Recurrent noncoding U1 snRNA mutations drive cryptic splicing in SHH medulloblastoma

Hiromichi Suzuki1,2,65, Sachin A. Kumar1,2,65, Shimin Shuai4,5, Ander Diaz-Navarro6,7, Ana Gutierrez-Fernandez8, Pasqualino De Antonellis1,2, Florence M. G. Cavalli1,2, Kyle Juraskich1,2, Hamza Farooq1,2,3, Ichiyo Shibahara1,2, Maria C. Vladoiu1,2,3, Jiao Zhang1,2, Namal Abeysundara1,2, David Przelicki1,2,3, Patryk Skowron1,2,3,4, Nicole Gauer1,2, Betty Luu1,2, Craig Daniels1,2, Xiaochong Wu1,2, Antoine Forget8,9, Ali Momin1,2,5, Jun Wang10, Weifan Dong1,2,5, Seung-Ki Kim11, Wieslawa A. Grajkowska12, Anne Jouvet13, Michelle Fèvre-Montange14, Maria Luisa Garrè15, Amulya A. Nageswara Rao16, Caterina Giannini17, Johan M. Kros18, Pim J. French19, Nada Jabado20, Ho-Keung Ng21, Craig Daniels1,2, Xiaochong Wu1,2, Antoine Forget8,9, Ali Momin1,2,5, Jun Wang10, Weifan Dong1,2,5, Seung-Ki Kim11, Wieslawa A. Grajkowska12, Anne Jouvet13, Michelle Fèvre-Montange14, Maria Luisa Garrè15, Amulya A. Nageswara Rao16, Caterina Giannini17, Johan M. Kros18, Pim J. French19, Nada Jabado20, Ho-Keung Ng21, Cristina Gómez-González22, Charles G. Eberhart23,24,25, Ian F. Pollack26, James M. Olson27, William A. Weiss28,29,30, Toshihiro Kikuchi31, Enrique López-Aguilar32, Boleslaw Lach33,34, Maura Massimino35, Erwin G. Van Meir36,37,38, Joshua B. Rubin39,40, Rajeev Vadhvan41, Robert J. Wechsler-Reya42, Swneke D. Bailey43,44, Livia Garzia45,46, A. Sorana Morrissy45,59, Marco A. Marra47, Xi Huang1,2, David Malkin62, Olivier Ayrault8,9, Vijay Ramaswamy2,62, Xose S. Puente6,7, John A. Calarco63, Lincoln Stein4 & Michael D. Taylor1,2,3,64

In cancer, recurrent somatic single-nucleotide variants—which are rare in most paediatric cancers—are confined largely to protein-coding genes1–3. Here we report highly recurrent hotspot mutations (r.3A>G) of U1 spliceosomal small nuclear RNAs (snRNAs) in about 50% of Sonic hedgehog (SHH) medulloblastomas. These mutations were not present across other subgroups of medulloblastoma, and we identified these hotspot mutations in U1 snRNA in only <0.1% of 2,442 cancers, across 36 other tumour types. The mutations occur in 97% of adults (subtype SHHδ) and 25% of adolescents (subtype SHHα) with SHH medulloblastoma, but are largely absent from SHH medulloblastoma in infants. The U1 snRNA mutations occur in the 5′ splice-site binding region, and snRNA-mutant tumours have significantly disrupted RNA splicing and an excess of 5′ cryptic splicing events. Alternative splicing mediated by mutant U1 snRNA inactivates tumour-suppressor genes (PTCH1) and activates oncogenes (GLI2 and CCND2), and represents a target for therapy. These U1 snRNA mutations provide an example of highly recurrent and tissue-specific mutations of a non-protein-coding gene in cancer.

Medulloblastoma, a cerebellar neuronal cancer, comprises four molecular subgroups (WNT, SHH, group 3 and group 4), each of which has its own distinct clinical, transcriptomic and genetic make-up4–6. Each of these four molecular subgroups can be further subdivided into subtypes; for SHH medulloblastoma, these comprise SHHα, SHHβ, SHHγ and SHHδ7. Noncoding single-nucleotide variants have recently been discovered in the promoter regions of TERT and a handful of other loci, which has given impetus to the careful examination of noncoding segments8,9. Thus, we sought to explore the genomic landscape of medulloblastoma with a particular focus on noncoding regions. We analysed whole-genome sequencing data of 114 medulloblastomas, and observed a recurrent hotspot mutation of noncoding U1 snRNA genes in 10 out of 114 cases (8.8%) (Fig. 1a, Extended Data Fig. 1, Supplementary Tables 1, 2, Methods). Hotspot mutations of U1 snRNA genes occur in the third nucleotide (r.3A>G), and are restricted to SHH medulloblastoma. These hotspot mutations are localized within the 5′ splice-site recognition sequence, which has been ultra-conserved in eukaryotes through nearly one billion years of evolution (Fig. 1b, Extended Data Fig. 2a). The human reference genome (hg19) has four annotated U1 snRNA genes (RNU1-1, RNU1-2, RNU1-3 and RNU1-4) and three ‘pseudogenes’ (RNU1-27P, RNU1-28P and RNU1-18P), all of which encode completely identical 164-base-pair transcripts. In addition, there are over 100 U1 snRNA pseudogenes spread across the genome, which highly complicates their identification by mutation callers owing to the inability to align short reads to any individual U1 snRNA gene10 (Extended Data Fig. 3). We remapped sequence reads that permitted multimapping, and successfully detected the U1 snRNA mutation in five additional cases (Methods). We validated hotspot U1 snRNA mutations in an additional 40 out of 227 cases of
medulloblastoma from the International Cancer Genome Consortium (ICGC) (Supplementary Tables 2–4). We also detected recurrent hotspot mutations of the U1 snRNA gene (RNU11) at the fifth nucleotide (r.3A>G), in the highly conserved 5′ splice-site recognition sequence (in a total of 4 out of 341 cases) (Extended Data Fig. 2b–d, Supplementary Table 2). Taken together, 51% (56 out of 109) of SHH medulloblastomas have at least one mutation in U1 or U11 snRNA (Fig. 2). These snRNA mutations significantly co-occur with mutations of the TERT promoter and DDX3X (Supplementary Tables 3, 6). We assessed the U1 snRNA r.3A>G mutations across 2,442 samples drawn from 36 cancer histologies of the ICGC, and found such a mutation in only one sample (0.04%)—a lone sample of pancreatic ductal adenocarcinoma (Supplementary Table 7). We conclude that U1 snRNA r.3A>G mutations are both highly recurrent in and extremely specific to, SHH medulloblastoma.

We validated the U1 snRNA r.3A>G mutations in an additional 159 cases of SHH medulloblastoma using allele-specific PCR. We detected mutations in the RNU1/27P and/or RNU1/28P genes (which were confirmed by Sanger sequencing) that were not identified by whole-genome sequencing (Extended Data Fig. 4a, b, Supplementary Table 8, Methods). Combining the results of whole-genome sequencing and allele-specific PCR, we found that U1 snRNA r.3A>G mutations were largely restricted to cases of SHH medulloblastoma in adulthood (SHHδ, present in 97% of cases) and adolescence (SHHβ, present in 25% of cases), and absent to cases of SHH medulloblastoma in adulthood (SHHδ, present in 97% of cases) and adolescence (SHHβ, present in 25% of cases), and absent from those in infancy (Fig. 3a, b). This remains true if accounting only for age, and not for molecular subtype. Indeed, most patients with SHHδ with TP53 mutations also have U1 snRNA r.3A>G mutations (Fig. 3c). Both broad and focal somatic copy-number variations are divergent between SHHδ with wild-type U1 snRNA, SHHβ with mutant U1 snRNA

**Fig. 1** Highly recurrent mutations of the U1 snRNAs in SHH medulloblastoma. a, Cartoon illustrating the number and subgroup-specific distribution of somatic mutations in the U1 snRNA genes. U1 snRNA sequence conservation scores, as determined using the Rfam database (http://rfam.xfam.org/). b, Secondary structure of the mutant U1 snRNA. The red circle identifies the location of the hotspot mutation. The yellow and green rectangles indicate the 5′ splice-site recognition site and the Sm protein-binding site, respectively. Numerals I to IV indicate stem loops.

**Fig. 2** Mutational repertoire of snRNA-mutant SHH medulloblastomas. Genomic landscape of mutations in SHH medulloblastomas (n = 109), with and without U1 or U11 mutations. Odds ratios (red dots) of coexistence of U1 and U11 snRNA mutations with other somatic events are shown with their 95% confidence intervals. Arrowheads represent values that are out of the axis range. Significantly correlated mutations are denoted in red (false discovery rate [FDR] < 0.1, asymptotic P values from odds-ratio tests (H0: odds ratio = 1, Methods) with Benjamini–Hochberg adjustment for multiple testing.
medulloblastoma with mutant U1 snRNA demonstrate enrichment of 6a–c, Supplementary Table 11). The U1 snRNA r.3A>G mutations would medulloblastomas with wild-type U1 snRNA11 (Extended Data Figs. 5a, b, 2.5–3 times more alternative 5′ cryptic splicing events than do SHH with an extremely poor prognosis. of both a mutation and a U1 snRNA r.3A>G mutation is associated TP53 0.46–15.88, P 0.11; TP53 mutation: hazard ratio 3.72, 95% confidence interval 0.74–18.87, P = 0.27). This suggests that, within SHHα, the combination of both a TP53 mutation and a U1 snRNA r.3A>G mutation is associated with wild-type U1 snRNA have an increased incidence of copy-number variations that encompass several oncogenes and tumour-suppressor genes, including MYCN, CCND2 and PPMID.

A univariate log-rank analysis of both progression-free survival and overall survival revealed that, within SHHα tumours, both U1 snRNA r.3A>G mutations confer a very strong risk of relapse (U1 snRNA r.3A>G mutation: hazard ratio 5.51, 95% confidence interval 1.15–26.35, P = 0.03; TP53 mutation: hazard ratio 3.01, 95% confidence interval 0.55–16.65, P = 0.21). A similar trend was observed for overall survival (U1 snRNA r.3A>G mutation: hazard ratio 3.72, 95% confidence interval 0.74–18.87, P = 0.11; TP53 mutation: hazard ratio 2.70, 95% confidence interval 0.46–15.88, P = 0.27). This suggests that, within SHHα, the combination of both a TP53 mutation and a U1 snRNA r.3A>G mutation is associated with an extremely poor prognosis.

Intron-centric alternative splicing analysis using LeafCutter confirms that mutant U1 snRNA variants of both SHHα and SHHδ have 2.5–3 times more alternative 5′ cryptic splicing events than do SHH medulloblastomas with wild-type U1 snRNA15 (Extended Data Figs. 5a, b, 6a–c, Supplementary Table 11). The U1 snRNA r.3A>G mutations would be predicted to affect the recognition of the sixth intronic nucleotide from the 5′ splice site; indeed, cryptic 5′ splice sites recognized in SHH medulloblastoma with mutant U1 snRNA demonstrate enrichment of a dominant C base, as opposed to the T base observed in tumours with wild-type U1 snRNA (Extended Data Figs. 5c, 6d, e). Pathway analysis of differentially expressed transcripts between SHH medulloblastoma with wild-type versus mutant U1 snRNA demonstrates an increase in nonsense-mediated decay, which is consistent with the destruction of aberrantly spliced transcripts (Extended Data Fig. 7a). To validate the effect of U1 snRNA mutations, we transfected wild-type or mutant U1 snRNA r.3A>G vectors into human embryonic kidney 293T cells, and examined the effects on splicing. Intron-centric analysis clearly demonstrates an enrichment of a C base at the sixth intronic position when using the mutant U1 snRNA vector, and a considerable increase in the incidence of cryptic 5′ splicing events that do not overlap with those of SHH medulloblastoma with wild-type U1 snRNA (Extended Data Fig. 7b–d, Supplementary Tables 12, 13).

Clustering on the basis of significant alternative splicing events is clearly driven by U1 snRNA mutational status (Extended Data Fig. 7e, Methods), and tumours with mutant U1 snRNA segregated distinctly from those with wild-type U1 snRNA. We conclude that the U1 snRNA r.3A>G mutation has a marked effect on alternative splicing in affected tumours.

As a complementary approach, we conducted exon-centric alternative splicing analysis using rMATS22. We observed that SHH medulloblastoma tumours with mutant U1 snRNA have a higher incidence of cassette exons than do wild-type U1 snRNA controls (Extended Data Figs. 8a–c, 9a, b, Supplementary Table 14). Similar to cryptic 5′ alternative splicing events, the dominant base at the sixth intronic base is C (Extended Data Figs. 8d, 9c, Supplementary Table 15). In addition, an increase of retained introns is observed in tumours with mutant U1 snRNA. The 5′ splice-site sequences of missed splice sites in retained introns do not have a dominant C at the sixth nucleotide, but instead have the canonical T. This latter result suggests a mechanism in which
Fig. 4 | Aberrant splicing of Hedgehog signalling genes in SHH medulloblastoma with mutant U1 snRNA. a, Overview of cryptic alternative splicing of PTCH1, demonstrating the position of a cryptic cassette exon with the 5′ splice-site sequence. b, Top, Sashimi plots of splicing of PTCH1 in representative cases. The bar plot shows counts per million reads (CPM). Numbers refer to the number of junctional reads. Bottom, annotated exon tracks with genomic positions marked. Junctural reads specific to U1 snRNA mutants are in red. c, Scatter plot comparing detected alternatively spliced read and total junction reads that shared the 5′ splice site. Jittering was performed for both values. d, Per cent spliced in values by SHHα with mutant U1 snRNA, SHHδ with mutant U1 snRNA and SHH medulloblastoma with wild-type U1 snRNA (all subtypes of SHH medulloblastoma). P values were calculated using two-sided Wilcoxon rank-sum test. e, Box plot of fold changes in expression of the alternatively spliced isoform of PTCH1 compared to the wild-type isoform of PTCH1 in subsets of SHH medulloblastoma, as determined by real-time qPCR. Data are mean ± s.d. P values were calculated using two-sided Wilcoxon rank-sum test. f, Illustration of canonical isoforms and the cryptic alternative isoform of PTCH1. Putative translation start sites are indicated with an arrow. Resulting proteins (and size) are displayed for each isoform. aa, amino acid; UTR, untranslated region. In the box plots (d, e), the centre lines show data median; box limits indicate the IQR from the 25th to the 75th percentiles; and lower and upper whiskers extend 1.5× the IQR. Outliers are represented by individual points.

mutant U1 snRNA r.3A>G not only recognizes alternative 5′ splice sites but also inhibits the wild-type U1 snRNA from detecting canonical splice sites, which results in their aberrant splicing. The retained intron event with the highest per cent spliced in, validated by real-time quantitative (q)PCR, occurs in the gene PAX6, (which undergoes frequent somatic mutation in SHH medulloblastoma) and a chromatin remodelling gene TOX443,444 (Extended Data Figs. 8e–h, 9d, Supplementary Table 16). The retained introns in both genes result in a frameshift, which leads to loss of function. These data may support a model in which U1 snRNA r.3A>G mutations impede normal splicing, which leads to intron retention and to an mRNA frameshift.

To detect pathogenic alternative splicing, we identified cryptic 5′ splicing events with a C base at the sixth intronic position that are shared by mutant U1 snRNA variants of both SHHα and SHHδ tumours (Extended Data Fig. 9e, Supplementary Table 17, 18). Using both RNA sequencing and real-time qPCR, we detected cryptic splicing events with high effect sizes in both PTCH1 and GLI2; these events were highly specific to U1 snRNA r.3A>G mutation (in both SHHα and SHHδ tumours), as compared to wild-type U1 snRNA controls (Fig. 4a–e). PTCH1 is known to have at least three different initial exons. Splicing mediated by the U1 snRNA r.3A>G mutant results in the inclusion of a cassette exon between exon 2 and exon 3, which causes a frameshift, and therefore predicted translation from the ATG in exon 3 (Fig. 4f). It has previously been reported that loss of expression of the 1,447-amino-acid isoform of PTCH1 results in derepression of Hedgehog signalling. Similarly, the U1 snRNA r.3A>G cassette exon in GLI2 is spliced between exon 4 and exon 5, which results in a putative GLI2 protein that lacks the repressor domain (Extended Data Fig. 10a–f). Physiological GLI2 protein has a repressor domain at its amino terminus, and constructs that lack the amino terminus are much more potent at activating Hedgehog signalling than the full-length protein45.

Alternative splicing of the cell-cycle gene CCND2, a known downstream target of SHH signalling that is recurrently amplified in SHH medulloblastoma, is detected in U1 snRNA r.3A>G mutants of SHHδ but not in SHHα46,47 (Extended Data Fig. 10g–l). Focal amplifications of CDK6 are highly recurrent in SHHα U1 snRNA r.3A>G mutants, but not in SHHδ with wild-type U1 snRNA or SHHδ U1 snRNA r.3A>G mutants, which suggests convergence on dysregulation of the G1–S cell-cycle checkpoint. The CCND2 alternative isoform is prematurely terminated, which results in N-terminal sequences in which the PEST domain is predicted to be deleted. Deletion of the PEST domain causes resistance to protein degradation and impaired export from the nucleus, which results in CCND2 accumulating in the nucleus to promote cell-cycle progression48. PAX5, another known tumour-suppressor gene, is affected by cryptic 5′ alternative splicing in U1 snRNA r.3A>G mutants (Extended Data Fig. 10m–q). SHH medulloblastomas with wild-type and mutant U1 snRNA express distinct cryptic isoforms. The cryptic isoform present in SHH medulloblastomas with wild-type U1 snRNA translates the complete DNA-binding domain of PAX5. However, the cryptic exon (also known as a poison exon49,50) that is present in SHH medulloblastomas with mutant U1 snRNA results in a stop codon before the DNA-binding domain. Mutations of PAX5 in cancer are typically concentrated in the DNA-binding site51. Together, the data relating to the alternative splicing of PTCH1, GLI2, CCND2 and PAX5 support a model in which cryptic alternative splicing mediated by mutant U1 snRNA r.3A>G functions as a driver in subsets of SHH medulloblastoma.

A U1 snRNA r.3A>G mutation is the most common single-nucleotide variant in medulloblastoma. The restriction of these mutations not
just to SHH medulloblastoma but to the SHHα and SHHδ subtypes suggests a model in which the specific cell of origin, the temporally specific microenvironment or co-occurring mutations (that is, of TP53) are necessary for U1 snRNA to contribute to oncogenesis. Although the almost universal occurrence of U1 snRNA mutations in SHHδ highly supports their role in tumour initiation, proof of the ongoing role of mutant U1 snRNA r.3A-G in tumour maintenance will await its knockdown in a tumour in which it was the initiating genetic event.

Patients with SHHδ with a U1 snRNA r.3A-G mutation are an extremely high-risk population who should be prioritized for the development of targeted therapies. Drugs that are under development directly target the spliceosome, which may show anti-tumour effects in cancers with spliceosomal mutations. Loss of expression of specific genes through cryptic splicing or intron retention could create opportunities for synthetic lethal approaches. Finally, cryptic splicing in SHH medulloblastoma with mutant U1 snRNA leads to a unique form of post-transcriptional hypermutation; this would be predicted to result in the expression of numerous cell-surface neo-epitopes that are never seen in healthy tissues, and which could be targeted using immunotherapies.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1650-0.
Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Subjects and materials
The study included two large cohorts of medulloblastomas, one from Toronto and the other from the ICGC (Extended Data Fig. 1). The Toronto cohort consisted of 293 cases (whole-genome sequencing, 114 cases; RNA sequencing (RNA-seq) 225 cases; 46 cases overlapped), collected at diagnosis after informed consent was obtained from subjects as part of the Medulloblastoma Advanced Genomics International Consortium. All patient recruitment and tumour-sample collection was approved and in compliance with the ethical regulations of each of the following institutions: The Hospital for Sick Children, Seoul National University Children’s Hospital, The Children’s Memorial Health Institute, Mayo Clinic, The Chinese University of Hong Kong, John Hopkins University School of Medicine, Seattle Children’s Hospital, University of California San Francisco, McMaster University, Erasmus University Medical Center, Kitasato University School of Medicine, Fondazione IRCCS Istituto Nazionale Tumori, Emory University, Osaka National Hospital, Washington University School of Medicine, University of Calgary, Children’s Hospital of Pittsburgh, Hospital Pediatria CentroMédico Nacional Century XXI, University of Debrecen, McGill University, Vanderbilt Medical Center, University of Colorado Denver, Istituto Giannina Gaslini and Université de Lyon. The whole-genome sequences consisted of 109 published and 5 unpublished (WNT, n = 2; SHH, n = 37; group 3, n = 26; and group 4, n = 49). Samples were obtained as freshly frozen tissue from the time of diagnosis and stored at −80 °C until processed for the purification of nucleic acids. Genomic DNA was isolated by incubation with proteinase K overnight at 55 °C followed by 3 sequential phenol extractions and ethanol precipitation. Messenger RNA library construction and sequencing were performed as previously described. The ICGC cohort consisted of 227 cases, which were downloaded from ICGC under accession DACO-1036229.

Whole-genome sequencing
Whole-genome sequencing was performed at Canada’s Michael Smith Genome Science Centre at the BC Cancer Agency, using the Illumina HiSeq 2000/2500 platform as previously described.

Sequence alignment of whole-genome sequencing data
Whole-genome sequencing reads were aligned to the human reference genome ‘hs37d5’ by 1000 Genomes Project Phase II, using Burrows–Wheeler aligner (BWA) – MEM version 0.7.8 with the ‘-T 0’ parameter. Duplicates were marked using biobambam v.0.0.148. Sequencing coverages were calculated using GenomonQC software, which was downloaded from Genomon Project, and are shown in Supplementary Table 1.

Somatic variant calling
Somatic variants were called using eight variant callers: MuTect2, EBCall, Varscan2, Strelka2, SomaticSniper, Virmid, Platypus and Seurat. MuTect2 was run using GATK v.3.5.0 with the default setting. Candidate variants were discarded if the variants were detected in a panel of normal that was made by MuTect2 with ‘-artefact_detection_mode’ and GATK ‘CombineVariants’ function with ‘-minN 2’. EBCall v.0.2.1 was run with the default setting. We used the following criteria: requiring P value (by EBCall) < 10−3, variant reads in tumour ≥ 2 and variant reads in normal ≤ 1. Varscan2 v.2.4.3 was run with parameters ‘-strand-filter 1 -min-var-freq 0.08’. The results were filtered using the ‘fpfilter’ function with the option ‘dream3-settings’. Strelka v.1.0.15 was run with default parameters. Virmid v.1.1.0 was run with the option ‘-q 10’. Somatic Sniper v.1.0.5.0 was run with the parameters ‘-Q 15 -q 1 -G L’ and the results were filtered by the author’s recommended filter using bam-readcount. The candidates with more than 0.03 of variant allele frequency in the matched normal sample were discarded. Platypus v.0.8.1 was run with a default setting. Detected variants that passed the standard Platypus filtering criteria or showed ‘allelic bias’ were used. We used the following additional criteria: requiring likelihood(reference allele)/likelihood(variant allele) < 10−1 in tumour, likelihood(variant allele)/likelihood(reference allele) < 10−3 in matched control, variant reads in tumour ≥ 2 and variant reads in normal ≤ 1. Seurat v.2.5 was run with the option ‘-indels’. We used variants that were called by at least two callers. Obtained results were filtered by ≤ 2 variant reads in the matched normal control, calculated using the realignment function of Genomon-MutationFilter v.0.2.1. Variants were annotated using ANNOVAR. The correlations of U1 and U11 snRNA mutations with other somatic events were analysed using R package ‘Epi’ v.2.30. Asymptotic P values from odds-ratio tests were calculated using the two-by-two function, followed by Benjamini–Hochberg adjustment for multiple testing.

Copy-number calling for whole-genome sequencing
Copy-number alterations were detected using Control-FREEC v.10.3 with the following parameters: breakPointType = 4, ploidy = “2,3,4”, step = 10000, window = 50000 (ref. 36).

Variant calling of U1 and U11 snRNA genes
To explore mutations in low-mappability regions, we first picked up reads from whole-genome sequencing data on U1 and U11 snRNA genes and pseudogenes using samtools and biobambam. To accept multiplexing, we used STAR aligner. To prevent gaps, we set the setting with ‘-scoreGap -20 -alignEnds Type EndToEnd’. Mutations were called by EBCall with the same setting as for whole-genome sequencing, except for the acceptance of secondary alignment. We used the following criteria: requiring P value (by EBCall) < 10−3, variant reads in tumour ≥ 4 and variant reads in matched control ≤ 1.

To evaluate the exact loci of variant reads and multiple mutations of U1 snRNAs, we mapped variant reads to case-specific references. First, we extracted all variant reads of U1 snRNA mutations (r.3A>G) with mate-paired reads. Then, we constructed a case-specific reference that included U1 snRNA hotspot mutations (r.3A>G) and case-specific germline variants detected from extracted variant reads, using the samtools mpileup function. Variant reads were mapped again on the case-specific reference, using bwa-mem with the same setting as was used in the whole-genome sequencing analysis. Using bam files with the case-specific reference, we called variants in flanking regions of the U1 snRNA hotspot mutation (r.3A>G), using samtools mpileup function to evaluate multiple mutations. No samples had recurrent variant reads. Therefore, we concluded that U1 snRNA mutation occurs in one allele. To interpret the mutated genes, we extracted consecutive consensus sequences of upstream U1 snRNA sequences with two or more supported reads. Then, the consensus sequence was mapped using BLAST software to U1 snRNA genes and pseudogenes with 1,000-bp upstream sequences from the hg19 reference. Because there were many variants and a high level of similarity in the upstream sequences, we cannot detect the exact positions of mutated reads except for RNU17-18 mutations. Therefore, we classified U1 snRNA mutations into (1) RNU1 genes (RNU1-1, RNU1-2, RNU1-3 or RNU1-4), (2) RNU1-18 and (3) RNU1 pseudogenes (RNU1-27P or RNU1-28P) on the basis of the similarity of sequences of the flanking region. Finally, we performed a manual review of detected mutations with the Integrative Genome Viewer. Detected mutations are shown in Supplementary Tables 2–4.

Secondary structure of U1 and U11 snRNAs
The conservation scores of U1 (RF000003) and U11 (RF00548) snRNAs were downloaded from Rfam. U1 and U11 sequences of other species were downloaded from seed sequences from Rfam. The secondary structures are described on the basis of the consensus structure in Rfam using
shown in Supplementary Table 20.

skipped exon. We filtered the events with FDR < 0.01 and change of alternative 3′ splice site, alternative 5′ splice site, retained intron and FAM (wild-type) dyes, as well as ROX dye as a passive reference. Plates were read on the StepOnePlus (Applied Biosystems) RT-PCR machine, and genotypes were using the StepOne v.2.3 software. The primer sequences are available in Supplementary Table 19.

RNA-seq

Sequencing reads were mapped by STAR v.2.5.1b on fasta, which includes the human reference genome hs37d5v1000 Genomes Project Phase II, spike-in sequences of profile C1_2 ERCC spike-in concentrations used for C1 fluidigm and Caltech profile 3 spike-ins by Encyclopedia of DNA elements (ENCODE), with the options `-outFilterMultimapNmax 20 -alignSJoverhangMin 8 -alignMateGapMax 2000000 -alignSJstitchMismatchNmax 5 -l 5 -s OUTSAMmultNmax 20 -twopassMode Basic`35. Mapping results are distributed stochastic neighbour embedding (t-SNE) analysis was performed using DESeq2 v.1.16.1 with the default setting, after extracting genes expressed at ~5 counts per million in at least 20% of cases. We performed two comparisons: SHHδ with mutant U1 snRNA (n = 30) versus other SHH subtypes with wild-type U1 snRNA (n = 90) and SHHα with mutant U1 snRNA (n = 13) versus SHHα with wild-type U1 snRNA (n = 39). Gene-set enrichment analysis (GSEA) for differentially expressed genes was performed using pre-ranked gene lists ordered by -log2(1-Pvalue) multiplied by +1 for upregulation or −1 for downregulation with gse v3.0. We used two datasets for a pathway of nonsense-mediated decay, 'GO nuclear transcribed mrna catabolic process nonsense mediated decay' from the C5 gene set and 'Reactome nonsense mediated decay enhanced by the exon junction complex' from the C2 gene set.

TP53 mutation status

Germline mutations of TP53 were analysed using EBCall v.0.2.1. EBCall was run with the default setting. We used the following criteria: requiring Pvalue (by EBCall) < 10−3, 90% posterior quantile calculated by EBCall > 0.3. The results were annotated using ANNOVAR.

The mutation call from RNA-seq was run using GATK v.3.8.0. Adding read groups and flagging duplicate reads were performed using Picard tool v.2.18.0. Then, we split reads into exon segments using GATK with the setting `--rf ReassignOneMappingQuality -RMQF 255 -RMQT 60 -U ALLOW_N_CIGAR_READS`. Base recalibration was performed using GATK. Mutation call was performed using the ‘HaplotypeCaller’ function of GATK with the setting `--dontUseSoftClippedBases -stand_call_conf 20.0`. Variants were filtered using the ‘VariantFilter’ function of GATK with the setting ‘--window 35 -cluster 3 -filterName FS -filter “FS > 30.0” -filterName QD -filter “QD < 2.0”’. The variants were discarded if they were also detected in any RNA-seq data generated from nine normal brain samples (five adult brains and four fetal brains). Sanger sequencing was as performed in a previous study44. We discarded the mutations that showed a frequency of 0.01 or higher in 1000 Genomes v5b or ESP-6500, or 0BSNP138.

Survival analysis

Overall survival and progression-free survival were evaluated using the log-rank with R package ‘survival’ v.2.40.1. Overall survival was defined as the time from date of surgery to death or date of last follow-up, and progression-free survival as the time from date of surgery to first event (progression or relapse) or date of last follow-up.

Pan-cancer analysis

We analysed 2,442 cases of cancer across 36 tumour types from ICGC. The hotspot mutations were analysed with the method described in ‘Variant calling of U1 and U11 snRNA genes’ , except for the use of the mapping tool. For pan-cancer data, we used bowtie aligner instead of STAR45.

SNP6 copy-number analysis

Array files were downloaded from the Gene Expression Omnibus (GEO) under accession number GSE37385, and the relevant Affymetrix SNP6 arrays were extracted. Affymetrix Power Tools v.1.18.2 was used to process and normalize the probe intensities to generate log R ratio (LRR) and B-allele frequency (BAF) using the PennCNV-Affy pipeline46. The affy6g.6h.19.pib file was used to map the probes onto the hg19 genome. All other parameters were left as default.

The resulting probe-level LRR and BAF values were taken into ASCAT v.2.4.3. GC wave correction was then performed, followed by predicting germline genotypes; this finally led to running the ASCAT algorithm to determine the copy-number values for each genomic region, as well as
the overall ploidy and purity of the sample. Samples with a model fit that was less than 80% failed their ASCAT processing stage. Log ratios for each segment were calculated by using the copy number of each segment, as well as the average ploidy of the sample, according to the equation:

$$\text{ratio} = \log_{2} \left( \frac{\text{copy number}}{\text{ploidy}} \right)$$

Adjacent segments with log ratios that differed by less than 0.25 were then merged using their size-weighted mean:

$$\text{new ratio} = \frac{\text{length1} \times \text{ratio1} + \text{length2} \times \text{ratio2}}{\text{length1} \times \text{ratio1}}$$

Copy-number states were assigned to each segment on the basis of their log ratio and their ploidy values, according to the Supplementary Table 21. Broad copy-number changes are defined as occurring in 75% or more of chromosome arm in size. Focal copy-number variants were analysed using GISTIC v.2.0.2348. GISTIC was run with the settings ‘ta 0.25 -t 0.3 -js 10 -brl 0.7 -gcm ‘extreme’-armpeel’.

**PCR with reverse transcription and qPCR analysis**

RNA was obtained for samples from 18 patients that had fragments per kilobase of transcript per million mapped reads (FPKM) values of more than 2 for targeted genes from our larger cohort (6 SHHx with wild-type U1 snRNA, 6 SHHx with mutant U1 snRNA and 6 SHHx with mutant U1 snRNA). cdNA was synthesized using SuperScript III (Thermo Fisher 18080400). PCRs were performed with cdNA and Taq polymerase using 35 cycles, and products run on a 2% agarose gel. qPCRs were performed using SYBR-Green with ROX (Thermo Fisher Scientific). The primer sequences are shown in Table 2. Broad copy-number changes are defined as occurring in 75% or more of chromosome arm in size. Focal copy-number variants were analysed using GISTIC v.2.0.2348. GISTIC was run with the settings ‘ta 0.25 -t 0.3 -js 10 -brl 0.7 -gcm ‘extreme’-armpeel’.

**Exogenous expression of U1 snRNA r.3A>G mutation**

The pLKO.1-puro U6 sgRNA BfuAI stuffer lentiviral vector (Addgene no. 50920) was modified by removing the internal U6 promoter (between Ndel and EcoRI), and it was replaced by the U1 snRNA locus, including 393 bases of internal native U1 promoter, the U1 sequence, and 39 bases of 3′-flanking region using the following oligonucleotides (5′-GTCGAGAATCTTGCCGACTGTTGTTTTT and 5′-CTATCATATGAAGGCCAAGCTTCTTTGGAGA). The PCR products were digested with NdeI and EcoRI, and cloned in the modified pLKO.1 plasmid. The r.3A>G mutation was introduced by site-directed mutagenesis. All plasmids were verified by Sanger sequencing.

**Verification of the expression of U1 snRNA r.3A>G mutation**

Rapid amplification of cdNA ends (RACE) was performed using 1 μg of total RNA from HEK-293T cells transfected with either pLKO.1-U1wt or pLKO.1-U1r.3A>G following the recommendations of the manufacturer (Sigma-Aldrich 353621001), and the following specific oligonucleotides (U1-RACE-SP1: 5′-CAGGGGAAGGCAGCAGT and U1-RACE-SP2: 5′-CCCATCACACAAATATAGC). A single amplification band of the expected size (160 bp) was excised from the gel, purified and sequenced with the internal oligonucleotide U1-RACE-SP2.

**Sequence analyses of exogenous expression analyses**

Messenger RNA library construction was performed based on oligo dt-based mRNA isolation using NEBNext Poly(A) mRNA Magnetic Isolation Module. RNA sequencing was performed on NextSeq 550 using 100-bp paired-end mode. Mapping and intron clustering were performed with the methods described in ‘RNA-seq’ and ‘Intron-centric alternative splicing analysis’. LeafCutter was run with the option ‘-g 0 -i 2’ and the obtained results were filtered by a q value of each cluster < 0.1, in which at least 1 absolute effect size calculated by LeafCutter was more than 1.5.

**Reporting summary**

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

**Data availability**

Sequencing data have been deposited in the European Genome-Phenome Archive (EGA) and Gene Expression Omnibus (GEO): RNA-seq (EGAD00001001899 and EGAD00001004958), whole-genome sequences (EGAD00001003125 and EGAD00001004347) and RNA-seq of exogenous expression analyses (GSE128005). Materials used in this study are available from the following: GENCODE (ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_19/gencode.v19.annotation.gtf.gz), ICGC (https://igcg.org/), hs37d5 reference (ftp://ftp1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2a-reference-assembly_sequence), Burrows–Wheeler aligner (bwa) (http://bio-bwa.sourceforge.net/), Mutect2 (https://software.broadinstitute.org/gatk), EBCall (https://github.com/friendlsw/EBCall), VarScan 2 (http://dkoldbdt.gitlab.io/varsan2/), Strelka (https://github.com/Illumina/strelka), Somatic Sniper (http://gmt.genome.wustl.edu/packages/somatic-sniper/), Virmid (https://sourceforge.net/p/virmid/wiki/Home/), Platypus (http://www.well.ox.ac.uk/platypus), Seurat (https://sites.google.com/site/seuratomatic/), ENCODE (https://wwwencodeproject.org/), PennCNV (http://penncnv.openbioinformatics.org/en/latest/), Database of Genomic Variants (http://dgv.tcag.ca/dgv/app/home), Genomon Project (https://github.com/Genomon-Project), SpliceRack (http://katahdin.mssm.edu/splice/index.cgi?database=spliceNew) and GEO (https://www.ncbi.nlm.nih.gov/geo/).

---

24. Morrisey, A. S. et al. Divergent clonal selection dominates medulloblastoma at recurrence. Nature 529, 353–357 (2016).
25. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. 31, 213–219 (2013).
26. Shirahashi, Y. et al. An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. Nucleic Acids Res. 41, e89 (2013).
27. Koboldt, D. C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 22, 568–576 (2012).
28. Saunders, C. T. et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics 28, 1811–1817 (2012).
29. Larson, D. E. et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics 28, 311–317 (2012).
30. Kim, S. et al. Virmid: accurate detection of somatic mutations with sample impurity inference. Genome Biol. 14, R90 (2013).
31. Rimmer, A. et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat. Genet. 46, 912–918 (2014).
32. Christoforides, A. et al. Identification of somatic mutations in cancer through Bayesian-based analysis of sequenced genome pairs. BMC Genomics 14, 302 (2013).
33. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38, e164 (2010).
34. Boeva, V. et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics 29, 213–219 (2013).
35. Robinson, J. T. et al. Integrative genomics viewer. Nat. Biotechnol. 29, 24–26 (2011).
36. Kallwass, L. et al. Raman 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res. 46, D335–D342 (2018).
37. Darty, K., Denise, A. & Ponty, Y. VARNA: interactive drawing and editing of the RNA secondary structure. Bioinformatics 25, 1974–1975 (2009).
38. Wang, K. et al. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38, e164 (2010).
39. Boeva, V. et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics 29, 213–219 (2013).
40. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013).
41. Robinson, J. T. et al. Integrative genomics viewer. Nat. Biotechnol. 29, 24–26 (2011).
42. Kallwass, L. et al. Raman 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res. 46, D335–D342 (2018).
43. Darty, K., Denise, A. & Ponty, Y. VARNA: interactive drawing and editing of the RNA secondary structure. Bioinformatics 25, 1974–1975 (2009).
Acknowledgements M.D.T. is supported by the NIH (RO1CA148699 and RO1CA159859), The Pediatric Brain Tumour Foundation, The Terry Fox Research Institute, The Canadian Institutes of Health Research, The Cure Search Foundation, b.r.a.i.n.child, Meagan’s Walk, Genome Canada, Genome BC, Genome Quebec, the Ontario Research Fund, Worldwide Cancer Research, V-Foundation for Cancer Research and the Ontario Institute for Cancer Research, through funding provided by the Government of Ontario. M.D.T. is also supported by a Canadian Cancer Society Research Institute Impact grant and by a Stand Up To Cancer (SU2C) St. Baldrick’s Pediatric Dream Team Translational Research Grant (SU2C- AACR-071113) and SU2C Canada Cancer Stem Cell Dream Team Research Funding (SU2C-AACR-0719-13) provided by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research, with supplementary support from the Ontario Institute for Cancer Research through funding provided by the Government of Ontario. SU2C is a programme of the Entertainment Industry Foundation administered by the American Association for Cancer Research. M.D.T. is also supported by the Garron Family Chair in Childhood Cancer Research at the Hospital for Sick Children and the University of Toronto. E.G.V.M. is supported by the NIH (R01-NS096236 and RO1CA235162) and the CURE Childhood Cancer Foundation. X.S.P. was supported by 2017-1.2.1-NKP-2017-00002 National Brain Research Program NAP 2.0. M.L.G. is supported by AIRC (Italian Association for Cancer Research) and by Fondazione Berlucchi. H.S. is a recipient of a Research Fellowship (Astellas Foundation for Research on Metabolic Disorders). S.A.K. is a recipient of funding from the Restracomp Research Fellowship (SickKids Research Institute) and the MD/PHD Studentship Award (Canadian Institute of Health Research). A.D.N. is a recipient of the Department of Education of the Basque Government (PRE_2017_1_0100). J.R. is supported by Genome Canada Genome Technology Platform Grant 12505, Canada Foundation for Innovation Project 33408. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics and on the Niagara supercomputer at the SciNet HPC Consortium. SciNet is funded by the Canada Foundation for Innovation under the auspices of Compute Canada; the Government of Ontario; Ontario Research Fund - Research Excellence; and the University of Toronto. The authors would also like to thank M. Kamal and J. Loukides, and recognize the Labatt Brain Tumour Research Centre.

Author contributions M.D.T. led the study. H.S., S.S. and S.D.B. performed whole-genome sequencing analysis (Figs. 1, 2, Extended Data Figs. 1–3). F.M.G.C., N.G., J.R. and A.S.M. contributed to the pre-processing of RNA-seq data. H.S. and H.F. contributed to SNP6 copy-number analyses (Extended Data Figs. 1, 4c, d). H.S., S.S., M.F.G.C., I.S. and S.S. contributed to RNA expression analyses (Extended Data Fig. 7a). H.S., S.S., I.S., A.F., S.D.B. and O.A. contributed to alternative splicing analyses (Fig. 4a–d, Extended Data Figs. 5–10). H.S. and V.R. performed clinical analysis (Fig. 3, Extended Data Fig. 4e–i). A.G. and M.A.M. helped with bioinformatics analyses and provided expert advice. S.A.K., P.D.A., K.J. and M.C.V. performed real-time PCR and qPCR analyses (Figs. 4e, f, Extended Data Figs. 9d, 10e, 10f). S.A.K., A.D.N., A.G.-F., P.D.A., K.J., I.S., N.A., D.P., A.M., J.W., W.D., R.I.W.-R. and X.S.P. contributed to exogenous expression experiments (Extended Data Fig. 7b–d). S.A.K., K.J. and S.S. performed rhAMP SNP experiments (Figs. 1, 3a–c, Extended Data Figs. 1, 4a, b). P.S. and B. Luu contributed to the collection and processing of human tissue samples. C.D., X.W., R.J.W.-R., L.G., X.S.P., J. A. Chan and L.S. provided expert advice for experiments. S.K.K., W.A.G., A.J., M.F.M., M.L.G., A.A.N.R., C.G., J.M.K., P.J.F., N.J., H.-K.N., W.S.P., C.G.E., I.F.P., J.M.O., W.A.W., T.K., E.L.-A., B. Lach, M.M., E.G.V.M., J.B.R., R.V., L.B.C., N.K., A.K., L.B., J. A. Calarco, C.C.F., S.M.P., I.G. and D.M. provided patient material and helped design the study. H.S., S.A.K., S.S., J.A. Calarco, L.S. and M.D.T. prepared the manuscript and figures.

Competing interests The authors declare no competing interests.

Additional information Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1650-0.

Correspondence and requests for materials should be addressed to M.D.T.

Peer review information Nature thanks Rotem Karni, Brandon Wainwright and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.
Extended Data Fig. 1 | Overview of analysed cohorts and methods. a, The detection methods for U1 snRNA mutations by each cohort, and comparison methods for alternative splicing analysis. b, Cohort specification. c, Subgroup distribution of whole-genome sequencing (WGS) cohorts.
Extended Data Fig. 2 | See next page for caption.
Extended Data Fig. 2 | U11 snRNA mutations, and conservation of U1 and U11 snRNA genes across evolution.

a, Seed sequences of the U1 snRNA obtained from the Rfam database demonstrate high-level conservation across a variety of eukaryotic species, particularly at the site of the SHH medulloblastoma mutation. The consensus sequence and first 50 nucleotides of reference sequences are included for comparison. Grey indicates nucleotide differences, and red identifies the SHH medulloblastoma hotspot mutation.

b, Cartoon illustrating the number of somatic mutations in the U11 snRNA genes. Sequence conservation scores for U11 snRNA, as determined using the Rfam database.

c, Secondary structure of the mutant U11 snRNA. The red circle identifies the location of the hotspot mutation. The yellow and green rectangles indicate the 5’ splice-site recognition site and the Sm protein-binding site, respectively. Numerals I to IV indicate stem loops.

d, Seed sequences of the U11 snRNA obtained from the Rfam database demonstrate high-level conservation across a variety of eukaryotic species, particularly at the site of the SHH medulloblastoma mutation. The consensus sequence and first 30 nucleotides of reference sequences are included for comparison. Grey indicates nucleotide differences, and red identifies the SHH medulloblastoma hotspot mutation.
Extended Data Fig. 3 | See next page for caption.
Extended Data Fig. 3 | High levels of genomic conservation surrounding human U1 snRNAs complicate the specific PCR amplification of any individual locus. a, Genomic locations of the 4 expressed U1 snRNA genes (on chromosome 1p, red) and 136 pseudogenes across the Homo sapiens genome, as indicated. Three pseudogenes with sequences that are identical (hg19) to the expressed U1 snRNA genes are indicated in orange. b, Average mapping quality of bwa-mem and coverage of each of the expressed U1 and U11 snRNA genes from whole-genome sequencing of germline samples from patients with medulloblastoma are illustrated (n = 341 patients). Blue bars represent the alignability of 100-mers using the GEnome Multitool (GEM) mapper from ENCODE and Centre for Genomic Regulation (CRG). Regions that are >1,000 bases upstream and downstream are shown on a log10 scale. Red bar indicates the gene body. c, Average number of multimapped reads overlapped for each gene pair using STAR aligner. The heat map shows the average number of mapped reads across the whole-genome sequencing of germline samples from patients with medulloblastoma (n = 341 patients). d, Sequence similarity of U1 snRNA genes, U1 snRNA pseudogenes with 164 identical base pairs and the U11 snRNA gene. The numbers in each square and heat map indicate identity scores and bit scores calculated using blast software. A blank square indicates that no hit was found.
Extended Data Fig. 4  See next page for caption.
Extended Data Fig. 4 | Allele-specific rhAmp SNP PCR of RNU1 loci, copy-number changes in SHH medulloblastoma with mutant U1 snRNA versus wild-type U1 snRNA, and prognostic analysis. a, The frequency of any U1 snRNA mutation in the RNU1_batch primer set (RNU1-1, RNU1-2, RNU1-3, RNU1-4 and RNU1-18) (left) and the RNU1_pseudo primer set (RNU1-27P and RNU1-28P) (right). b, Hotspot mutations of RNU1-27P or RNU1-28P U1 snRNA pseudogenes, as confirmed by Sanger sequencing. c, Broad copy-number aberrations in SHHα with wild-type U1 snRNA (n = 25), SHHα with mutant U1 snRNA (n = 8) and SHHβ with mutant U1 snRNA (n = 41). Dark blue and dark red bars, as well as asterisks, identify statistically significant regions, comparing SHHα with mutant versus wild-type U1 snRNA (P < 0.05, two-sided Fisher’s exact test). d, Significant focal copy-number aberrations in SHHα with wild-type U1 snRNA (n = 25), SHHα with mutant U1 snRNA (n = 8) and SHHβ with mutant U1 snRNA (n = 41) illustrate significant genomic differences between cases with wild-type and mutant U1 snRNA. Candidate target genes within the corresponding loci are indicated. q values were calculated using GISTIC (Methods). e–g, Overall survival of patients with SHHα, stratified by mutational status of U1 snRNA mutation (n = 10 for mutant, n = 27 for wild type) (e), TP53 (n = 15 for mutant, n = 22 for wild type) (f) or both (n = 9 for both mutant, n = 1 for U1 snRNA mutation only, n = 6 for TP53 mutation only, n = 21 for both wild type) (g). P values were determined using the two-sided log-rank test. h, i, Progression-free survival (h) and overall survival (i) stratified by U1 snRNA mutation and SHH subtype (n = 10 for SHHα with mutant U1 snRNA, n = 27 for SHHα with wild-type U1 snRNA, n = 23 for SHHβ with wild-type U1 snRNA, n = 24 for SHHγ with wild-type U1 snRNA, n = 46 for SHHδ with mutant U1 snRNA). P values were determined using the two-sided log-rank test. +, censored case.
Extended Data Fig. 5 | Intron-centric analysis of SHHδ medulloblastomas.

**a**, Quantification of alternative splicing events by SHH subtype, as detected by intron-centric alternative splicing analysis (n = 30 of each subtype). Bar plot shows adjusted standardized residual of included alternative splicing events. Positive values indicate a relatively higher number, and negative values indicate a relatively lower number among subtypes.

**b**, Volcano plots of alternative splicing events (n = 30 of each subtype). Significant events (FDR < 0.01 and absolute log effect size > 1.5 calculated using LeafCutter (Methods)) are illustrated by colour. Alternative splicing events of PTCH1 and GLI2 with the highest effect size are annotated.

**c**, Splice-site sequences of included alternative splicing events by subtype (n = 30 of each subtype). Asterisk denotes nucleotide sites with q-value < 10⁻² (χ² test and Benjamini–Hochberg method).
Extended Data Fig. 6 | See next page for caption.
Extended Data Fig. 6 | Intron-centric analysis of SHHα medulloblastomas.

a, b, Quantification (a) and proportion (b) of alternative splicing events between SHHα medulloblastoma with mutant U1 snRNA \( (n = 13) \), and SHHα medulloblastoma with wild-type U1 snRNA \( (n = 39) \) as detected by intron-centric alternative splicing analysis. \( P \) value was calculated by \( \chi^2 \) test.

c, Volcano plots of alternative splicing events \( (n = 13 \text{ for SHHα with mutant U1 snRNA, } n = 39 \text{ for SHHα with wild-type U1 snRNA}) \). The \( x \) axis shows the difference of per cent spliced in calculated using LeafCutter. Significant events (FDR < 0.01 and absolute log effect size > 1.5, calculated by LeafCutter (Methods)) are illustrated by colour.

d, Splice-site sequences of included alternative splicing events in SHHα with mutant U1 snRNA \( (n = 13 \text{ for SHHα with mutant U1 snRNA, } n = 39 \text{ for SHHα with wild-type U1 snRNA}) \). Size and colour for each circle indicate the \( q \) values and Cramer’s \( V \) values for each nucleotide position. \( q \) values were calculated by \( \chi^2 \) test and Benjamini–Hochberg method; the precise values are given in Supplementary Table 11.

e, Residual analysis of 5′ splice-site sequences of annotated and cryptic 5′ alternative splicing \( (n = 13 \text{ for SHHα with mutant U1 snRNA, } n = 39 \text{ for SHHα with wild-type U1 snRNA}) \). The size and colour of each circle denote the two-sided \( P \) value, and adjusted standardized residual calculated by Haberman’s method. The precise values are given in Supplementary Table 11.
Extended Data Fig. 7 | Nonsense-mediated decay pathway in SHH medulloblastoma with mutant U1 snRNA, and exogenous expression analyses. a, Enrichment plots of GO nuclear transcribed mRNA catabolic process nonsense mediated decay by GSEA between SHHδ with mutant U1 snRNA (n = 30) and other subtypes of SHH medulloblastoma with wild-type U1 snRNA (n = 90), and SHHα with mutant U1 snRNA (n = 13) and SHHα with wild-type U1 snRNA (n = 39). P values were calculated using gsea v.3.0 (Methods). b, Quantification of alternative splicing events between HEK-293T cells with mutant U1 snRNA and HEK-293T cells with wild-type U1 snRNA, as detected by intron-centric alternative splicing analysis. c, Splice-site sequences of included alternative splicing events in HEK-293T cells with mutant U1 snRNA. Asterisk denotes nucleotide sites with q value < 10⁻² (χ² and Benjamini–Hochberg method). d, Comparison of the extent of overlap between detected alternative splicing events by SHH medulloblastoma (either of SHHα or of SHHδ) with mutant U1 snRNA, SHH medulloblastoma (either of SHHα, SHHβ or SHHγ) with wild-type U1 snRNA and HEK-293T cells with mutant U1 snRNA exogenous expression. Left, alternatively spliced events with cryptic 5′ sites. Right, alternatively spliced events with cryptic 3′ sites and C base at the 6th intron. e, Alternative splicing signatures by r-SNE analysis. Left, the per cent spliced-in values of detected cryptic 5′ alternative splicing events, with a C nucleotide at the 6th base in the intron from the 5′ splice site. Top right, per cent spliced-in values of all cryptic 5′ alternative splicing events. Bottom right, per cent spliced-in values of all alternative splicing events.
Extended Data Fig. 8 | See next page for caption.
Extended Data Fig. 8 | Retained introns inactivate tumour-suppressor genes in tumours with U1 snRNA r.3A>G mutation. a, Illustration of the different types of alternative splicing events analysed using rMATS (n = 30 of each subtype). Red arrows indicate expected 5′ prime sites recognized by the mutant U1 snRNA. b, Quantification of alternative splicing events by subtype of SHH medulloblastoma, as detected by exon-centric alternative splicing analysis. c, Scatter plots of alternative splicing events (n = 30 of each subtype). The x axis shows the difference of per cent spliced in (psi) calculated using rMATS. Different types of significant event (FDR < 0.01 and absolute differential psi > 0.05 calculated using rMATS (Methods)) are illustrated by different colours, as annotated. d, Splice-site sequences of alternative 5′ splice site (ASSS), included cassette exon (CE) and included retained intron (RI) events in SHHδ with mutant U1 snRNA (n = 30). Each event corresponds to a red arrow cartoon in a. Asterisk denotes nucleotide sites with q value < 10^{-2} (χ^2 test and Benjamini–Hochberg method). e, Distribution of per cent spliced in for PAX6 based on U1 snRNA mutation status (n = 13 for SHHα with mutant U1 snRNA, n = 30 for SHHδ with mutant U1 snRNA, n = 99 for SHH medulloblastoma with wild-type U1 snRNA (n = 90) and normal brain tissue (n = 9)). Dashed line defines threshold that divides the dataset into two groups (k-means method). The table displays the number of samples above the threshold (high) or below (low) based on mutational status. P value was calculated using two-sided Fisher’s exact test compared to samples with wild-type U1 snRNA. Samples with mutant U1 snRNA are indicated in pink, and wild-type samples in blue. f, Sashimi plot of splicing of PAX6 based on mutational status determined by exon-centric alternative splicing analysis (rMATS). The bar plot shows modified FPKM. Numbers refer to average junctional reads across all samples. Annotated exon tracks are shown below, with genomic positions marked. g, Distribution of per cent spliced in for TOX4 based on U1 snRNA mutation status (n = 13 for SHHα with mutant U1 snRNA, n = 30 for SHHδ with mutant U1 snRNA, n = 99 for SHH medulloblastoma with wild-type U1 snRNA (n = 90) and normal brain tissue (n = 9)). Dashed line defines threshold that divides the dataset into two groups (k-means method). The table displays the number of samples above the threshold (high) or below (low) based on mutational status. P value was calculated using two-sided Fisher’s exact test compared to samples with wild-type U1 snRNA. Samples with mutant U1 snRNA are indicated in pink, and wild-type samples in blue. h, Sashimi plot of splicing of TOX4 based on mutational status determined by exon-centric alternative splicing analysis (rMATS). The bar plot shows modified FPKM. Numbers refer to average junctional reads across all samples. Annotated exon tracks are shown below, with genomic positions marked.
Extended Data Fig. 9 | See next page for caption.
Extended Data Fig. 9 | Exon-centric analysis of SHHα medulloblastomas and overlapped splicing events. a, Quantification of alternative splicing events between SHHα medulloblastoma with mutant U1 snRNA and SHHα medulloblastoma with wild-type U1 snRNA, as detected by exon-centric alternative splicing analysis. b, Scatter plots of alternative splicing events (n = 13 for SHHα with mutant U1 snRNA, n = 39 for SHHα with wild-type U1 snRNA). The x axis shows the difference of per cent spliced in calculated using rMATS. Different types of significant event (FDR < 0.01 and absolute differential psi > 0.05 calculated using rMATS (Methods)) are illustrated by different colours, as annotated. c, Splice-site sequences of alternative 5′ splice sites, included cassette exon and included retained intron events in SHHα medulloblastoma with mutant U1 snRNA and SHHα medulloblastoma with wild-type U1 snRNA. Each event corresponds to a red arrow cartoon in Extended Data Fig. 8a. Asterisk denotes nucleotide sites with q value < 10^{-2} (χ^2 test and Benjamini–Hochberg method). d, Box plot of fold changes in expression of the alternatively spliced isoform as compared to the wild-type isoform in subsets of SHH medulloblastoma, as determined by real-time qPCR. In the box plots, the centre lines show data median; box limits indicate the IQR from the 25th and 75th percentiles; lower and upper whiskers extend 1.5× the IQR. Outliers are represented by individual points. P values were calculated using two-sided Wilcoxon rank-sum test. e, Comparison of the extent of overlap between splicing events by subtype of SHH medulloblastoma and U1 snRNA mutational status. Effect sizes are calculated by LeafCutter with an absolute effect-size threshold of 1.5.
Extended Data Fig. 10 | See next page for caption.
Extended Data Fig. 10 | Aberrant splicing of oncogenes and tumour-suppressor genes in SHH medulloblastoma with mutant U1 snRNA.

a, Overview of cryptic alternative splicing of GLI2, demonstrating the position of a cryptic cassette exon with the 5′ splice-site sequence. b, Sashimi plot of splicing of GLI2 in representative cases. The bar plot shows counts per million reads. Numbers are of junctional reads; reads for the mutant U1 snRNA isoform are in red. c, Scatter plot comparing detected alternatively spliced read and total junction reads that share a 3′ splice site. Jittering was performed for both values. d, Per cent spliced in values for SHHα with mutant U1 snRNA, SHHδ with mutant U1 snRNA and SHH medulloblastoma with wild-type U1 snRNA (all subtypes of SHH medulloblastoma). e, Box plot of fold changes in expression of the alternatively spliced isoform as compared to the wild-type isoform of GLI2 in subsets of SHH medulloblastoma, determined by real-time qPCR. f, Illustration of canonical and cryptic isoforms of GLI2. Translation start sites are indicated by an ATG arrow. Resulting proteins (and sizes) are displayed for each isoform. Repression and activation domains are indicated in blue and orange respectively. g, Overview of cryptic alternative splicing of CCND2, illustrating the position of a cryptic cassette exon with the 5′ splice-site sequence. h, Sashimi plot of representative cases demonstrates alternative splicing at the CCND2 locus. Numbers illustrate junctional reads. Junctional reads specific to U1 snRNA mutants are in red. i, The canonical isoform and the cryptic isoform of CCND2. j, Scatter plot comparing detected alternatively spliced read and total junction reads that share a 3′ splice site. Jittering was performed for both values. k, Per cent spliced in values for U1-mutant SHHα (n = 13), U1-mutant SHHδ (n = 58), and U1-wildtype SHH (all SHH subtypes, n = 104). l) Real-time qPCR comparing the expression of the cryptic isoform of CCND2 demonstrates high levels of expression of CCND2 restricted to SHHδ cases (n = 8 for SHHα with mutant U1 snRNA, n = 6 for SHHδ with mutant U1 snRNA, n = 6 for SHHα with wild-type U1 snRNA). m, Overview of cryptic alternative splicing of PAX5, illustrating the position of a cryptic cassette exon with the 5′ splice-site sequence. n, Sashimi plot of representative cases demonstrates alternative splicing at the PAX5 locus. Numbers denote junctional reads. Junctional reads specific to U1 snRNA mutants are in red. o, The canonical isoform and the cryptic isoform of PAX5. p, Scatter plot comparing detected alternatively spliced read and total junction reads that shared a 3′ splice site. Jittering was performed for both values. q, Per cent spliced in values for SHHα with mutant U1 snRNA (n = 5), SHHδ with mutant U1 snRNA (n = 27) and SHH medulloblastoma with wild-type U1 snRNA (all subtypes of SHH medulloblastoma, n = 7). In all box plots, centre lines show data median; box limits indicate the IQR from the 25th and 75th percentiles; lower and upper whiskers extend 1.5× the IQR. Outliers are represented by individual points. P values were calculated using two-sided Wilcoxon rank-sum tests.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- [ ] The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- [ ] A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- [ ] The statistical test(s) used AND whether they are one- or two-sided
- [ ] A description of all covariates tested
- [ ] A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- [ ] A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- [ ] For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted Give P values as exact values whenever suitable.
- [ ] For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- [ ] For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- [ ] Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

No software was used to collect data.

Data analysis

Burrows-Wheeler Aligner (v0.7.8), biobambam (v0.0.148), GATK (v3.5.0 and v.3.8.0), EBScan2 (v0.2.1), Varscan2 (v2.4.3), Strelka (v1.0.15), SomaticSniper (v1.0.5.0), Virmid (v1.1.0), Platypus (v0.8.1), Seurat (v2.5), Control-FREEC (v10.3), STAR (v2.5.1b), LeafCutter (v.0.2.7), rMATS (v4.0.1), MISO (v0.5.4), htseq (v0.6.0), gsea (v3.0), Picard tool (v2.18.0), bowtie (v2.3.4.1), Affymetrix Power Tools (v1.18.2), PennCNV (v1.0.3), ASCAT (v2.4.3), GISTIC (v2.0.23), R (v.3.6.0), R packages - Ttne (v0.13), ggseqlogo (v0.157), survival (v2.40.1), Epi(v.2.30).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequencing data have been deposited in the European Genome-Phenome Archive (EGA) and Gene Expression Omnibus (GEO) : RNA-seq (EGA00001001899, and EGA00001004958), whole genome sequence (EGA00001003125 and EGA00001004347) and RNA-seq of exogenous expression analyses (GSE128005).
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- [ ] Life sciences
- [ ] Behavioural & social sciences
- [ ] Ecological, evolutionary & environmental sciences

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

Life sciences study design

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

| Sample size | Sample size was determined by the availability of the human samples. We used all sequencing data we have. |
|-------------|---------------------------------------------------------------------------------------------------|
| Data exclusions | In ICGC medulloblastoma data, we excluded three samples. The reasons are a truncated file, obvious tumor contamination in the matched control sample, and single-end sequencing data. The exclusion criteria was not pre-established. These three files cannot be analyzed with our mutation call pipeline described in Methods section. |
| Replication | rhAmp SNP genotyping were done three times. All attempts at replication were successful. All qPCR was done in technical triplicate aside from biological replicates. |
| Randomization | For alternative splicing analysis, we choose samples randomly using sample function in R software. |
| Blinding | Blinding was not relevant for our study since this is an exploratory study and blinding is impossible or unlikely to affect the results or interpretation of the results. |

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

| Materials & experimental systems | Methods |
|---------------------------------|---------|
| n/a                             | n/a     |
| [x] Antibodies                 | [ ] ChIP-seq |
| [ ] Eukaryotic cell lines      | [ ] Flow cytometry |
| [ ] Palaeontology              | [ ] MRI-based neuroimaging |
| [ ] Animals and other organisms|         |
| [ ] Human research participants|         |
| [ ] Clinical data              |         |

Eukaryotic cell lines

Policy information about cell lines

- Cell line source(s): HEK293T was obtained from American Type Culture Collection (ATCC)
- Authentication: HEK293T cell was not authenticated.
- Mycoplasma contamination: Mycoplasma negative
- Commonly misidentified lines (See ICLAC register): No commonly misidentified cell lines were used in this study.

Human research participants

Policy information about studies involving human research participants

- Population characteristics: Samples were collected at diagnosis after informed consent was obtained from subjects as part of the Medulloblastoma Advanced Genomics International Consortium. Research participants were patients with confirmed diagnosis of medulloblastoma at local centers. All cases used in this study were primary cases.
- Recruitment: Patients were recruited by investigators from each local center. There is no bias of recruitment because patients are not prescreened. Potential self-selection bias or other biases were not identified.
Ethics oversight  The Hospital for Sick Children

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