SUPPLEMENTARY INFORMATION FOR

# The genomic landscape of diffuse intrinsic pontine glioma and pediatric non-brainstem high-grade glioma

Gang Wu[1]*, Alexander K Diaz[2,3]*, Barbara S Paugh[2], Sherri L Rankin[2], Bensheng Ju[4], Yongjin Li[1], Xiaoyan Zhu[2], Chunxu Qu[1], Xiang Chen[1], Junyuan Zhang[2], John Easton[5], Michael Edmonson[1], Xiaotu Ma[1], Charles Lu[6], Panduka Nagahawatte[1], Erin Hedlund[1], Michael Rusch[1], Stanley Pounds[7], Tong Lin[7], Arzu Onar-Thomas[7], Robert Huether[1], Richard Kriwacki[8], Matthew Parker[1], Pankaj Gupta[1], Jared Becksfort[1], Lei Wei[9], Heather L Mulder[5], Kristy Boggs[5], Bhavin Vadodaria[5], Donald Yergeau[5], Jake C Russell[2], Kerri Ochoa[6], Robert S Fulton[6], Lucinda L Fulton[6], Chris Jones[10,11], Frederick A Boop[12], Alberto Broniscer[13], Cynthia Wetmore[13], Amar Gajjar[13], Li Ding[6], Elaine R Mardis[6], Richard K Wilson[6], Michael R Taylor[4], James R Downing[14], David W Ellison[14], Jinghui Zhang[1#] and Suzanne J Baker[2,3#] for the St Jude Children's Research Hospital – Washington University Pediatric Cancer Genome Project.

[1]Departments of Computational Biology, [2]Developmental Neurobiology, [4]Chemical Biology and Therapeutics, [5]Pediatric Cancer Genome Project, [7]Biostatistics, [8]Structural Biology, [12]Surgery, [13]Oncology, and [14]Pathology, St. Jude Children's Research Hospital, Memphis, TN 38105

[3]Integrated Biomedical Sciences Program, University of Tennessee Health Science Center, Memphis, TN 38163

[6]The Genome Institute, Washington University, 633108

[9]Biostatistics and Bioinformatics, Roswell Park Cancer Institute, Buffalo, NY 14263

Divisions of [10]Molecular Pathology and [11]Cancer Therapeutics, Institute for Cancer Research, London, UK SM2 5NG

*These authors contributed equally
#To whom Correspondence should be addressed:
Suzanne Baker; Suzanne.Baker@stjude.org
Jinghui Zhang; Jinghui.Zhang@stjude.org

## Table of Contents

# 1. Supplementary Note

## 1.1 Mutation profile

There is a wide range of mutation rates for pediatric HGGs, ranging from only a few SNVs or SVs to over 800,000 somatic mutations. For example, the tumor SJHGG082_D from the youngest patient (approximately1 month old) has only one missense mutation in TPSG1 (p.A77S), together with an ETV6-NTRK3 translocation. By contrast, we have identified four hypermutator samples (three with matched normal samples: SJHGG003_D, SJHGG030_D, SJHGG111_D, and one unpaired tumor: SJHGG119_D) with over 10,000 somatic mutations in gene coding regions (tier1), resulting in an estimated background mutation rate on the order of $10^{-4}$. The background mutation rate for HGG hypermutator tumors is higher than what has been reported in the literature so far [1] [2]. Mutations in these tumors were not included in calculations of mutation frequency for specific target genes. However, there is no difference between the background mutation rate of DIPG and NBS-HGG tumors (p-value=0.09), as well as the number of non-silent SNVs, SVs and copy number alterations (Supplementary Figure 1).

The tier 1 mutation rate was computed as (no. tier 1 mutations detected +1)/(no. tier 1 bases with 10x coverage +2)/3. The small constants 1 and 2 were added in the numerator and denominator, respectively, to stabilize the log-transformed data that were used for subsequent outlier detection analysis. Least median squares was applied to the log10-transformed tier 1 mutation rate data to identify the densest bulk of observations and outlying observations [3]. Outliers were identified by modeling the densest bulk of observations as the inter-quartile range of a normal distribution to determine an approximate 95% prediction interval for individual observations. Observations falling outside the approximate 95% prediction interval were flagged as outliers (Supplementary Figure 7).

## 1.2 Multiple somatic mutations for the same gene or homologous gene

We observed examples of two different mutations of the same cancer gene or functionally important gene in non-hypermutator HGG tumors . For example, two mutations in *TP53* were observed in multiple cases: SJHGG073_A (p.R273C, p.K120M), SJHGG102_D (p.R273C, p.Q144*), SJHGG105_A (p.E339*, p.R248W) and SJHGG021 (p.Q167*, p.R273H). For *H3F3A*, in addition to the known p.G34R mutation, we found a nonsense mutation (p.K56*) on the other allele in the same sample, SJHGG031_D. In the autopsy DIPG SJHGG007_A, in addition to a nonsense mutation (p.R816*) in NF1, we identified a somatic SV that disrupts *NF1*, and did not detect any *NF1* expression by RNA-seq in this sample. *BCOR* and *BCORL1* loss-of-function mutations are another example. SJHGG018_D contained a *BCORL1* nonsense mutation p.R609*, in addition to another frameshift mutation in *BCOR* (p.P152fs). In SJHGG129_D, there are two somatic mutations (p.Y1692*, p.E1255_E1256>E) in *BCORL1,* with only the pY1692* allele showing expression in RNA-seq data. In these examples of genes targeted by two mutations within the same tumor, at least one of the mutations is often a nonsense or frameshift mutation, suggesting loss of function of the gene.

## 1.3 Germline mutations in HGG

We identified interesting germline variations in pediatric HGGs (Supplementary Table 11). Strikingly, two hypermutator tumors (SJHGG003_D and SJHGG111_D) contain a *PMS2*

germline nonsense mutation p.Q643*. In addition, SJHGG111_D also contains a second germline frameshift indel in *PMS2* (p.K431fs). In two independent tumors arising in patient SJHGG003, different somatic mutations in the second allele of *PMS2* were identified; a frameshift p.K413fs in SJHGG003_D and structural variant in SJHGG003_A. An additional hypermutator SJHGG030_D contained a germline structural variant in *PMS2*, and a somatic mutation p.S624F of *PMS2* in the tumor.

We also identified germline mutations with a second somatic hit in the same gene. Three patients had germline mutations in *TP53* (Supplementary Table 11). SJHGG089_A inherited a germline *NF1* loss-of-function mutation (p.R416*) then gained a second somatic frameshift mutation in *NF1* (p.F2176fs). SJHGG034_D contains a germline frameshift mutation (p.E546fs) and a somatic missense mutation (p.R1334Q) in *MSH6*, which is involved in DNA mismatch repair.

We observed two germline mutations occurring in a single patient, SJHGG027_D. This patient had ataxia telangiectasia (A-T), caused by two germline SNVs (p.L2544_E52splice and p.L2312fs) in *ATM* (Supplementary Table 11). The RNA-seq data confirms that the splice variant introduces a novel exon junction (chr11:108202284→chr11:108202617), leading to a frameshift due to 11bp deletion. A-T is associated with profound neurodegeneration and hematopoietic malignancies, but it is not frequently associated with brain tumors, although there are a few reports of brain tumors arising in A-T patients [4] [5].

### 1.4 EGFR vIII analysis in WGS and RNA-seq

To determine if the EGFR vIII variant commonly found in adult glioblastoma is also present in pediatric HGGs, we evaluated EGFR structural variations in the WGS data as well as the RNA-seq data. EGFRvIII was found in one tumor (SJHGG019_E) analyzed by WGS. Interestingly, the relapsed tumor from the same patient, SJHGG019_S, did not have the EGFRvIII in WGS or RNASeq. The low frequency of EGFRvIII (1 out of 85 tumors combining WGS and RNASeq data) highlights a significant difference between adult and pediatric HGGs.

### 1.5 Chromothripsis

Twenty-two HGG tumors analyzed by whole-genome sequencing have >20 SVs, 15 of which were accompanied by oscillating copy number states at SV breakpoints, raising the possibility that the structural variations in these tumors were generated by a single catastrophic event known as chromothripsis [6] . Applying the statistical test outlined by Korbel and Campbell [7], 13 tumors met the criteria for chromothrispsis (Supplementary Table 12, Supplementary Figure 8-9). TP53 mutations were found in 10 of these 13 tumors with chromothripsis compared to 13 out of 27 tumors without chromothripsis (Odds ratio=3.5). However, the association betweenTP53 mutation status and chromothripsis is only marginal (Fisher exact test, one-tailed p-value=0.082). Notably, chromothripsis contributed to the rearrangements in oncogenes EGFR (SJHGG019_S and SJHGG019_E), NTRK3 (SJHGG004_D, Supplementary Figure 8a-d) and PDGFRA (SJHGG003_A, Supplementary Figure 8e-h)

To gain further understanding about the complexity of the chromosomal re-arrangements in contributing to oncogenic fusion genes, we selected two tumors to construct derivative chromosomes resulting from chromothripsis. The first case, SJHGG004_D, has multiple SV breakpoints joining oscillating CNV segments on chromosome 15 (Supplementary Figure 8a-d).

These segments were re-arranged resulting in a 3-gene fusion involving BTBD1, CPEB1, NTRK3, which were verified by RNA-seq with the high expression of BTBD1-NTRK3 fusion transcript.

The second case SJHGG003_A has a low-level 6-fold amplification interspersed with a high-level >50-fold amplification with multiple SV breakpoints and CNV segments on chromosomes 4, 10 and 12 (Supplementary Figure 8e). We were able to construct a derivative chromosome comprised of highly amplified segments connected by 17, 3 and 5 SVs on chromosomes 4, 10, 12, respectively (Supplementary Figure 8f). The segment order and orientation was reshuffled and the high SV junction read count indicates formation of an episomal through chromothripsis followed by replication of the episome. On this derivative chromosome, intron 1 of DIP2C is connected to intron 10 of PDGFRA (chr10: 710870→chr4: 55140417), resulting in a fusion gene that is comprised of exon 1 (chr10:735434) of DIP2C to exon 11 (chr4:55140698) of PDGFRA (Supplementary Figure 8g). High-level expression of this fusion transcript was detected in RNASeq (Supplementary Figure 8h). The resulting fusion protein lacks the Ig-like domain but retains the intact TK domain of PDGFRA, which is similar to the FIP1L1-PDGFRA fusion protein previously identified in idiopathic hypereosinophilic syndrome [8].

### 1.6 Tumor heterogeneity analysis

In order to accurately estimate tumor heterogeneity, the tumor purity should be taken into account. For germline heterozygous SNPs, loss of heterozygosity (LOH) measures the absolute difference between the mutant allele fraction in tumor and that in germline sample (0.5).LOH is the result of copy number alterations and/or copy neutral-LOH in tumor cells.
Compared to copy number gains (a single copy gain in 100% tumor results in a LOH value of 0.167), regions with copy number loss showed stronger LOH (a single copy loss in 100% tumor result in a LOH value of 0.5). Consequently, we used LOH signals in copy neutral or heterozygous copy number loss regions (CNA value between [-1, 0]) to estimate tumor purity for all WGS samples. Briefly, a single copy loss in $x$% tumor cells resulted in an estimated CNA value of $-\frac{x}{100}$ and a LOH value of $\frac{x}{400-2x}$. Assuming the remaining LOH signal came from CN-LOH (CN-LOH in $x$% tumor cell resulted in a LOH value of $\frac{x}{200}$, the tumor content in a region could be estimated as the sum of the fraction with copy number loss and the fraction with CN-LOH by: $-CNA + 2* \left(LOH - \frac{-CNA}{4-2CNA}\right)$. Using tumor content estimates from various regions within the genome, we performed an unsupervised clustering analysis using the *mclust* package (version 3.4.8) in R (version 2.11.1). The tumor purity of the sample was defined as the highest cluster center value among all clusters.

To avoid assessing low tumor purity due to subclonal LOH/CNV, distribution of mutant allele fraction (MAF) for validated SNVs in regions with no CNV nor LOH is analyzed for tumors with ≥50 such qualified SNVs. If the highest MAF peak exceeds the purity estimate from CNV/LOH, the tumor purity is adjusted to 2 * highest MAF peak.

Once we obtained the estimation on tumor purity, we adjusted the mutant allele fraction (MAF) by tumor purity using deep sequencing data (capture validation, which provides more accurate assessment of mutant allele fraction), as follows

$$\frac{\#Mutant\ reads}{(\#Total\ reads)\times(tumor\ purity)}$$

We used all validated autosomal SNVs satisfying the following criteria in heterogeneity analysis:

1) In copy neutral region (Log2ration between (-0.1, 0.1) in CNV analysis).
2) Not in regions with LOH (LOH value < 0.1).
3) With MAF > 0.05 or mutant allele count > 2.

We drew the kernel density estimate plot for MAFs of the qualifying SNVs using the density function in the stat package in R. We also estimated the number of significant peaks and the relative MAF component for each peak (peaks with less than 5 SNVs, peaks with less than 1% SNVs, and peaks with excessive variance were ignored). A sample with heterogeneity shows density peaks at a MAF smaller than 0.5 (the expected MAF assuming heterogeneous SNVs).

As shown in Supplementary Figure 10a, subclonal mutations were found in all except for two tumors (SJHGG003_D and SJHGG012_D). Although multiple distinct peaks were found in 50% of the tumors (Supplementary Figure 10a), the broad peak shoulder or the presence of a plateau instead of a peak in the remaining tumors suggest the presence of multiple subclones with overlapping MAF density distribution that cannot be separated by a density plot. Tumors with a single plateau included SJHGG019_E, SJHGG005_A, SJHGG112_D and SJHGG112_E.

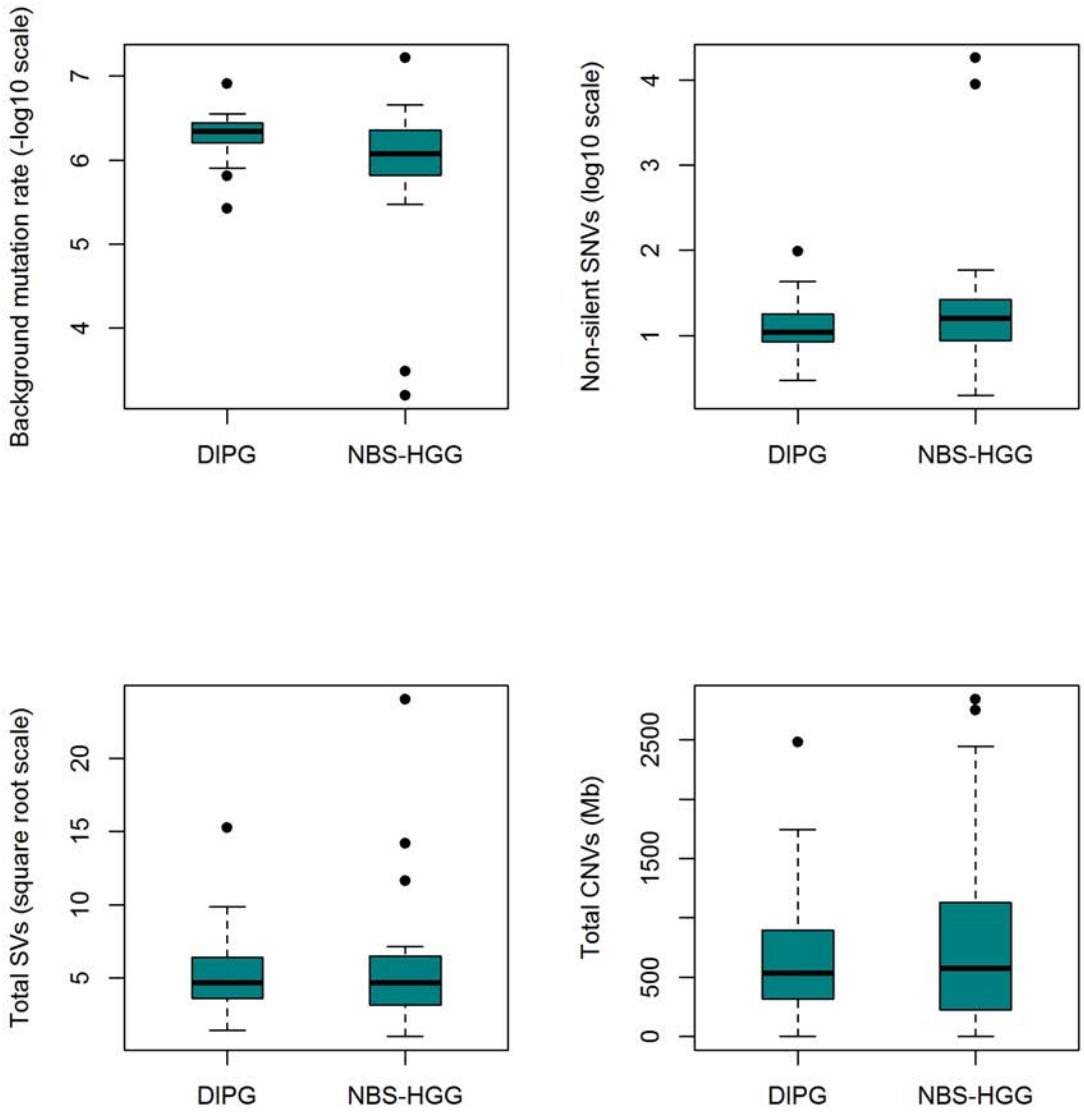
### 1.7 Tumor evolution analysis

To study clonal evolution, we pooled tier 1-3 SNVs from the same tumor. Also, we combined read counts from WGS and capture validations for the validated SNVs to increase accuracy of MAF estimation. Further, we required an SNV to have at least 20x coverage to be included in clonal estimation. To account for copy number variations and copy neutral loss of heterogeneity (LOH), we excluded SNVs in CNV regions (log2 ratio >0.1 or <-0.1) as well as sex chromosomes. We then estimated population sizes of subclones using Gaussian mixture model by R package *mclust*.

To illustrate the clonal architecture and tumor evolution from diagnosis to relapse (or autopsy), we selected two cases, SJHGG002 and SJHGG112, as there are sufficient SNVs (≥50) in the diploid regions for assessing the lineage of subclones. The results are shown in Supplementary Figure 10b-10c. SJHGG002 diagnostic tumor has a founder clone that harbors a *TP53* truncation mutation (TP53 R213*). The founder clone has two descendant subclones that acquired additional mutations; each constitutes approximately 1/3 of the total tumor population. Only one subclone from the diagnosis (labeled as chr21.14659745.C.G in Supplementary Figure 10b) was present in autopsy and it co-exists with its own descendant subclone that acquired additional autopsy-specific mutations. The founder clone of SJHGG122 harbors key mutations; *H3F3A* K27M and *PIK3CA* Q546K (Supplementary Figure 10c). A descend subclone acquired additional mutations including a nonsense mutation in *PPM1D* (W427*). Interestingly, only the founder clone was present in relapse but not the subclone with *PPM1D* mutation. In relapse, the founder clone gave rise to a subclone that acquired additional mutations including HMGXB3 L611W. This descendant subclone becomes dominant clone in relapse as the founder clone constitutes only ¼ of the total tumor population.

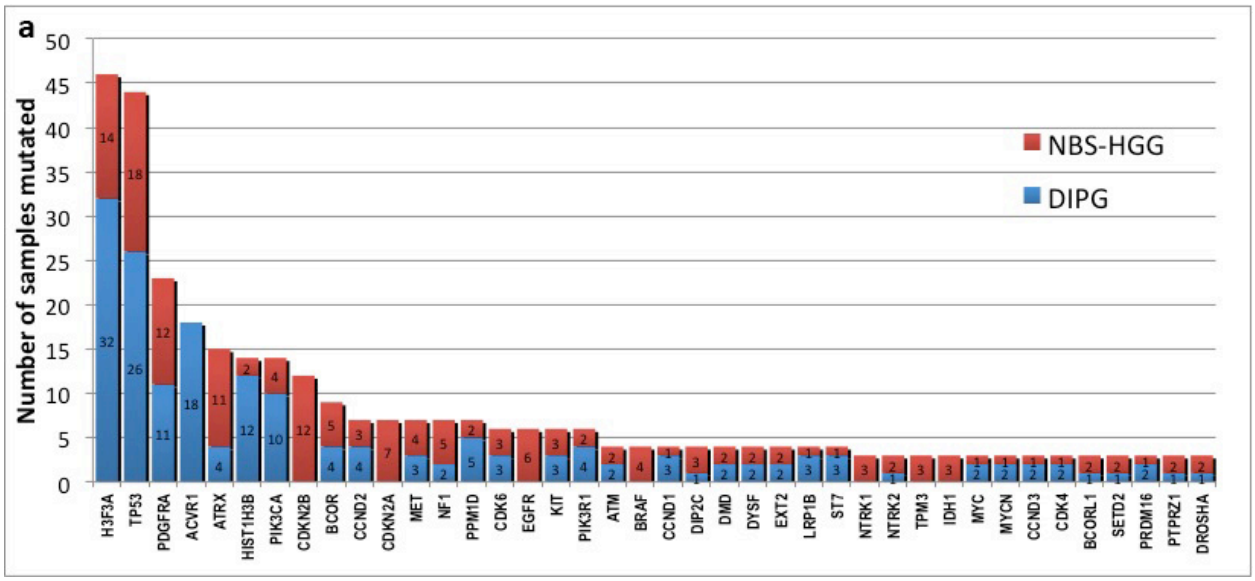
## 2. Supplementary Figures

**Supplementary Figure 1.** The distribution of background mutation rate, non-silent SNVs, total SVs, and total CNVs.

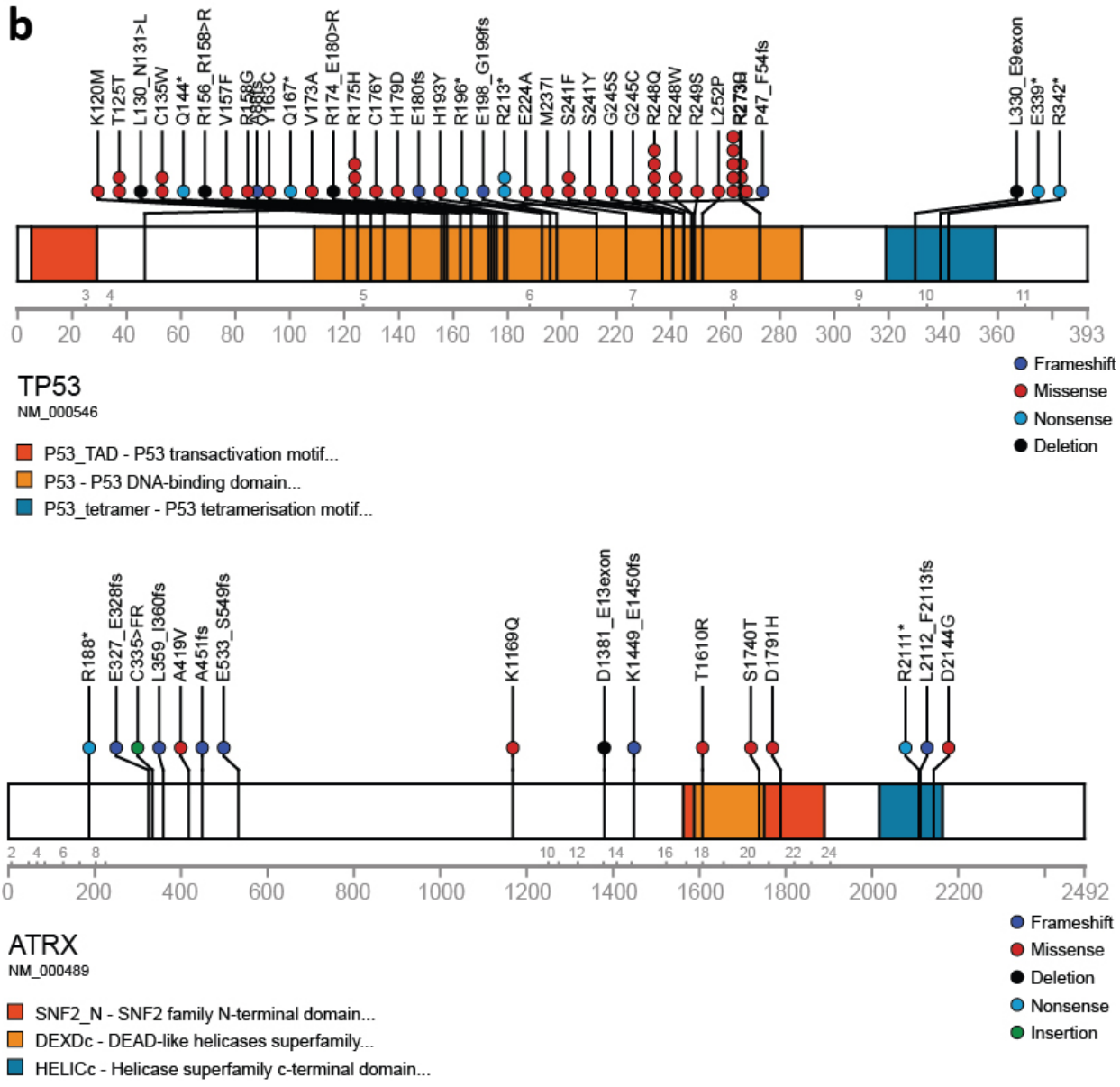


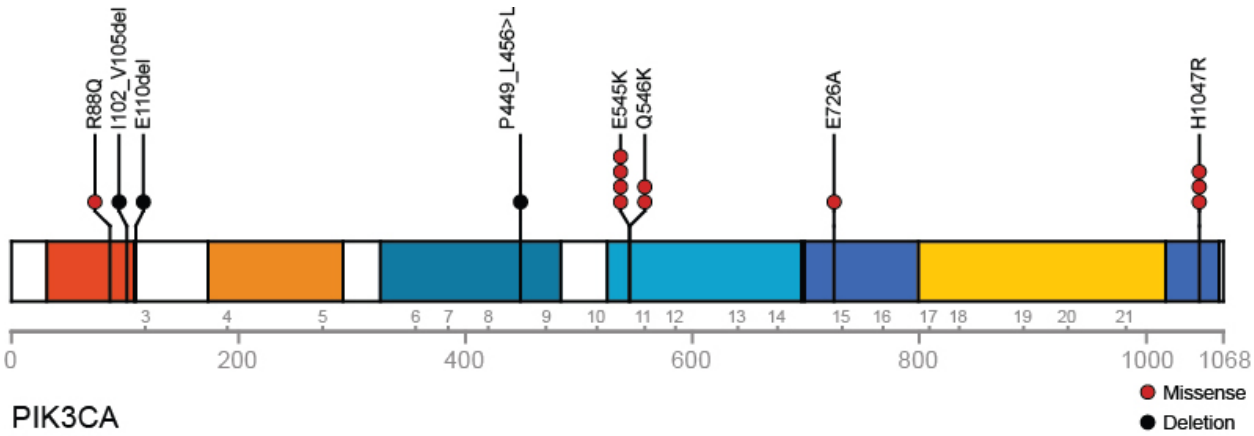

There is no significant difference between the background mutation rate and the number of non-silent SNVs, SVs and CNVs between DIPG tumors and NBS-HGG tumors

**Supplementary Figure 2.** Genes frequently mutated in DIPG or NBS-HGG tumors (a) and high frequency non-silent point mutations or indels in *TP53*, *ATRX, PIK3CA, PIK3R1, BCOR, BCORL1*, *PPM1D* and *NF1* (b).



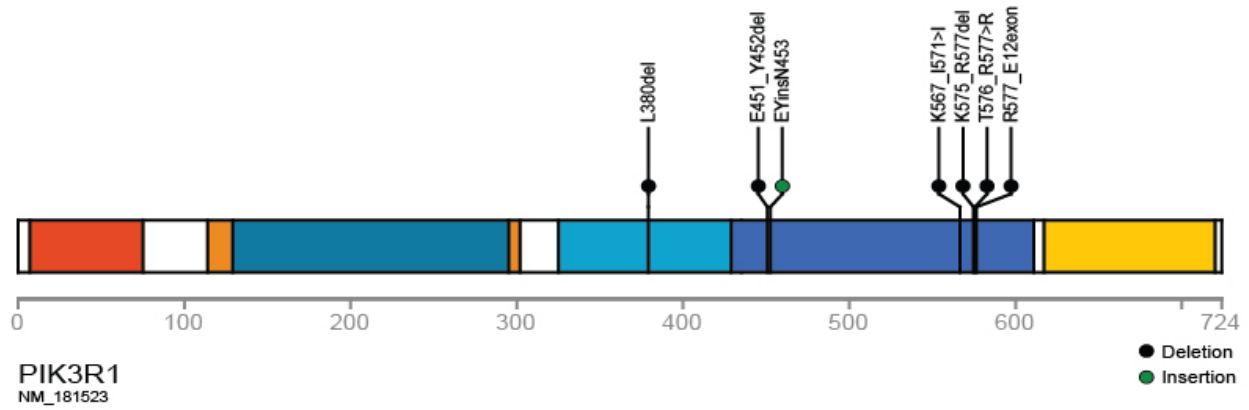

All somatic alterations including non-silent SNVs, indels, SVs, and focal CNVs were included in the frequency calculation. Hypermutator samples were excluded for the frequency summary. The top 40 most frequently mutated genes were included in the plot. We excluded *TTN, MUC16, CSMD1, RYR2* because these genes have been shown to be likely false cancer related genes [9]. For genes with frequency count of 3, we included the genes known to associate with cancer.


b
K120M
T125T
L130_N131>L
C135W
Q144*
R156_R158>R
V157F
Q167*
V173A
R174_E180>R
R175H
C176Y
H179D
E180fs
H193Y
R196*
E198_G199fs
R213*
E224A
M237I
S241F
S241Y
G245S
G245C
R248Q
R248W
R249S
L252P
P47_F54fs
L330_E9exon
E339*
R342*
0 20 40 60 80 100 120 140 160 180 200 220 240 260 280 300 320 340 360 393
TP53
NM_000546
P53_TAD - P53 transactivation motif...
P53 - P53 DNA-binding domain...
P53_tetramer - P53 tetramerisation motif...
Frameshift
Missense
Nonsense
Deletion
R188*
E327_E328fs
C335>FR
L359_I360fs
A419V
A451fs
E533_S549fs
K1169Q
D1381_E13exon
K1449_E1450fs
T1610R
S1740T
D1791H
R2111*
L2112_F2113fs
D2144G
0 200 400 600 800 1000 1200 1400 1600 1800 2000 2200 2492
ATRX
NM_000489
SNF2_N - SNF2 family N-terminal domain...
DEXDc - DEAD-like helicases superfamily...
HELICc - Helicase superfamily c-terminal domain...
Frameshift
Missense
Deletion
Nonsense
Insertion



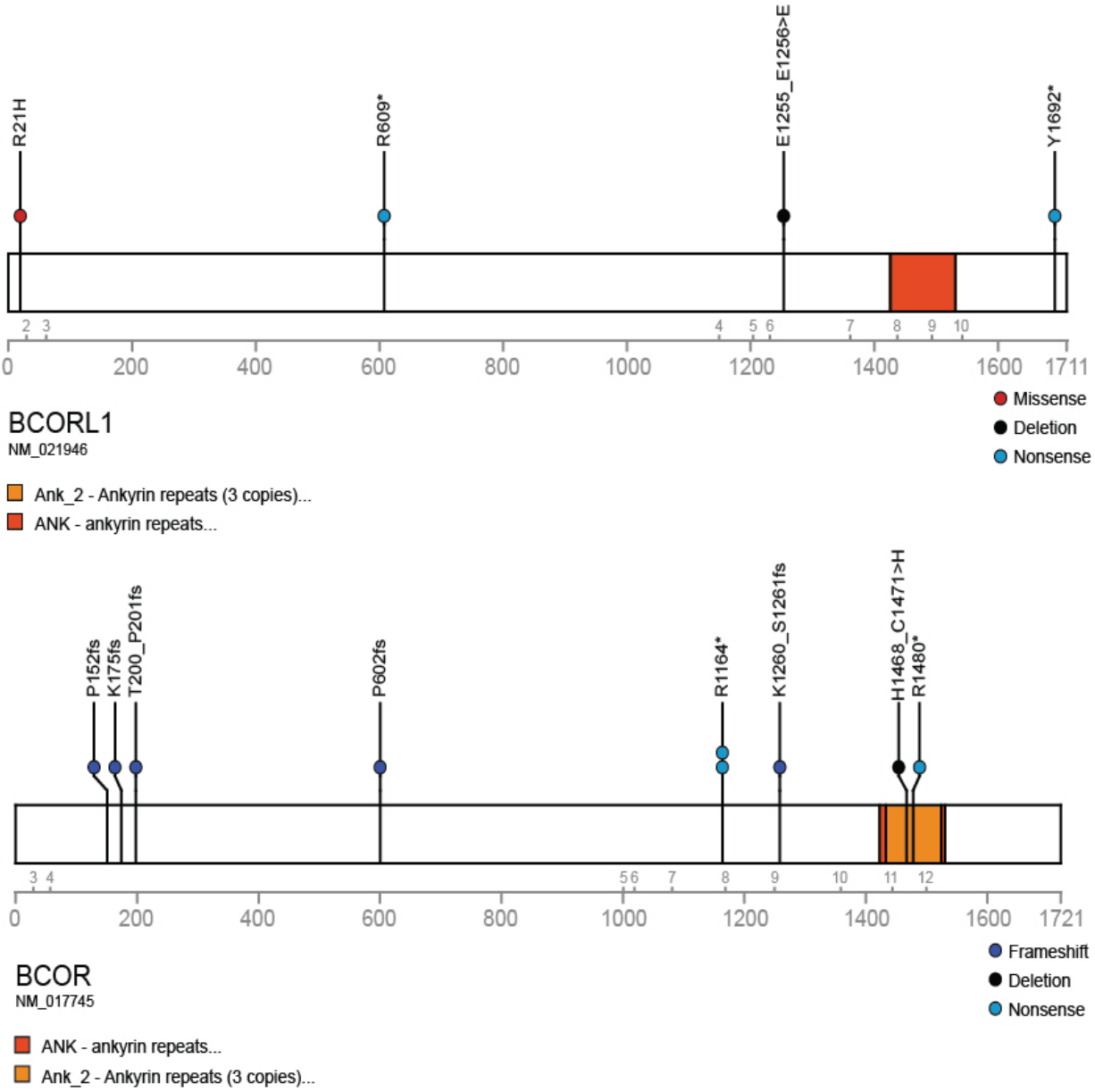

PIK3CA
NM_006218

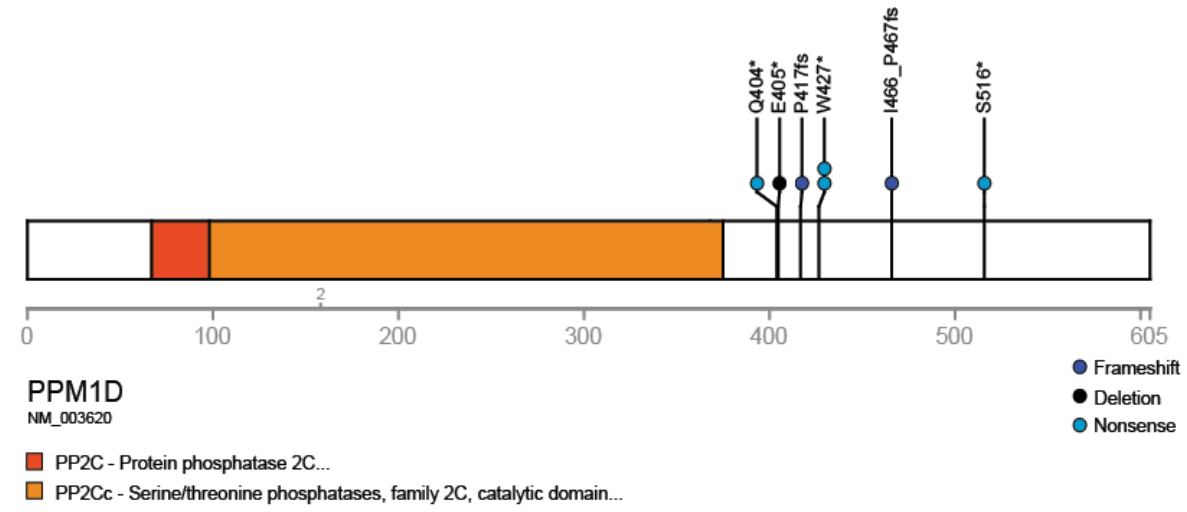
- PI3K_p85B - PI3-kinase family, p85-binding domain...
- PI3K_rbd - PI3-kinase family, ras-binding domain...
- C2_PI3K_class_I_alpha - C2 domain present in class I alpha phosphatidylinositol 3-kinases (PI3...
- PI3Ka_I - Phosphoinositide 3-kinase (PI3K) class I, accessory domain ...
- PI3Kc_IA_alpha - Phosphoinositide 3-kinase (PI3K), class IA, alpha isoform, catalytic d...
- PI3Kc - Phosphoinositide 3-kinase, catalytic domain...

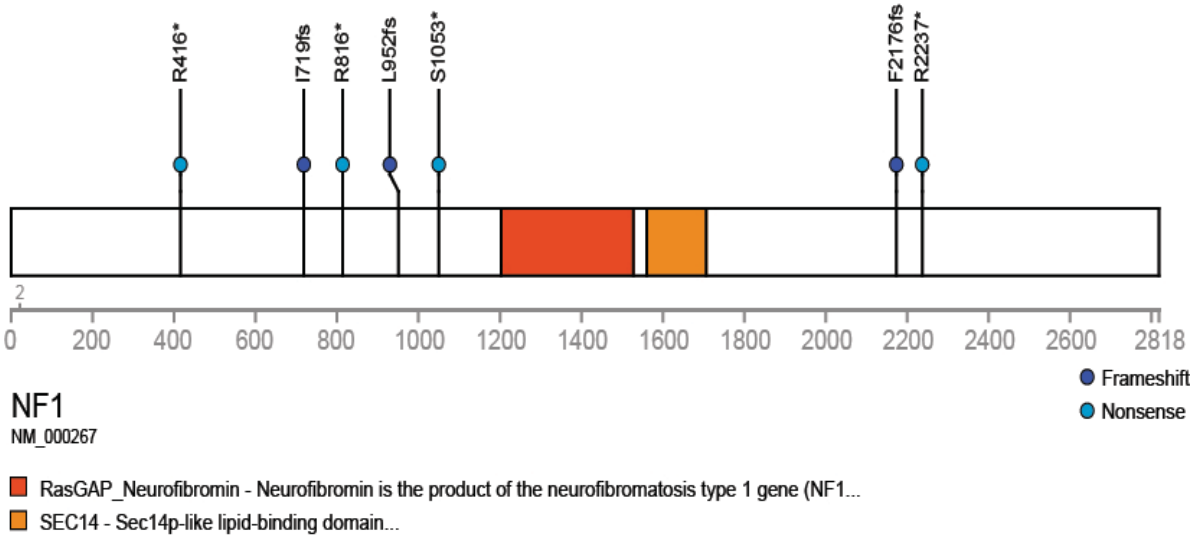



PIK3R1
NM_181523

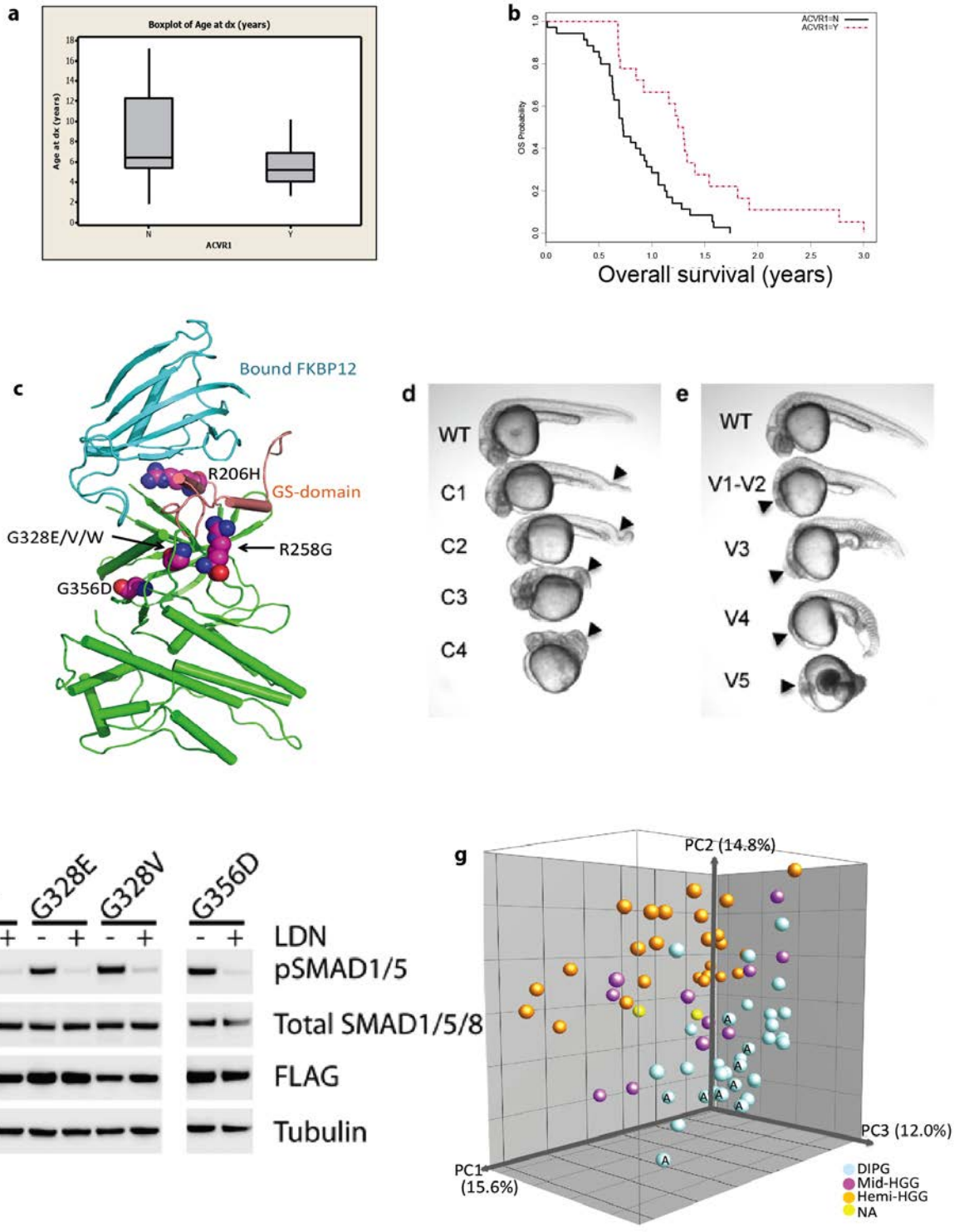
- SH3 - Src homology 3 domains...
- RhoGAP_p85 - RhoGAP_p85...
- RhoGAP - GTPase-activator protein for Rho-like GTPases...
- SH2_nSH2_p85_like - N-terminal Src homology 2 (nSH2) domain found in p85...
- COG4942 - Membrane-bound metallopeptidase [Cell division and chromosome partitio...
- SH2_cSH2_p85_like - C-terminal Src homology 2 (cSH2) domain found in p85...


R21H
R609*
E1255_E1256>E
Y1692*
2 3
4 5 6
7 8 9 10
0 200 400 600 800 1000 1200 1400 1600 1711
Missense
Deletion
Nonsense
BCORL1
NM_021946
Ank_2 - Ankyrin repeats (3 copies)...
ANK - ankyrin repeats...
P152fs
K175fs
T200_P201fs
P602fs
R1164*
K1260_S1261fs
H1468_C1471>H
R1480*
3 4
5 6 7 8 9 10 11 12
0 200 400 600 800 1000 1200 1400 1600 1721
Frameshift
Deletion
Nonsense
BCOR
NM_017745
ANK - ankyrin repeats...
Ank_2 - Ankyrin repeats (3 copies)...


Q404*
E405*
P417fs
W427*
I466_P467fs
S516*
0
100
200
300
400
500
605
Frameshift
Deletion
Nonsense
PPM1D
NM_003620
PP2C - Protein phosphatase 2C...
PP2Cc - Serine/threonine phosphatases, family 2C, catalytic domain...


R416*
I719fs
R816*
L952fs
S1053*
F2176fs
R2237*
0
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
2400
2600
2818
Frameshift
Nonsense
NF1
NM_000267
RasGAP_Neurofibromin - Neurofibromin is the product of the neurofibromatosis type 1 gene (NF1...
SEC14 - Sec14p-like lipid-binding domain...

**Supplementary Figure 3.** *ACVR1* mutations in DIPG patients



a. DIPG patients with *ACVR1* mutations are younger at diagnosis on average (p=0.010 by Kruskal-Wallis test)

b. *ACVR1* mutation is associated with longer survival of DIPGs. Kaplan-Meier plot for 53 DIPG patients showing longer overall survival of patients with *ACVR1* mutation (red dotted line) compared to those with wild-type *ACVR1* (black line), quantified by the log rank test; p=0.0015.

c. Structure of the ACVR1 G/S domain and kinase domain, bound to FKPB12. Taken from the structure reported by Chaikuad et al, 2012 [10]. The inhibitory protein FKBP12 is shown in turquoise, GS domain in salmon, Kinase domain in green. The residues mutated in DIPG cluster in the GS domain and kinase domain, and make interactions that are important for stabilizing the inactive conformation of ACVR1. Loss of this stabilizing effect would be predicted to allow leady activity of the kinase.

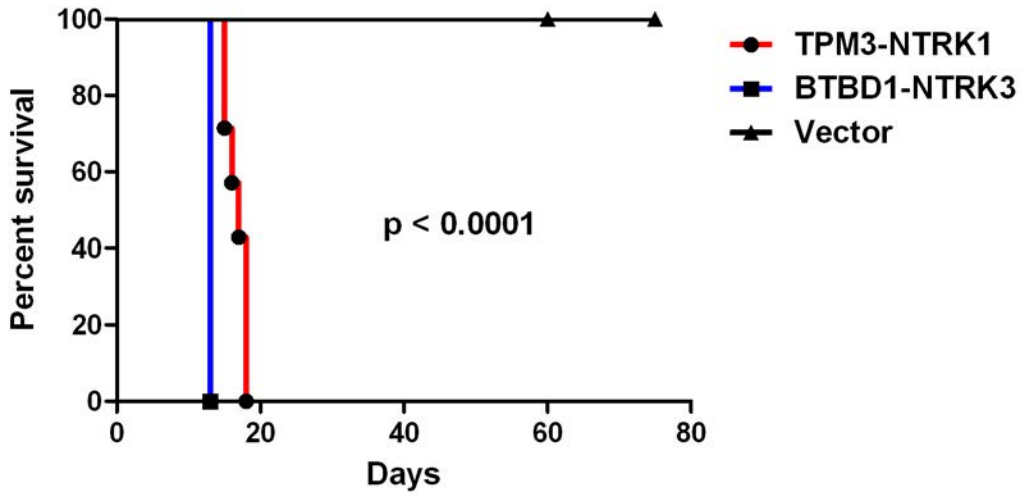
d. The range of dorsalized phenotypes (C1-C4) for embryos at 24 hpf. The arrowheads indicate the loss of ventral structures with C1 being the least and C4 being the most severe. Scoring criteria for dorsalized and ventralized zebrafish embryos based on previous studies [11,12].

e. The range of ventralized phenotypes (V1-V5) for embryos at 24 hpf. The arrowheads indicate the loss of dorsal structures with V1-V2 being the least and V5 being the most severe.

f. LDN-193189 blocks SMAD1/5 phosphorylation downstream of mutant ACVR1. Primary astrocytes isolated from brainstem of neonatal Tp53 conditional knockout mice were transduced with retroviruses expressing FLAG-tagged ACVR1 wild-type, G328E, G328V or G356D and treated with or without 1 μM LDN-193189 for 24 hours. Western blots for indicated antibodies show that LDN blocks p-SMAD1/5. Lysates for G356D with and without LDN-193189 treatment were run together on a separate gel from the other lysates.

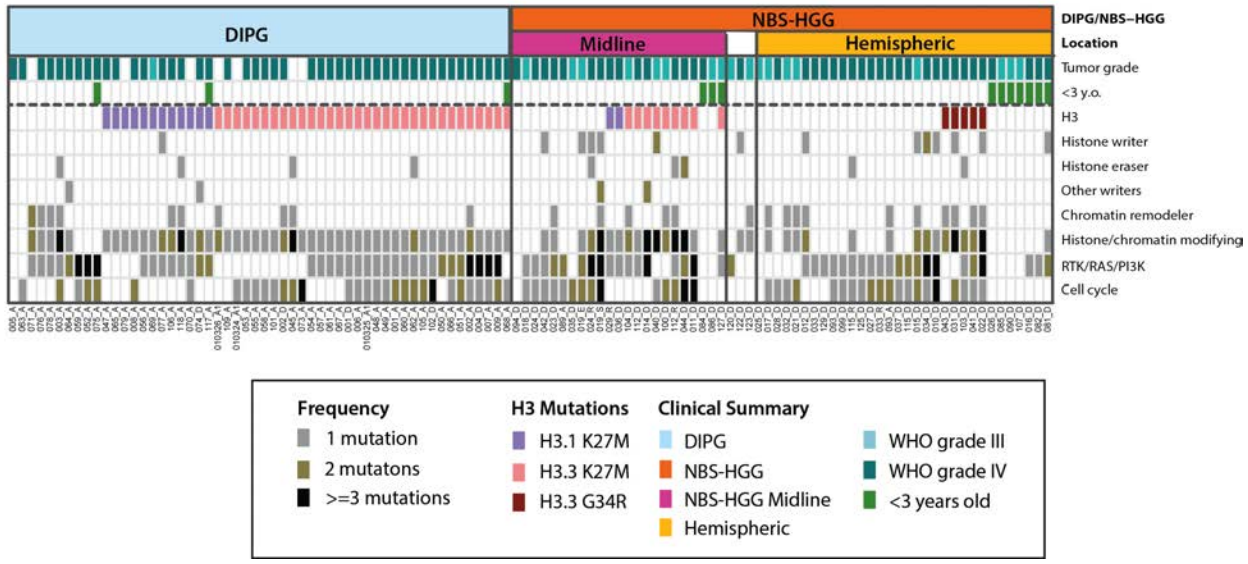
g. Gene expression of 71 DIPG and non-brainstem HGG were profiled using Affymetrix HG-U133 plus2 arrays. Using top 1000 probe-sets selected based on median absolute deviation (MAD) score, principal component analysis was performed using GeneMaths XT (Version 2.12). A total of 42.4% variance can be explained by the first three components. The mutation status of *ACVR1* was indicated using letter “A”.

**Supplementary Figure 4.** *NTRK* fusion genes drive gliomagenesis with short latency and complete penetrance



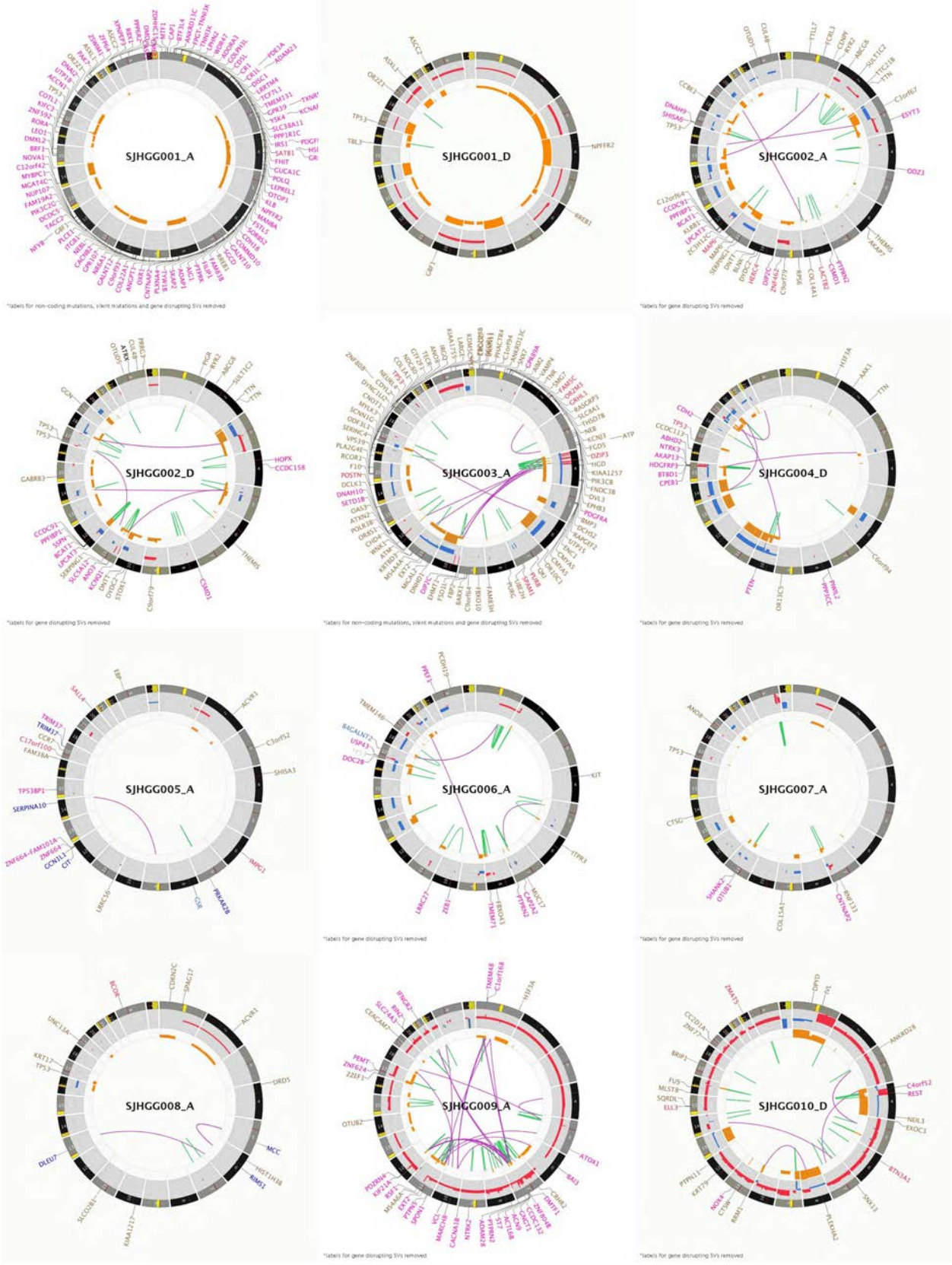

Kaplan–Meier curves showing time to morbidity due to brain tumors of mice following intracranial implantation of *Tp53* null cortical PMAs expressing empty vector (n=7) or TPM3-NTRK1 (n=7), or *Tp53* null brainstem PMAs expressing empty vector (n=7) or BTBD1-NTRK3 (n=7).

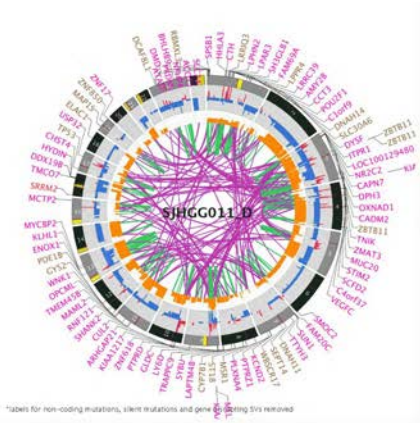
**Supplementary Figure 5:** Proportion of tumors with mutations in common pathways



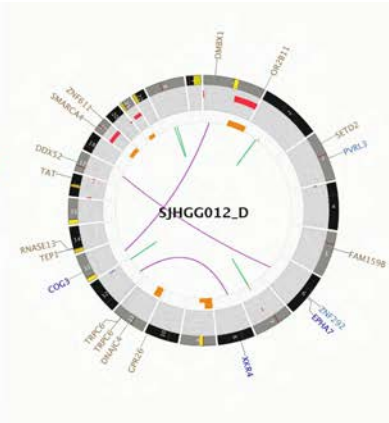

Genetic alterations in commonly targeted pathways for 112 HGGs are shown. Tumor subgroup (DIPG or NBS-HGG), location of NBS-HGGs (midline versus tumors in cerebral hemispheres), and tumor grade are indicated. White boxes for location or tumor grade indicates information not available. < 3 y.o denotes less than 3 years of age. Gene lists for pathways are in Online Methods section 4.

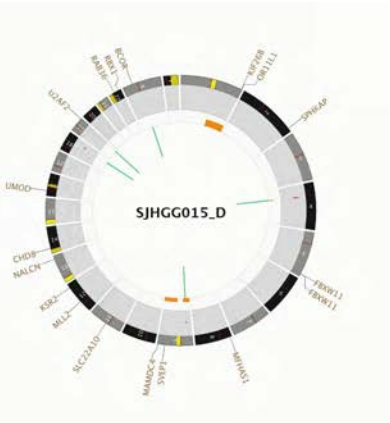
**Supplementary Figure 6.** CIRCOS plots of WGS tumors analyzed

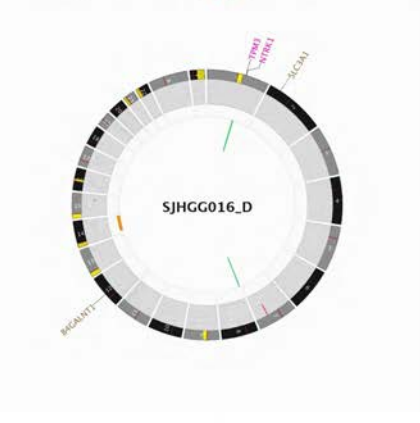


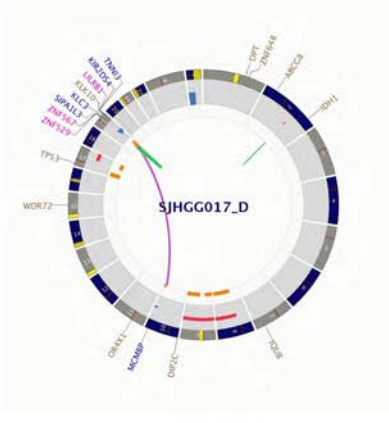


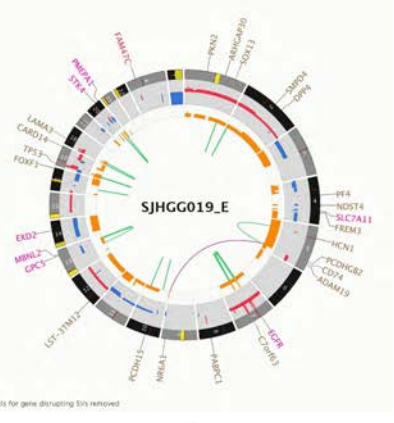

*labels for non-coding mutations, silent mutations and gene disrupting SVs removed



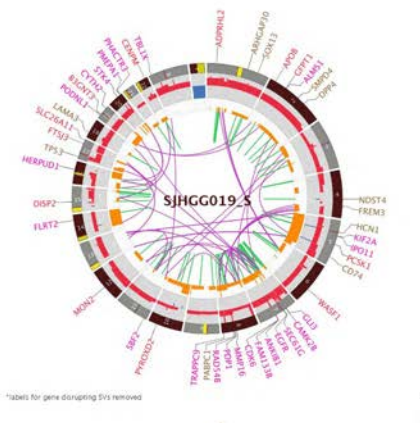



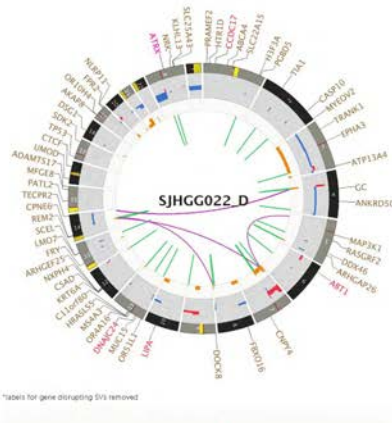



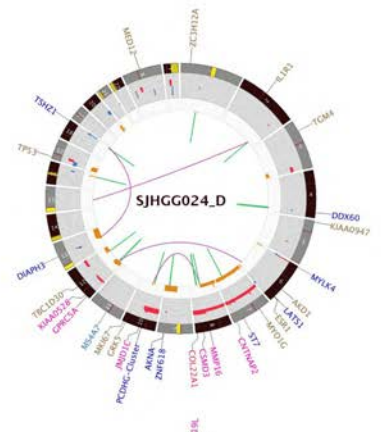

*labels for gene disrupting SVs removed

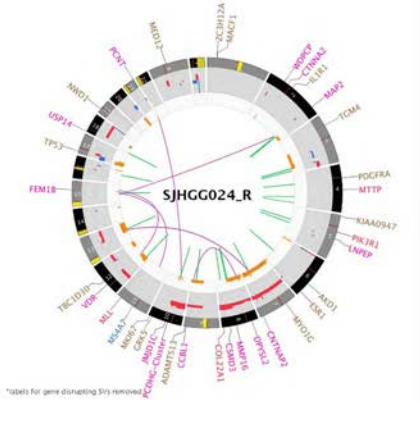



*labels for gene disrupting SVs removed

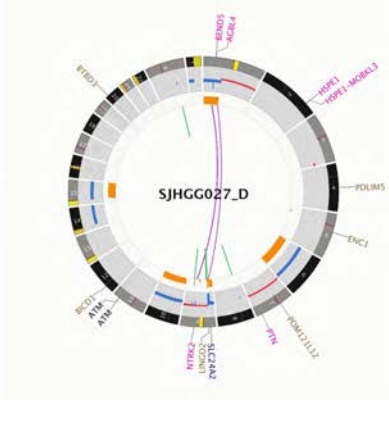


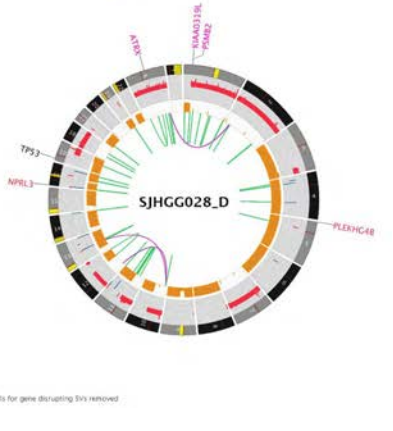

*labels for gene disrupting SVs removed



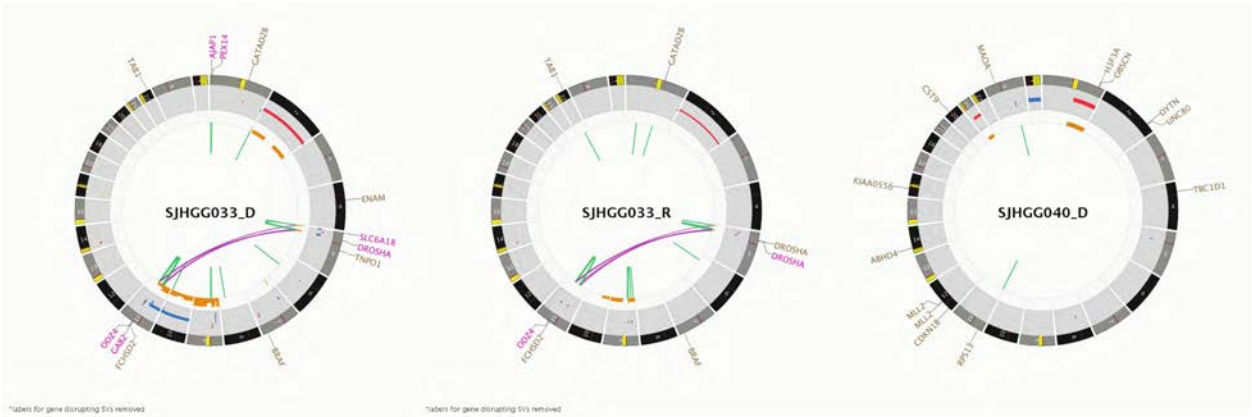



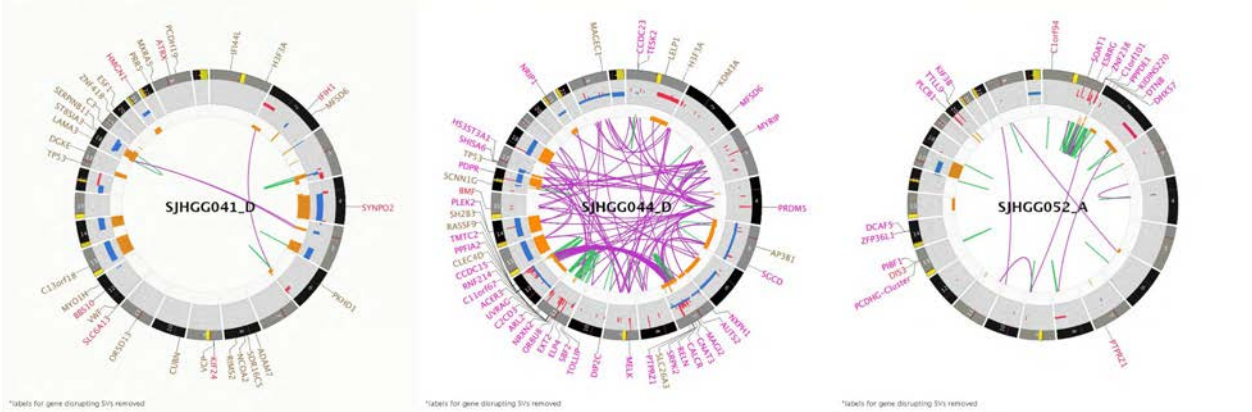

*labels for gene disrupting SVs removed

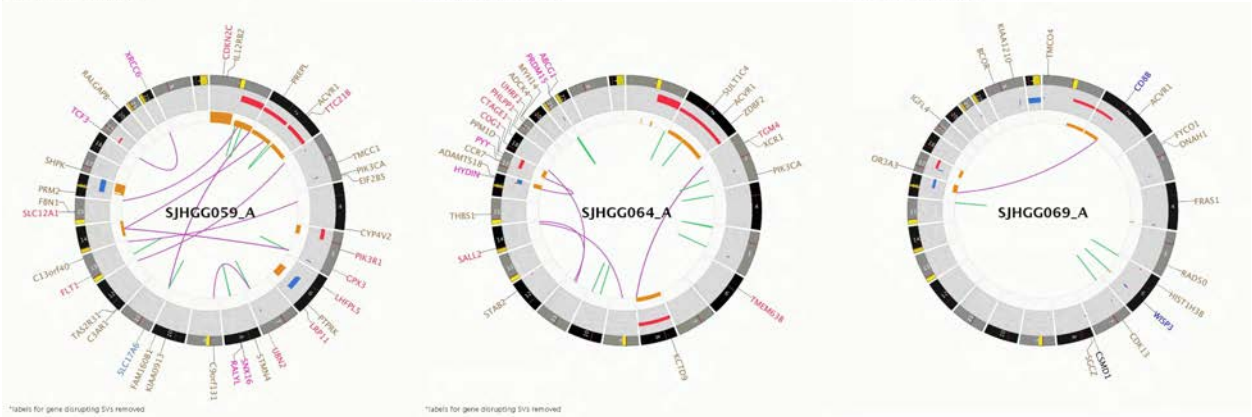


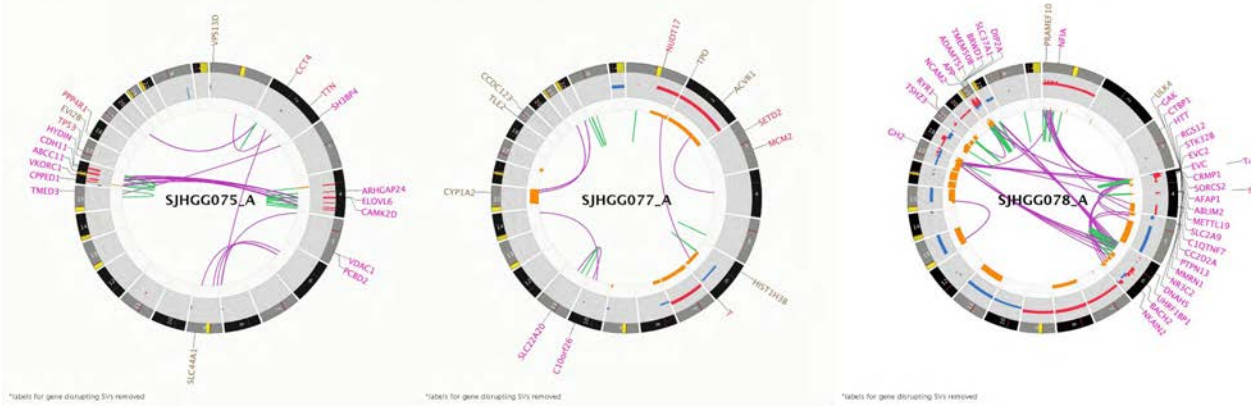




*labels for gene disrupting SVs removed


SJHGG033_D
SJHGG033_R
SJHGG040_D
*labels for gene disrupting SVs removed


SJHGG041_D
SJHGG044_D
SJHGG052_A
*labels for gene disrupting SVs removed


SJHGG059_A
SJHGG064_A
SJHGG069_A
*labels for gene disrupting SVs removed


SJHGG075_A
SJHGG077_A
SJHGG078_A
*labels for gene disrupting SVs removed

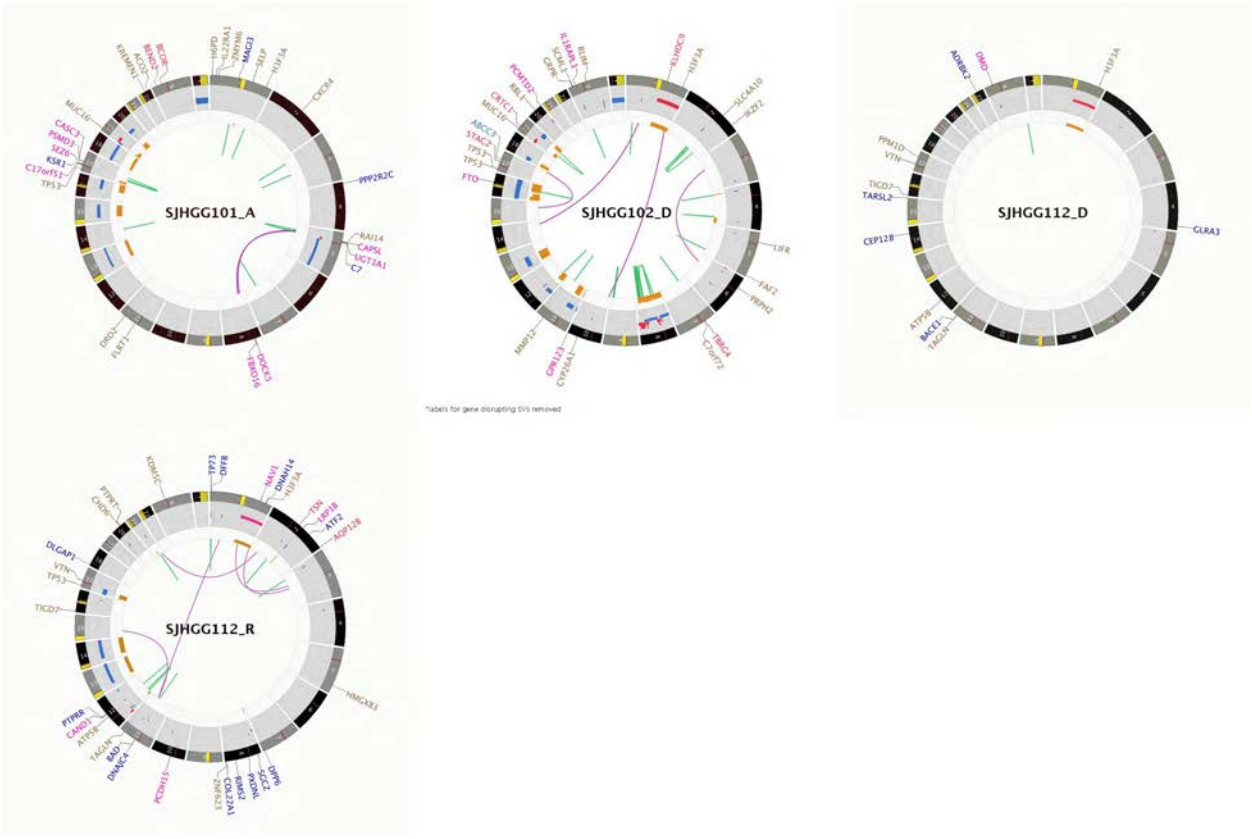
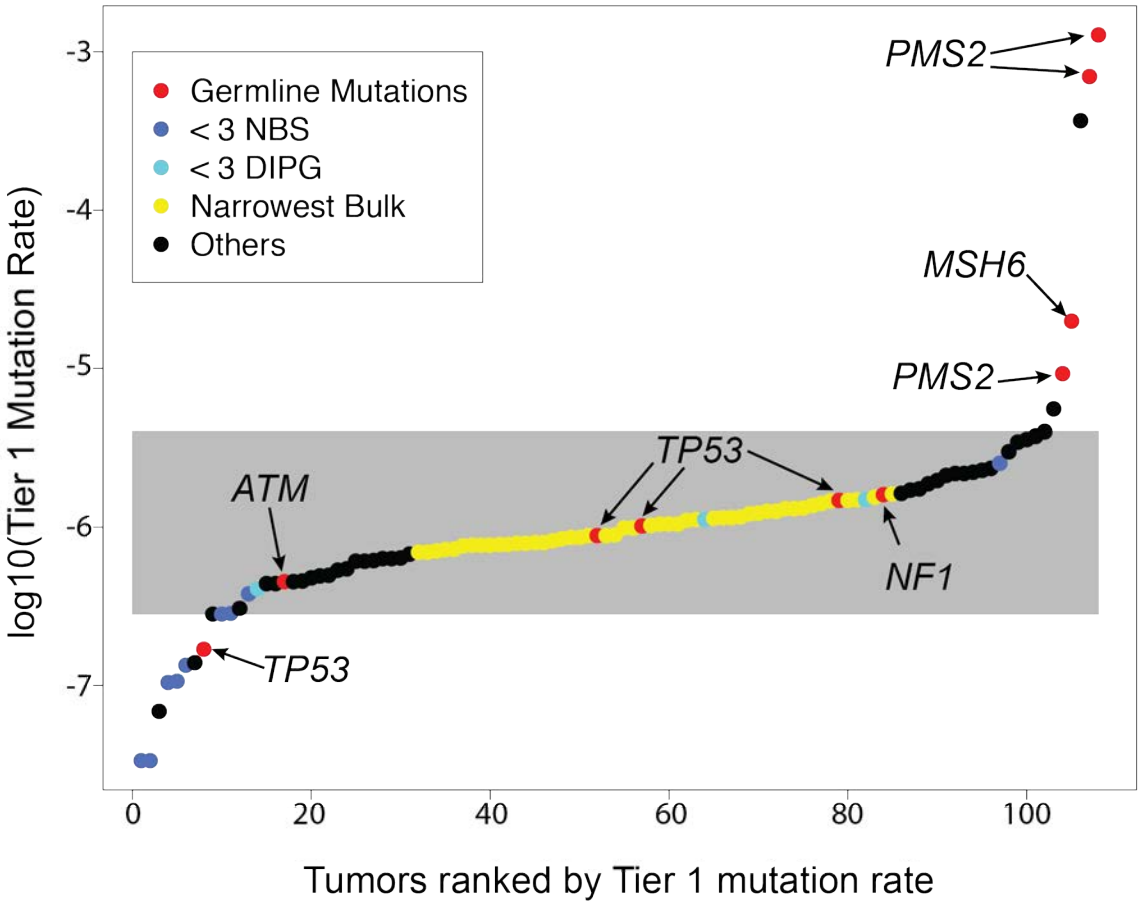


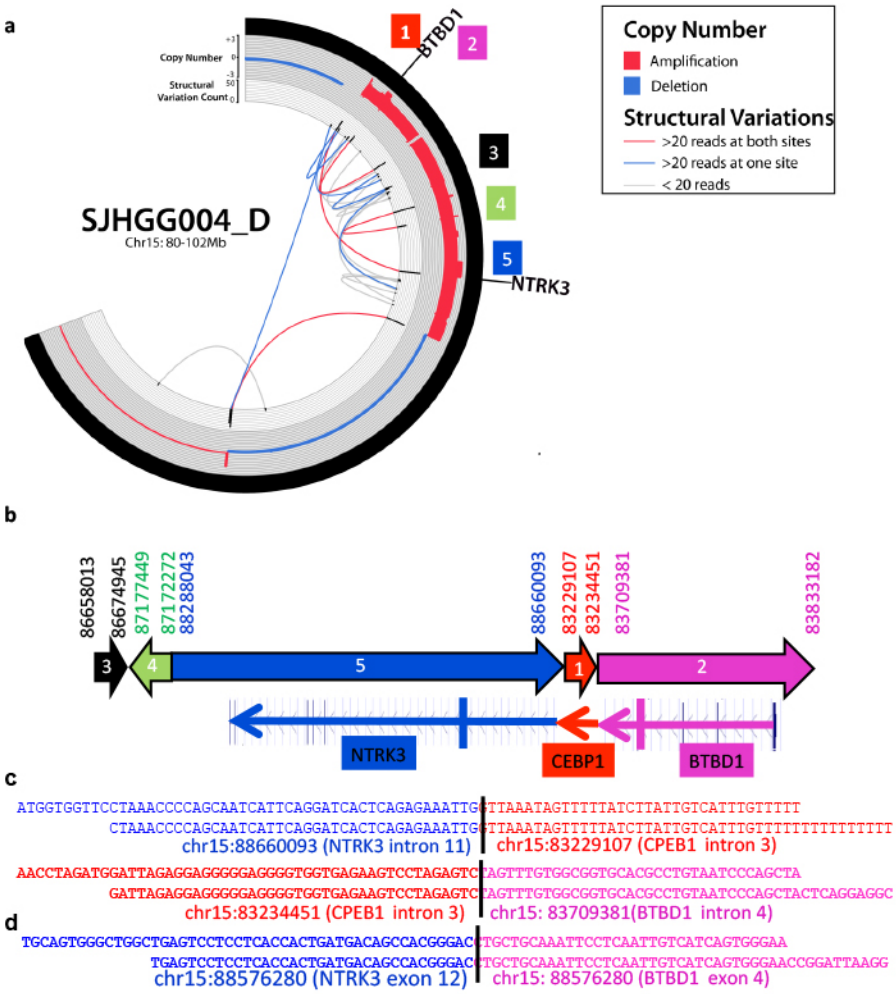

Figure legends are the same as Figure 5.

**Supplementary Figure 7.** Outlier analysis of the tier1 mutation rate in all pediatric HGGs




Least median squares was used to find the narrowest interval containing at least 50% of the samples (shown in yellow) in terms of their tier1 mutation rate. An approximate 95% prediction interval was determined by modeling this interval as the inter-quartile range of a normal distribution (shown by the gray box). Only samples with matched germline DNA were used for this analysis. The mutation rate is significantly lower in tumors from children less than 3 years of age than in the others ($p = 0.0005823$, Wilcoxon rank-sum test), with NBS-HGG tumors from children less than three shown in blue and DIPGs from children less than 3 years of age shown in turquoise. Germline mutations are shown in red, with specific mutated genes labeled. Hypermutator samples were associated with germline *PMS2* mutations and outlier mutation rates were also seen in samples with germline *MSH6* or *PMS2* mutation. In contrast, *TP53* germline mutations show a range of mutation rates, including an outlier for low mutation rate. Germline ATM mutation was associated with a lower mutation rate than the narrowest bulk. A tumor from a patient with germline *NF1* mutation fell within the narrowest bulk.

**Supplementary Figure 8.** Complex re-arrangement resulting in *BTBD1-NTRK3* fusion in tumor SJHGG004_D (a-d) and DIP2C-PDGFRA in tumor SJHGG003_A (e-h).

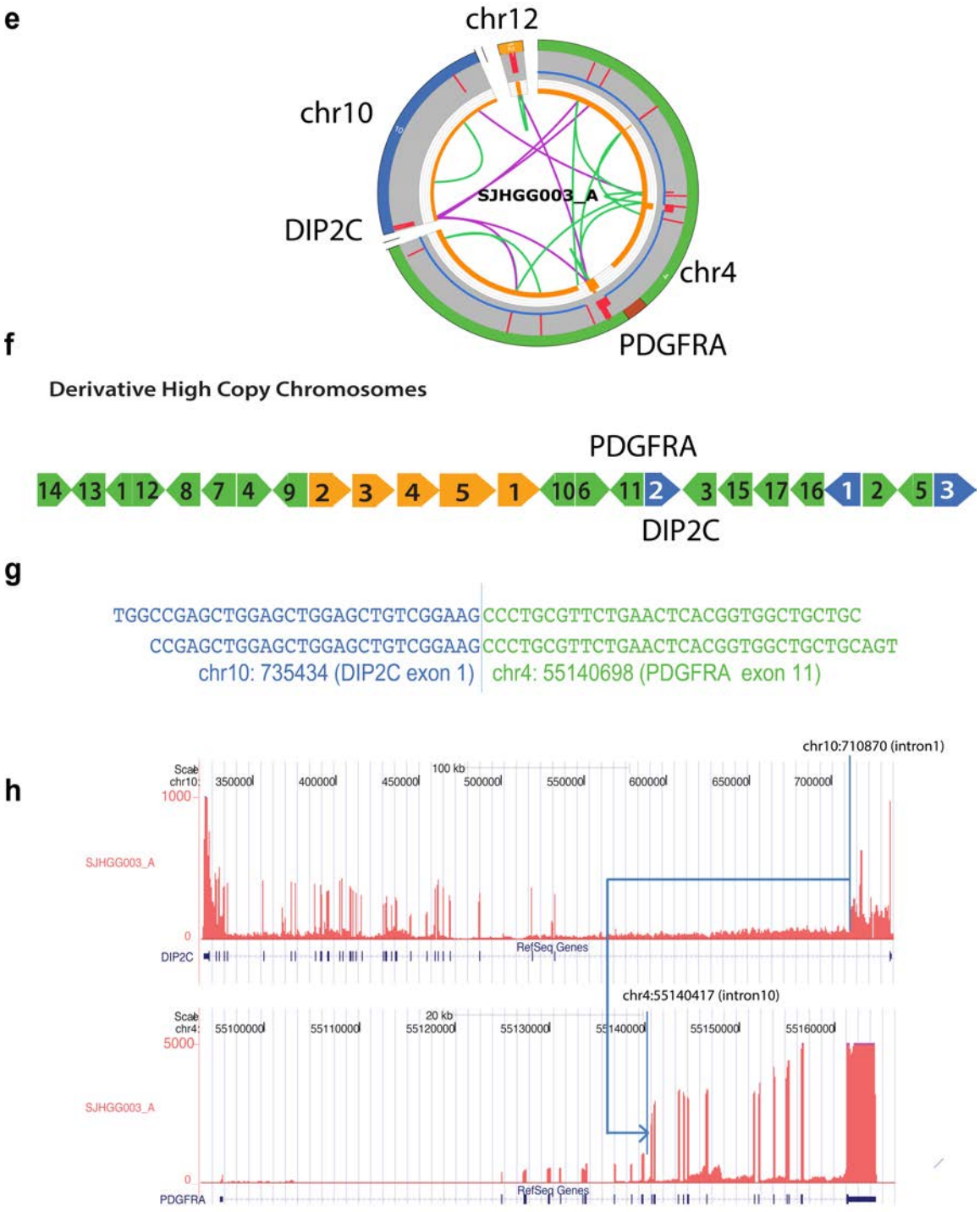



a) A total of 21 structural variations (SVs) were found in a 20Mb region on chromosome 15. The SV breakpoints match CNV segments with oscillating copy number gain of 2 and 4 in this region. Five CNV segments with the highest amplification are marked as segment 1 to 5 by their genomic location.

b) Re-arrangement of segments 1-5 forms a 525kb region generating a chimeric gene comprised of *BTBD1, CPEB1* and *NTRK3*. SV breakpoints marked in base-pair resolution connect intron 4 of *BTBD1* to intron 3 of *CPEB1* and intron 3 of *CPEB1* to intron 11 of *NTRK3*. There is no SV that directly connects *BTBD1* to *NTRK3* as the two are separated by a 5kb intron 3 segment of CPEB1. Fusion of *BTBD1-NTRK3* can be formed only if all SVs occur in cis on the same haplotype.

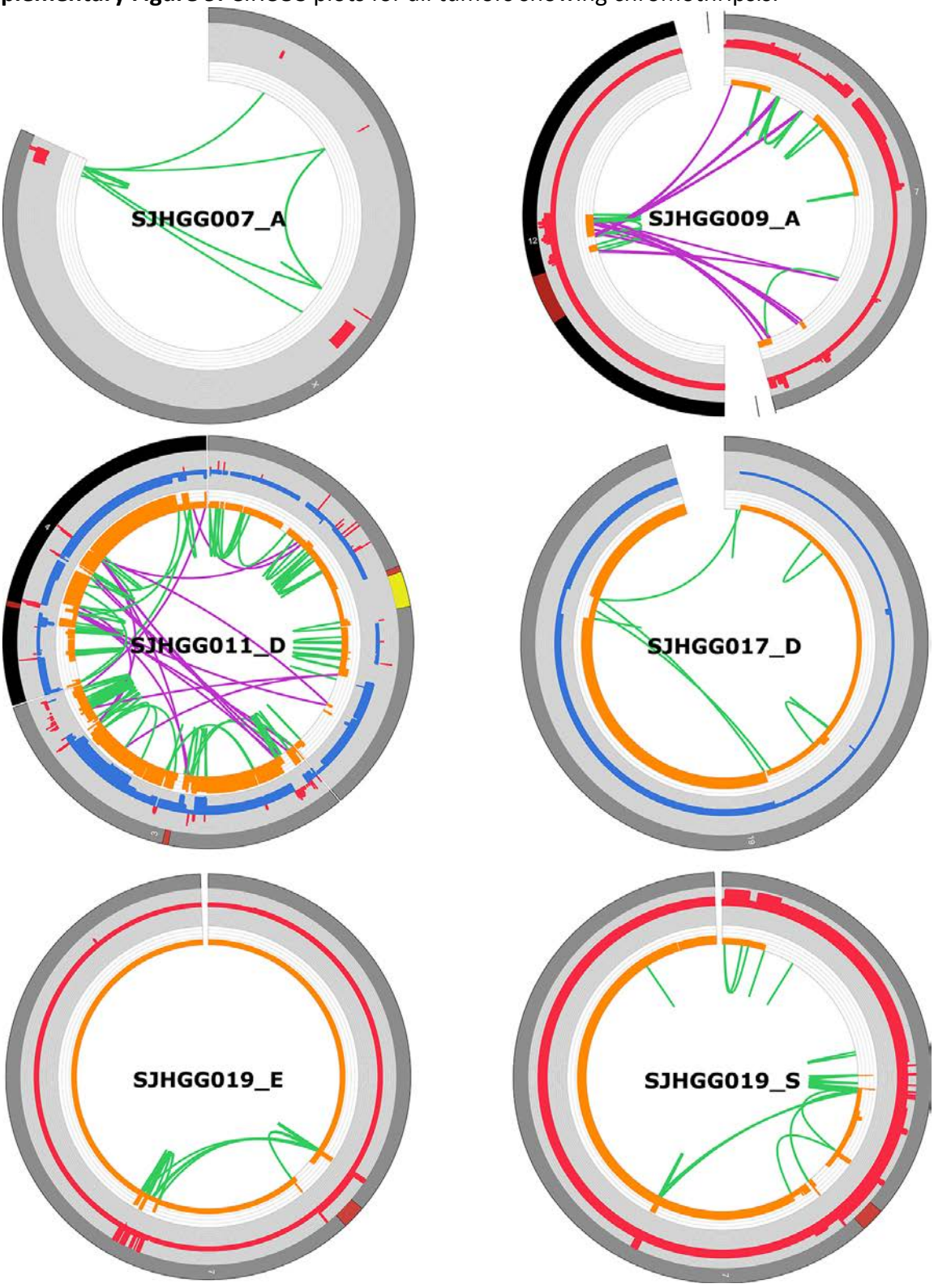
c) DNA junction reads show *BTBD1* is connected to *CPEB1* while *CPEB1* is connected to *NTRK3*.

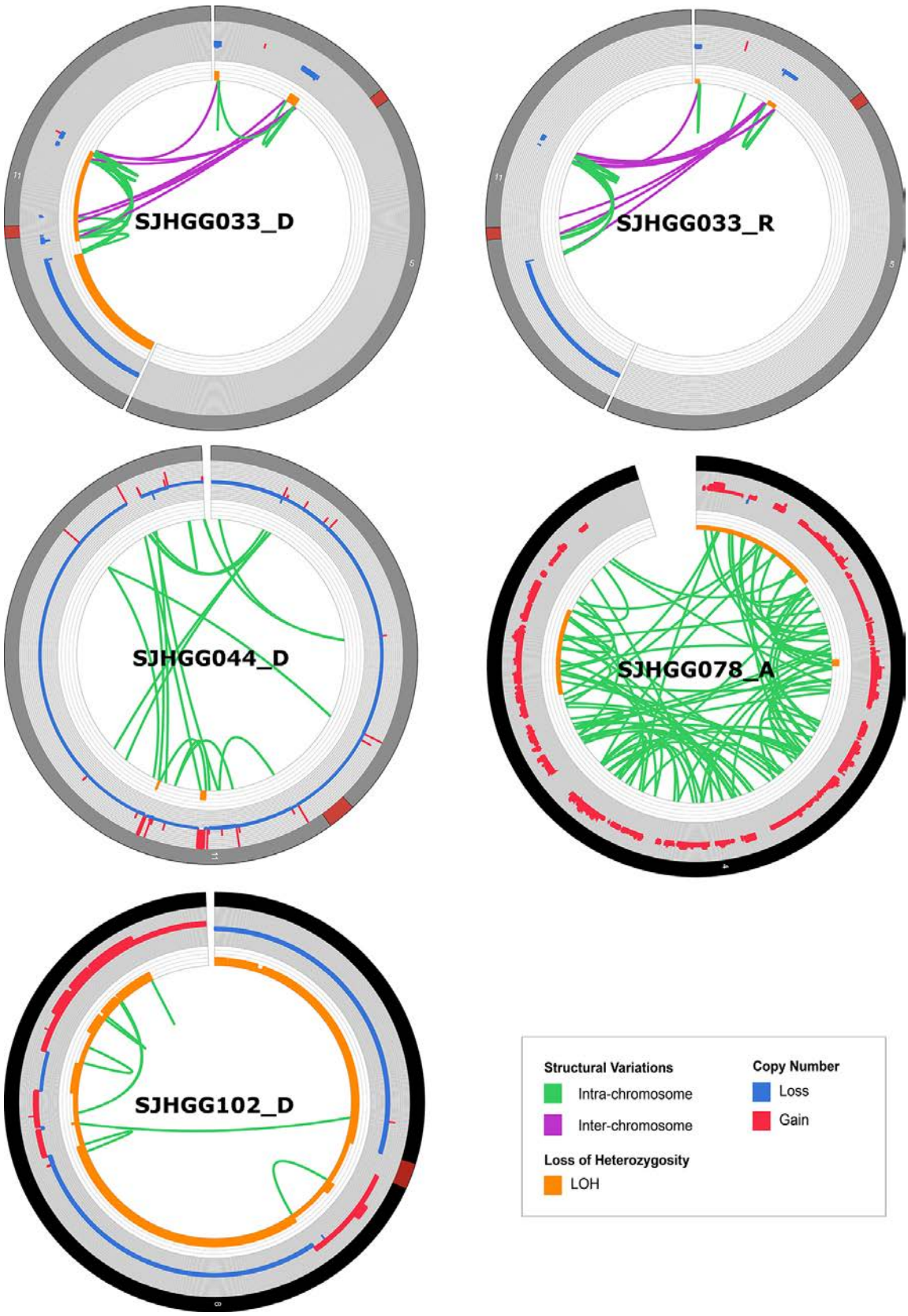
d) RNA-seq junction reads show exon 4 of *BTBD1* is joined to exon 12 of *NTRK3* forming an in-frame fusion protein. No additional fusion products were detected indicating that all SVs occur *in cis* on the same haplotype.



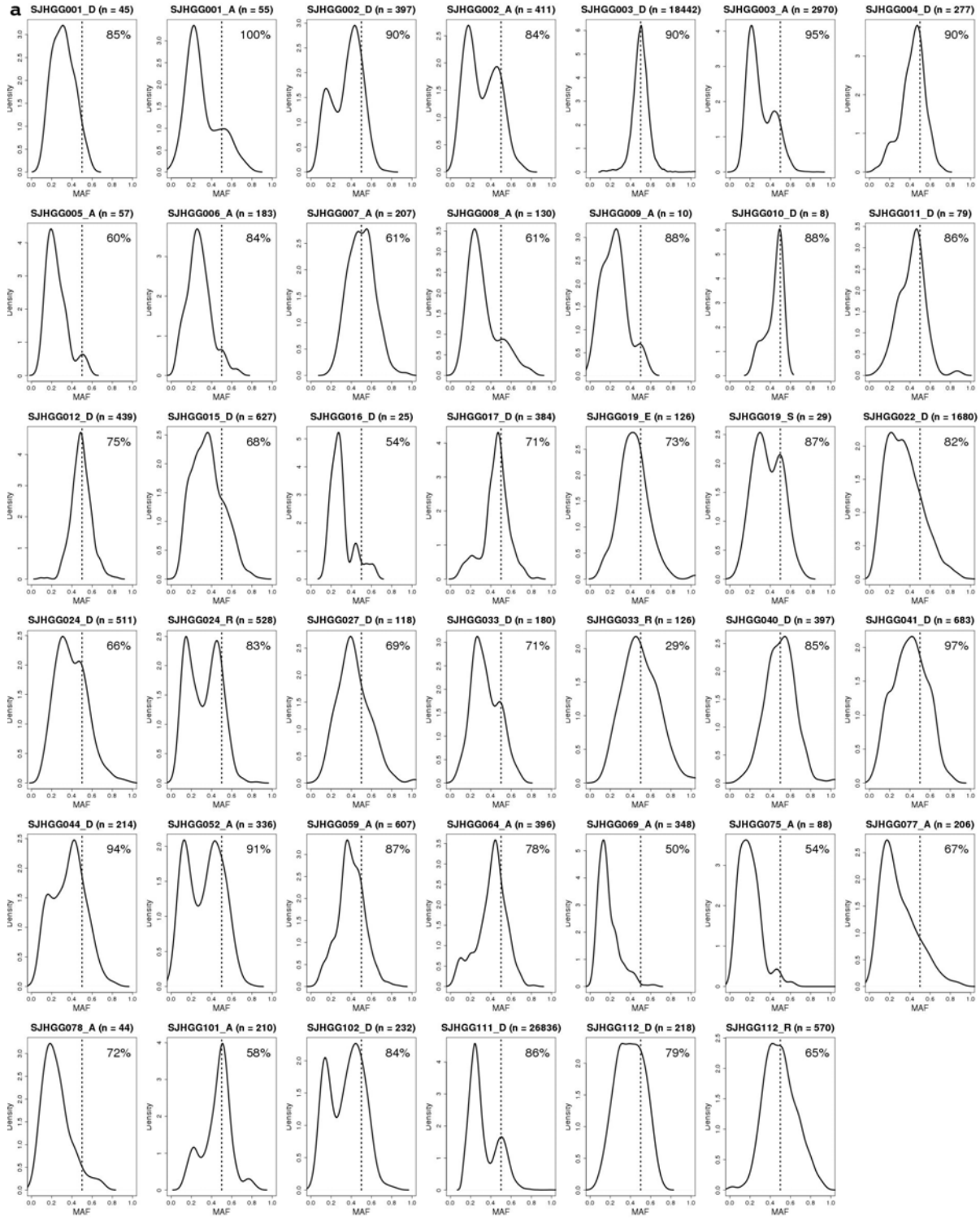

e) A CIRCOS plot showing multiple segments from chromosomes 4, 10 and 12 amplified and inter-connected. f) The segments from three different chromosomes are re-arranged in random order and random orientations. g) The reads illustrating the junction of the cDNA fusion product. h) The coverage wiggle plots at *DIP2C* and *PDGFRA* showing the elevated expression of the fusion product compared with their natural form.

**Supplementary Figure 9.** CIRCOS plots for all tumors showing chromothripsis.

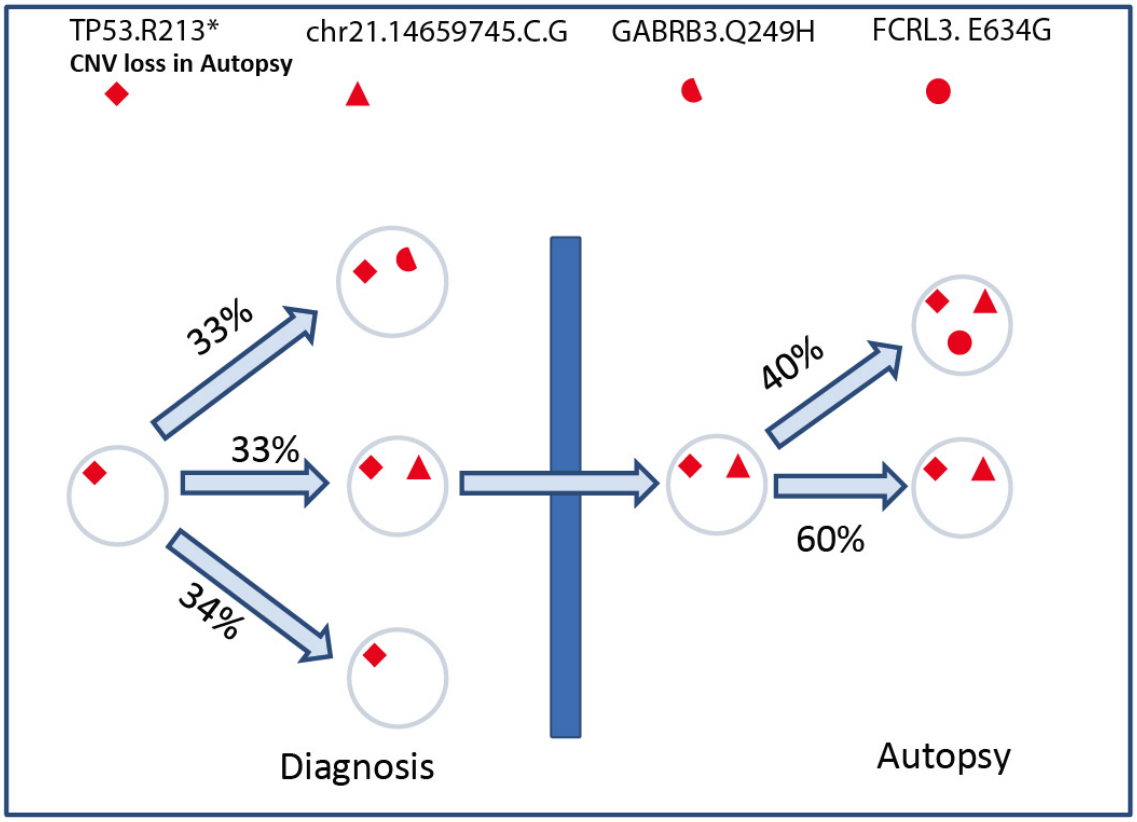



SJHGG033_D
SJHGG033_R
SJHGG044_D
SJHGG078_A
SJHGG102_D
Structural Variations
Intra-chromosome
Inter-chromosome
Loss of Heterozygosity
LOH
Copy Number
Loss
Gain

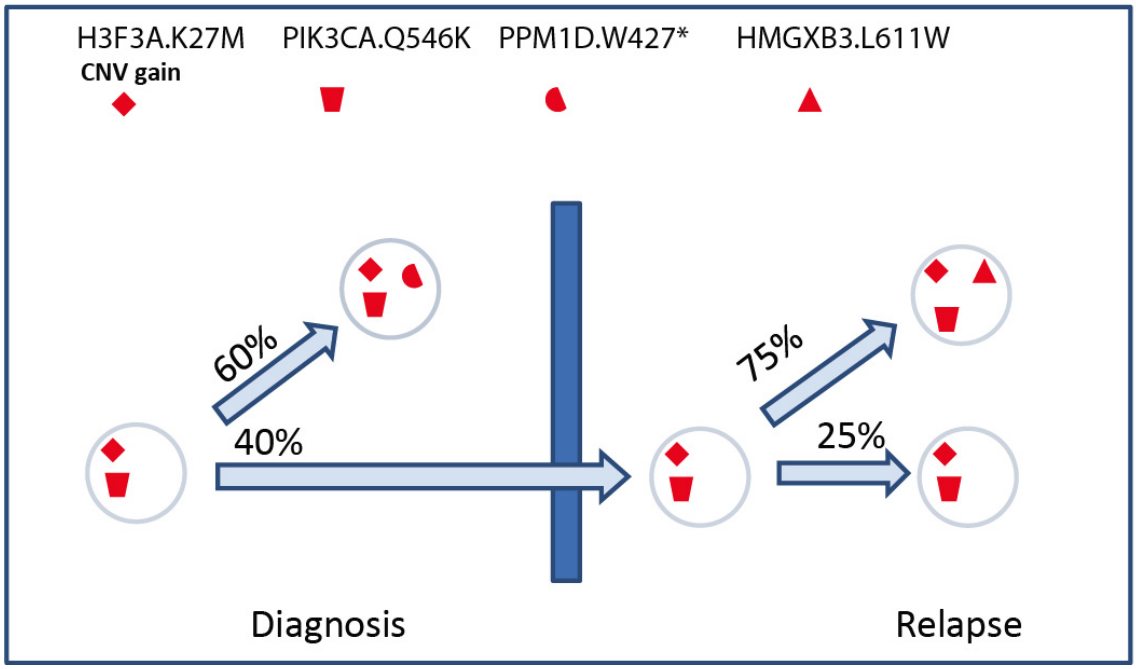
**Supplementary Figure 10.** Tumor heterogeneity analysis of all WGS tumors (a) and clonal evolution analysis for diagnosis to autopsy of SJHGG002 (b) and for diagnosis to relapse of SJHGG112 (c)**.**



b




c




Each sub-clone in a tumor is labeled by a reprehensive SNV mutation selected by biological relevance (e.g. *TP53, H3F3A*) when possible. Population size (i.e., percentage of all tumor cells) of a sub-clone is derived from the mean MAF of all SNVs in this sub-clone multiplied by 2 to account for a heterozygous allele in a diploid region.

## 3. Supplementary Tables; Titles and Legends

**Supplementary Table 1.** Clinical, pathological, sequencing information and genetic lesions on all 127 tumors in this study

Suffix '_D' or '_E' denotes a biopsy from first diagnostic surgery. The relapse tumors have the suffix ' _R' or '_S'. Tumors with the suffix ' _A' or '_A1' were collected at autopsy.

**Supplementary Table 2.** Coverage and mapping status of WGS (a), WES (b) and RNA-seq (c) data

**Genomic Coverage**: the average coverage of all non-ambiguous bases in GRCh37-lite.
**Exon Coverage**: the average coverage at all exonic bases (including all noncoding RNAs annotated in RefSeq).
**% Genomic bases covered**: the percentage of all non-ambiguous bases covered at least 10x.
**% Exonic bases covered**: the percentage of all bases in RefSeq annotated exons covered at least 10x.
**% Coding bases covered**: the percentage of all RefSeq protein coding bases covered at least 10x.
**% SNP detection**: concordance of genotype calls derived by WGS and those of Affymetrix SNP 6.0

**Supplementary Table 3.** Summary of somatic sequence mutations, somatic SVs and somatic copy number abnormalities identified in 42 WGS HGGs (a) and 66 paired WES HGGs (b).

* **Non-silent SNVs** include missense, nonsense and splice
** **Other SNVs** include SNVs in UTRs, noncoding exons and introns.
^ **Tiers 2-3** include only high-quality or validated mutations.
**Tier1 covered**: the number of gene coding bases covered with at least 10x in both tumor and matched normal samples
**Tier1_mutation_rate**: mutation rate computed as described in the Supplementary Methods
**BMR_tier1_silent**: background mutation rate computed based on the number of silent mutation and covered silent bases
**Tier3 covered**: the number of tier3 bases covered with at least 10x in both tumor and matched normal samples
**BMR_tier3**: background mutation rate computed based on the number of Tier3 mutations and Tier3 covered bases.

**Supplementary Table 4.** Sequence mutations (a: Tier1, b: Tier2, c: Tier3) identified for 40 non-hypermutator WGS cases

Both somatic SNVs and indels are included

Column definition is listed below:

**GeneName**: HUGO gene symbol
**Sample**: name of the sample
**Chromosome**: chromosome
**WU_HG19_Pos**: chromosome position in hg19 coordinates.
**Class**: classification based on amino acid change pattern. 'exon' refers to mutations in non-coding RNA genes; 'splice_region' refers to mutations not directly affecting the canonical splice sites but located within 10bp of the canonical splice sites.
**AAChange**: predicted amino acid change for the mutation
**ProteinGI**: NCBI protein GI number
**mRNA_acc**: RefSeq accession number
**# Mutant_In_Tumo**r: number of WGS reads containing mutant allele (tumor)
**# Total_In_Tumor**: number of WGS reads covering the site (tumor)
**Reference Allele**: the allele represented in the reference human genome. Reference allele is marked as '–' for an insertion.
**Mutant Allele:** non-reference allele
**Flanking**: 20bp [reference allele/mutant allele] 20bp
**Validation_status**: is this a putative mutation or an experimentally validated mutation

**Supplementary Table 5.** Sequence mutations (including both SNVs and indels) identified for non-hypermutator WES tumors

The column headers are the same as above.

**Supplementary Table 6.** Copy number alterations identified in WGS data

**Seg.mean** is the mean of the difference (adjusted for GC content) between the normalized read depth in paired Diagnosis (D) and Germline (G) samples in each genomic segment.
**$Log_2$Ratio** is the mean of the GC-corrected $\log_2$ (read depth of normalized D/G) in each genomic segment.
**Label**: whether it's supported by structural variant breakpoints.
**No. Overlapping Genes**: how many genes are in the regions
**Gene List**: genes in the regions. When there are the more than 10 genes, the first 5 and the last 5 genes are showed.

**Supplementary Table 7.** Structural variations identified in WGS samples

| Column Title | Column Definition |
|---|---|
| Sample | Sample name |
| ChrA | Chromosome for breakpoint A |
| PosA | Position of breakpoint A |
| OrientationA | + Region to the left of PosA is included in mutant genotype<br>- Region to the right of PosA is included in mutant genotype |

| Column Title | Column Definition | |
|---|---|---|
| ChrB | Chromosome for breakpoint B | |
| PosB | Position of breakpoint B | |
| OrientationB | + Region to the right of PosA is included in mutant genotype<br>- Region to the left of PosA is included in mutant genotype | |
| Type | INS Insertion<br>DEL Deletion<br>INV Inversion<br>ITX Intrachromosomal translocation<br>CTX Interchromosomal translocation | |
| Usage | GENIC | Both endpoints were in genes: checked for fusion |
| | HALF_INTERGENIC | One endpoint was in a gene: checked for truncation |
| | CO_GENIC | Both endpoints were in genes: checked for and found fusion that involved multiple events |
| | INTERGENIC / INTRONIC | Neither endpoint was in a gene or both were in the same intron of a gene; no gene fusion or truncation |
| | INVERTED_REPEAT | Both endpoints were in the same gene, but in opposite orientations: checked for truncation |
| Gene | Fusion or truncated gene that would result from structural variation | |
| Chromosomes | Chromosomes involved in the rearrangement | |
| Valid CDS | Number of predicted fusion transcripts with an annotated CDS start and stop | |
| In-Frame CDS | Number of “Valid CDS” transcripts with a CDS length divisible by three. | |
| Modified In-Frame CDS | Number of “In-Frame CDS” transcripts that are not identical to an existing annotated transcript. | |
| mutA | Number of reads supporting the structural variation at breakpoint A | |
| mutB | Number of reads supporting the structural variation at breakpoint B | |
| Validation Status | Valid | The SV has been experimentally validated |
| | Putative | The SV has yet to be validated |

**Supplementary Table 8.** Structural variations identified in RNAseq samples

**Sample**: PCGP tumor sample ID
**geneA**: fusion gene A
**chrA**: chromosome for fusion gene A
**posA**: hg19 genomic location for the breakpoint in fusion gene A
**ortA**: orientation for fusion gene A
**geneB**: fusion gene B.
**chrB**: chromosome for fusion gene B

**posB**: hg19 genomic location for the breakpoint in fusion gene B
**ortB**: orientation for fusion gene B
**readsA**: the number of reads supporting breakpoint in gene A
**readsB**: the number of reads supporting breakpoint in gene B
**contig**: the consensus sequence surrounding the breakpoint
**validation_status**: if it's validated or putative (not validated yet)
**validation_source**: the validation method (MiSeq or Sanger sequencing)

**Supplementary Table 9.** Recurrent Genetic Alterations shown in Figure 1

The samples are arranged in the same order as in Figure 1 from left to right.

**Supplementary Table 10.** Gene expression analysis of DIPGs with and without ACVR1 mutations (a) and Gene Ontology enrichment analysis of differentially expressed genes (b).

**Supplementary Table 11.** Germline SNVs and indels found in SJHGG patients

The column header is the same as supplementary table 4, with the additional columns as below:

**#Mutant_In_Normal**: number of WGS reads containing mutant allele in the normal sample
**#Total_In_Normal**: number of WGS reads covering the site in the normal sample

**Supplementary Table 12.** Statistical analysis of chromothripsis events in pediatric HGGs

**Supplementary Table 13.** Primer sequences used for recurrence screening and validation.

## 4. Supplementary Information References


1. Pleasance, E.D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* 463, 191-6 (2010).
2. Lee, W. *et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* 465, 473-7 (2010).
3. Rousseeuw, P. Least Median Squares Regression. *Journal of the American Statistical Association* 79, 871-880 (1984).
4. McKinnon, P.J. ATM and the molecular pathogenesis of ataxia telangiectasia. *Annu Rev Pathol* 7, 303-21 (2012).
5. DeWire, M.D. *et al.* Radiation therapy and adjuvant chemotherapy in a patient with a malignant glioneuronal tumor and underlying ataxia telangiectasia: a case report and review of the literature. *J Clin Oncol* 31, e12-4 (2012).
6. Stephens, P.J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* 144, 27-40 (2011).
7. Korbel, J.O. & Campbell, P.J. Criteria for inference of chromothripsis in cancer genomes. *Cell* 152, 1226-36 (2013).
8. Cools, J. *et al.* A tyrosine kinase created by fusion of the PDGFRA and FIP1L1 genes as a therapeutic target of imatinib in idiopathic hypereosinophilic syndrome. *N Engl J Med* 348, 1201-14 (2003).
9. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214-8 (2013).
10. Chaikuad, A. *et al.* Structure of the bone morphogenetic protein receptor ALK2 and implications for fibrodysplasia ossificans progressiva. *J Biol Chem* 287, 36990-8 (2012).
11. Cannon, J.E., Upton, P.D., Smith, J.C. & Morrell, N.W. Intersegmental vessel formation in zebrafish: requirement for VEGF but not BMP signalling revealed by selective and non-selective BMP antagonists. *Br J Pharmacol* 161, 140-9 (2010).
12. Shen, Q. *et al.* The fibrodysplasia ossificans progressiva R206H ACVR1 mutation activates BMP-independent chondrogenesis and zebrafish embryo ventralization. *J Clin Invest* 119, 3462-72 (2009).