer (R294T), MB (F234L), skin cancer (A502T) and ovarian cancer (R548T and N551H); these mutations occur within the helicase domain, ATP binding site, and phosphorylation sites, and potentially disrupt ATPase activity or regulation of the protein11. Three of the MB-associated mutations (T275M, G302V, and G325E) are clustered together (a mutation "hot spot") on the surface of the DEAD domain and would likely affect interactions with nucleic acid substrates. Unlike the previously described cancer-associated missense mutations, these MB-associated mutations are unlikely to disrupt ATPase activity but rather are likely to interfere with the binding of nucleic acid substrates to the protein. The residue Thr 323, an amino acid located in the mutation "hot spot", was shown to be phosphorylated in vitro by the mitotic kinase, Cdc2/cyclin B. In yeast complementation assays, mutation of Thr 323 to Glu was associated with impaired growth, suggesting that the mutation caused loss of DDX3X function and that phosphorylation of Thr 323 regulates function in vivo[63]. The MB-associated mutations might mimic this putative regulatory mechanism for DDX3X by inhibiting nucleic acid binding. The mutations identified here appear to be associated with loss of DDX3X function through altered binding of nucleic acid substrates or destabilization of the protein core.

KDM

Several members of the KDM protein family were found to harbor missense mutations in samples from medulloblastoma (Mb) patients. The three-dimensional structures of two of these mutated proteins, KDM1a and KDM4c, have previously been structurally characterized, allowing analysis of the structural and functional implications of the MB-associated mutations.

Missense mutation of Glu 375 to Lys (E375K) in the KDM1a protein was observed in an Mb case. Glu 375, which is conserved within many mammalian orthologs, occurs within α-helix C of the substrate-binding pocket[64,65] of the amine oxidase domain responsible for demethylase function. The side chains of Glu 375 and Asp 379, which form one side of the substrate recognition site, bind to the side chain guanidinium group of a conserved Arg8 residue within the tail of the histone peptide substrate (Supplementary Figure 16A). Substitution of Glu 379 with Lys, as observed in Mb samples, is likely to cause steric and electrostatic clashes with Arg 8 of the histone, potentially reducing substrate affinity and specificity. Mutation of several residues within an acidic patch on the opposite side of the pocket (Asp 553 Lys, Asp 555 Lys, and Asp 556 Lys) is associated with significantly decreased demethylase activity, demonstrating the biochemically disruptive effects of mutations within this substrate binding pocket[66].

The second recurrent, Mb-derived mutation for which relevant structural data is available occurs within the KDM4c protein and corresponds to substitution of Asn 282 with Ile (N282I). This mutation also occurs within the catalytic oxidase domain. Residue Asn 282 (N282) is conserved in many mammalian orthologs and within many other Fe2+-binding KDM proteins[65]. The activity of these enzymes requires the binding of a ferrous iron ion (Fe2+) and the cofactor molecule α-Ketoglutaric acid[67]. Initially, the Fe2+ion is stabilized by the α-Ketoglutaric acid and, during catalytic turnover, the α-Ketoglutaric acid is converted to succinate and released from the active site[68]. Cofactor binding and coordination is

mediated by several polar side chains of the KDM4c protein (Supplementary Figure 16B). The 5-carboxylic acid moiety of the α-Ketoglutaric acid cofactor, opposite the Fe2+ion, interacts with two conserved residues, Lys 208 and Tyr 134. The side chain amine of Lys 208 and hydroxyl group of Tyr 134 each form bifurcated hydrogen-bonds with i) the carboxylate of α-Ketoglutaric acid and ii) the side chain amide and carbonyl of Asn 282. Mutation of Asn 282 to an Ile residue would disrupt this hydrogen bonding network, which we predict would destabilize interactions with the cofactor that is critical for catalysis and would inhibit demethylase activity.

In summary, two recurrent, Mb-associated mutations occurred within genes for the KDM family of Lysine demethylase enzymes and could be analyzed in terms of their possible effects on protein structure and function. One of these mutations, E375K in KDM1a, is predicted, based on structural analysis, to affect histone substrate binding and indirectly affect lysine demethylation. A second missense mutation, N282I in KDM4c, is predicted to affect binding of a critical co-factor with in the enzyme active site and therefore is likely to directly affect lysine demethylase function.

Additional analysis of WGS data

Complex copy number variations of Chromosome 12 in SJMB004

SJMB004 has a total of 107 validated structural variations, the highest of all tumors analyzed in this study and the only tumor with long telomeres (Supplementary Figure 24). Inter- and intra-chromosomal breakpoints cluster at chromosomes 12, 13, 17 and 18 in amplified regions with varying copy number. For example, a 30Mb region at 12p has a total of 12 (may consider add the 3 high peaks) copy number segments ranging from 1 to 6 (highest spike) copy. Connecting the 40 SV breakpoints with CNVs in this region shows that the amplification is likely to be formed by a 4-step re-arrangements dominated by fold-back inversions in both haplotypes, but no evidence of chromothripsis could be found in this sample even though it does have a TP53 somatic mutation

Analysis of the complex genomic profile for SJMB008

SJMB008 is the sample with the highest number of somatic sequence mutations with a total of 5,872 validated somatic mutations. Our automated CNV analysis selects chromosomes with no LOH as the reference for 2-copy DNA. Using this process, 10 chromosomes which include chromosome 6 and 8 of SJMB008 were projected to be diploid. Although the resulting CNV profile is consistent with the SNP array analysis, comparison of CNV and LOH revealed a 8.5Mb region on chromosome 15 with conflicting results. This region was predicted to have a 1-copy deletion but lacks LOH signal of the 4,747 germline heterozygous SNPs. A re-analysis using this 8.5Mb region as the reference for 2-copy DNA projects 86% of the tumor genome including chromosome 6 and 8 has amplification while the rest (excluding the 8.5Mb reference segment) has copy-neutral LOH. Detection of amplification of chromosome 6 and 8 by FISH confirms the gross aneuploidy of this tumor genome. In addition 17p has copy-neutral LOH while 17q has 2-copy amplification, indicating the presence of isochromosome 17 coupled with 1-copy amplification of the normal chromosome 17. Consistent with the FISH result, MYCN is projected to have >80x amplification. Interestingly, SV analysis indicates that this amplification is achieved by replication of an 809Kb episome consisting 7 discontinuous regions on chromosome 2 re-arranged by a single catastrophic event, i.e. “chromothripsis” [69] (Supplementary Figure 13).

Deep sequence coverage resulting from custom capture validation allows accurate assessment of mutant allele fraction, i.e. the percentage of reads representing the mutant allele in tumor. For a pure, near diploid tumor, the mutant allele fraction for most of the mutations is expected to be 50%. For example, the median of the 306 validated tier3 mutations in SJMB002, a tumor in which copy number alteration is found in <6% of the genome, is 46.3%, a close match to the 50% expectation. By contrast, the median of mutant allele fraction of 4,902 validated tier3 sequence mutations in SJMB008 is 25.0%, a 50% reduction from the expected 50% fraction. The reduced mutant allele fraction in SJMB008 can be explained either by hyperploidy or by low tumor purity.

We evaluated tumor purity by assessing the extent of LOH in tumor. Germline SNPs with non-reference allele fraction (G_non_ref_allele_fraction) in [0.4,0.6] were selected and the difference between non-reference allele fraction in tumor (T_non_ref_allele_fraction) and germline using Allelic Imbalance (AI= G_non_ref_allele_fraction – T_non_ref_allele_fraction|). Complete LOH result in AI close to 0.5 because T_non_ref_allele_fraction is expected to be either 0 or 1 depending on whether the non-reference allele is lost in LOH. We compared the AI of chromosomes 3 and 8 in SJMB008 with chromosome 13 AI distributions of two retinoblastoma samples, SJRB003 (copy neutral LOH at 13q) and SJRB002 (copy-neutral LOH across chr13) known to have 95% and 75% tumor purity based on presence of wild-type RB1. The median of AI distribution of SJMB008 for both chromosomes are at 0.45, close to the 0.46 observe in SJRB003, suggesting high tumor purity in this sample (Supplementary Figure 9).

Chromothripsis in Medulloblastoma

We discovered evidence of chromosome shattering and rejoining, so called “chromothripsis” in two samples, SJMB008 and SJMB038 (5%; Supplementary Figure 13). Both samples have small regions of high copy number gain dispersed over a single chromosome (chr2 and ch17) linked by numerous intra-chromosomal translocations. It is likely, therefore that the chromosome in each of these cases was shattered into many pieces, a subset of these pieces were re-joined forming an episome which was repeatedly replicated. In the case of SJMB008 we were able to determine which of the pieces of chr2 were contained within the episome using SV beakpoints detected by CREST, this episome contains the *MYCN* locus, causing 80x amplification of this oncogene. In the case of SJMB038, the amplification involves exons 1-8 of GAS7, a gene known to be involved in neuronal development and was focally amplified (CN=5.0 in a 529kb segment) in SJMB004. No common SNV or indel is shared by these samples therefore it is unlikely that a mutation lead to this event. TP53 mutations, somatic or germline, were not found in these two samples.

Telomere Analysis

Like most cancers, the majority of medulloblastomas sequenced in this cohort have significantly shorter telomeres (less telomeric reads) compared to their matched normal counterpart (p=0.0006, t-test). Telomeres are maintained above a critical threshold to maintain genome stability and retain replicative capacity. SHH-09 (SJMB004), however, contains much more telomeric DNA than normal cells from this patient, possibly indicating greater genome instability in this sample (Supplementary Figure 25). This sample also has one of the most structurally altered genomes in this cohort and has a TP53 mutation.

Comparison to Parsons et al., Science 2011: "The Genetic Landscape of Childhood Medulloblastoma" (Ref. 45 main manuscript).

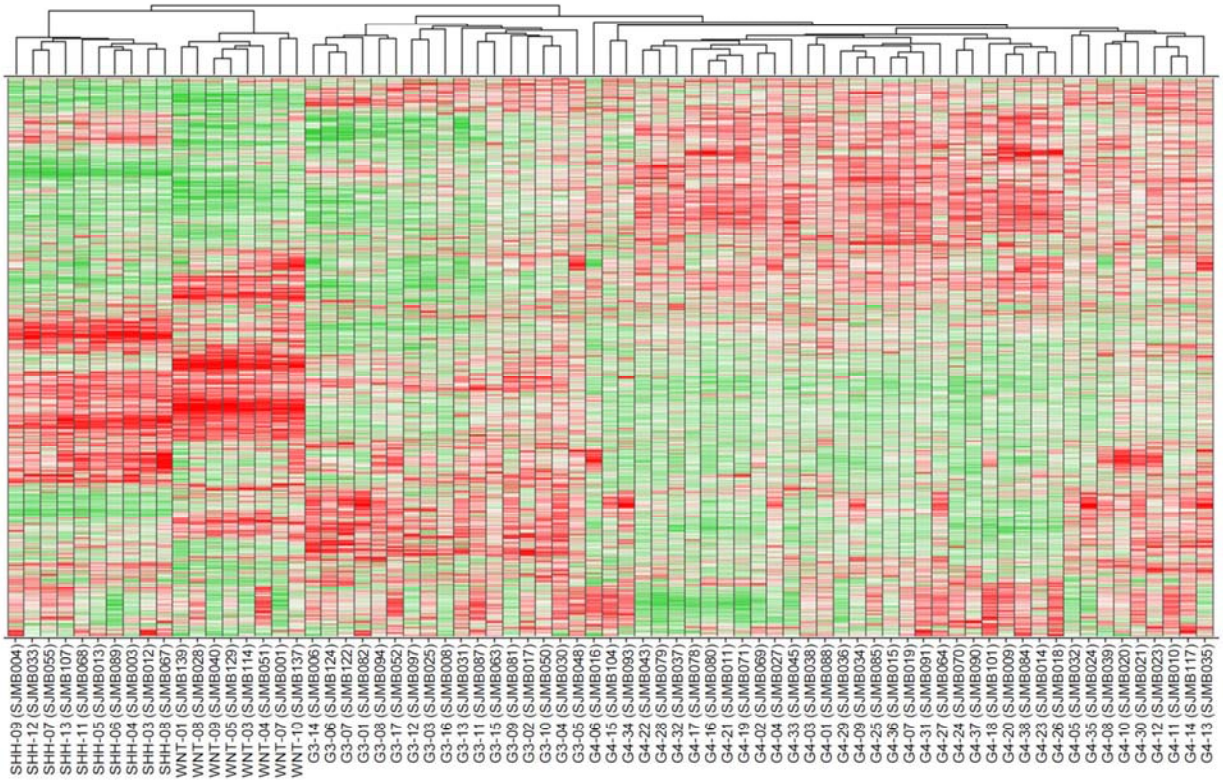
In the study "The Genetic Landscape of Childhood Medulloblastoma" the authors sequenced the whole exome and regulatory regions of 22 childhood medulloblastomas by Sanger sequencing, finding on average 8 non-silent mutations per sample. In this study, however, we report a mean of 11.8 non-silent coding mutations per sample. There are a number of contributing factors that could go some way to explain this modest increase. G3-16 (SJMB008), G4-06 (SJMB016), G3-02 (SJMB017), and G3-13 (SJMB031) had > 20 total non-silent coding mutations. G3-16 in particular has a large amount of coding mutations (64). Excluding these samples results in a mean of 8.3 total non-silent coding mutations per sample which is in agreement with the mean reported by Parsons *et. al.* Another contributing factor is the sensitivity of next generation sequencing as opposed to PCR and Sanger sequencing based technologies, NGS is much more sensitive and allows for the detection of mutations present at a frequency lower than the threshold detectable by Sanger. Age is also a major difference between the two datasets (Kruskal-Wallis p-value = 2.2e-16), although the discovery datasets are similar, the Parson's prevalence dataset (Supplementary Figure 26) has a broad patient age range when compared to SJMB (Parsons; Mean = 15.7, Max = 48.0, SJMB; Mean = 7.9, Max = 18.2). Interestingly the majority of the *PTCH1* mutations discovered by Parsons *et al.* in the prevalence screen are in older patients (Supplementary Figure 26), Red points indicate one or more PTCH1 mutation(s).

To assess the degree of overlap between the data presented here and that of Parsons *et. al.* We analyzed the sequence mutations present in both datasets and determined any overlap. The dataset presented here is significantly enriched for genes which were found in the Parsons study (Fishers Exact Test p = 0.0008, odds ratio: 2.55) but there are a number of key differences highlighted by this comparison. We found *DDX3X* as the most recurrent novel gene in Medulloblastoma found in 8 samples is only present in 1 case in the Parsons dataset. The KDM genes are a prominent feature of this dataset, yet there is no overlap with Parson's *et al.* who report a single *KDM6B* nonsense mutation. Also striking is the difference in the frequency of *PTCH1* mutations, approximately 25% of cases in the Parsons study harbored a *PTCH1* aberration however we found only 2 cases containing sequence mutations. This highlights the differences in the clinical cohorts selected for both studies, in the Parsons et al. dataset subgroup information is not available/unknown for the patients sequenced, but with this information it seems likely (taking into consideration the *PTCH1* frequency) that the Parson's cohort was heavily skewed towards patients with SHH tumors. *MLL2* and *MLL3* were classified as "tumor-suppressor genes" by Parson's *et al*. due to the high frequency (17%) of inactivating mutations in their cohort. Like *PTCH1* the frequency of mutations in the MLL genes in our cohort is much lower (6%). These genes are well covered by our sequencing and therefore a lack of mutations is not likely due to systematic sequencing errors or low coverage of *MLL2* and *MLL3*. Other differences of note are *SMARCA4* (3% vs 8%), and *CHD7* (1% vs 5%).

As described previously we discovered novel recurrent copy number alterations in O*TX2* and *DDX31*. *OTX2* amplification was reported in two cases, both originating from xenograft samples in the Parson's study which contrasts to this study where we see 6/37 (16%) samples harboring focal amplification of *OTX2*. *DDX31* copy number loss or gain was not detected by Parsons *et. al.*but here we report *DDX31*

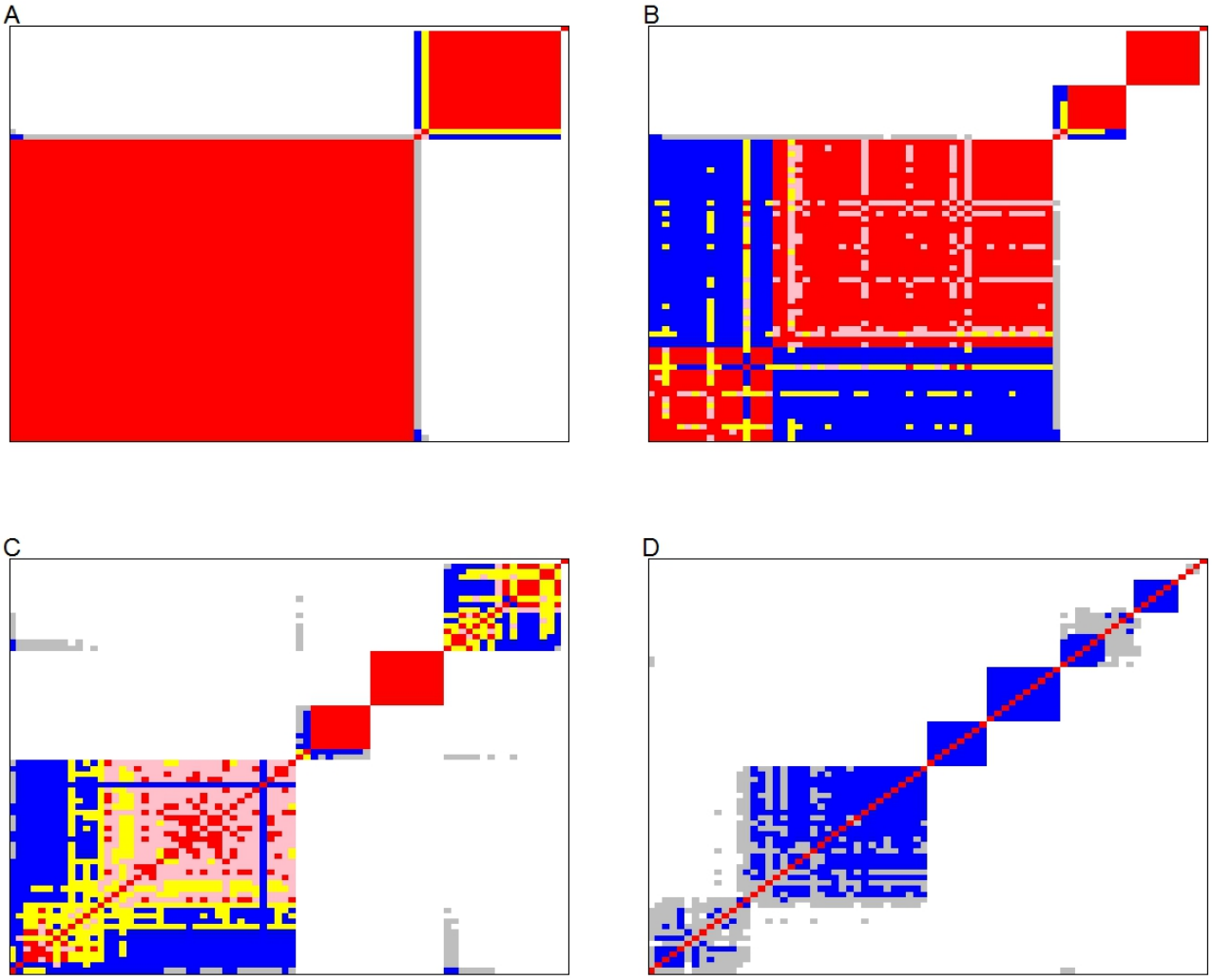
loss in 3/37 (8%) and gain in 1/37 (3%) cases. In addition to these novel copy number alterations MYCN amplification is a known feature of some medulloblastomas. We find *MYCN* amplification in 9/91 cases (10%) but this aberration was less prevalent in the Parson's dataset where *MYCN* was amplified in 3 xenograft cases.

SUPPLEMENTARY FIGURES



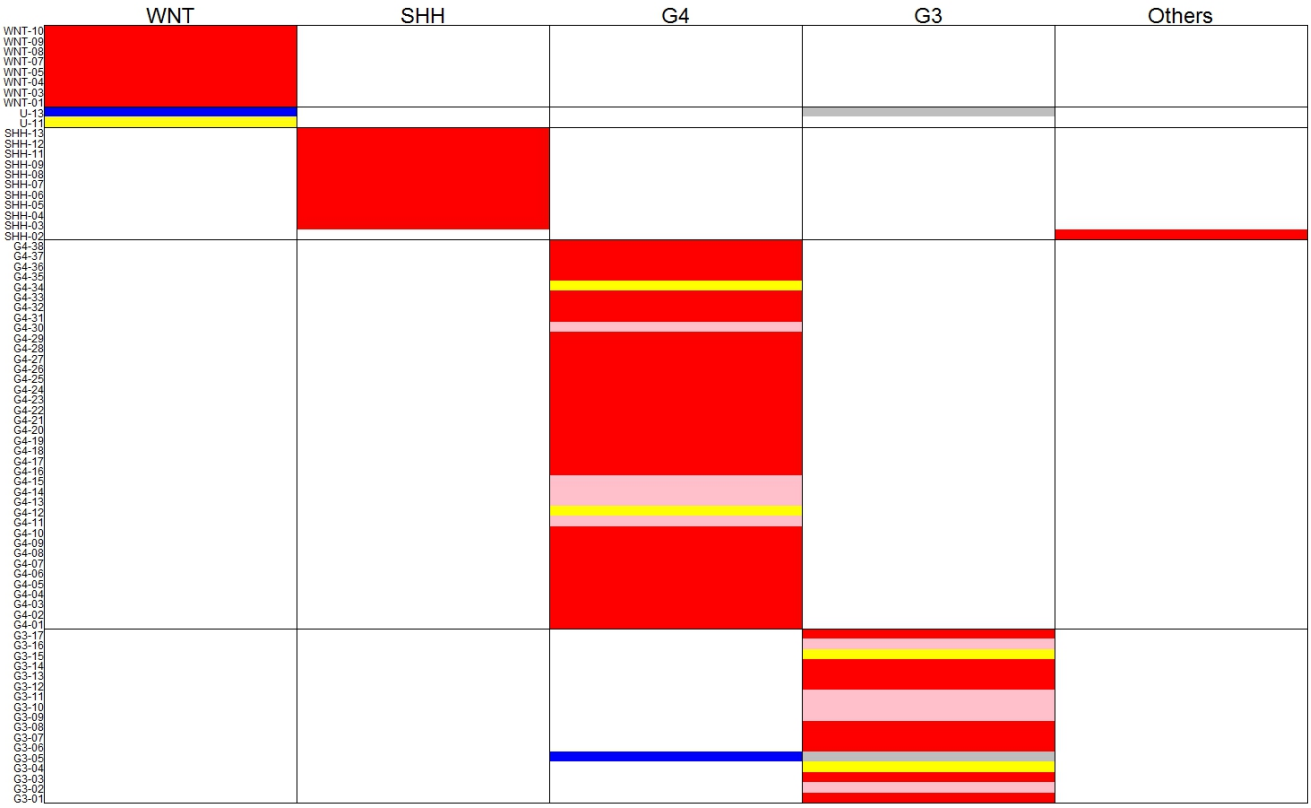
Supplementary Figure 1. Unsupervised Hierarchical Clustering Dendrogram.

The log-expression data of the 2,750 probe-sets with the greatest MAD score.



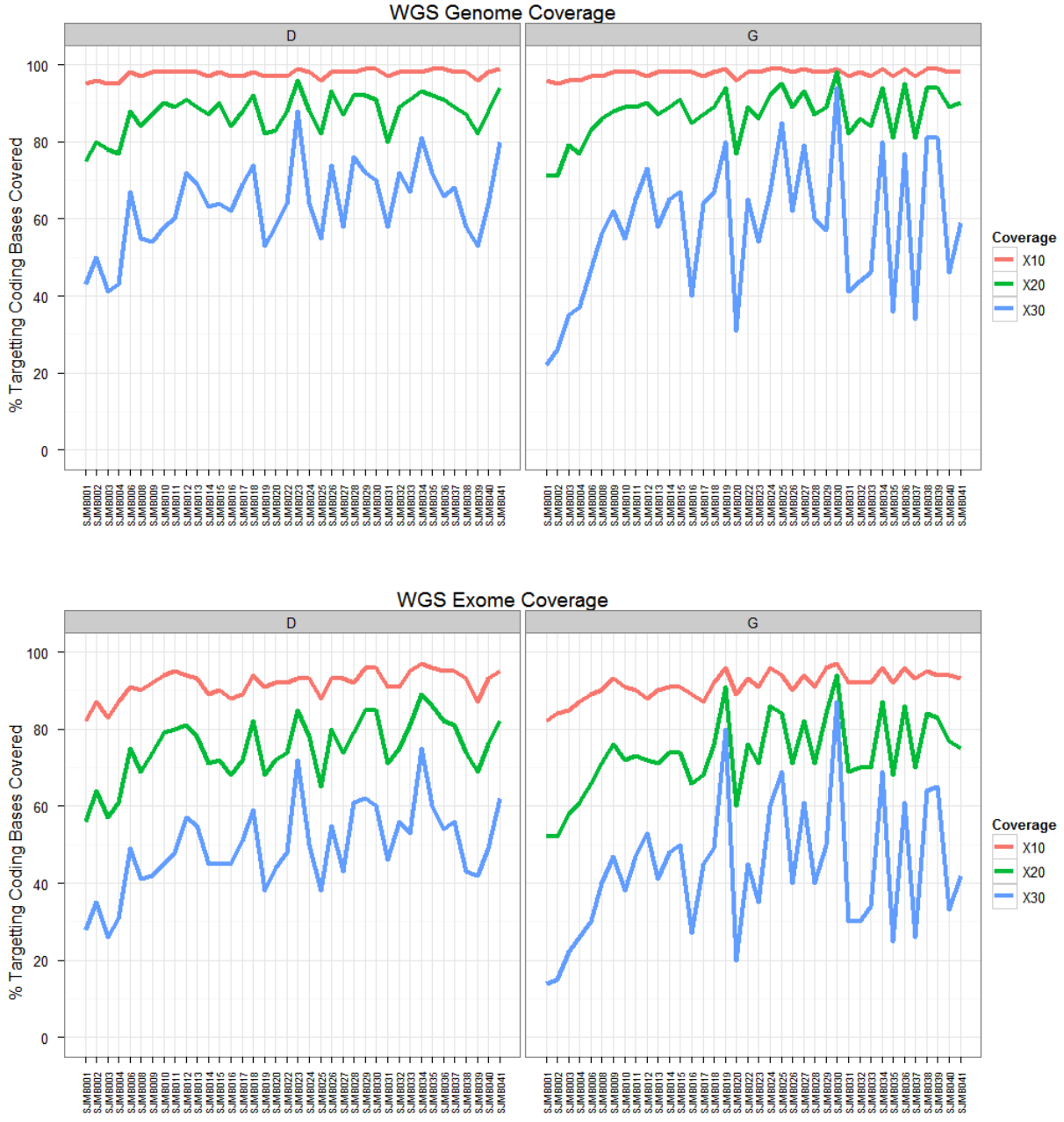
Supplementary Figure 2. Bootstrap co-assignment probability estimates for hierarchical cluster analysis.

Figure shows the assignments of subjects into two, three, four, or five subgroups with at least five subjects (panels A-D, respectively). Each heatmap is a matrix of bootstrap estimates of the probability that each pair of subjects is assigned to the same subgroup. The color scheme is 0 ≤ white ≤ 0.25 < gray ≤ 0.5 < blue ≤ 0.75 < yellow ≤ 0.9 < pink ≤ 0.95 < red ≤ 1.



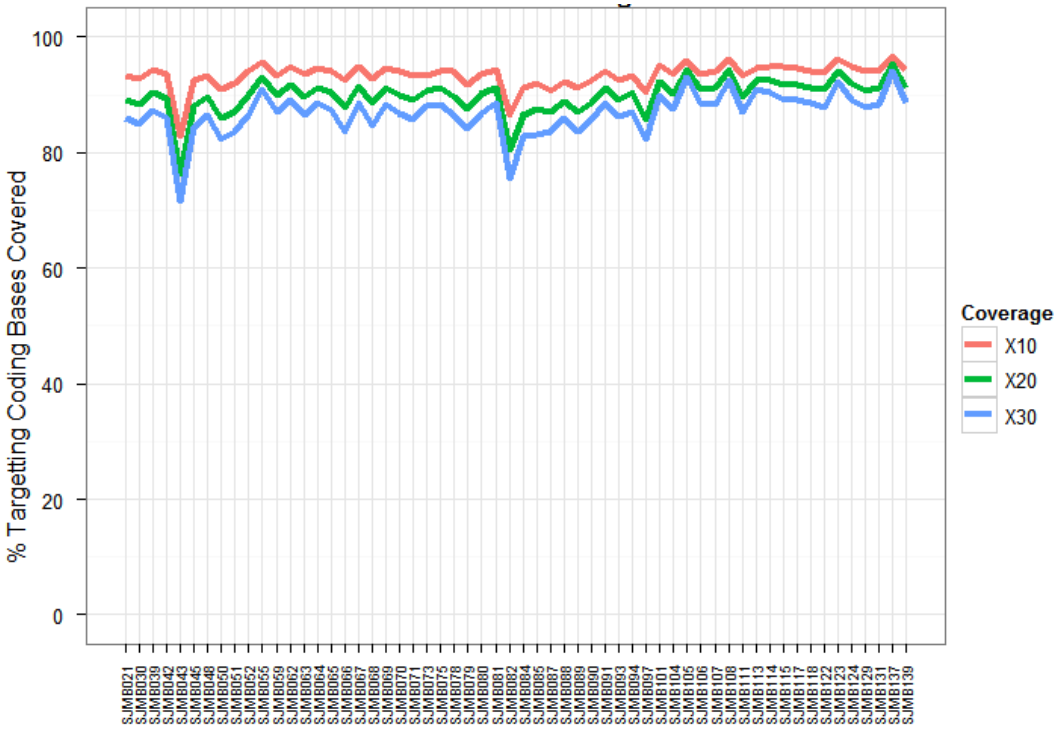
Supplementary Figure 3. Bootstrap subgroup assignment probability estimates.

Assigns subjects into four subgroups with at least five subjects. Each row of the heatmap shows the bootstrap probability estimate that the subject is assigned to the subgroup indicated by the column heading. The color scheme is 0 ≤ white ≤ 0.25 < gray ≤ 0.5 < blue ≤ 0.75 < yellow ≤ 0.9 < pink ≤ 0.95 < red ≤ 1.

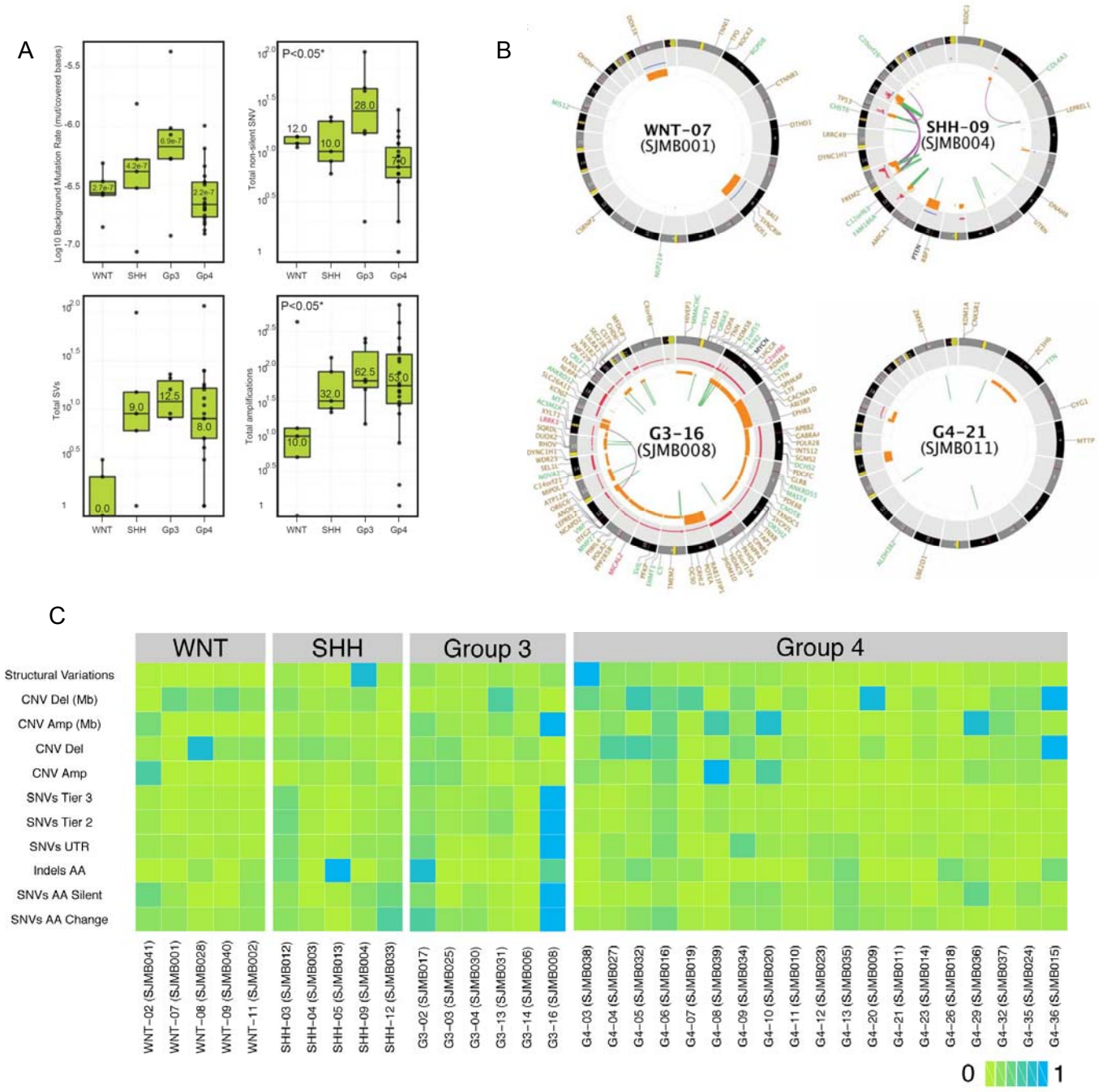


Supplementary Figure 4. Genome coverage.

Percent of whole-genome (top) and whole-exome (bottom) covered at 10x, 20x and 30x high-quality reads for the 37 WGS cases.

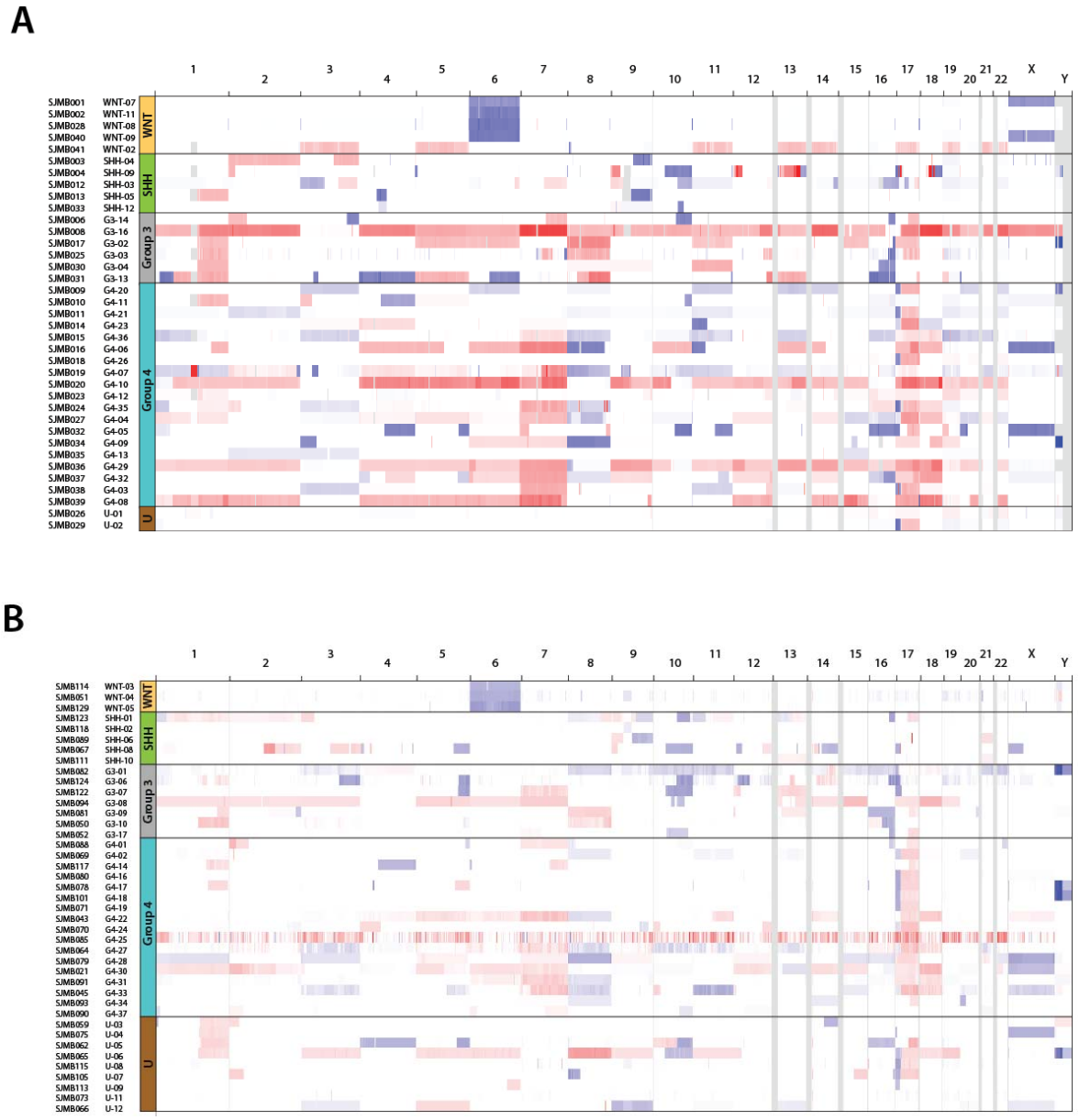


Supplementary Figure 5. Coverage of coding exons of the 136 genes selected for recurrence screening



Supplementary Figure 6. Relative abundance of somatic mutations by subgroup.

(a) Box plots (median [in box] ±2SD) of sequence variables in medulloblastoma subgroups. P-values are age and sex adjusted Kruskal Wallis test. (b) CIRCOS plots of exemplary medulloblastomas. Loss of heterozygosity (orange), amplifications (red), and deletions (blue). Interchromosomal translocations (green lines), intrachromosomal translocations (pink lines). Genes shown contained silent SNVs (green), missense SNVs (brown), nonsense SNVs (dark blue), and splice-site mutations (pink). (c) Heatmap view showing the relative abundance of somatically acquired sequence mutations (including SNVs and Indels), structural variations and copy number variations (including amplifications and deletions) in the 37 MB WGS samples. Each row represents one type of somatic lesion. The count for each class of somatic lesion across the 37 discovery cases was scaled to the interval of [0-1] to show the relative abundance in the discovery cohort.

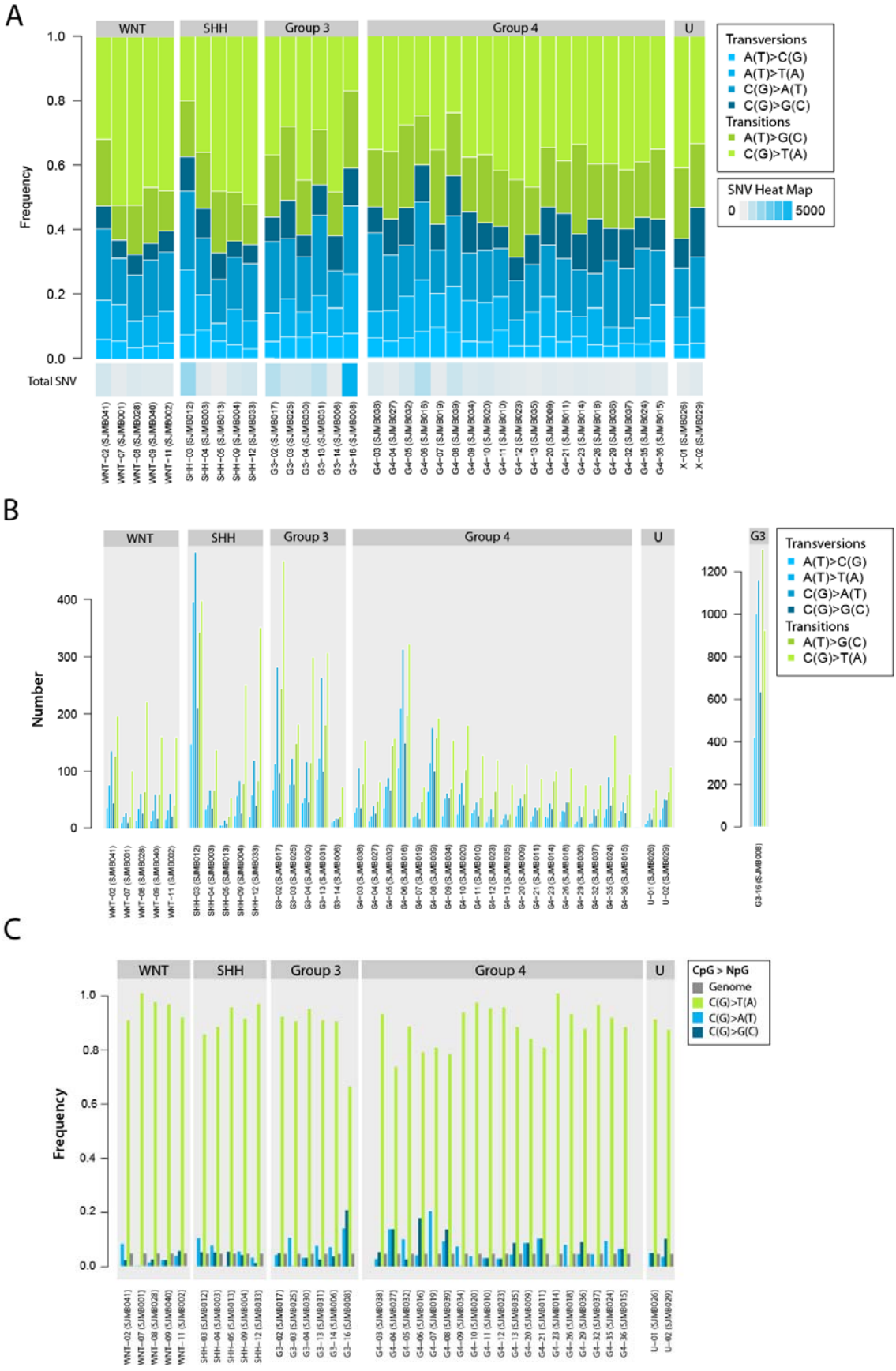


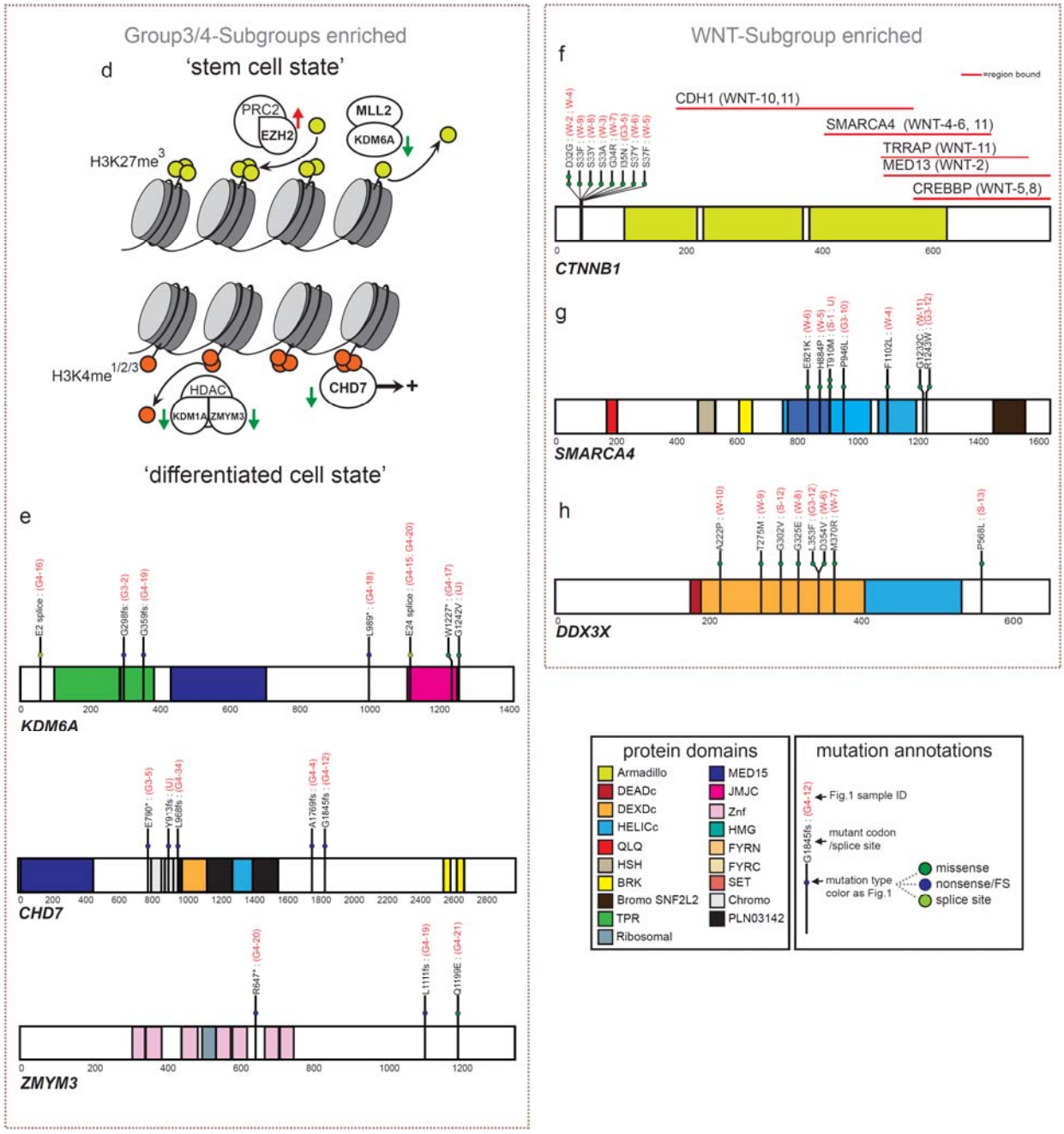
Supplementary Figure 7. Copy number heatmaps generated by for WGS and SNP cases

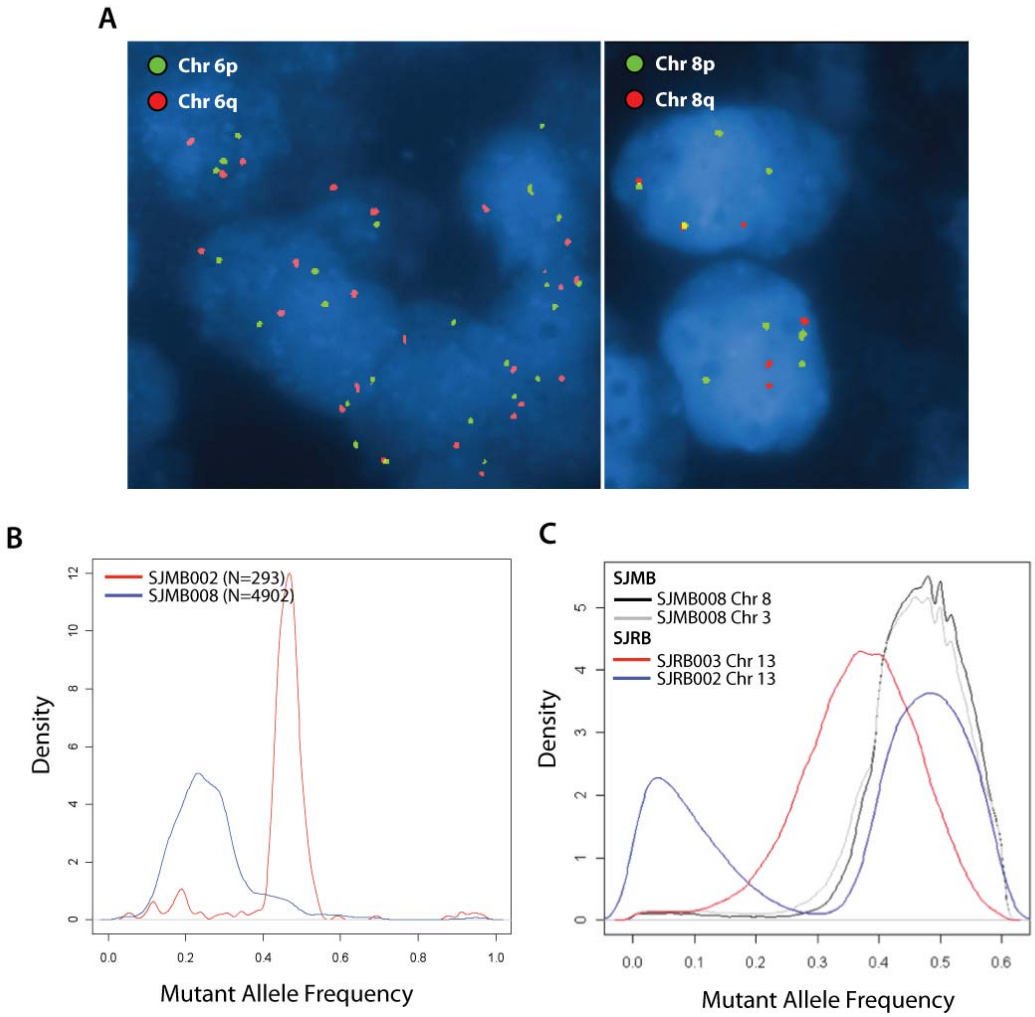
Copy number log2 ratio (i.e. log2 Tumor/Normal) derived from WGS and SNP 6.0 array shown in the IGV viewer (**http://www.broadinstitute.org/igv/**) with the color scale set to -2 to 2. (A) Copy number from MB discovery cases using WGS. (B) Copy number of MB validation cases using SNP 6.0 array.

Supplementary Figure 8. The mutation spectrum of Medulloblastoma

See next two pages: (a) The mutation spectrum of MB normalized (percentage) to the total for each sample shows a pattern typical of a spontaneously arising tumor, with no real evidence of mutations caused by ionizing radiation and alkylating agents. A heatmap of the number of total SNVs is displayed below the main graph. This highlights that those samples with a larger number of SNVs tend to have a different substitution pattern with more transversions than other samples. **(b)** Raw numbers of mutations (G3-16 (SJMB008) is separated to the right). **(c)** The fraction of mutations in CpGs to NpGs (Where N is ATCG). The genome fraction in is obtained from[70]. **(d)** Alterations predicted to increase (red arrow) or decrease (green arrow) the function of genes that read, write and erase histone lysine methylation marks in subgroups-3 and 4 (see text for details). **(e)** Domain structure of KDM6A, CHD7, ZMYM3 showing mutations. **(f)** WNT-subgroup tumours were enriched for mutations in CTNNB1, and CDH1, SMARCA4, MED13 and CREBBP that bind CTNNB1 (red lines). Mutated tumours are indicated. **(g,h)** Domain structure of SMARCA4 and DDX3X showing mutations.


A
WNT
SHH
Group 3
Group 4
U
Frequency
Total SNV
Transversions
A(T)>C(G)
A(T)>T(A)
C(G)>A(T)
C(G)>G(C)
Transitions
A(T)>G(C)
C(G)>T(A)
SNV Heat Map
0
5000
B
Number
G3
C
CpG > NpG
Genome
C(G)>T(A)
C(G)>A(T)
C(G)>G(C)

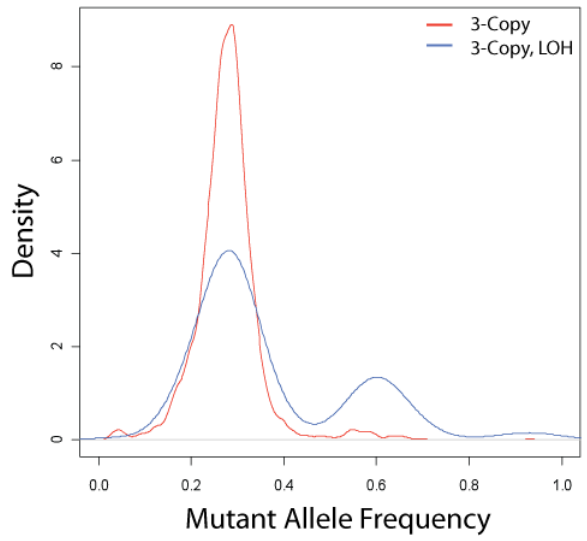

Group3/4-Subgroups enriched
d
'stem cell state'
PRC2
EZH2
MLL2
KDM6A
H3K27me3
H3K4me1/2/3
HDAC
KDM1A
ZMYM3
CHD7
'differentiated cell state'
e
KDM6A
CHD7
ZMYM3
WNT-Subgroup enriched
f
CDH1 (WNT-10,11)
SMARCA4 (WNT-4-6, 11)
TRRAP (WNT-11)
MED13 (WNT-2)
CREBBP (WNT-5,8)
CTNNB1
g
SMARCA4
h
DDX3X
protein domains
Armadillo
DEADc
DEXDc
HELICc
QLQ
HSH
BRK
Bromo SNF2L2
TPR
Ribosomal
MED15
JMJC
Znf
HMG
FYRN
FYRC
SET
Chromo
PLN03142
mutation annotations
Fig.1 sample ID
mutant codon /splice site
mutation type color as Fig.1
missense
nonsense/FS
splice site



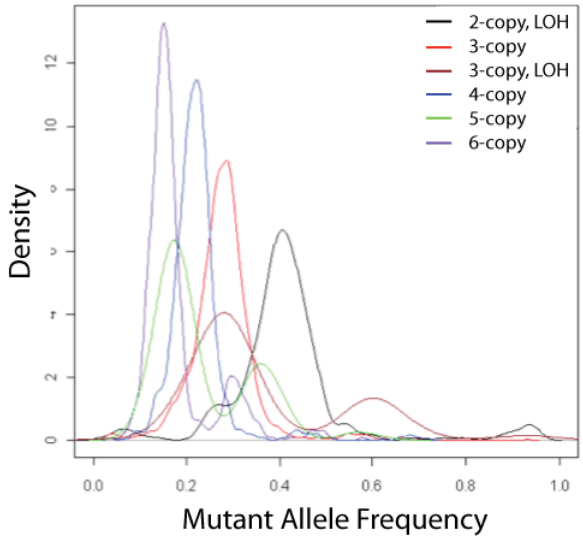
Supplementary Figure 9. Copy number alteration, mutant allele fraction and LOH in SJMB008 (G3-16).

(a) FISH analysis for chromosomes 6 (left) and 8 (right) shows amplification of both chromosomes. Green and red colors are for probes of chromosome p and q arms. **(b)** Density plot of mutant allele fraction (i.e. # of reads with the mutant allele/# of total reads) for validated tier3 autosomal SNV mutations found in SJMB008 (n=4902, blue) and SJMB002 (n=306, red). The highest peak for SJMB002 is at 0.5 while that of SJMB008 is at 0.25. **(c)** AI (allelic imbalance) distribution of chromosomes 3 and 8 for SJMB008 (black and gray) and chromosome 13 for two retinoblastoma (RB) samples: SJRB003 (red) and RJB002 (blue). SJRB003 is expected to have 95% tumor purity based on mutant allele read count in RB1 (61 out of 64) while the purity for SJRB002 is expected to be 77% (111 out of 144). SJRB003 has copy neutral LOH at 13q while SJRB002 has copy neutral LOH at the entire chromosome 13.

A



B

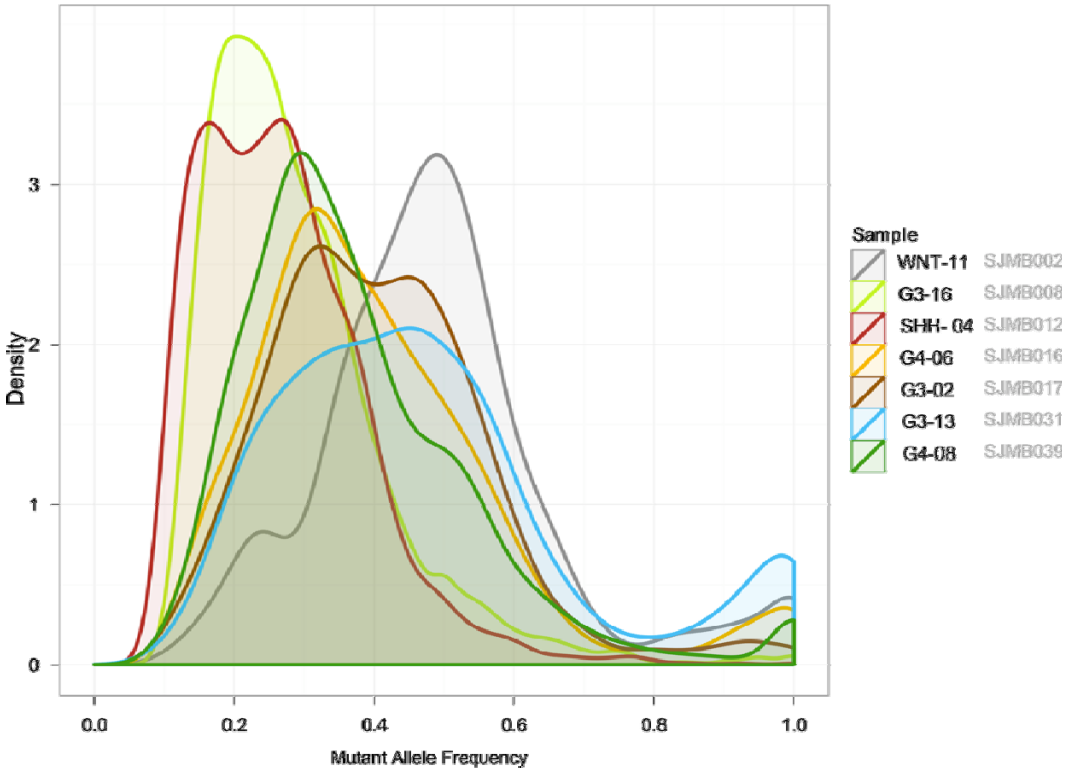


C

Color	CN A	# SNVs	% of Total SNVs	Chromosomes	% of Autosomes	Peak*	Interpretation
Black	2 & LOH	300	6.10%	3, 16q, 17p	9.00%	0.41	1/2
Red	3	1860	37.90%	1p,5,6,9,10,11,12,13,15,16,20	44.70%	0.29	1/3
Brown	3 & LOH	189	3.90%	8	5.20%	0.29, 0.60*	1/3, 2/3
Blue	4	1485	30.30%	1q,2,4,9,11,12,13,17q,19,21	27.00%	0.22	1/4
Green	5	179	3.60%	14	3.40%	0.18, 0.36*	1/5, 2/5
Purple	6	522	10.60%	7,18	8.40%	0.15,0.30*	1/6, 2/6

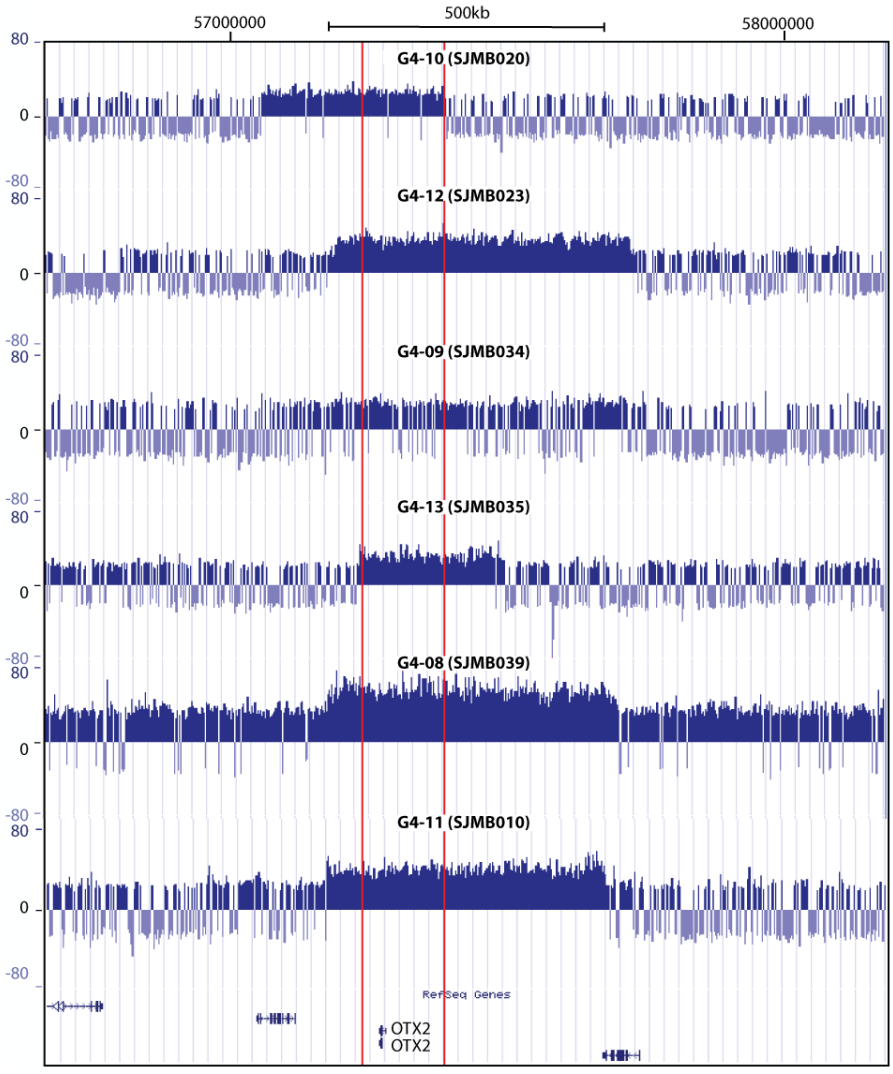
Supplementary Figure 10. Relationship of copy number alteration and mutant allele fraction in SJMB008 (G3-16)

Mutant allele fraction was derived from deep sequencing result of custom capture. **(a)** Density plot of mutant allele fraction for somatic sequence mutations in genomic regions with estimated 3 copy number. Mutations in regions of LOH are shown in blue and there are two peaks: a dominant peak at 0.29 and a minor peak at 0.60. **(b)** Density plot for all validated somatic sequence mutations colored by their chromosome CNA status. **(c)** Summary for CNA and mutant allele fraction for plot (b). The minor peak of a double-peak plot is marked by a *. The column “Interpretation” provides the closest rounding for mutant allele fraction observed at each peak.



Supplementary Figure 11. Comparison of mutant allele frequencies across the samples with the highest number of total mutations

Samples with high numbers of total sequence mutations have unique mutant allele frequencies when compared to other samples in the cohort like WNT-11 (SJMB002) (in gray). Mutant allele frequencies are lower than would be expected, 0.5, if one allele was affected. The mutant allele frequencies were calculated using WGS data since not all samples were subject to capture validation. In the case of these top 6 cases, it appears that a genome wide amplification event has occurred before the acquisition of the majority of sequence mutations resulting in a lower overall frequency of these aberrations.



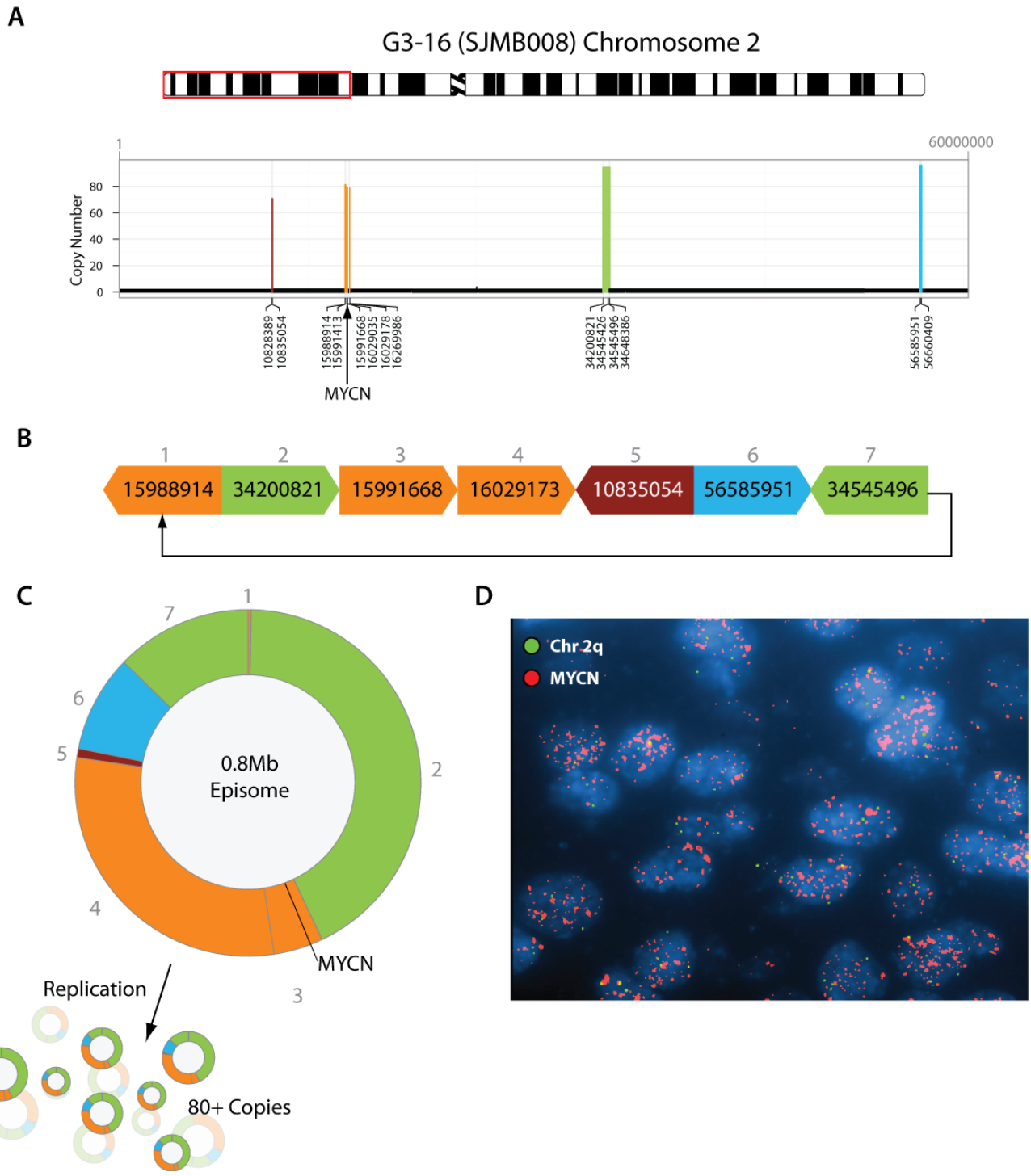
Supplementary Figure 12. OTX2 copy number variations

OTX2 has been previously identified as an oncogene in Medulloblastoma and accordingly amplification was detected in 6 cases all belonging to Group 4 of the disease. Amplification of OTX2 in G4-09 (SJMB034) was detected by manual review of this region.

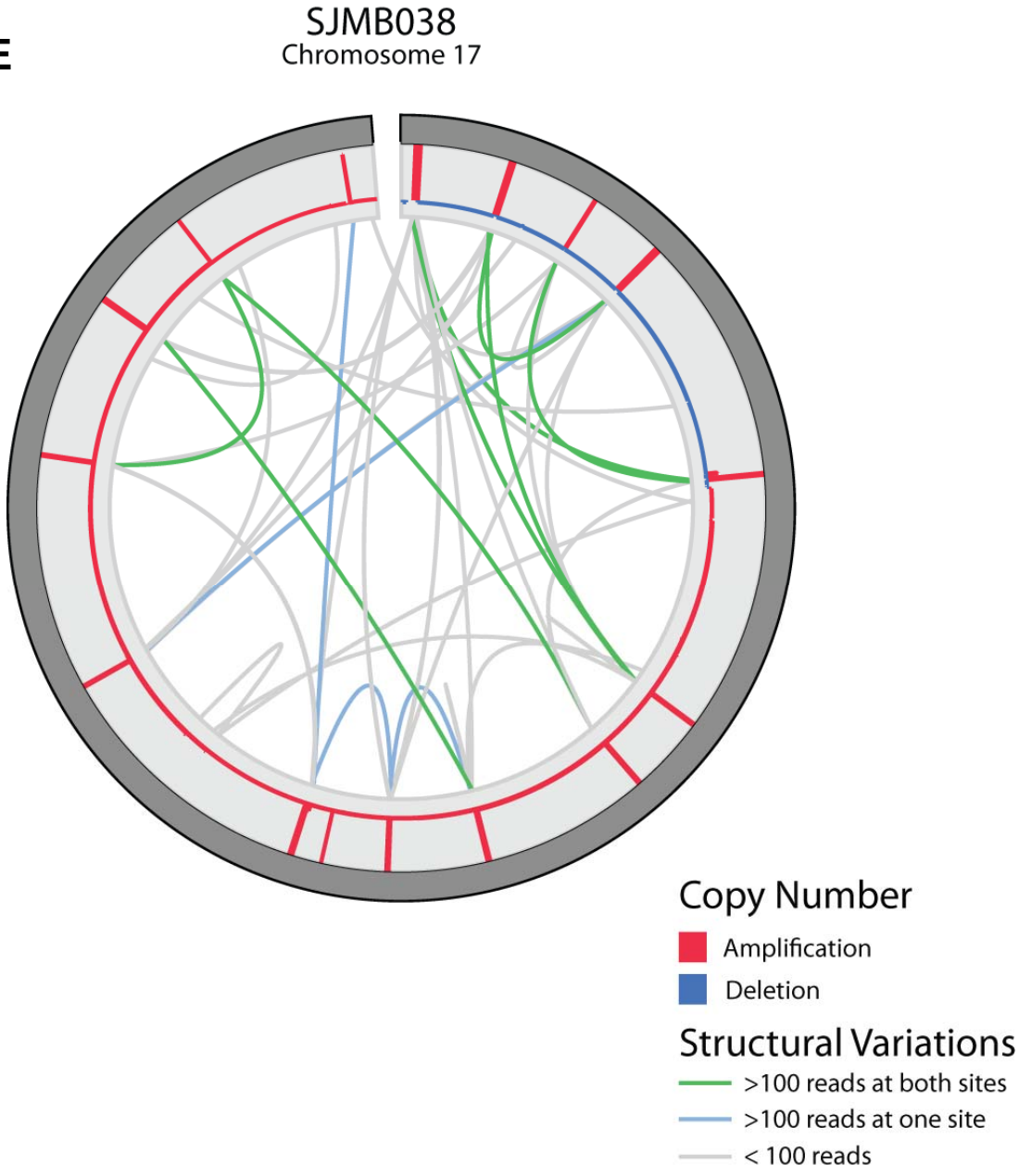
Supplementary Figure 13. Chromothripsis in SJMB008 and 38.

See next two pages: Evidence for chromothripsis in SJMB008. **(a)** High level copy number gains and structural variations indicate a "shattering" of chromosome 2 which are subsequently pieced together in the pattern shown in **(b).** The pattern of these breakages and rejoins implies the formation of an episome which is replicated repeatedly **(c)** giving rise to the FISH pattern seen in **(d)**. **(e)** Chromosome 17 in SJMB038 shows a pattern consistent with a chromothripsis event. A large number of intra-chromosomal translocations were detected in this chromosome in concert with large copy number "spikes" (2^{nd} track - red) and >100 reads supporting some of the structural variations (Green). This pattern is similar to that shown in SMB008, shattering of chromosome 17 caused the formation of an episome which is subsequently replicated many times.

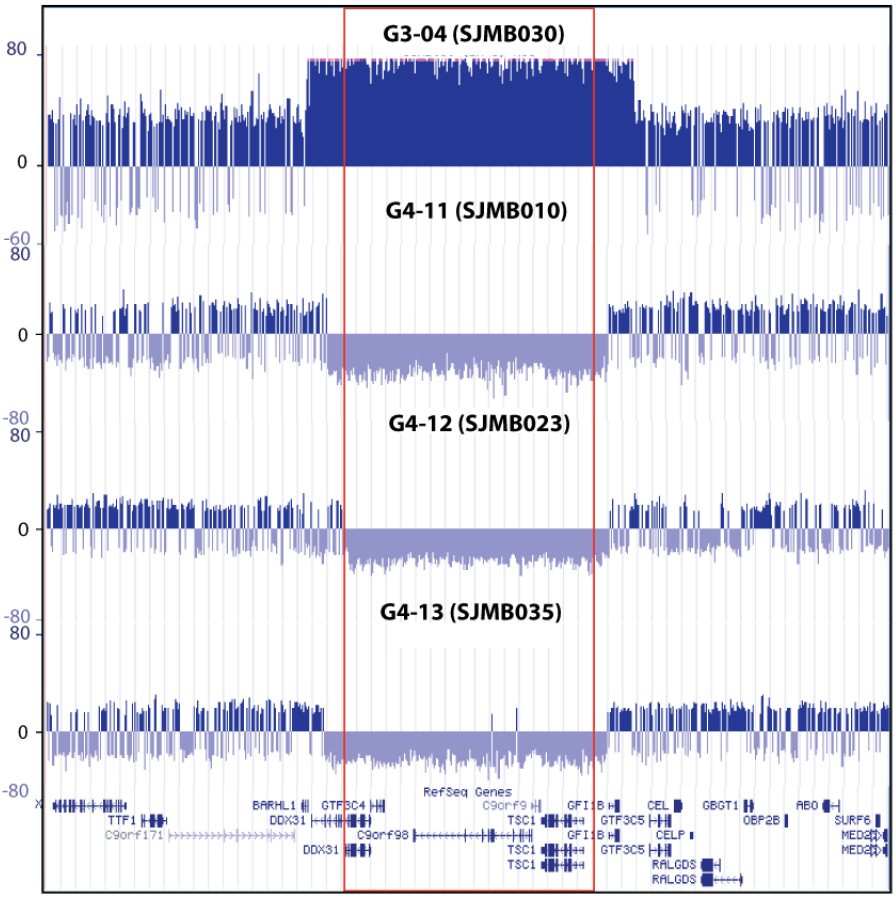
Chromosome 2 image By David Adler (University of Washington, Pathology, 1994)


A
G3-16 (SJMB008) Chromosome 2
1
60000000
Copy Number
80
60
40
20
0
10828389
10835054
15988914
15991413
15991668
16029035
16029178
16269986
MYCN
34200821
34545426
34545496
34648386
56585951
56660409
B
1
2
3
4
5
6
7
15988914
34200821
15991668
16029173
10835054
56585951
34545496
C
0.8Mb
Episome
MYCN
Replication
80+ Copies
D
Chr 2q
MYCN

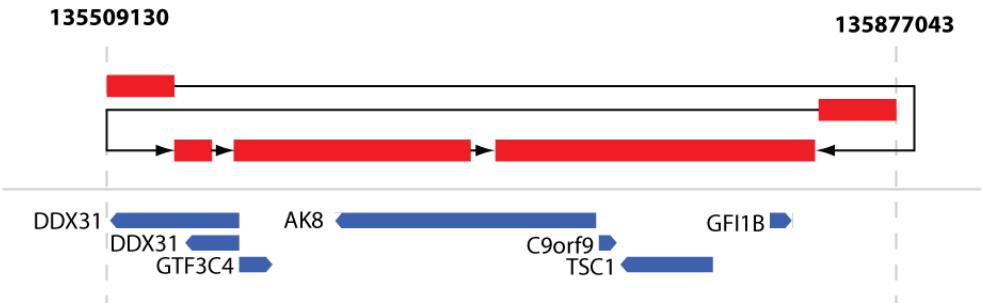
E


SJMB038
Chromosome 17
Copy Number
Amplification
Deletion
Structural Variations
>100 reads at both sites
>100 reads at one site
< 100 reads

A



B



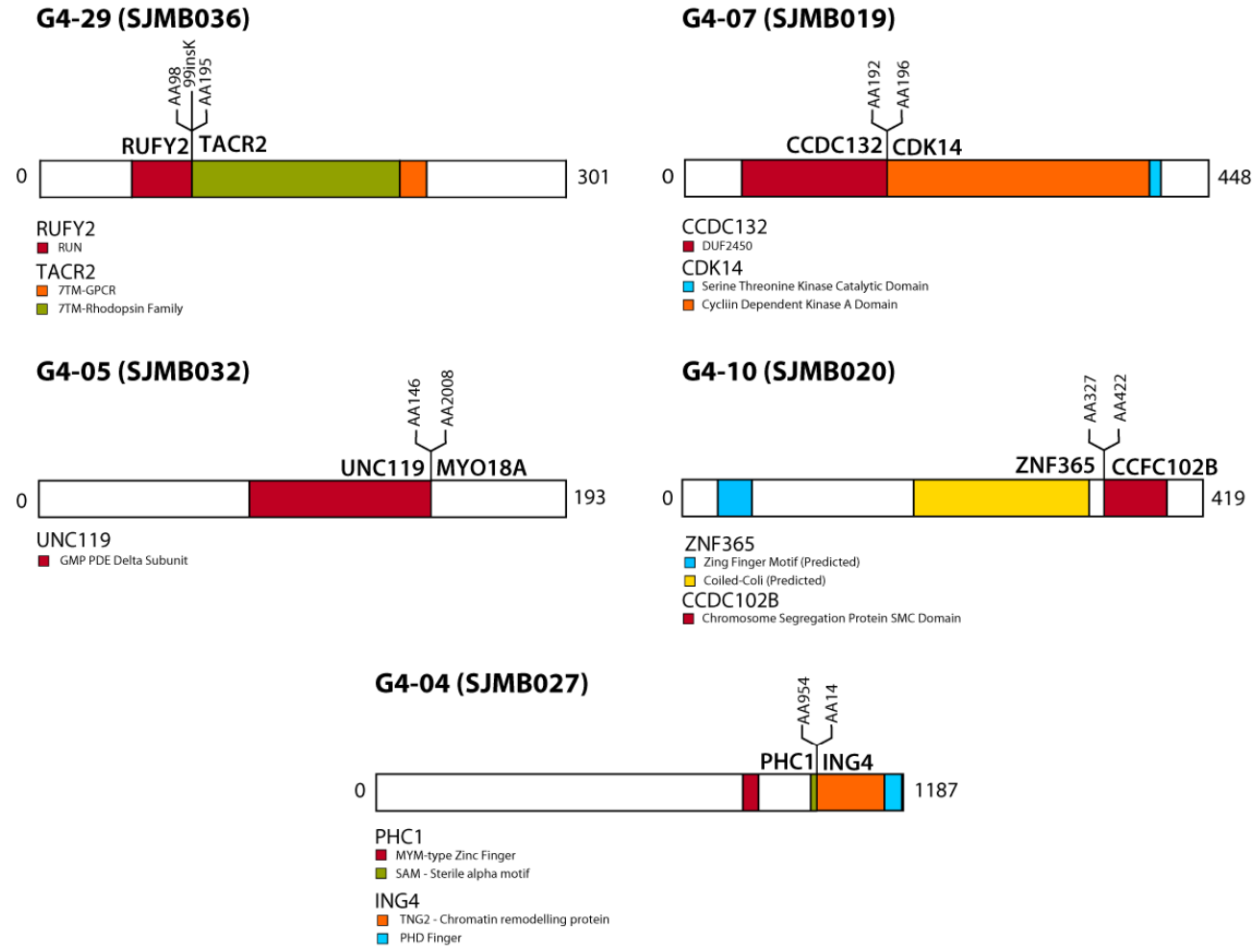
Supplementary Figure 14. *DDX31* copy number variations

Recurrent copy number alterations in DDX31, AK8, and TSC1. **(a)** 3 samples from Group 4 have copy number loss and SJMB030 (Group 3) has a gain in this region. The MOR in each of these samples and encompasses these three genes and in all cases the breakpoint falls within DDX31. **(b)** Breakpoints indentified by CREST in DDX31 in an additional sample; SJMB026 (U-01).

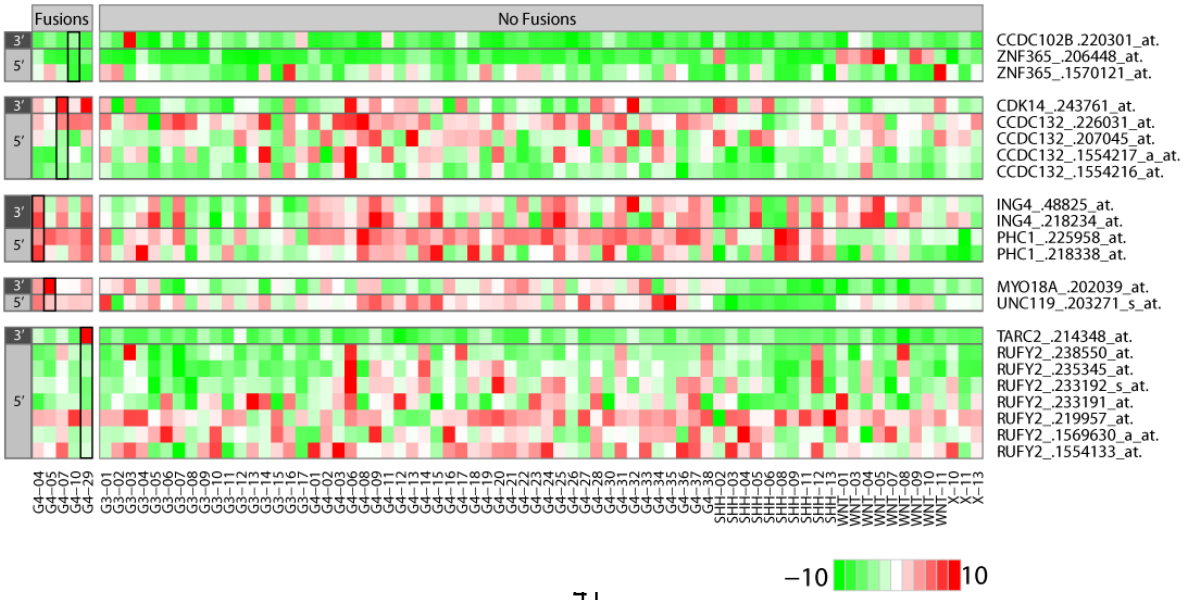
Supplementary Figure 15. Genomically validated in-frame fusions proteins detected by CREST

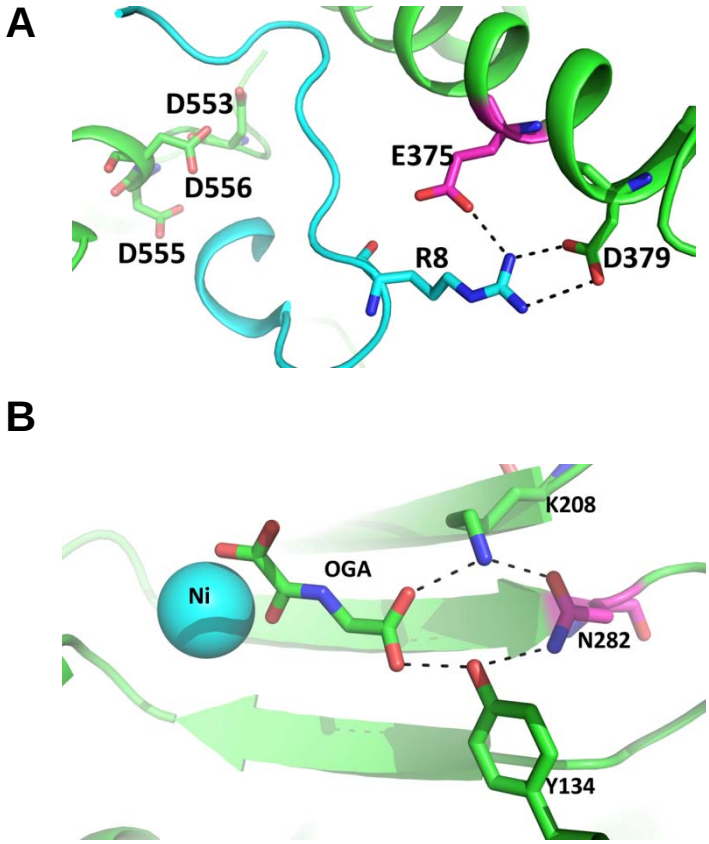
See next page: (a) 5 In-frame fusion proteins were discovered in the SJMB cohort, all of which were validated at the genomic level by sanger sequencing. **(b)** Expression of the gene involved in the fusions (log2 expression values are rescaled within each gene), in each case 5' and 3' partners are grouped and are outlined in black in the corresponding sample with that fusion. In two cases (G4-29 and G4-05) the 3' fusion partner was overexpressed with respect to a number of other MB samples in this cohort, indicating the fusion is causing aberrant expression of these genes.

A


G4-29 (SJMB036)
AA98
99insK
AA195
RUFY2
TACR2
0
301
RUFY2
RUN
TACR2
7TM-GPCR
7TM-Rhodopsin Family
G4-07 (SJMB019)
AA192
AA196
CCDC132
CDK14
0
448
CCDC132
DUF2450
CDK14
Serine Threonine Kinase Catalytic Domain
Cycliin Dependent Kinase A Domain
G4-05 (SJMB032)
AA146
AA2008
UNC119
MYO18A
0
193
UNC119
GMP PDE Delta Subunit
G4-10 (SJMB020)
AA327
AA422
ZNF365
CCFC102B
0
419
ZNF365
Zing Finger Motif (Predicted)
Coiled-Coli (Predicted)
CCDC102B
Chromosome Segregation Protein SMC Domain
G4-04 (SJMB027)
AA954
AA14
PHC1
ING4
0
1187
PHC1
MYM-type Zinc Finger
SAM - Sterile alpha motif
ING4
TNG2 - Chromatin remodelling protein
PHD Finger

B


Fusions
No Fusions
3′
5′
CCDC102B .220301_at.
ZNF365_.206448_at.
ZNF365_.1570121_at.
CDK14_.243761_at.
CCDC132_.226031_at.
CCDC132_.207045_at.
CCDC132_.1554217_a_at.
CCDC132_.1554216_at.
ING4_.48825_at.
ING4_.218234_at.
PHC1_.225958_at.
PHC1_.218338_at.
MYO18A_.202039_at.
UNC119_.203271_s_at.
TARC2_.214348_at.
RUFY2_.238550_at.
RUFY2_.235345_at.
RUFY2_.233192_s_at.
RUFY2_.233191_at.
RUFY2_.219957_at.
RUFY2_.1569630_a_at.
RUFY2_.1554133_at.
G4–04
G4–05
G4–07
G4–10
G4–29
G3–01
G3–02
G3–03
G3–04
G3–05
G3–06
G3–07
G3–08
G3–09
G3–10
G3–11
G3–12
G3–13
G3–14
G3–15
G3–16
G3–17
G4–01
G4–02
G4–03
G4–06
G4–08
G4–09
G4–11
G4–12
G4–13
G4–14
G4–15
G4–16
G4–17
G4–18
G4–19
G4–20
G4–21
G4–22
G4–23
G4–24
G4–25
G4–26
G4–27
G4–28
G4–30
G4–31
G4–32
G4–33
G4–34
G4–35
G4–36
G4–37
G4–38
SHH–02
SHH–03
SHH–04
SHH–05
SHH–06
SHH–08
SHH–09
SHH–11
SHH–12
SHH–13
WNT–01
WNT–03
WNT–04
WNT–05
WNT–07
WNT–08
WNT–09
WNT–10
WNT–11
X–10
X–11
X–13
−10
10



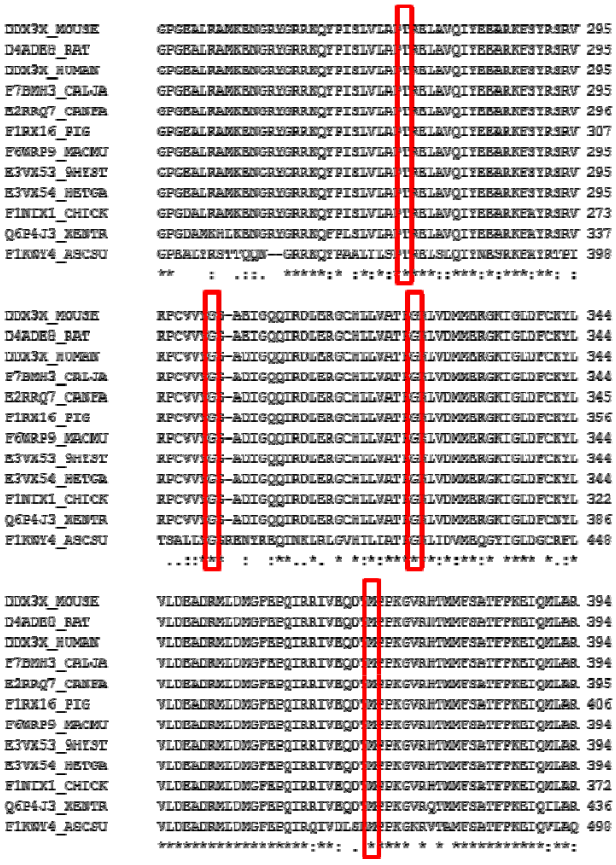
Supplementary Figure 16. Structural analysis of KDM mutations.

Recurrent mutations were identified in genes of the Lysine demethylase family (termed KDM genes) from 37 Medulloblastoma (Mb) whole-genome data sets. Two occur in KDM proteins (*KDM1A* and *KDM4C*) for which structural data exists to analyze the functional consequences of these mutations. **(a)** The missense mutation in *KDM1A* (E375K) occurs within the substrate binding pocket and the mutation is likely to cause steric and electrostatic clashes that inhibit substrate binding. **(b)** The mutation in *KDM4C* (N282I) occurs in the active site and is predicted to disrupt binding of a critical co-factor needed for catalysis. These mutations appear to potentially destabilize key substrate and cofactor interactions within the active sites of both structures, adversely affecting Lysine demethylase function.

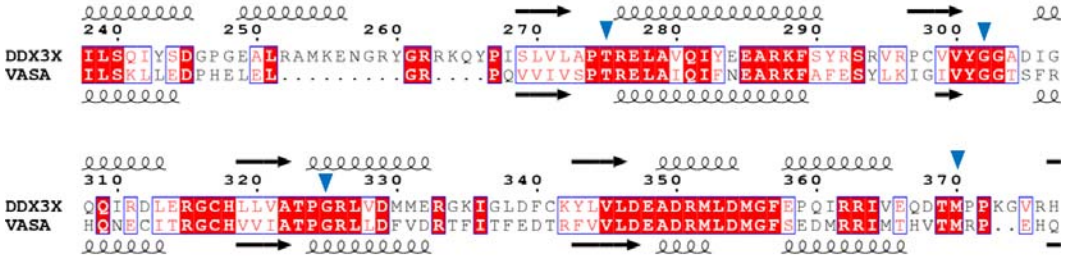
Supplementary Figure 17. DDX3X additional mutational analysis

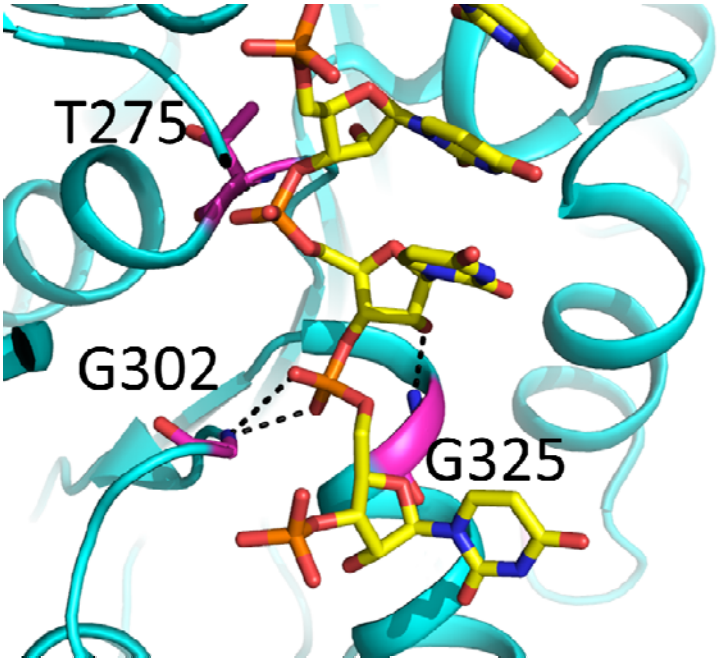
See next page: (a) Sub-section of DDX3X representing mutations in an amino acid sequence alignment of 12 orthologs of human DDX3X. Residues identified as missense mutations in the WGS cohort (T275M, G302V, G325E, and M370R) are indicated with a red box around them. The sequences are annotated with Uniprot entry name and species respectively: O00571_HUMAN Homo sapiens (Human), Q62167_MOUSE Mus musculus (Mouse), F1KWY4_ASCSU Ascarissuum (Pig roundworm), F7BMH3_CALJA Callithrix jacchus (White-tufted-ear marmoset), E2RRQ7_CANFA Canis familiaris (Dog) (Canis lupus familiaris), F6WRP9_MACMU Macacamulatta (Rhesus macaque)F1RX16_PIG Suss crofa (Pig), E3VX53_9HYST Fukomysanselli (Ansell's mole-rat), E3VX54_HETGA Heterocephalusglaber (Naked mole rat), D4ADE8_RAT Rattus norvegicus (Rat), F1NIX1_CHICK Gallus gallus (Chicken), Q6P4J3_XENTR Xenopus tropicalis (Western clawed frog). **(b)** Sequence alignment of a section of the helicase domain in DDX3X (PDB ID: 2I4I) and a structural homolog Vasa (PDB ID: 2BD3). To focus on region where the mutations are present the entire alignment is not shown. Secondary structures α-helices (curls) and β-sheets (arrows) are indicated above the alignment for DDX3X and below for Vasa. Residues that are white with red background are identities and similar residues are in red with a white background. Residues identified as missense mutations (T275M, G302V, G325E, and M370R) are indicated by a blue triangle above the position in the sequence. Arrows represent the mutation sites.

A



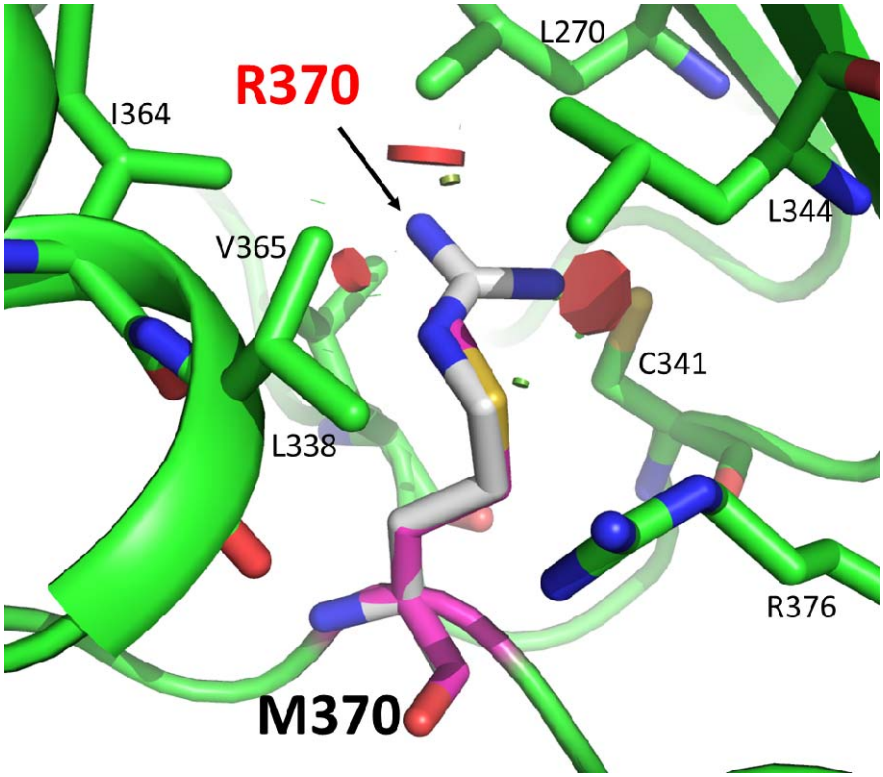
B


240
250
260
270
280
290
300
DDX3X
VASA
310
320
330
340
350
360
370
DDX3X
VASA



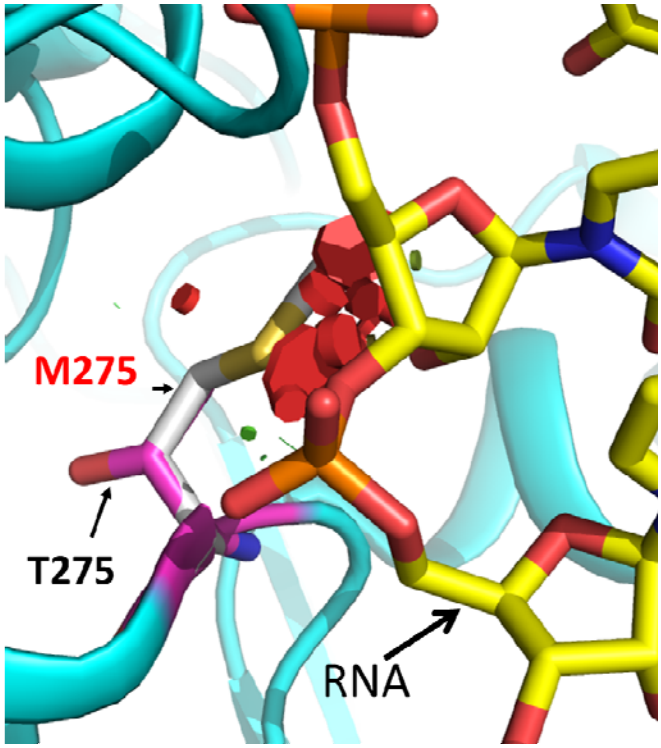
Supplementary Figure 18. T275, G302 and G325 of DDX3X stabilize bound RNA.

Locations of observed missense mutations Thr 275, Gly 302, and Gly 325 (T275, G302, and G325 respectively; purple sticks) in DDX3X are involved in stabilizing bound RNA (yellow sticks). Dotted lines indicate hydrogen bonds between amino acids and RNA.



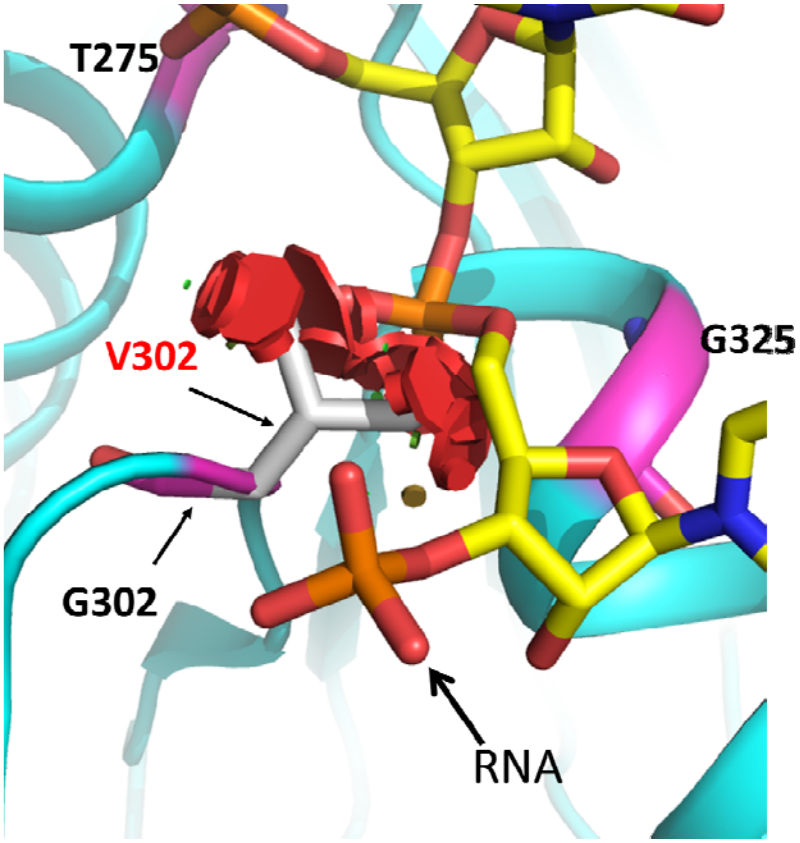
Supplementary Figure 19. DDX3X mutation M370R.

Protein DDX3X (PDB ID: 2I4I; green) is modeled with a focus on residue Met 370 (M370; purple stick). Methionine 370 is surrounded by the hydrophobic residues Ile 364 (I364), Val 365 (V365), Leu 338 (L338), Cys 341 (C341), Leu 270 (L270), Leu 344 (L344) and the polar residue Arg 376 (R376). Mutation of Met 370 to Arg 370 (R370; white stick and red label) creates a number of steric (red disks) clashes within the core of the protein. The presence of a polar residue in a hydrophobic pocket would be very destabilizing.



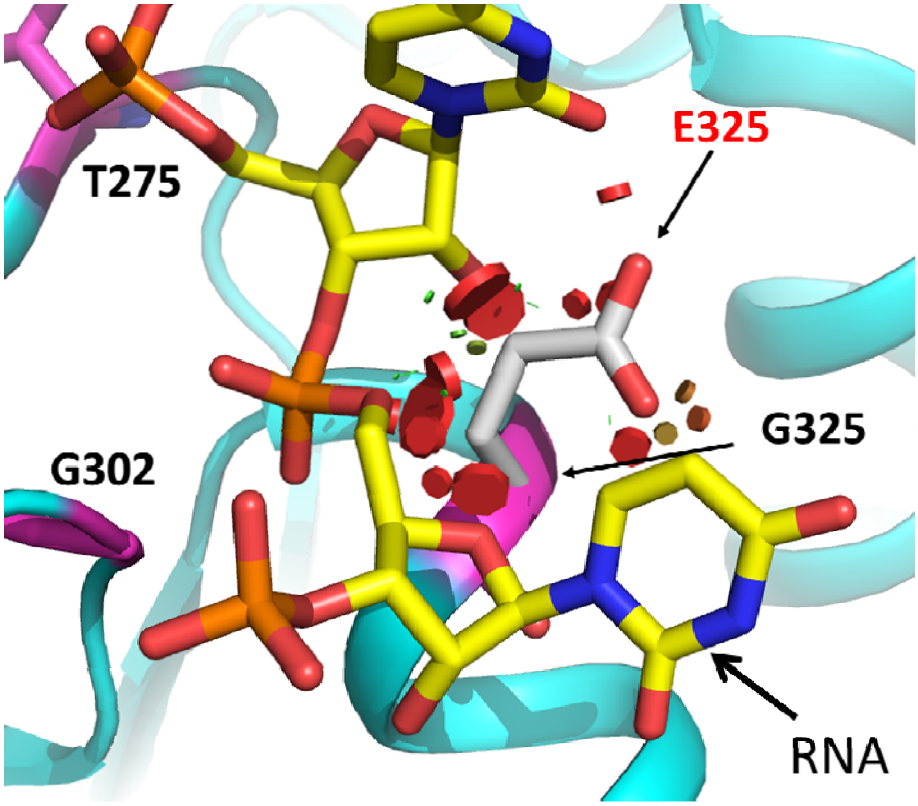
Supplementary Figure 20. DDX3X mutation T275M

The residue Thr275 (T275; purple stick) does not directly hydrogen bond with the RNA but stabilizes the amino termi of its helix. Mutation to Met275 (M275; white stick and red label) would cause large steric clashes (red disks) with the bound RNA (yellow sticks).



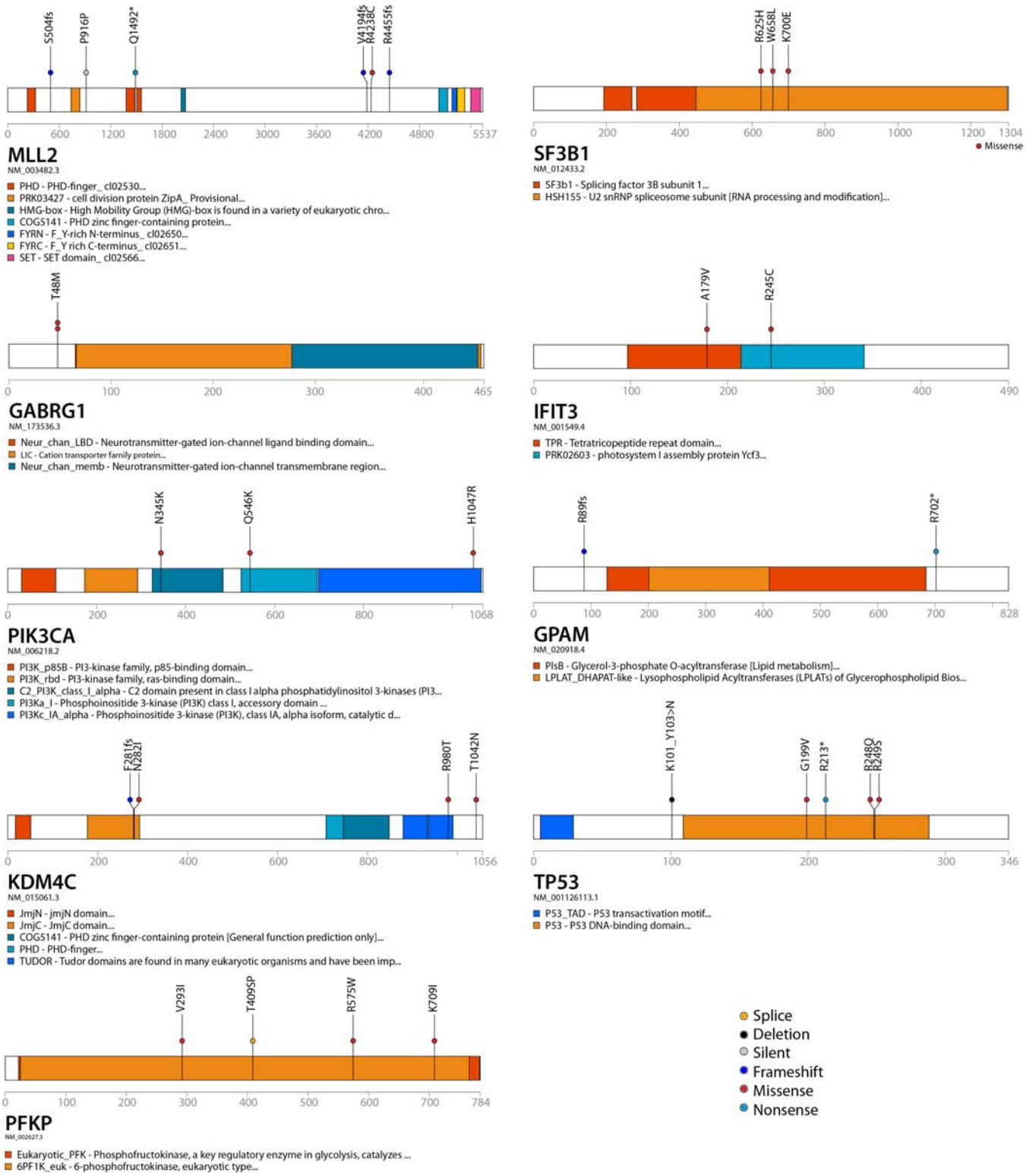
Supplementary Figure 21. DDX3X mutation G302V.

Mutation of the Gly302 (G302; purple stick) to a Val (V302; white stick and red label) mutant in white would cause a loss of the hydrogen bond from the amide of Gly302 as well as form steric clashes(red disks) with the bound RNA (yellow sticks). Additional positions of missense mutations also are present in purple (T275 and G325).



Supplementary Figure 22. DDX3X mutation G325E.

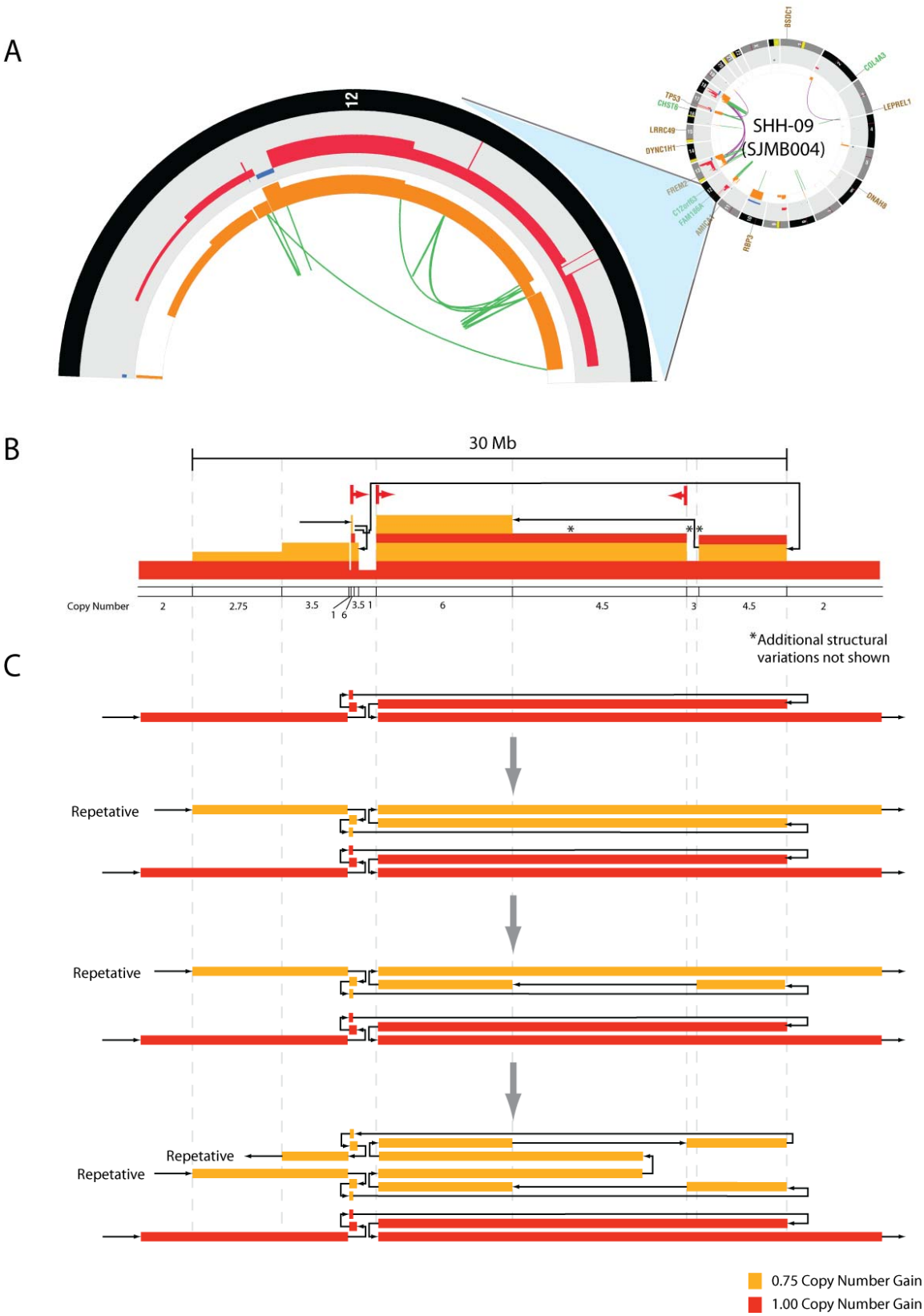
Mutation of Gly 325 (G325; purple stick) to Glu 325 (E325; white stick and red label) creates a number of steric (red disks) and electrostatic clashes with the RNA (yellow sticks). Additional positions of missense mutations also are present in purple (T275 and G325).

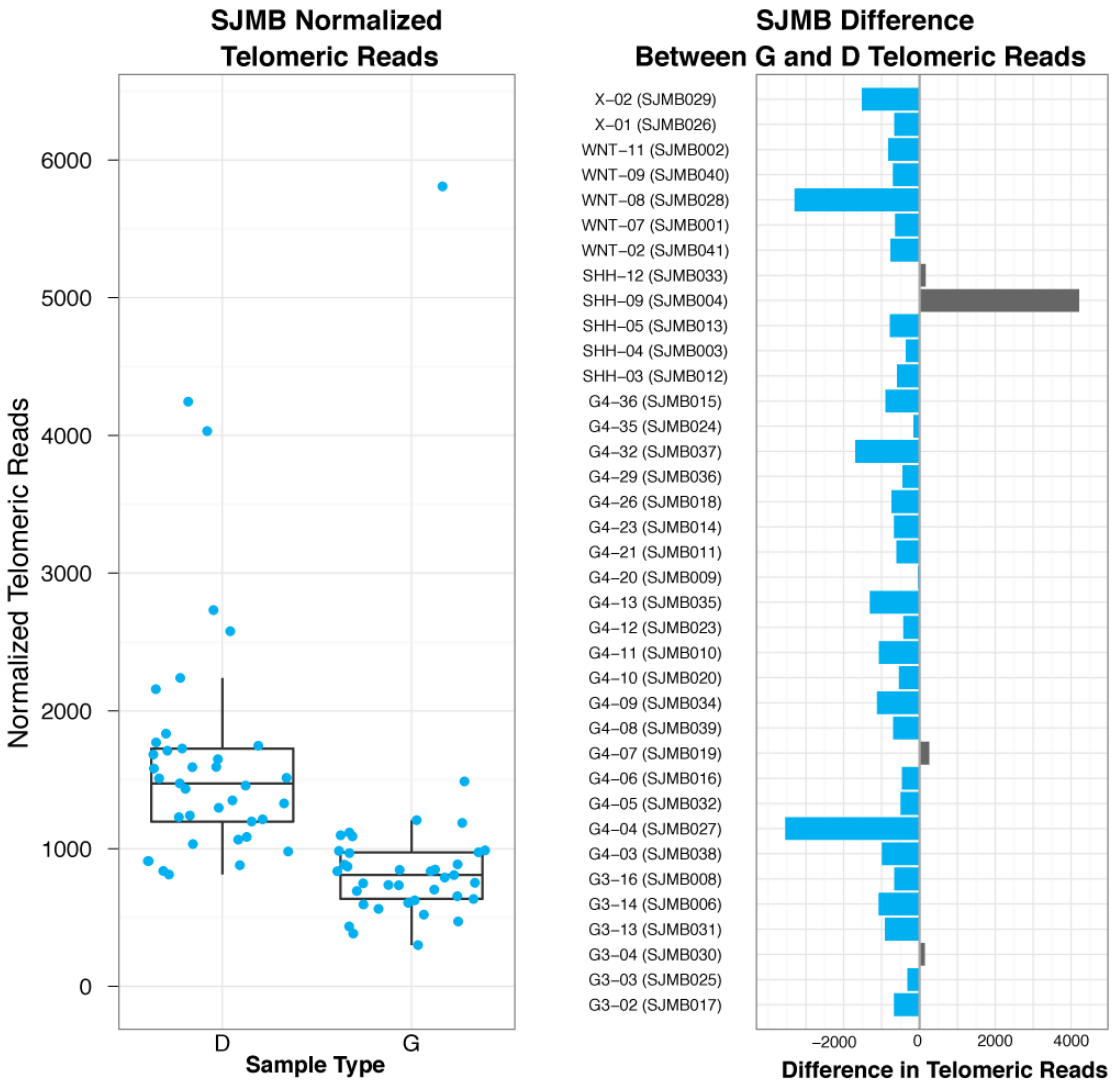


Supplementary Figure 23. Additional protein plots for recurrently mutated genes in Figure 1.

Supplementary Figure 24. Complex copy number variation in SHH-09 (SJMB004)

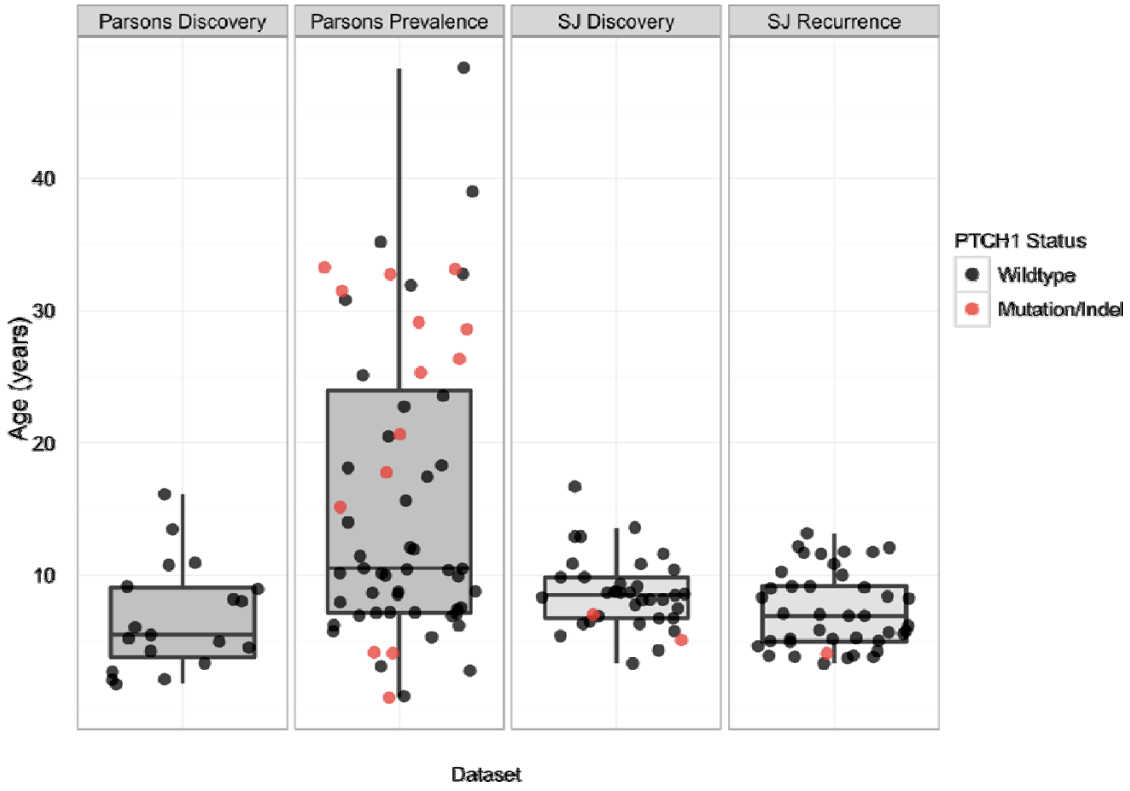
(a). Chromosome 12 in SHH-09 has multiple copy number states ranging from 1 to 6 copies with supporting structural variations **(b)** . In an attempt to understand the copy number alterations that lead to this interesting pattern we hypothesized a sequence of events that could lead to this pattern, shown in **(c): Step 1:** A initial set of complicated structural variations (including the fold-back type SV) occurred on one copy of chr12.
Step 2: The modified chromosome was amplified (in 75% of the tumor cells).
Step 3: A small segment was deleted on the amplified chromosome.
Step 4: A new round of fold-back SV occurred on the amplified chromosome and the final amplified chromosome was predicted in 75% of the tumor cells.


A
12
SHH-09
(SJMB004)
BSDC1
COL4A3
LEPREL1
DNAH8
RBP3
AMICA1
FREM2
LRRC49
DYNC1H1
TP53
CHST6
B
30 Mb
Copy Number
2
2.75
3.5
1 6
3.5 1
6
4.5
3
4.5
2
*Additional structural
variations not shown
C
Repetative
Repetative
Repetative
Repetative
0.75 Copy Number Gain
1.00 Copy Number Gain



Supplementary Figure 25. Estimation of total telomere content using WGS data

A. In general telomeric sequences appear reduced in the tumour sample as expected. **B.** Comparing the reads of the germline and diagnostic sample in each patient reveals an outlier sample – SHH-09 (SJMB004) which has over-represented telomeric sequence when compared to germline control.



Supplementary Figure 26. Comparison of ages of patients sequenced.

Comparison of patient ages in this cohort and the cohort sequenced by Parson's *et. al.,* (main manuscript ref. 42). The broad age range of the Prevalence samples used by Parsons may account for the differences in mutations discovered (median from left to right; 5.5, 10.5, 8.5, 6.9, Kruskal-Walls p-value=2.2e-16). Red points indicate those samples with one or more PTCH1 mutations. (The two mutations in PTCH1 found in the Parson's discovery cohort were in the same sample and this sample has no age information available).

SUPPLEMENTARY REFERENCES

51 Johnson, R. A. *et al.* Cross-species genomics matches driver mutations and cell compartments to model ependymoma. *Nature.* **466**, 632-636. (2010).

52 McShane, L. M. *et al.* Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics* **18**, 1462-1469, doi:10.1093/bioinformatics/18.11.1462 (2002).

53 Dudoit, S. & Fridlyand, J. Bagging to improve the accuracy of a clustering procedure. *Bioinformatics* **19**, 1090-1099, doi:10.1093/bioinformatics/btg038 (2003).

54 Ellison, D. *et al.* Medulloblastoma: clinicopathological correlates of SHH, WNT, and non-SHH/WNT molecular subgroups. *Acta Neuropathologica* **121**, 381-396, doi:10.1007/s00401-011-0800-8 (2011).

55 Schmittgen, T. D. & Livak, K. J. Analyzing real-time PCR data by the comparative C(T) method. *Nat Protoc.* **3**, 1101-1108. (2008).

56 Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999-1005, doi:http://www.nature.com/nature/journal/v464/n7291/suppinfo/nature08989_S1.html (2010).

57 Mardis, E. R. *et al.* Recurring Mutations Found by Sequencing an Acute Myeloid Leukemia Genome. *New England Journal of Medicine* **361**, 1058-1066, doi:doi:10.1056/NEJMoa0903840 (2009).

58 Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol.* **132**, 365-386. (2000).

59 Castle, J. *et al.* DNA copy number, including telomeres and mitochondria, assayed using next-generation sequencing. *BMC Genomics* **11**, 244 (2010).

60 Hogbom, M. *et al.* Crystal structure of conserved domains 1 and 2 of the human DEAD-box helicase DDX3X in complex with the mononucleotide AMP. *J Mol Biol.* **372**, 150-159. Epub 2007 Jun 2026. (2007).

61 Umate, P., Tuteja, N. & Tuteja, R. Genome-wide comprehensive analysis of human helicases. *Commun Integr Biol.* **4**, 118-137. (2011).

62 Sengoku, T., Nureki, O., Nakamura, A., Kobayashi, S. & Yokoyama, S. Structural Basis for RNA Unwinding by the DEAD-Box Protein Drosophila Vasa. *Cell* **125**, 287-300, doi:10.1016/j.cell.2006.01.054 (2006).

63 Sekiguchi, T., Kurihara, Y. & Fukumura, J. Phosphorylation of threonine 204 of DEAD-box RNA helicase DDX3 by cyclin B/cdc2 in vitro. *Biochemical and Biophysical Research Communications* **356**, 668-673, doi:10.1016/j.bbrc.2007.03.038 (2007).

64 Stavropoulos, P., Blobel, G. & Hoelz, A. Crystal structure and mechanism of human lysine-specific demethylase-1. *Nat Struct Mol Biol* **13**, 626-632, doi:http://www.nature.com/nsmb/journal/v13/n7/suppinfo/nsmb1113_S1.html (2006).

65 Klose, R. J., Kallin, E. M. & Zhang, Y. JmjC-domain-containing proteins and histone demethylation. *Nat Rev Genet* **7**, 715-727, doi:http://www.nature.com/nrg/journal/v7/n9/suppinfo/nrg1945_S1.html (2006).

66 Chen, Y. *et al.* Crystal structure of human histone lysine-specific demethylase 1 (LSD1). *Proceedings of the National Academy of Sciences* **103**, 13956-13961, doi:10.1073/pnas.0606381103 (2006).

67 Metzger, E. & Schule, R. The expanding world of histone lysine demethylases. *Nat Struct Mol Biol* **14**, 252-254 (2007).

68 Ng, S. S. *et al.* Crystal structures of histone demethylase JMJD2A reveal basis for substrate specificity. *Nature* **448**, 87-91, doi:http://www.nature.com/nature/journal/v448/n7149/suppinfo/nature05971_S1.html (2007).

69 Stephens, P. J. *et al.* Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. *Cell* **144**, 27-40, doi:10.1016/j.cell.2010.11.055 (2011).

70 Pleasance, E. D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184-190, doi:http://www.nature.com/nature/journal/v463/n7278/suppinfo/nature08629_S1.html (2010).