OPEN
doi:10.1038/nature25795

# Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours

Xiaotu Ma[1]*, Yu Liu[1]*, Yanling Liu[1], Ludmil B. Alexandrov[2], Michael N. Edmonson[1], Charles Gawad[1], Xin Zhou[1], Yongjin Li[1], Michael C. Rusch[1], John Easton[1], Robert Huether[3]†, Veronica Gonzalez-Pena[4], Mark R. Wilkinson[1], Leandro C. Hermida[5], Sean Davis[6], Edgar Sioson[1], Stanley Pounds[7], Xueyuan Cao[7], Rhonda E. Ries[8], Zhaoming Wang[1], Xiang Chen[1], Li Dong[1], Sharon J. Diskin[9], Malcolm A. Smith[10], Jaime M. Guidry Auvil[5], Paul S. Meltzer[6], Ching C. Lau[11,12], Elizabeth J. Perlman[13], John M. Maris[9], Soheil Meshinchi[8], Stephen P. Hunger[9], Daniela S. Gerhard[5] & Jinghui Zhang[1]

**Analysis of molecular aberrations across multiple cancer types, known as pan-cancer analysis, identifies commonalities and differences in key biological processes that are dysregulated in cancer cells from diverse lineages. Pan-cancer analyses have been performed for adult[1–4] but not paediatric cancers, which commonly occur in developing mesodermic rather than adult epithelial tissues[5]. Here we present a pan-cancer study of somatic alterations, including single nucleotide variants, small insertions or deletions, structural variations, copy number alterations, gene fusions and internal tandem duplications in 1,699 paediatric leukaemias and solid tumours across six histotypes, with whole-genome, whole-exome and transcriptome sequencing data processed under a uniform analytical framework. We report 142 driver genes in paediatric cancers, of which only 45% match those found in adult pan-cancer studies; copy number alterations and structural variants constituted the majority (62%) of events. Eleven genome-wide mutational signatures were identified, including one attributed to ultraviolet-light exposure in eight aneuploid leukaemias. Transcription of the mutant allele was detectable for 34% of protein-coding mutations, and 20% exhibited allele-specific expression. These data provide a comprehensive genomic architecture for paediatric cancers and emphasize the need for paediatric cancer-specific development of precision therapies.**

Paired tumour and normal samples from 1,699 patients with paediatric cancers enrolled in Children's Oncology Group clinical trials were analysed, including 689 B-lineage acute lymphoblastic leukaemias (B-ALL), 267 T-lineage ALLs (T-ALL), 210 acute myeloid leukaemias (AML), 316 neuroblastomas (NBL), 128 Wilms tumours and 89 osteosarcomas (Extended Data Fig. 1a–c). All tumour specimens were obtained at initial diagnosis, and 98.5% of patients were 20 years of age or younger (see Methods, Extended Data Fig. 1d).

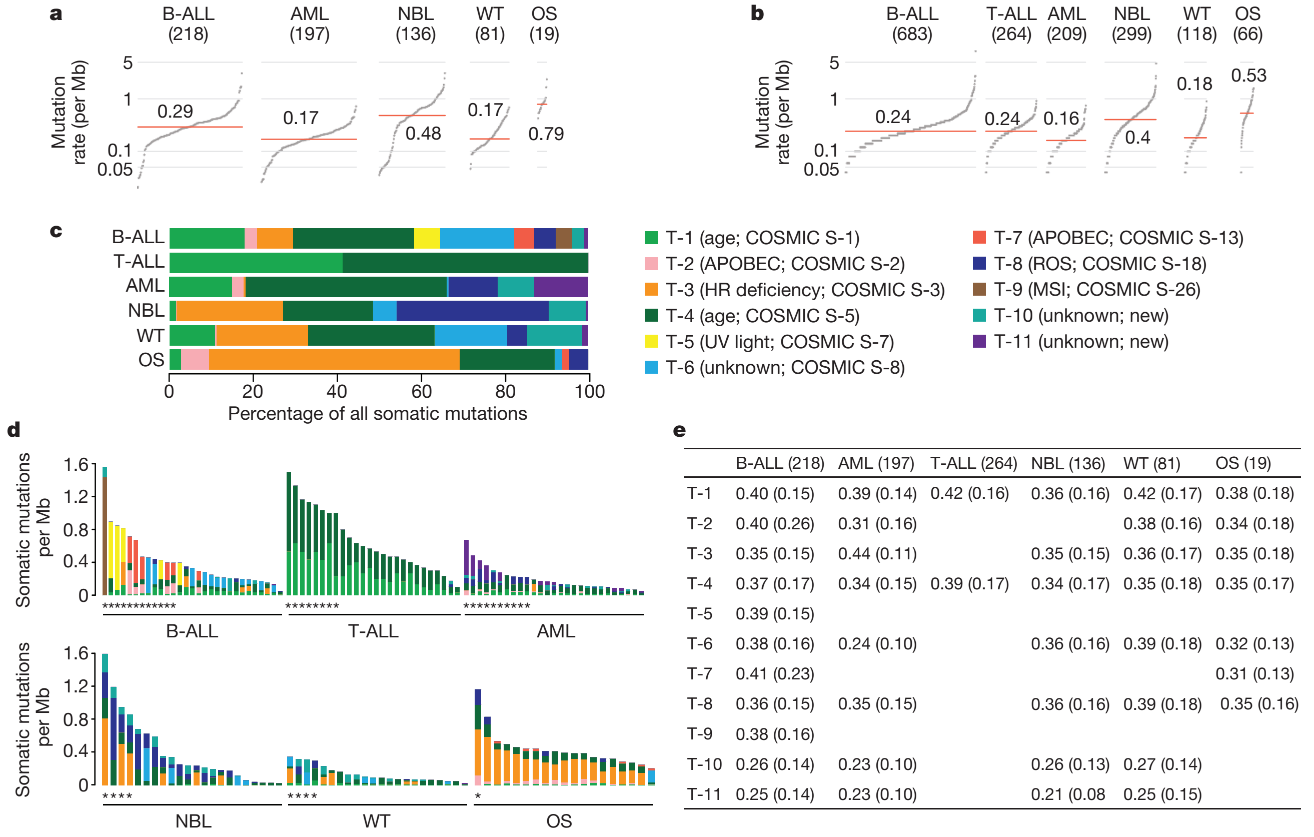
The median somatic mutation rate ranged from 0.17 per million bases (Mb) in AML and Wilms tumours to 0.79 in osteosarcomas (Fig. 1a, b), lower than the 1–10 per Mb found in common adult cancers[6]. Genome-wide analysis (see Methods) identified 11 mutational signatures (T-1 through T-11; Fig. 1c–e and Supplementary Table 1a–c). Signatures T-1 through T-9 corresponded to known COSMIC signatures[7], whereas T-10 and T-11 were novel but enriched in mutations with a low (<0.3) mutant allele fraction (MAF).

Signatures T-1 and T-4 (clock-like endogenous mutational processes) were present in all samples and contributed to large proportions of all mutations in T-ALL (97%), AML (63%), B-ALL (36%), and Wilms tumours (28%). T-2 and T-7 (APOBEC (apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like)) were highly enriched in B-ALLs with *ETV6-RUNX1* fusions (15-fold and 9-fold enrichment for T-2 and T-7, respectively; Supplementary Table 1e). T-3 (homologous recombination deficiency) was present in many childhood cancers, including osteosarcomas (18 of 19), NBLs (59 of 137), Wilms tumours (28 of 81), and B-ALL (47 of 218). T-8 (8-oxoguanine DNA damage) was present in a small proportion (4.5–12%) of AML, B-ALL, osteosarcoma, and Wilms tumour samples. T-8 was also present in many (36%) NBL samples and was associated with age at diagnosis (Supplementary Table 1d). T-9 (DNA repair deficiency) was present in two B-ALLs, including one (sample PARJSR) with a somatic *MSH6* frameshift mutation. T-2, T-3, T-5, T-7, T-8, and T-9 were enriched among the 39 samples with elevated mutation rates in each histotype (Fig. 1d).

The T-5 ultraviolet-light (UV)-exposure signature was unexpectedly present in eight B-ALL samples (Extended Data Fig. 2a–c). Although its mutation rate in B-ALL, ranging from 0.06 to 0.72 per Mb, was 100-fold lower than the average rate in adult (15.8 per Mb)[8] and paediatric (14.4 per Mb)[9] skin cancer, T-5 exhibited other features associated with UV-related DNA damage. Specifically, CC>TT dinucleotide mutations were enriched 110-fold in these eight B-ALL samples when compared with other samples ($P = 1.07 \times 10^{-7}$), which is consistent with pyrimidine dimer formation. Moreover, transcriptional strand bias in T-5 indicated that photodimer formation contributed to cytosine damage. The validity of T-5 was further confirmed by analysis of the mutation clonality, cross-platform concordance, genomic distribution and mutation spectrum of each sample (see Methods, Extended Data Fig. 2d–i), indicating that UV exposure or other mutational processes[10,11] may contribute to paediatric leukemogenesis. Notably, all T-5 B-ALLs had aneuploid genomes ($P = 3 \times 10^{-5}$; two-sided binomial test; cohort frequency 24%) without any oncogenic fusions.

By analysing the enrichment[12,13] of somatic alterations within each histotype or the pan-cancer cohort (see Methods), we identified 142 significantly mutated driver genes (Fig. 2a, Supplementary Table 2, Extended Data Fig. 3a). Somatic alterations in *CDKN2A*, which were predominantly deletions, occurred at the highest frequency, affecting

[1]Computational Biology, St Jude Children's Research Hospital, Memphis, Tennessee 38105, USA. [2]Department of Cellular and Molecular Medicine and Department of Bioengineering and Moores Cancer Center, University of California, San Diego, La Jolla, California 92093, USA. [3]Independent Researcher, Chicago, Illinois 60654, USA. [4]Oncology, St Jude Children's Research Hospital, Memphis, Tennessee 38105, USA. [5]Office of Cancer Genomics, National Cancer Institute, Bethesda, Maryland 20892, USA. [6]Genetics Branch, Center for Cancer Research, National Cancer Institute, NIH, Bethesda, Maryland 20892, USA. [7]Department of Biostatistics, St Jude Children's Research Hospital, Memphis, Tennessee 38105, USA. [8]Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA. [9]Division of Oncology and Center for Childhood Cancer Research, Children's Hospital of Philadelphia and the Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [10]Cancer Therapy Evaluation Program, National Cancer Institute, Bethesda, Maryland 20892, USA. [11]Division of Hematology-Oncology, Connecticut Children's Medical Center, Hartford, Connecticut 06106, USA. [12]The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut 06032, USA. [13]Department of Pathology and Laboratory Medicine, Ann & Robert H. Lurie Children's Hospital of Chicago, Robert H. Lurie Cancer Center, Northwestern University, Chicago, Illinois 60208, USA.
†Present address: Tempus Labs Inc., Chicago, Illinois 60654, USA.
*These authors contributed equally to this work.



| | B-ALL (218) | AML (197) | T-ALL (264) | NBL (136) | WT (81) | OS (19) |
|---|---|---|---|---|---|---|
| T-1 | 0.40 (0.15) | 0.39 (0.14) | 0.42 (0.16) | 0.36 (0.16) | 0.42 (0.17) | 0.38 (0.18) |
| T-2 | 0.40 (0.26) | 0.31 (0.16) | | | 0.38 (0.16) | 0.34 (0.18) |
| T-3 | 0.35 (0.15) | 0.44 (0.11) | | 0.35 (0.15) | 0.36 (0.17) | 0.35 (0.18) |
| T-4 | 0.37 (0.17) | 0.34 (0.15) | 0.39 (0.17) | 0.34 (0.17) | 0.35 (0.18) | 0.35 (0.17) |
| T-5 | 0.39 (0.15) | | | | | |
| T-6 | 0.38 (0.16) | 0.24 (0.10) | | 0.36 (0.16) | 0.39 (0.18) | 0.32 (0.13) |
| T-7 | 0.41 (0.23) | | | | | 0.31 (0.13) |
| T-8 | 0.36 (0.15) | 0.35 (0.15) | | 0.36 (0.16) | 0.39 (0.18) | 0.35 (0.16) |
| T-9 | 0.38 (0.16) | | | | | |
| T-10 | 0.26 (0.14) | 0.23 (0.10) | | 0.26 (0.13) | 0.27 (0.14) | |
| T-11 | 0.25 (0.14) | 0.23 (0.10) | | 0.21 (0.08 | 0.25 (0.15) | |

**Figure 1 | Somatic mutation rate and signature.** Sample size of each histotype is shown in parentheses. Mutation rate using non-coding SNVs from WGS (**a**) and coding SNVs from WGS and WES (**b**). Red line, median. **a** and **b** are scaled to the total number of samples with WGS ($n = 651$), WGS or WES ($n = 1,639$), respectively. **c**, Mutational signatures identified from WGS and T-ALL WES data and their contribution in each histotype. **d**, Mutation spectrum of representative samples in each histotype. Hypermutators (three s.d. above mean rate of corresponding histotype) are labelled with an asterisk. **e**, Mean and s.d. of MAF of each signature in each histotype.

207 of 267 (78%) T-ALLs, 91 of 218 (42%) B-ALLs and 2 of 19 (11%) osteosarcomas (Extended Data Fig. 3b). More than half (73) of the driver genes were specific to a single histotype, such as *TAL1* for T-ALL and *ALK* for NBL (Extended Data Fig. 3c). Genes that were mutated in both leukaemias and the three solid tumour histotypes accounted for only 17% of driver genes (Extended Data Fig. 3e), of which some genes had various types of somatic alteration. For example, *STAG2*, a known driver gene for Ewing's sarcoma[14] and adult AML[15], exhibited five different types of somatic alteration (single nucleotide variants (SNVs), small insertions or deletions (indels), copy number alterations (CNAs), structural variants and internal tandem duplications (ITDs)) across five histotypes (Extended Data Fig. 4a–d). Nine *STAG2* variants were predicted to cause protein truncation, including four predicted by aberrant transcripts in RNA sequencing (RNA-seq). Notably, 78 of 142 driver genes (Supplementary Table 2) were not found in adult pan-cancer studies[1–4], and 43 (Fig. 2a and Extended Data Fig. 3a) were not found in the Cancer Gene Census (v81)[16]. Thirty-seven were absent from both sources, although mutations in cancer have been reported for 29 of these genes, such as *NIPBL*[17–19] and *LEMD3*[20] (Extended Data Fig. 4p, q). Nearly half (40–50%) of point mutations in leukaemia and NBL driver genes had low MAFs (<0.3), indicative of subclonal mutations contributing to tumorigenesis (Extended Data Fig. 3f).

Three hundred and four gene-pairs exhibited statistically significant ($P < 0.05$, two-sided Fisher's exact test; Fig. 2b, Supplementary Table 3) co-occurrence (for example, *USP7* and *TAL1* in T-ALL[21]) or mutual exclusivity (for example, *MYCN* and *ATRX* in NBL[22]). The analysis also unveiled novel co-occurrences (for example, *ETV6* and *IKZF1* in AML and *CREBBP* and *EP300* in B-ALL) and mutual exclusivities (for example, *SHANK2* and *MYCN* in NBL and *PAX5* and *TP53* in B-ALL).

Because of reduced power for detecting low-frequency drivers[2] (detection limits were 1% for the entire cohort and 3% for individual histotypes with more than 200 samples; Extended Data Fig. 5 and Methods), we performed subnetwork analyses[3] and variant pathogenicity classification[23] (see Methods), identifying 184 variants in 82 additional genes (Supplementary Table 4 and Extended Data Fig. 4e, f). A notable example is the *MAP3K4* G1366R mutation, which was found in one T-ALL, two B-ALLs, and one Wilms tumour. *MAP3K4* is a member of the MAPK family[24] and structural modelling indicates that the G1366R mutation is likely to cause disruption of normal inhibitory domain binding and kinase dynamics[24] (Extended Data Fig. 4l, m). Several genes in which structural variants were found (*PDGFRA*, *CDK4*, *YAP1*, *UBTF*) are listed in Extended Data Fig. 4.

While the percentage of tumours with point mutations in driver genes was highly consistent between whole-genome sequencing (WGS) and whole-exome sequencing (WES) (Fig. 3a), WGS makes it possible to detect CNAs and structural variants, which are frequently driver events for paediatric cancers. For example, 72% of NBL tumours analysed by WGS had at least one driver variant compared to 26% of those analysed by WES (Fig. 3a and Extended Data Fig. 4j, k). Furthermore, integrative analyses of CNAs and structural variants with WGS data revealed chromothripsis (that is, massive rearrangements caused by a single catastrophic event) in 11% of all samples (13 in osteosarcomas, 15 in Wilms tumours, 22 in NBL, 14 in B-ALL, and 6 in AML; Extended Data Fig. 1f). We next performed pathway analyses (see Methods) on 654 samples analysed by WGS and 264 T-ALL samples analysed by both WES and single nucleotide polymorphism (SNP) arrays, totaling 682 leukaemias and 236 solid tumours.

The 21 biological pathways that were disrupted by driver alterations were either common (for example, cell cycle and epigenetic regulation)

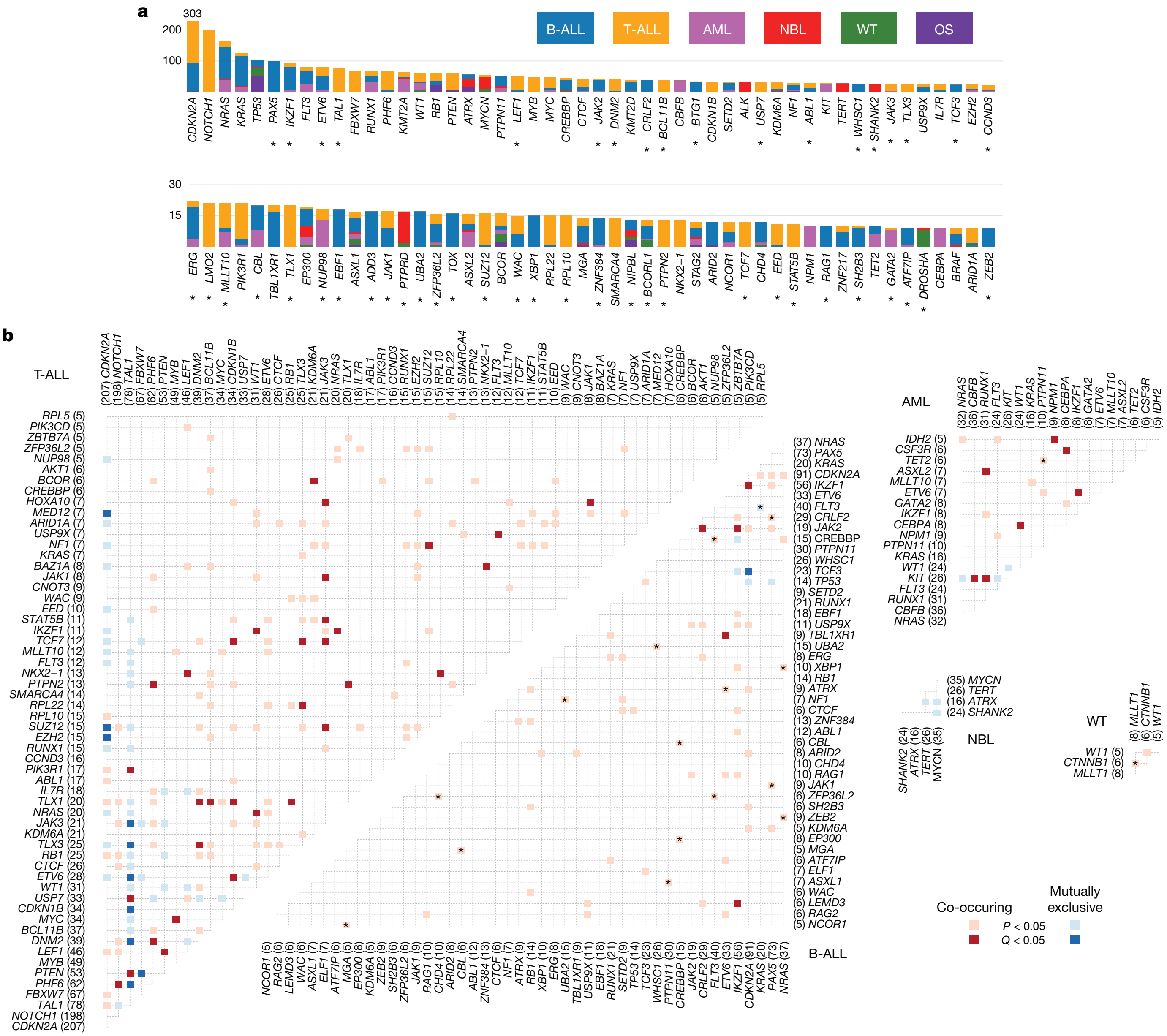
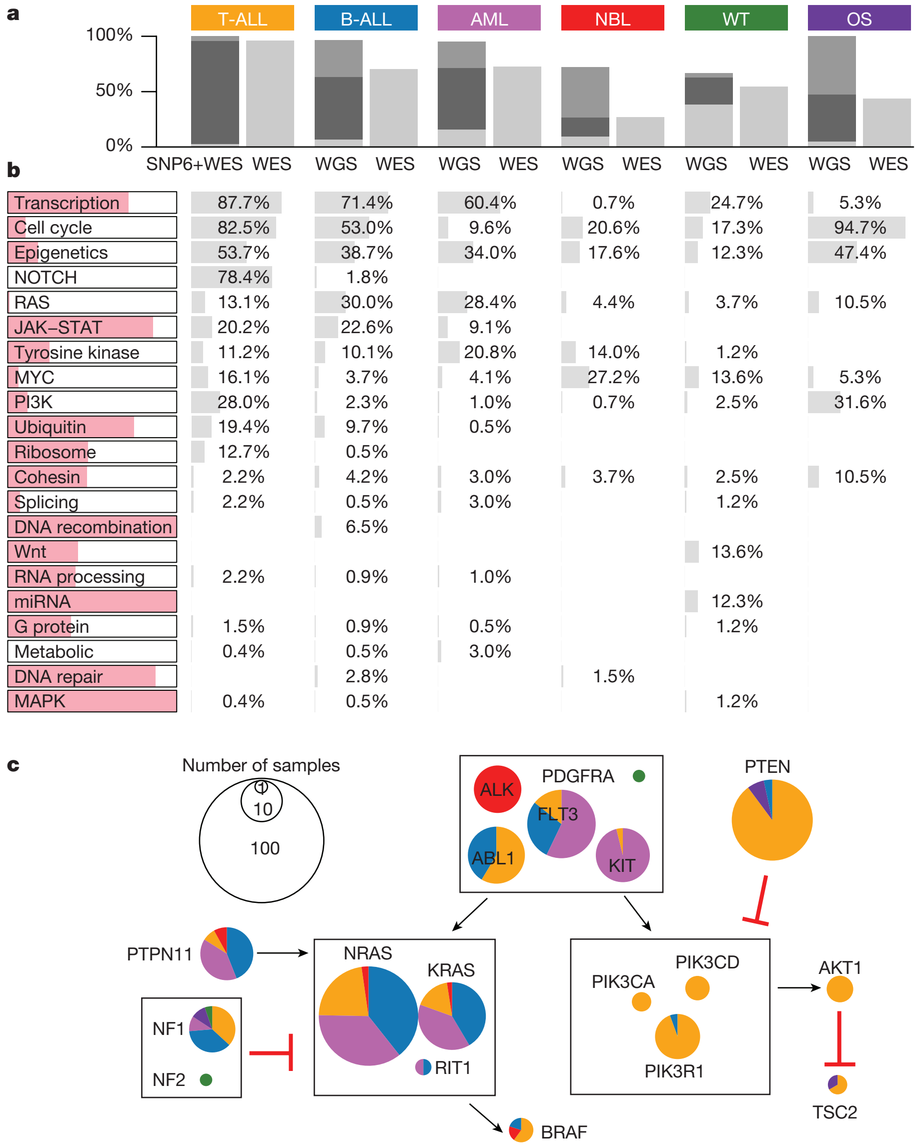


**Figure 2 | Candidate driver genes in paediatric cancer. a**, Top 100 recurrently mutated genes: case count for each histotype is shown in the same colour as the legend. Asterisk indicates gene not reported in prior adult pan-cancer analyses. **b**, Statistically significant pairwise relationships ($P < 0.05$; two-sided Fisher's exact test) for co-occurrence (red) or exclusivity (blue) in each histotype. Gene pairs with $Q < 0.05$ are coloured dark red (co-occurring) or dark blue (exclusive) to account for false discovery rate. Significance detected only in WGS + WES samples is marked with an asterisk. Shown in parentheses are number of mutated samples.

or histotype-specific (for example, JAK–STAT, Wnt/β-catenin, and NOTCH signalling) (Fig. 3b). More importantly, the genes that were mutated in each pathway differed between histotypes. One example is signalling pathways such as RAS, JAK–STAT and PI3K (Fig. 3c). For genes in these pathways, somatic alterations in solid tumours primarily occurred in *ALK*, *NF1*, and *PTEN*, whereas nearly all mutations in *FLT3*, *PIK3CA*, *PIK3R1*, and *RAS* were found in leukaemias. Although many biological processes are dysregulated in both paediatric and adult cancers[1,2,4], the affected genes may be either paediatric-specific (for example, transcription factors and JAK–STAT pathway genes) or common to both (for example, cell cycle genes and epigenetic modifiers). Notably, two novel *KRAS* isoforms were detected in 70% of leukaemias but rarely in solid tumours (Extended Data Fig. 6).

Evaluation of mutant allele expression makes it possible to assess the effects on the gene product and to detect potential epigenetic regulation that may cause allelic imbalance. Here we present this analysis on 6,959 coding mutations with matching WGS and RNA-seq data. RNA-seq expression clusters confirmed the tissue of origin of each histotype (Extended Data Fig. 7). Mutant alleles were expressed for 34% of these mutations, which is consistent with previous reports[25–27]. The expression of mutant alleles is generally associated with corresponding DNA MAF and the expression levels of host genes (Fig. 4a); however, exceptions can be found due to X-inactivation, imprinting, nonsense-mediated decay or complex structural re-arrangements (Extended Data Fig. 8a).

Allele-specific expression (ASE) was evaluated for 2,477 somatic point mutations with sufficient read-depth in DNA and RNA-seq (see Methods). Of 486 candidate ASE mutations (Supplementary Table 5), 279 had no detectable expression of the mutant allele, and a comparable DNA MAF distribution was found for truncating and non-truncating mutations ($P = 0.5$, two-sided Wilcoxon rank-sum test, Extended Data



| | T-ALL | B-ALL | AML | NBL | WT | OS |
|---|---|---|---|---|---|---|
| Transcription | 87.7% | 71.4% | 60.4% | 0.7% | 24.7% | 5.3% |
| Cell cycle | 82.5% | 53.0% | 9.6% | 20.6% | 17.3% | 94.7% |
| Epigenetics | 53.7% | 38.7% | 34.0% | 17.6% | 12.3% | 47.4% |
| NOTCH | 78.4% | 1.8% | | | | |
| RAS | 13.1% | 30.0% | 28.4% | 4.4% | 3.7% | 10.5% |
| JAK–STAT | 20.2% | 22.6% | 9.1% | | | |
| Tyrosine kinase | 11.2% | 10.1% | 20.8% | 14.0% | 1.2% | |
| MYC | 16.1% | 3.7% | 4.1% | 27.2% | 13.6% | 5.3% |
| PI3K | 28.0% | 2.3% | 1.0% | 0.7% | 2.5% | 31.6% |
| Ubiquitin | 19.4% | 9.7% | 0.5% | | | |
| Ribosome | 12.7% | 0.5% | | | | |
| Cohesin | 2.2% | 4.2% | 3.0% | 3.7% | 2.5% | 10.5% |
| Splicing | 2.2% | 0.5% | 3.0% | | 1.2% | |
| DNA recombination | | 6.5% | | | | |
| Wnt | | | | | 13.6% | |
| RNA processing | 2.2% | 0.9% | 1.0% | | | |
| miRNA | | | | | 12.3% | |
| G protein | 1.5% | 0.9% | 0.5% | | 1.2% | |
| Metabolic | 0.4% | 0.5% | 3.0% | | | |
| DNA repair | | 2.8% | | 1.5% | | |
| MAPK | 0.4% | 0.5% | | | 1.2% | |

**Figure 3 | Biological processes with somatic alterations in paediatric cancer. a**, Percentage of tumours with at least one driver alteration are shown for each histotype. WGS-analysed tumours may have point mutations (light grey), CNAs or structural variations (SV) (dark grey), or both (black). For T-ALL, CNAs were derived from SNP array. **b**, Percentage of tumours within each histotype that have somatic alterations in 21 biological pathways; histotype ordering is as in **a**. The coloured portion of each pathway indicates the percentage of variants in genes that are absent in three TCGA pan-cancer studies. **c**, Mutation occurrence by histotype in RAS, tyrosine kinase, and PI3K pathways.

Fig. 8b). Of the remaining 207 candidate ASE mutations, 76% of truncating mutations exhibited suppression of the mutant allele ($P=7\times10^{-5}$; two-sided binomial test), while 87% of hotspot mutations showed the opposite trend of elevated expression ($P=6\times10^{-5}$; two-sided binomial test; Fig. 4b, Extended Data Fig. 8c). Excluding hotspot mutations resulted in equal distribution of suppression versus elevation (66 versus 55) for the remaining 121 non-truncating ASE mutations ($P=0.4$; two-sided binomial test).

Subclonal loss-of-heterozygosity (LOH) in tumours is a confounding factor for ASE analysis. For example, significant allelic imbalance between tumour DNA and RNA MAF of *WT1* D447N in an AML that also harboured a subclonal 11p copy-neutral LOH (Fig. 4c) could be attributed to ASE or *WT1* expression of a subclone with a double-hit of D447N mutation and 11p LOH. To address this, we performed single-cell DNA sequencing on 63 germline variants on 11p and the somatic point mutations. We confirmed ASE by establishing that *WT1* D447N and 11p LOH occurred in separate subclones (Fig. 4c and Extended Data Fig. 9a, b). The resulting genotype data projected that one *WT1* allele was silenced in a common ancestor and the other was lost in the three descendant subclones by 11p LOH, acquisition of the *WT1* D447N mutation, or focal deletion. Two additional AMLs with *WT1* D447N also exhibited ASE (Extended Data Fig. 9c), implying that loss of *WT1* expression by epigenetic silencing or mutations in *cis*-regulatory elements is not rare in AML. Similarly, single-cell sequencing of an ALL sample confirmed ASE of a *JAK2* hotspot mutation (Extended Data Fig. 9d).

The somatic variants used for this study are available at the National Cancer Institute TARGET Data Matrix and our ProteinPaint[28] portal, which provides an interactive heat map viewer for exploring mutations, genes, and pathways across the six histotypes (Extended Data Fig. 10). The portal also hosts the somatic variants analysed by the companion paediatric pan-cancer study of 961 tumours from 24 histotypes, including 559 central nervous system tumours[29]. We anticipate that these complementary pan-cancer datasets will be an important resource for investigations of functional validation and implementation of clinical genomics for paediatric cancers.

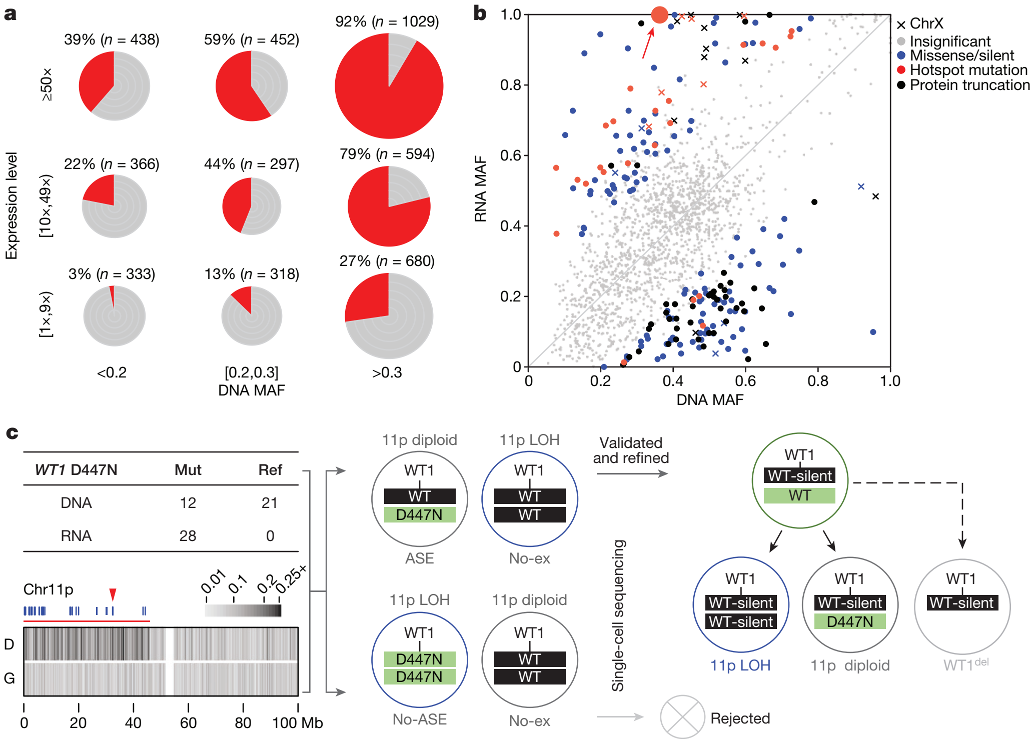


**Figure 4 | Mutant allele expression. a**, Percentage of expressed mutations (red) categorized by DNA MAF (*x* axis) and expression level (*y* axis). Circle size is proportional to mutation counts. **b**, Detection of ASE in expressed mutations by comparing DNA and RNA MAF in 443 samples (solid colours, statistically significant (two-sided Fisher's exact test $Q < 0.01$ and effect size $>0.2$); grey, not significant). **c**, Confirming ASE for *WT1* D447N (red arrow in **b**) by single-cell sequencing. Presence of subclonal 11p LOH leads to two possible outcomes: the mutant allele is in either non-LOH subclone (top) or LOH subclone (bottom): the former suggests ASE and the latter rejects ASE due to homozygosity. No-ex: *WT1* not expressed.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 6 February 2017; accepted 17 January 2018.**

**Published online 28 February 2018.**

1. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502,** 333–339 (2013).
2. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505,** 495–501 (2014).
3. Leiserson, M. D. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47,** 106–114 (2015).
4. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45,** 1134–1140 (2013).
5. Downing, J. R. *et al.* The Pediatric Cancer Genome Project. *Nat. Genet.* **44,** 619–622 (2012).
6. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).
7. Alexandrov, L. B. Understanding the origins of human cancer. *Science* **350,** 1175–1177 (2015).
8. Hayward, N. K. *et al.* Whole-genome landscapes of major melanoma subtypes. *Nature* **545,** 175–180 (2017).
9. Lu, C. *et al.* The genomic landscape of childhood and adolescent melanoma. *J. Invest. Dermatol.* **135,** 816–823 (2015).
10. Reid, T. M. & Loeb, L. A. Tandem double CC-->TT mutations are produced by reactive oxygen species. *Proc. Natl Acad. Sci. USA* **90,** 3904–3907 (1993).
11. Newcomb, T. G., Allen, K. J., Tkeshelashvili, L. & Loeb, L. A. Detection of tandem CC-->TT mutations induced by oxygen radicals using mutation-specific PCR. *Mutat. Res.* **427,** 21–30 (1999).
12. Pounds, S. *et al.* A genomic random interval model for statistical analysis of genomic lesion data. *Bioinformatics* **29,** 2088–2095 (2013).
13. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).
14. Tirode, F. *et al.* Genomic landscape of Ewing sarcoma defines an aggressive subtype with co-association of STAG2 and TP53 mutations. *Cancer Discov.* **4,** 1342–1353 (2014).
15. Ley, T. J. *et al.* Genomic and epigenomic landscapes of adult *de novo* acute myeloid leukemia. *N. Engl. J. Med.* **368,** 2059–2074 (2013).
16. Futreal, P. A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4,** 177–183 (2004).
17. Krantz, I. D. *et al.* Cornelia de Lange syndrome is caused by mutations in *NIPBL*, the human homolog of *Drosophila melanogaster Nipped-B*. *Nat. Genet.* **36,** 631–635 (2004).
18. Tonkin, E. T., Wang, T. J., Lisgo, S., Bamshad, M. J. & Strachan, T. *NIPBL*, encoding a homolog of fungal Scc2-type sister chromatid cohesion proteins and fly Nipped-B, is mutated in Cornelia de Lange syndrome. *Nat. Genet.* **36,** 636–641 (2004).
19. Barber, T. D. *et al.* Chromatid cohesion defects may underlie chromosome instability in human colorectal cancers. *Proc. Natl Acad. Sci. USA* **105,** 3443–3448 (2008).
20. Hellemans, J. *et al.* Loss-of-function mutations in LEMD3 result in osteopoikilosis, Buschke–Ollendorff syndrome and melorheostosis. *Nat. Genet.* **36,** 1213–1218 (2004).
21. Liu, Y. *et al.* The genomic landscape of pediatric and young adult T-lineage acute lymphoblastic leukemia. *Nat. Genet.* **49,** 1211–1218 (2017).
22. Cheung, N. K. & Dyer, M. A. Neuroblastoma: developmental biology, cancer genomics and immunotherapy. *Nat. Rev. Cancer* **13,** 397–411 (2013).
23. Zhang, J. *et al.* Germline mutations in predisposition genes in pediatric cancer. *N. Engl. J. Med.* **373,** 2336–2346 (2015).
24. Mita, H., Tsutsui, J., Takekawa, M., Witten, E. A. & Saito, H. Regulation of MTK1/MEKK4 kinase activity by its N-terminal autoinhibitory domain and GADD45 binding. *Mol. Cell. Biol.* **22,** 4544–4555 (2002).
25. Rashid, N. U. *et al.* Differential and limited expression of mutant alleles in multiple myeloma. *Blood* **124,** 3110–3117 (2014).
26. Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486,** 395–399 (2012).
27. Govindan, R. *et al.* Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* **150,** 1121–1134 (2012).
28. Zhou, X. *et al.* Exploring genomic alteration in pediatric cancer using ProteinPaint. *Nat. Genet.* **48,** 4–6 (2016).
29. Gröbner, S. N. *et al.* The landscape of genomic alterations across childhood cancers. *Nature* http://doi.org/10.1038/nature25480 (2018).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** This project was funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E to J.Z. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government. This project was also funded by a supplement to the Children's Oncology Group Chair's grant CA098543 and by funding from ALSAC to St Jude Children's Research Hospital and by a Cancer Center Support (Core) grant (CA21765) from the National Cancer Institute to St Jude Children's Research Hospital. S.P.H. is the Jeffrey E. Perelman Distinguished Chair in the Department of Pediatrics at the Children's Hospital of Philadelphia. We thank B. Raphael and M. Reyna for their help on HotNet2 analysis; K. Shannon for discussion on novel KRAS isoforms; L. Robison and C. Howell for providing SEER incidence data for childhood ALL; S. Newman for assistance in CDKN2A deletion analysis; P. Gupta for data download support; T. Pugh, P. Gesuwan, T. Davidsen, C. G. Mullighan, J. Farrar, V. Huff, and S. Gadd for their participation in the TARGET Analysis Working Group and contributions to data generation and analysis; and L. Tian, S. Wang and N. Badders for assistance in revising the manuscript.

**Author Contributions** J.Z., D.S.G., and L.B.A. designed all experiments. J.Z., X.M., Yu.L., Ya.L., and L.B.A. performed all experiments and analyses with the help of M.N.E., Y.L., X.Z., E.S., M.C.R., S.P., X.Ca., L.C.H., M.R.W., S.D., R.E.R., Z.W., X.Ch., L.D. and J.M.G.A. The structural modelling was performed by R.H. The single-cell assays were performed by C.G., V. G.-P., and J.E. WGS, WES, and RNA-seq data were provided by P.S.M., C.C.L., E.J.P., S.J.D., J.M.M., S.M., and S.P.H. S.M. and S.P.H. provided the leukaemia specimen for single-cell sequencing. D.S.G., J.M.G.A., M.A.S., and L.C.H oversaw the administrative and data management aspects of the TARGET project. The manuscript was written by J.Z., X.M., Y.L., L.B.A., J.M.M., and S.P.H. and was reviewed and edited by all authors.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to J.Z. (jinghui.zhang@stjude.org).

**Reviewer Information** *Nature* thanks M. Meyerson and the other anonymous reviewer(s) for their contribution to the peer review of this work.

## METHODS

**Patient samples.** Specimens were obtained through collaborations with the Children's Oncology Group (COG) and the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) project. Institutional review boards from the following institutions were responsible for oversight: Ann & Robert H. Lurie Children's Hospital, Fred Hutchinson Cancer Research Center, National Cancer Institute, St Jude's Children's Research Hospital, The Children's Hospital of Philadelphia, The University of New Mexico, Texas Children's Hospital, and The Hospital for Sick Children. In our cohort, osteosarcoma has a higher percentage of older patients because the age of onset has a bimodal distribution: the first peak occurs among adolescents and young adults, and the second (associated with Paget disease and with a different underlying biology[30]) occurs among the elderly. We used an age cutoff of 40 years, which is typical for COG-conducted osteosarcoma trials[31]. Informed consent was obtained from all subjects.

**Genomic datasets.** WGS, WES, and RNA-seq data were downloaded from dbGaP with study identifier phs000218 (including phs000463, phs000464, phs000465, phs000467, phs000471, and phs000468). Among the 1,699 cases analysed, 45 B-ALLs[32,33], 197 AMLs[34], 264 T-ALLs[21], 240 NBLs[35] and 115 Wilms tumours[36] have been included in published studies of individual histotypes. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

**WGS data analysis.** WGS data were generated with Complete Genomics Inc. (CGI) technology with an average genome-wide coverage of 50× using 31- to 35-bp mate-paired reads, which was powered for detecting mutations in 94% of mappable exonic regions[37,38]. Read pairs were mapped to hg19/GRCh37, and somatic SNVs, indels, and structural variants were analysed by comparing paired tumour and normal genomes using the CGI Cancer Sequencing service pipeline version 2[38,39].

For each case, we downloaded CGI-generated WGS files for somatic SNVs, indels, structural variants, and CNAs from the TARGET Data Matrix as the starting point for our analysis.

**Filtering of point mutations.** Putative somatic point mutations including SNVs and indels were extracted from Mutation Annotation Format files and run through a filter to remove false-positive calls. First, germline variants were filtered by using: (1) NLHBI Exome Sequencing Project (http://evs.gs.washington.edu/EVS/); (2) dbSNP (build 132); (3) St Jude/Washington University Paediatric Cancer Genome Project (PCGP); and (4) germline variants present in five or more TARGET CGI WGS cases in each cohort. Second, a variant was removed unless it met the following criteria: (1) at least three reads supported the mutant allele in the tumour; (2) the mutant read count in the tumour was significantly higher than normal ($P < 0.01$ by two-sided Fisher's exact test); and (3) the normal MAF was below 0.05. Finally, a BLAT search[40] was run on the mutant allele with 20-bp flanking to verify unique mapping.

A 'rescue' pipeline was implemented to avoid over-filtering, by using the customized AnnoVar annotation and pathogenicity identification tool Medal_Ceremony[23] (M.N.E. *et al.*, unpublished). Pathogenic variants were rescued and further curated with ProteinPaint[28].

This filtering has reduced the original 51 million SNVs and 38 million indels from the CGI files to a set of 711,490 SNVs and 57,700 indels. Of these, 9,397 SNVs and 1,000 indels are in protein coding regions. A comparison with gnomAD database (version r2.0.1; http://gnomad.broadinstitute.org/) indicated that 1.1% of our detected SNVs overlap with SNPs with population frequency greater than 0.1%. Verification of somatic point mutations after filtering is presented in Supplementary Note 1.

**Filtering of structural variation.** CGI structural variants were filtered to remove germline rearrangements, including those found in the Database of Genomic Variants, dbSNP, PCGP, recurrent germline rearrangements from CGI Mutation Annotation Format files, low-confidence somatic calls (>90% reference similarity to the assembled sequence) and those with both structural variant breakpoints falling into gap regions (hg19). Each structural variant was required to have an assembled contig length of at least 10 bp on each breakpoint. CNAs in each tumour were integrated into the structural variant analysis by matching breakpoints within a 5-kb window to rescue rearrangements with CNA support by manual curation. A comparison of CGI structural variants with the known oncogenic re-arrangement in AML and B-ALL is presented in Supplementary Note 2.

**Copy number alterations.** We adapted the CONSERTING algorithm[41] to detect CNAs from CGI WGS data. In brief, germline single nucleotide polymorphisms (SNPs) reported by CGI in Mutation Annotation Format files were extracted, and paralogous variants identified from 625 germline WGS cases generated by PCGP were removed. A coverage profile was constructed using the mean of SNP read counts within a sliding window of 100 bp, and the differences between tumour and normal samples were used as inputs for CONSERTING. To detect LOH, we used SNPs with variant allele fraction (VAF) in normal sample within an interval of (0.4, 0.6) and >15× coverage in tumour and normal samples. Allelic imbalance (AI; |Tumour_VAF-0.5|) was used to detect LOH. Regions with concomitant copy number changes (|log ratio|>0.2) or LOH (AI >0.1) were subjected to manual review. Finally, regions less than 2 Mb were considered focal and included in the GRIN[12] analysis to determine the significance of the somatic alterations. A comparison of our CNA detection with clinical information is provided in Supplementary Note 3.

For osteosarcomas, manual reviews of candidate genes affected by CNA were prioritized for the following three groups owing to the high number of rearrangements caused by chromothripsis in this histotype[42]: (1) gene expression change matched the CNA status; (2) genes with recurrent loss and gain; and (3) published osteosarcoma driver genes[42]. This resulted in the discovery of 13 focal CNAs affecting *CCNE1*, *CDKN2A*, *RB1*, *PTEN*, *TUSC7*, and *YAP1* in addition to *TP53*.

**WES data analysis.** Of the 1,131 tumour-normal WES pairs, all but 23 osteosarcoma pairs exhibited the expected binomial distribution of B-allele fraction for germline SNPs. The 23 outlier samples were therefore used neither for the discovery of driver genes nor for calculating mutation rate in coding regions (Fig. 1b). They were included only for determining driver mutation prevalence.

Somatic SNVs and indels were detected by the Bambino[43] program, followed by postprocessing and manual curation as previously described[44,45]. To address 8-oxo-G artefacts[35], we implemented the D-ToxoG filtering algorithm[46].

**Somatic mutation rate.** The median mutation rate of 651 CGI WGS samples (Fig. 1a) was calculated from tier3 non-coding SNVs[47]. This analysis did not include the T-ALL cohort as only three T-ALLs were analysed by the CGI platform. Mutations in coding regions were based on coding SNVs from 1,639 samples analysed by WGS or WES (Fig. 1b). Among these, 120 samples were analysed by both WGS and WES, and the union of coding SNVs from WGS and WES were used. Twenty-three osteosarcoma WES samples were excluded from coding mutation analysis owing to quality issues described in 'WES data analysis'. For osteosarcomas, the mutation rate in coding regions (0.53 per Mb) is lower than in non-coding regions (0.79 per Mb). Nineteen osteosarcoma samples were analysed by both CGI and WES. For these samples, the mutation rate in coding regions derived from either CGI or WES was 0.54 per Mb while the mutation rate in the non-coding regions was 0.79 per Mb, indicating a potential contribution of kataegis[42] in the elevated mutation rate in non-coding regions. Within each histotype, hypermutators were defined as having mutation rates 3 s.d. above the mean (trimming 5% outliers).

**Mutational signature analysis.** Mutational catalogues were generated for each sample by using a 96-bin classification (Supplementary Table 1b). These were examined for all samples with our previously established methodology[48] to decipher mutational signatures and to quantify their activities in individual samples. The correlation between age of diagnosis and mutational signature activities was computed by using robust regression[49]. We also compared the cosine similarity between original and reconstructed samples and found that samples with more than 100 mutations had cosine similarities greater than 0.85, whereas samples with less than 100 mutations mostly (93.5%) had cosine similarities less than 0.85.

To calculate the average MAF values for each signature (Fig. 1e), each of the 96 mutation types was assigned to the signature with the highest probability (the same result was obtained if we required the highest probability to be higher than the second (by $\Delta = 0.05$, 0.1, and 0.2; data not shown). This assignment was also used for Extended Data Fig. 2e–i.

The two novel signatures, T-10 and T-11, were enriched in low MAF mutations. T-11 was the only signature that was significantly correlated ($r^2 = 0.9$) with the presence of multi-nucleotide variations composed of co-occurring SNVs separated by 3 or 4 bp which were not verified by Illumina WES. Therefore, it is likely to be associated with platform-specific sequencing artefacts.

For the eight B-ALL cases identified with mutation signatures of UV-light exposure, only 0.96% of the somatic SNVs overlap with SNPs that have population allele frequencies (AFs) >0.1% in the gnomAD database (version r2.0.1; http://gnomad.broadinstitute.org/). The overlap is only 0.22% if using AF >1%. The overlap rate is comparable to the 1.1% observed for non-UV somatic SNVs across the entire cohort (0.27% match if using AF >1%).

For each of these eight B-ALL cases, UV- and non-UV-mutations were stratified according to the ploidy of their genomic locations (Extended Data Fig. 2e–g; cluster centres estimated using R package mclust). Inter-mutational distances were plotted for comparison of genomic distribution of UV- versus non-UV mutations. Chromosomal ploidy and tumour purity were obtained from TARGET clinical files and prior publications[50]. By adjusting for ploidy and corresponding tumour purity, we calculated expected MAFs for clonal mutations as follows: denoting the tumour purity as $\pi$, the expected MAF for clonal mutations was $\pi/(2-\pi)$ in the 1-copy loss region, $\pi/2$ in the diploid region, and $\pi/(2+\pi)$ in the 1-copy gain (wild-type allele) region.

Age-specific incidence rates for childhood ALL reported by the Surveillance, Epidemiology, and End Results (SEER) program show that the rate of incidence in African American children is half of that in white children (Extended Data Fig. 2h). While none of our eight patients is African American according to clinical information and genomic imputation, we were not able to test the significance of this observation as 6.6% of the children enrolled in the COG ALL trial are African American.

**Chromothripsis analysis.** To detect chromothripsis, we first assessed whether the distribution of structural variant breakpoints in each tumour departed from the null hypothesis of random distribution using Bartlett's goodness-of-fit test[42]. The distribution of structural variant types (deletion, tandem duplication, head-to-head and tail-to-tail rearrangements) was also evaluated using a goodness-of-fit test for chromosomes with a minimum of five structural variants. Chromosomes with $P < 0.01$ for Bartlett's test and with $P > 0.01$ for the structural variant type test were further reviewed for oscillation between restricted CNA states.

**Discovery of candidate driver genes.** For the 654 CGI samples, we ran GRIN[12] analysis with all somatic variants (structural variants, CNAs, SNVs and indels) for both individual histotypes and a combined pan-cancer cohort. Similarly, we combined coding SNVs and indels identified in both WGS and WES for MutSigCV[13]. Putative genes with $Q < 0.01$ by GRIN or MutSigCV were subjected to additional curation to determine their driver status. Only one candidate gene was included in this analysis for somatic alterations affecting multiple genes such as fusion pairs (Supplementary Note 4).

We discovered 142 candidate driver genes by this approach (Supplementary Table 2). Of these, 133 were significant by GRIN analysis (87 genes common to both GRIN and MutSigCV) and nine were significant only by MutSigCV.

**HotNet2 analysis.** We applied HotNet2[3] to somatic mutations using interaction data obtained from the HINT, HI2014, and KEGG databases. We reviewed all predicted sub-networks and identified the cohesin complex with three additional genes (*STAG1*, *PDS5A* and *PDS5B*; Extended Data Fig. 4e, f).

**Pathway analysis.** Biological pathways for candidate driver genes were assigned using public pathway databases (KEGG and version 2.0 of the NCI RAS Pathway, https://www.cancer.gov/PublishedContent/Images/images/nci/organization/ras/blog/ras-pathway-v2.__v60096472.jpg), literature reviews, and biological networks produced by HotNet2. For each pathway in each histotype, a tumour was counted if any genes of that pathway were mutated. The percentage of variants in genes unique to paediatric cancers was calculated by excluding genes reported in the three TCGA pan-cancer studies[1,2,4].

**Mutual exclusivity and co-occurrence of mutations.** We tested mutual exclusivity and co-occurrence of mutations for the 142 driver genes. For each histotype, we performed this analysis in two separate sample sets: (1) samples with WGS (T-ALL with WES and SNP6), and (2) WGS and WES together (only SNVs and indels considered to avoid detection bias due to platform differences for CNVs and structural variants). For a gene pair A and B (mutated in five or more samples), we performed two-sided Fisher's exact test according to their mutation status. The R package qvalue[51] was used to control for multiple testing. Although the co-occurrence test is well-powered for most gene pairs, we recognize that the mutual exclusivity test is not powered for most gene pairs, and pairs with $P < 0.05$ were reported even if $Q > 0.05$ (Supplementary Table 3).

**Saturation analysis.** To study the effect of sample size on detecting driver genes, we performed down-sampling analysis in the pan-cancer cohort and in each histotype[2], for GRIN and MutSigCV separately. For each combination, we repeated the statistical analysis for a series of subsets of cases from 1 to the total number of samples. The number of genes (of the 142 driver genes) with false discovery rate less than 0.01 were counted for the corresponding subset. Analysis for individual histotypes was limited to those with at least 200 samples (osteosarcomas and Wilms tumours excluded).

**Somatic variant pathogenicity analysis.** We implemented a somatic mutation classifier Medal_Ceremony[23] (M.N.E. *et al.*, unpublished) to identify additional driver variants in genes that did not pass the statistical testing. Pathogenic variants include (1) hotspot SNVs and indel mutations for known cancer genes in any cancer type; (2) pathogenic mutations in ClinVar; (3) truncation mutations in known tumour suppressor genes that were expressed in the cancer histotype; and (4) known recurrent gene fusions, focal deletions, truncations, and amplifications that affect key pathways of any cancer type and that were simultaneously corroborated by an aberrant expression profile. We identified 184 variants in another 82 genes (Supplementary Table 4). *BRAF* was the most frequently mutated, with nine variants.

We also reviewed novel hotspot mutations detected in three or more samples. After removing low-confidence mutations and those without expression, one hotspot was found (*MAP3K4* G1366R, $n = 4$). Recurrent internal tandem duplication (ITD) was also reviewed for evidence in both DNA and RNA, yielding the discovery of *UBTF*-ITD in AML.

**Tumour purity assessment.** We used regions with copy number loss or copy neutral LOH as well as SNVs (coding and noncoding) from diploid regions to estimate tumour purity. For regions with LOH, a previously described method was used[42]. For SNVs, an unsupervised clustering analysis was performed with the R package mclust. Tumour purity was defined to be two times the highest cluster centre that was <0.5. The maximal CNA and SNV purity was used.

We compared our estimates with blast counts for 197 AML and 9 B-ALL samples. Of the 135 tumours with blast count >70% (value 'many' in clinical file was mapped as >70%), we identified 127 (94%) with purities >70% (seven of the other eight tumours had purities >50%). An additional 40 tumours were estimated with purities >70%, although their blast count was below 70%. Thirty-one tumours were classified as low purity (<70%) by both our analysis and blast count.

**KRAS isoform analysis.** We investigated alternative splicing in *KRAS* (Extended Data Fig. 6), as differential oncogenic activity of mutant alleles expressed in KRAS 4a or 4b isoforms has been reported previously[52]. We detected splice junction reads connecting exon 3 to one of the two novel acceptor sites in the last intron (53 bp apart). This aberrant splicing is predicted to create two novel isoforms, each incorporating one of the two novel exons (40 bp and 93 bp, respectively) located 2.2 kb downstream of exon 4A (Extended Data Fig. 6b). These novel isoforms would form truncated KRAS proteins (154/150 amino acid), each retaining the GTPase domain but losing the hypervariable region that is critical for targeting KRAS to the plasma membrane[53].

One of the two novel isoforms (novel isoform 2) was detected in myeloid cells from three healthy donors (data not shown). Protein products of *KRAS* isoforms in AML cells were analysed by western blot (Supplementary Notes 5, 6).

**RNA-seq data analysis.** RNA-seq data were mapped with StrongArm[23], and rearrangements identified with CICERO[23], followed by manual review. We performed RNA-seq clustering to confirm the tissue of origin and analysed immune infiltration using ESTIMATE[54] and CIBERSORT[55] (Extended Data Fig. 7, Supplementary Notes 7, 8).

**Allele-specific expression (ASE) in RNA-seq.** CGI and WES allele counts were combined whenever possible. Point mutations were required to have DNA and RNA coverage ≥20×. Variants with $|\text{RNA\_MAF} - \text{DNA\_MAF}| > 0.2$ and a false discovery rate of <0.01 (calculated with R package qvalue[51] on two-sided Fisher's exact test *P*) were considered to show ASE. Within-sample analysis was performed to distinguish ASE from potential artefacts caused by normal-in-tumour contamination (Extended Data Fig. 8d, Supplementary Note 9, Supplementary Table 5).

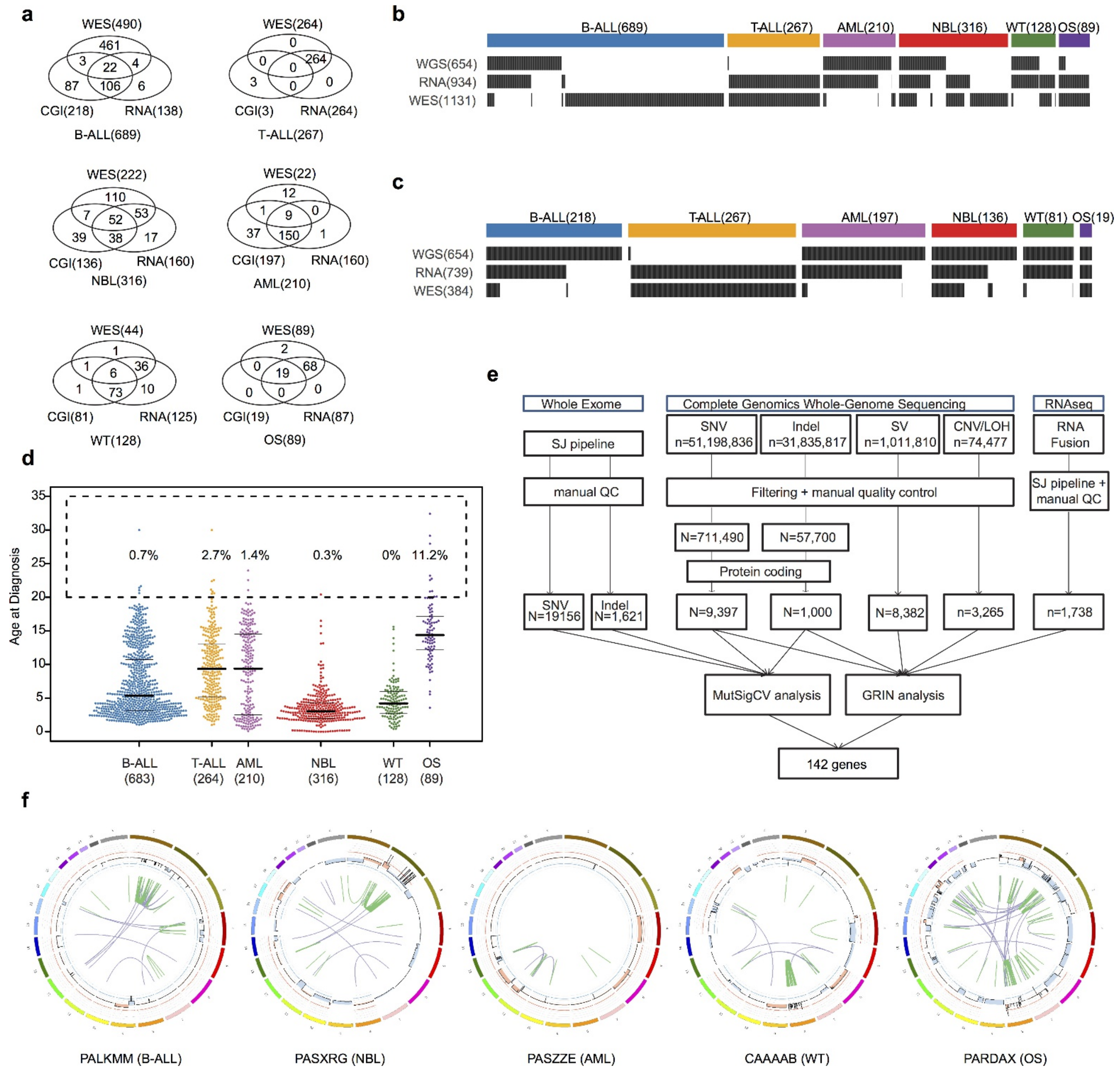
**Single-cell targeted re-sequencing.** One cryopreserved vial each from patients PAPWIU and PABLDZ was thawed using the ThawSTAR system (MedCision) and then diluted in RPMI supplemented with 1% BSA. The cells were then washed five times with C1 DNA-seq wash buffer according to the manufacturer's instructions (Fluidigm), counted and viability estimated using the LUNA-FL system (Logos Biosystems), then diluted to 300 cells per μl and loaded in a small C1 DNA-seq chip according to the manufacturer's instructions (Fluidigm, except the suspension buffer to cell ratio was changed from 4:6 to 6:4). The cells also underwent an on-chip LIVE/DEAD viability stain (Thermo Fisher). Each capture site was imaged using a Leica inverted microscope and phase contrast images, as well as fluorescent images with GFP and Y3 filters, were acquired to determine the number of cells captured and the viability of each. The cells then underwent lysis, neutralization, and MDA WGA according to the manufacturer's instructions (Fluidigm) using the GenomePhiv2 MDA kit (GE Life Sciences). One C1 chip was run per patient. Selected variants and germline SNPs then underwent microfluidic PCR-based targeted resequencing in the bulk sample or genomes amplified from the single cells using the Access Array System as previously described[56]. Target-specific assays were designed using primer3plus (https://probes.pw.usda.gov/batchprimer3/) and employed oligos purchased from Integrated DNA Technologies; multiplexing was performed according to guidelines in the Access Array manual (Fluidigm). All samples were loaded with the Access Array loader and underwent PCR cycling in an FC1 system, followed by sample-specific barcoding using standard PCR, all according to the manufacturer's instructions (Fluidigm). Amplicons were run on the MiSeq using v2 chemistry with 2 × 150-bp paired-end reads (Illumina), using custom sequencing primers, according to the Access Array manual (Fluidigm).

**Single-cell sequencing data analysis.** Mapped BAM files for each of the 96 single-cell assays were genotyped for all designed markers. Assays with two captured cells (6 assays for both cases) or assays with fewer than 50% of designed markers with coverage 10× or greater, were dropped, resulting in 48 assays for case PAPWIU (Supplementary Table 6) and 64 assays for case PAPEWB (Supplementary Table 7). The assays were called tumour cells if they had one or more somatic markers with MAFs greater than 0.05. Germline markers with MAFs greater than 0.05 were called positive. The R package pheatmap was used to visualize the single-cell data using hierarchical clustering with 'binary' distance and 'complete' agglomeration method.

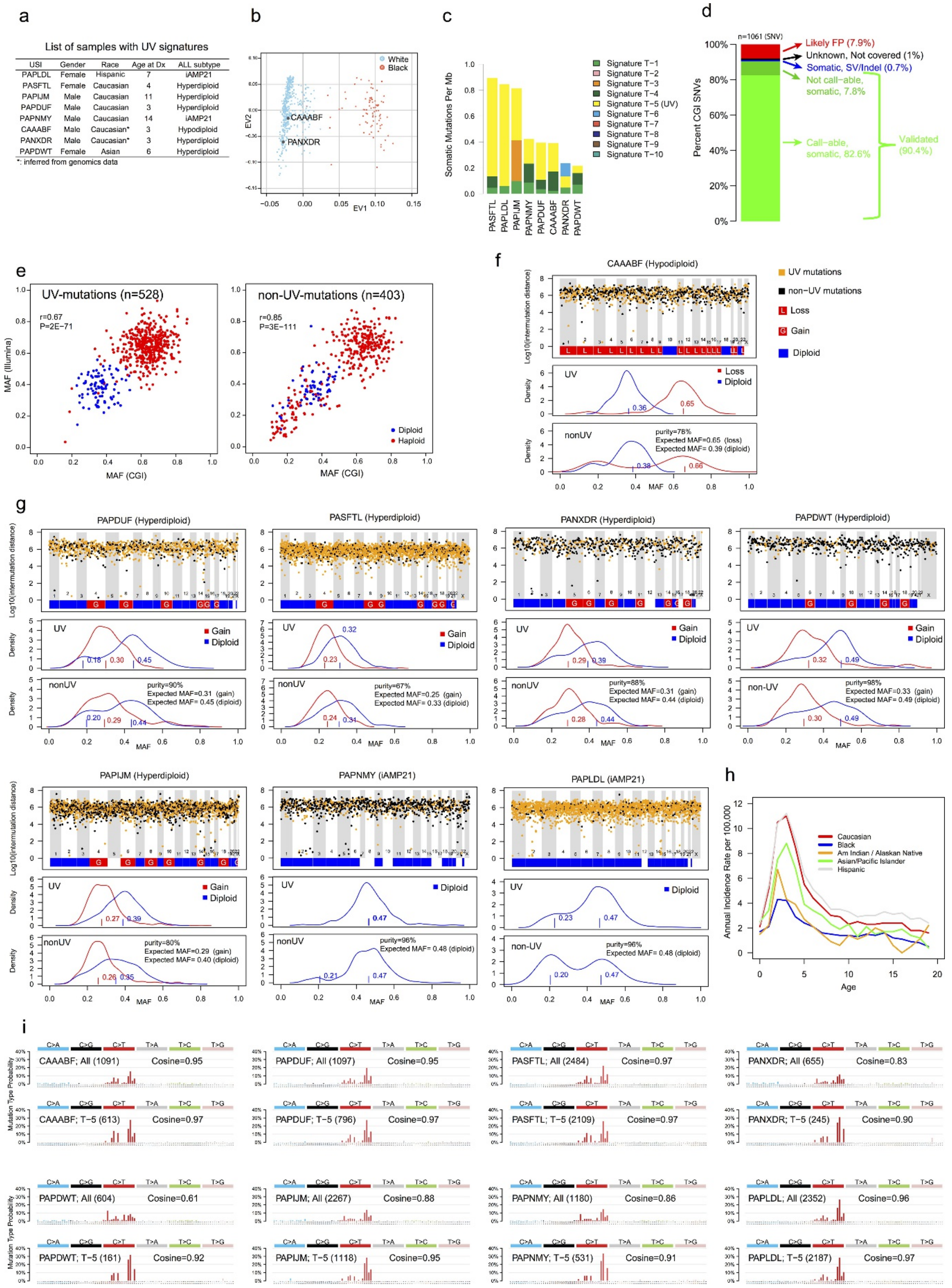
**Code availability.** Custom codes are available from the authors upon request.

**Data availability.** The somatic variants used for this study are available at the National Cancer Institute TARGET Data Matrix (https://ocg.cancer.gov/programs/target/data-matrix) and our ProteinPaint[28] portal (https://pecan.stjude.org/proteinpaint/study/pan-target), which also hosts variant data generated by Gröbner *et al.*[29] (https://pecan.stjude.org/proteinpaint/study/dkfz-ppc).

30. Mirabello, L., Troisi, R. J. & Savage, S. A. International osteosarcoma incidence patterns in children and adolescents, middle ages and elderly persons. *Int. J. Cancer* **125,** 229–234 (2009).
31. Behjati, S. *et al.* Recurrent mutation of IGF signalling genes and distinct patterns of genomic rearrangement in osteosarcoma. *Nat. Commun.* **8,** 15936 (2017).
32. Mullighan, C. G. *et al.* Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia. *N. Engl. J. Med.* **360,** 470–480 (2009).
33. Ma, X. *et al.* Rise and fall of subclones from diagnosis to relapse in pediatric B-acute lymphoblastic leukaemia. *Nat. Commun.* **6,** 6604 (2015).
34. Bolouri, H. *et al.* The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. *Nat. Med.* **24,** 103–112 (2018).
35. Pugh, T. J. *et al.* The genetic landscape of high-risk neuroblastoma. *Nat. Genet.* **45,** 279–284 (2013).
36. Gadd, S. *et al.* A Children's Oncology Group and TARGET initiative exploring the genetic landscape of Wilms tumor. *Nat. Genet.* **49,** 1487–1494 (2017).
37. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327,** 78–81 (2010).
38. Carnevali, P. *et al.* Computational techniques for human genome resequencing using mated gapped reads. *J. Comput. Biol.* **19,** 279–292 (2012).
39. Lee, W. *et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465,** 473–477 (2010).
40. Kent, W. J. BLAT–the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
41. Chen, X. *et al.* CONSERTING: integrating copy-number analysis with structural-variation detection. *Nat. Methods* **12,** 527–530 (2015).
42. Chen, X. *et al.* Recurrent somatic structural variations contribute to tumorigenesis in pediatric osteosarcoma. *Cell Rep.* **7,** 104–112 (2014).
43. Edmonson, M. N. *et al.* Bambino: a variant detector and alignment viewer for next-generation sequencing data in the SAM/BAM format. *Bioinformatics* **27,** 865–866 (2011).
44. Zhang, J. *et al.* The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature* **481,** 157–163 (2012).
45. Zhang, J. *et al.* A novel retinoblastoma therapy from genomic and epigenetic analyses. *Nature* **481,** 329–334 (2012).
46. Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **41,** e67 (2013).
47. Mardis, E. R. *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.* **361,** 1058–1066 (2009).
48. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3,** 246–259 (2013).
49. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47,** 1402–1407 (2015).
50. Holmfeldt, L. *et al.* The genomic landscape of hypodiploid acute lymphoblastic leukemia. *Nat. Genet.* **45,** 242–252 (2013).
51. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100,** 9440–9445 (2003).
52. To, M. D. *et al.* Kras regulatory elements and exon 4A determine mutation specificity in lung cancer. *Nat. Genet.* **40,** 1240–1244 (2008).
53. Eisenberg, S. & Henis, Y. I. Interactions of Ras proteins with the plasma membrane and their roles in signaling. *Cell. Signal.* **20,** 31–39 (2008).
54. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4,** 2612 (2013).
55. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12,** 453–457 (2015).
56. Gawad, C., Koh, W. & Quake, S. R. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proc. Natl Acad. Sci. USA* **111,** 17947–17952 (2014).



**Extended Data Figure 1 | Cohort description and workflow. a**, Venn diagram of samples analysed by whole-exome (WES), whole genome (CGI) and whole transcriptome (RNA-seq) sequencing in this cohort. **b**, **c**, Sample-level sequencing status of the entire cohort (**b**) and those with WGS data (**c**, SNP6 for T-ALL). **d**, Age distribution for each histotype. Median, first and third quartiles are indicated by horizontal bars. Sample sizes are indicated in parentheses. Percentage of cases with age over 20 years are indicated. **e**, Analytical workflow. The tumour/normal BAM files of WES data were analysed by our in-house pipeline followed by manual quality control. The mutation annotation format files generated by CGI were downloaded from TARGET Data Matrix (see Methods) and analysed by a pipeline developed for this dataset, including SNVs, indels and structural variants. CAN and LOH were analysed using read counts of germline SNPs in the mutation annotation format files. Manual quality control was also performed. For RNA-seq data, the FASTQ files were re-mapped and fusions and ITDs were analysed with CICERO. The resultant mutations were analysed by GRIN (SNVs, indels, CNAs, structural variants and fusions) and MutSigCV (SNVs and indels) to discover 142 recurrently mutated genes. **f**, One representative sample with chromothripsis for each histotype. CNAs are shown in the inner circle, orange indicates copy gain and blue indicates copy loss. Intra- and inter-chromosomal rearrangements are shown as green and purple curves, respectively.



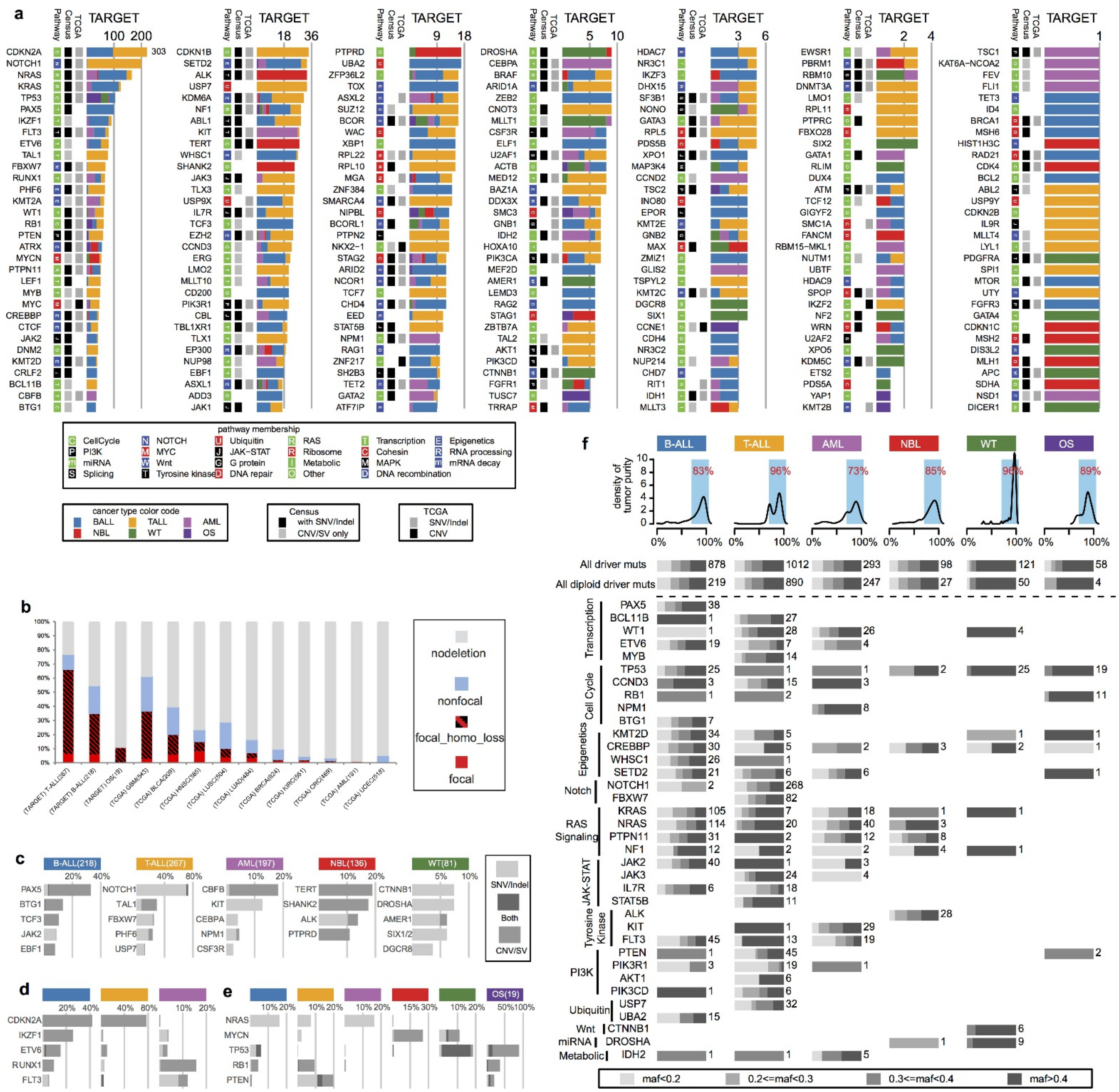
List of samples with UV signatures

| USI | Gender | Race | Age at Dx | ALL subtype |
|---|---|---|---|---|
| PAPLDL | Female | Hispanic | 7 | iAMP21 |
| PASFTL | Female | Caucasian | 4 | Hyperdiploid |
| PAPIJM | Male | Caucasian | 11 | Hyperdiploid |
| PAPDUF | Male | Caucasian | 3 | Hyperdiploid |
| PAPNMY | Male | Caucasian | 14 | iAMP21 |
| CAAABF | Male | Caucasian* | 3 | Hypodiploid |
| PANXDR | Male | Caucasian* | 3 | Hyperdiploid |
| PAPDWT | Female | Asian | 6 | Hyperdiploid |

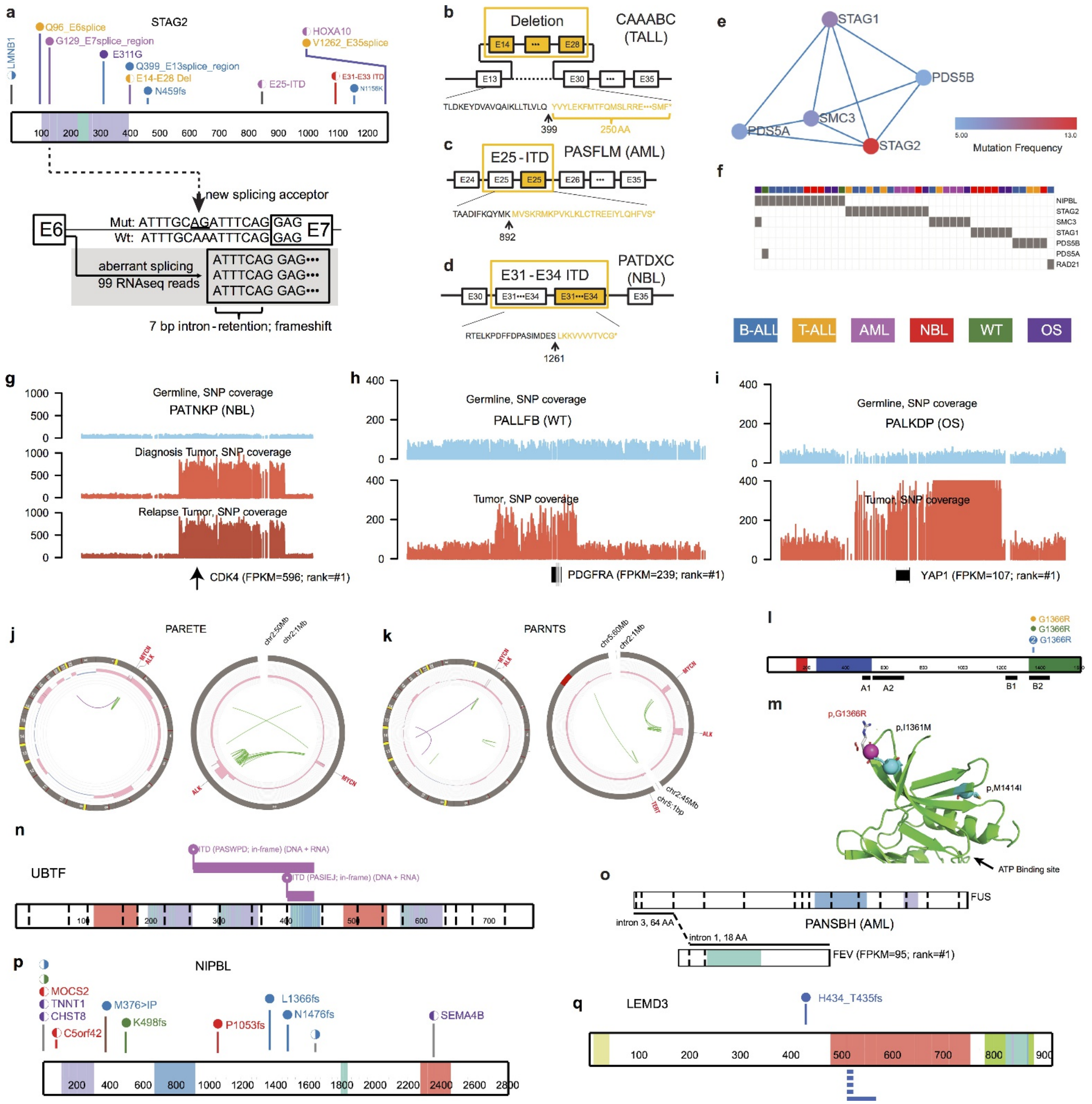
*: inferred from genomics data

**Extended Data Figure 2 |** See next page for caption.

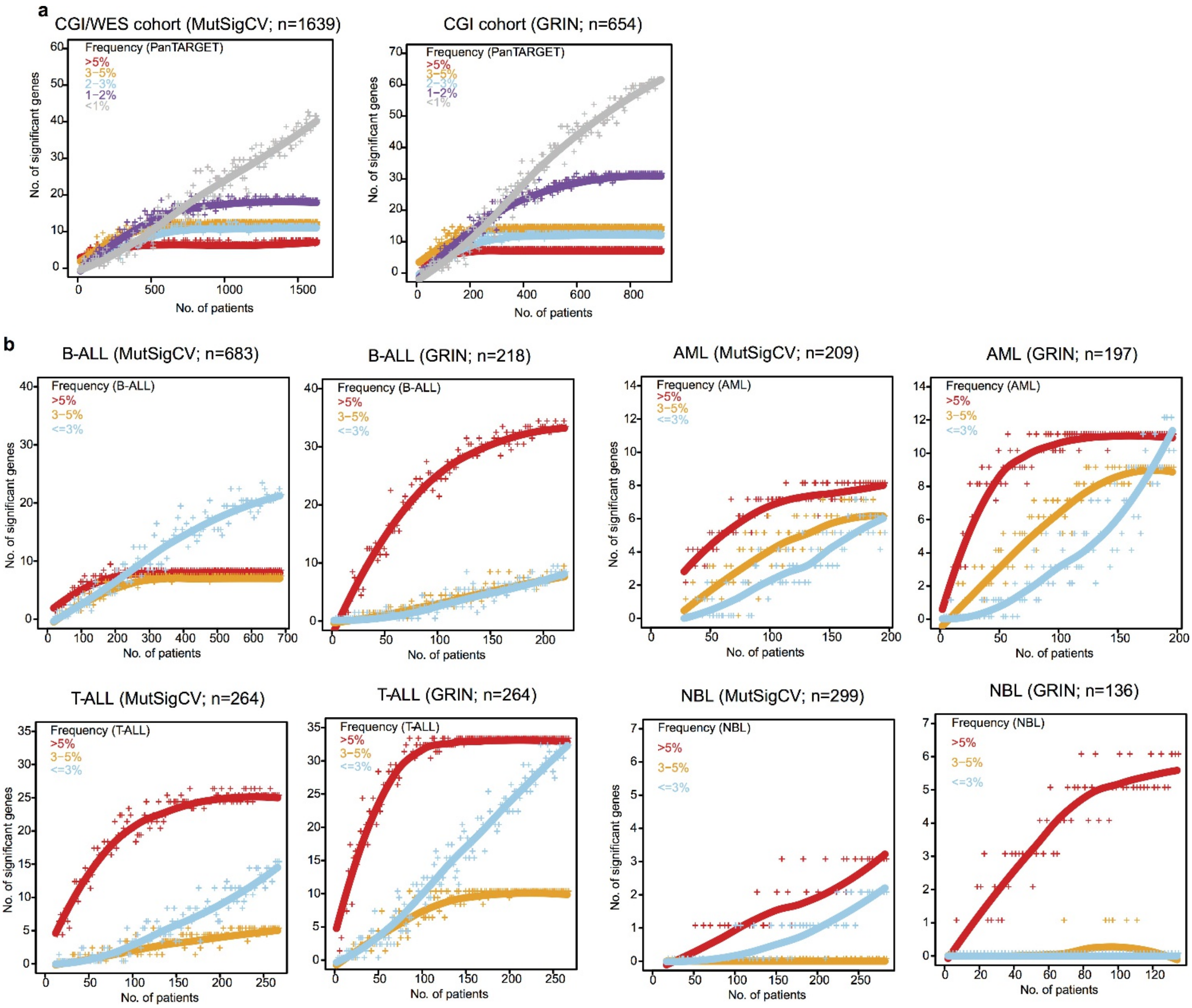
**Extended Data Figure 2 | Eight B-ALL samples with signatures of UV exposure.** **a**, List of samples with UV signatures detected. **b**, Inference of ethnicity for cases CAAABF and PANXDR from 654 TARGET CGI samples by principal component analysis (Supplementary Note 10). **c**, Total spectrum of mutational signatures of the eight UV mutation samples. **d**, SNVs of case CAAABF have a cross-validation rate of 90.4% with Illumina WGS data. **e**, High concordance of MAF values of SNVs derived from CGI and Illumina WGS, categorized by UV and non-UV mutations. Listed are Pearson's correlation coefficient (*r*) and *P* value derived from linear regression. Numbers of SNVs are indicated in parentheses. **f**, Inter-chromosomal distance and density plots for UV and non-UV mutations in case CAAABF. Top, inter-mutational distance ($\log_{10}$ scale) of UV (orange dots) and non-UV (black dots) mutations. Chromosomal level gain and loss statuses are indicated. The results indicate uniform distribution of mutations with or without UV signature across the genome. Middle and bottom panels show density plots of UV- and non-UV-mutations, respectively, categorized by chromosomal loss (red) and diploid (blue) status in corresponding tumour samples. Estimated cluster centres are indicated by corresponding colours. The expected MAFs for clonal mutations at given purity and chromosomal ploidy status of corresponding tumour are listed in the bottom panel. The density plots show that mutations with UV signatures are clonal after adjusting for ploidy. **g**, Inter-chromosomal distance and density plots for the other seven cases (key shown in **f**). **h**, ALL incidence by ethnicity obtained from the most recent registry (1973–2014) of SEER Program research data (Supplementary Note 11). **i**, Mutation spectrum for all SNVs (All) and for UV SNVs (T-5) for each of eight cases. Total number of SNVs and cosine similarity with COSMIC signature-7 are indicated in each panel.



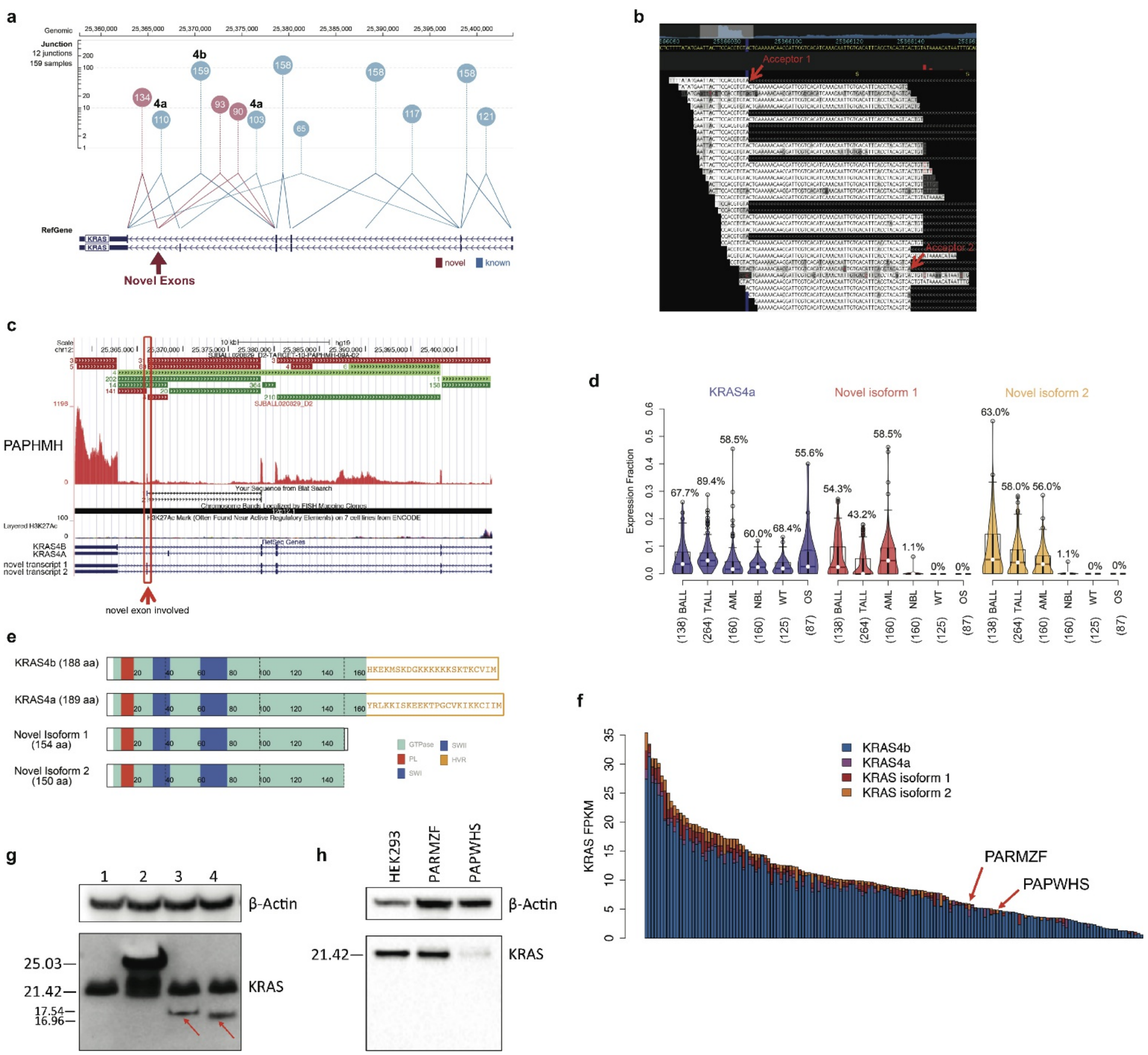
**Extended Data Figure 3 | Driver mutation landscape in paediatric cancers. a**, The number of samples mutated in each histotype is shown with colours coded as in Fig. 2. The presence of each gene in the Cancer Gene Census (Census) and prior pan-cancer studies of The Cancer Genome Atlas (TCGA) project are indicated. Pathway membership is also labelled for each gene. Somatic alterations in T-ALL were based on coding SNVs and indels from WES and CNAs from SNP array. **b**, Percentage of samples with focal (≤2 Mb) and non-focal (>2 Mb) deletions in *CDKN2A*. In the focal deletion category, samples with a second hit (either a second CNA or a copy neutral LOH) were categorized as 'focal_homo_loss'. For B-ALL, 27 of 218 (12%) non-focal samples had arm-level (such as hyperdiplod or hypodiploid B-ALL) CNAs on chromosome 9. Nine of 218 (4%) B-ALL cases had homozygous *CDKN2A* deletions with sizes from 2.1 Mb to 7.2 Mb and were counted as non-focal. TCGA data (no ALL data available) were downloaded in December 2015. The number of samples is indicated for each histotype. **c**, Top five genes mutated exclusively in each histotype. **d**, Top five genes mutated in leukaemias. **e**, Top five genes mutated in both leukaemias and solid tumours. **f**, MAF distribution of point mutations in driver genes. Top, density plot of tumour purity for each histotype. Percentages of samples with tumour purity >70% are indicated. Bottom, MAF distribution of point mutations in driver genes. Aggregated distribution for all driver genes is shown at the top ('All driver muts'), as well as all driver genes in diploid regions (for CGI data, CNA |seg.mean| < 0.2, |logRatio| < 0.2, and LOH seg.mean < 0.1; for T-ALL SNP array data, CNA |seg.mean| < 0.2). For each biological process defined in Fig. 3, the MAF distribution is shown for the genes with the five highest mutation frequencies that are mutated in more than five samples. The number of mutations in each histotype is shown.



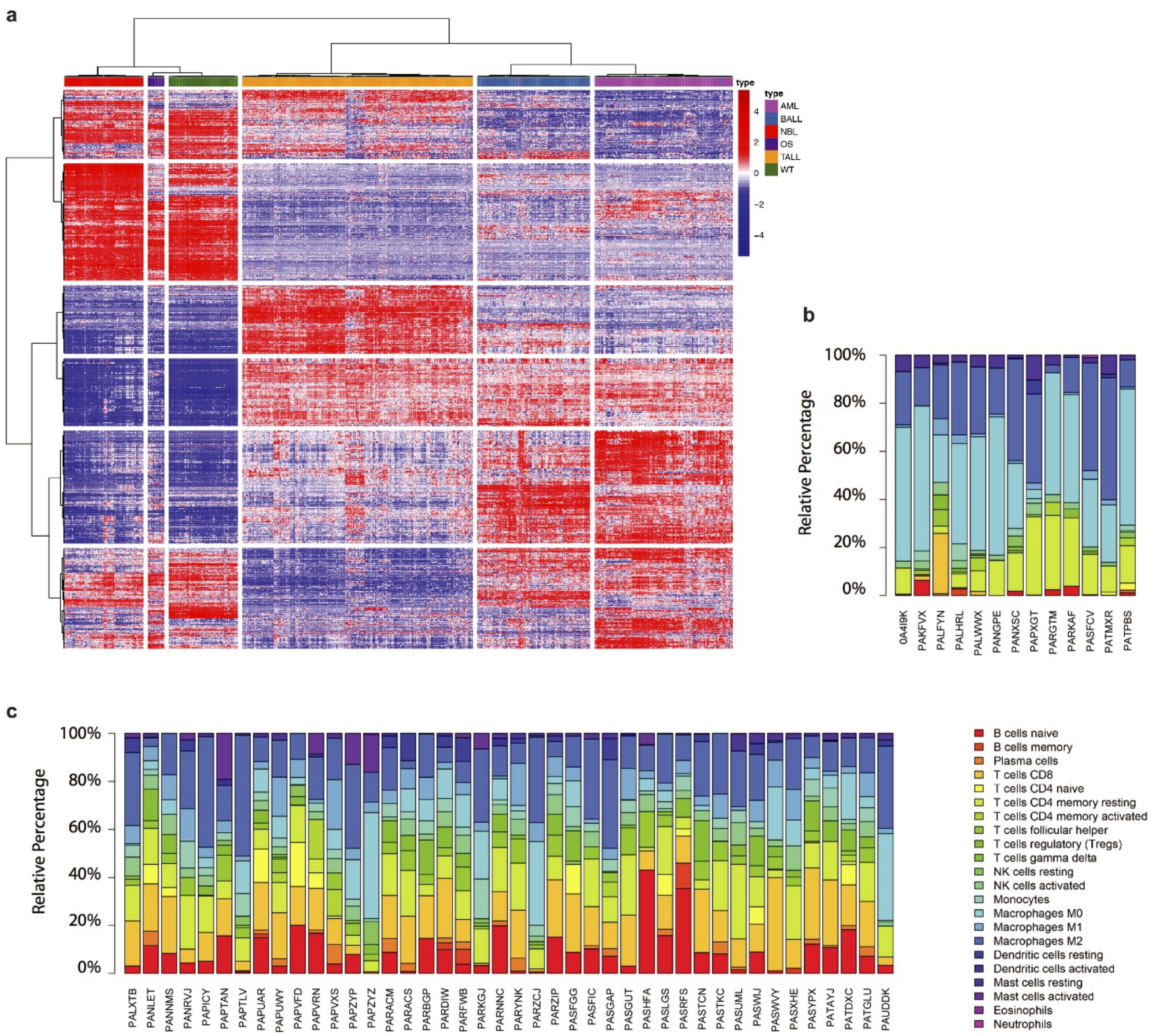
**Extended Data Figure 4 | Example driver mutations. a**, Diverse mutation types of *STAG2*. Variants are coloured by histotype as in Fig. 2. Circles and half-moons represent mutations and structural alterations, respectively. Bottom panel shows RNA-seq for an SNV at the −8 position of *STAG2* exon 7, which created a *de novo* splice site resulting in an out-of-frame transcript. **b**–**d**, Truncating mutations by deletion or ITD. **e**, Cohesin complex detected by HotNet2 analysis. **f**, Samples with mutations in cohesion complex. **g**–**k**, Selected examples of singleton oncogenic activation caused by high level amplifications including *CDK4* (**g**), *PDGFRA* (**h**), and *YAP1* (**i**) with FPKM and histotype-wise ranks indicated, as well as recurrent co-amplification of *MYCN-ALK* in two NBL samples (**j**, **k**). **l**, Recurrent *MAP3K4* mutation with structural model in N lobe (**m**). Location of the mutation p.G1366R is indicated by a magenta sphere and the alteration side chain is modelled as a stick. Known activating alterations (p.I1361M and p.M1415I) are shown as teal spheres. GADD45 binding (A1), kinase inhibitor (A2), and kinase domains (B1, B2) are indicated in **l**. **n**, ITD in *UBTF*. **o**, Fusion of *FEV*. **p**, **q**, Mutations in novel driver genes *NIPBL* and *LEMD3*.



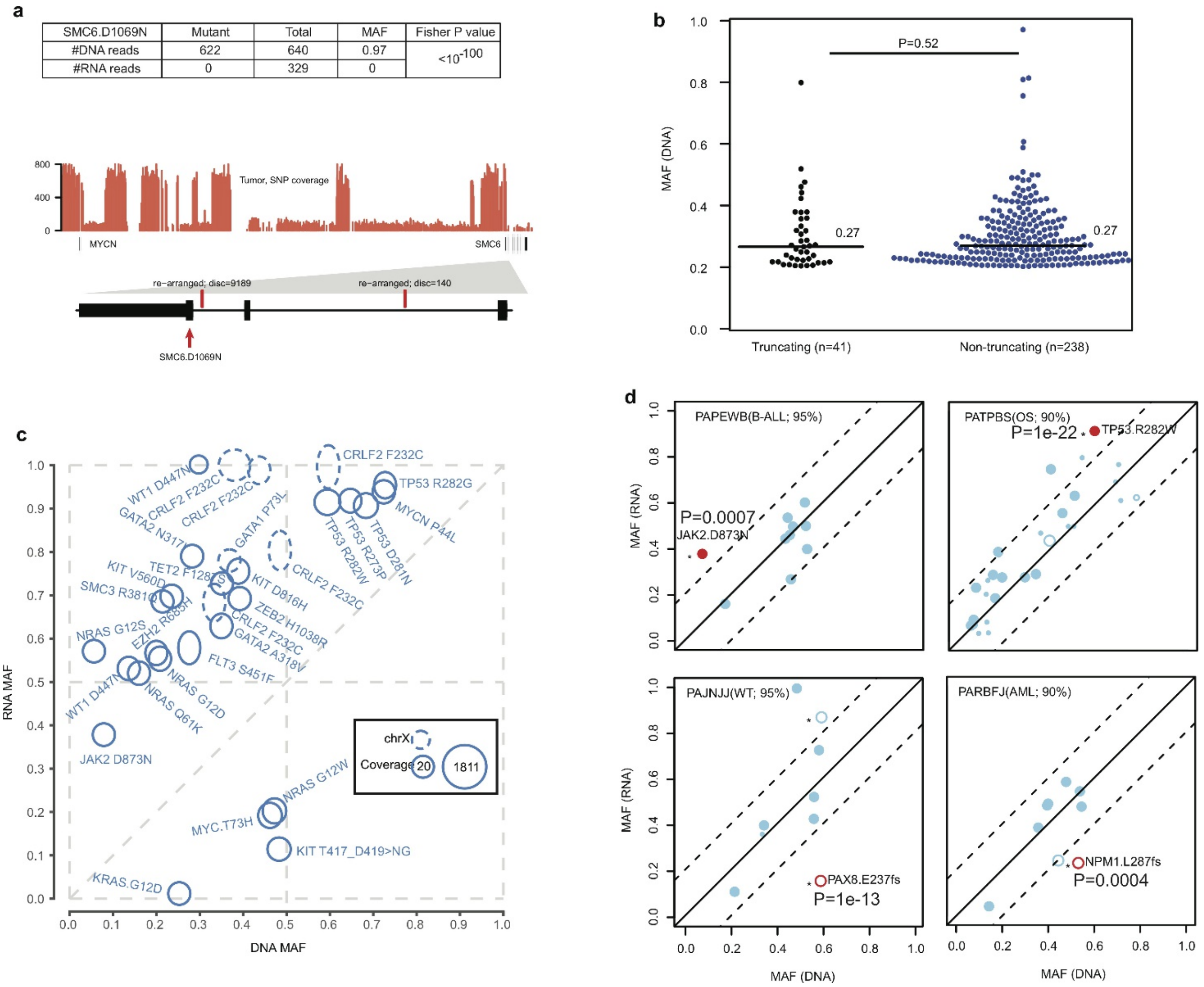
**Extended Data Figure 5 | Down-sampling analysis of gene discovery.** The analysis was performed on point mutations with MutSigCV and on SNVs, indels, structural variants, CNAs and fusion variants with GRIN (see Methods). The resulting candidate driver genes were categorized into five frequency bins indicated by different colours. Each point (+) represents a random subset of the pan-cancer cohort. The line is a smoothed fit. **a**, Analysis performed on entire CGI/WES cohort with MutSigCV (left) and CGI cohort with GRIN (right). **b**, Analysis performed with MutSigCV and GRIN for each histotype. Candidate driver genes were assigned to three frequency bins (according to corresponding histotypes). Sample sizes are indicated in parentheses in each panel.



**Extended Data Figure 6 | Expression of novel *KRAS* isoforms. a**, *KRAS* RNA-seq reads spanning splice junctions in AML samples. Each junction is shown as a circle labelled by counts of detected samples, with lines connecting the splice sites. The circle's *y*-axis position represents the median supporting read count. Canonical junctions are coloured blue and novel junctions red. **b**, RNA-seq reads in the last intron of *KRAS* illustrate the two novel exons detected in a B-ALL sample (PAPHMH). Novel splicing acceptor sites are indicated by red arrows. **c**, Junction reads for *KRAS* in the same B-ALL sample. Canonical *KRAS* exons are shown as green horizontal bars while novel exons are shown in red (top) and the RNA-seq coverage at the *KRAS* gene locus is shown below. The two novel exons are indicated with red arrows. **d**, Expression of two novel isoforms with *KRAS4a* as a control. Percentage of samples expressing these isoforms are indicated. Median, first and third quartiles are indicated by horizontal bars. Sample sizes are indicated in parentheses. **e**, Protein domains for *KRAS4a*, *KRAS4b* and two novel isoforms. **f**, *KRAS* expression (FPKM) in AML samples analysed in this study, categorized by the four isoforms. **g**, Western blot for *KRAS* in 293T cells. Cells were transfected with empty vector (lane 1), tagged wild-type *KRAS* (lane 2), novel isoform 1 (lane 3) and novel isoform 2 (lane 4). Protein products of the two novel *KRAS* isoforms are indicated by red arrows. **h**, Western blot for *KRAS* in two patient tumour samples (PARMZF and PAPWHS). Protein products of the two novel isoforms were not detected in these two samples. For **g** and **h**, the experiments were performed in duplicate and similar results were observed (see Supplementary Fig. 1 for gel source data).

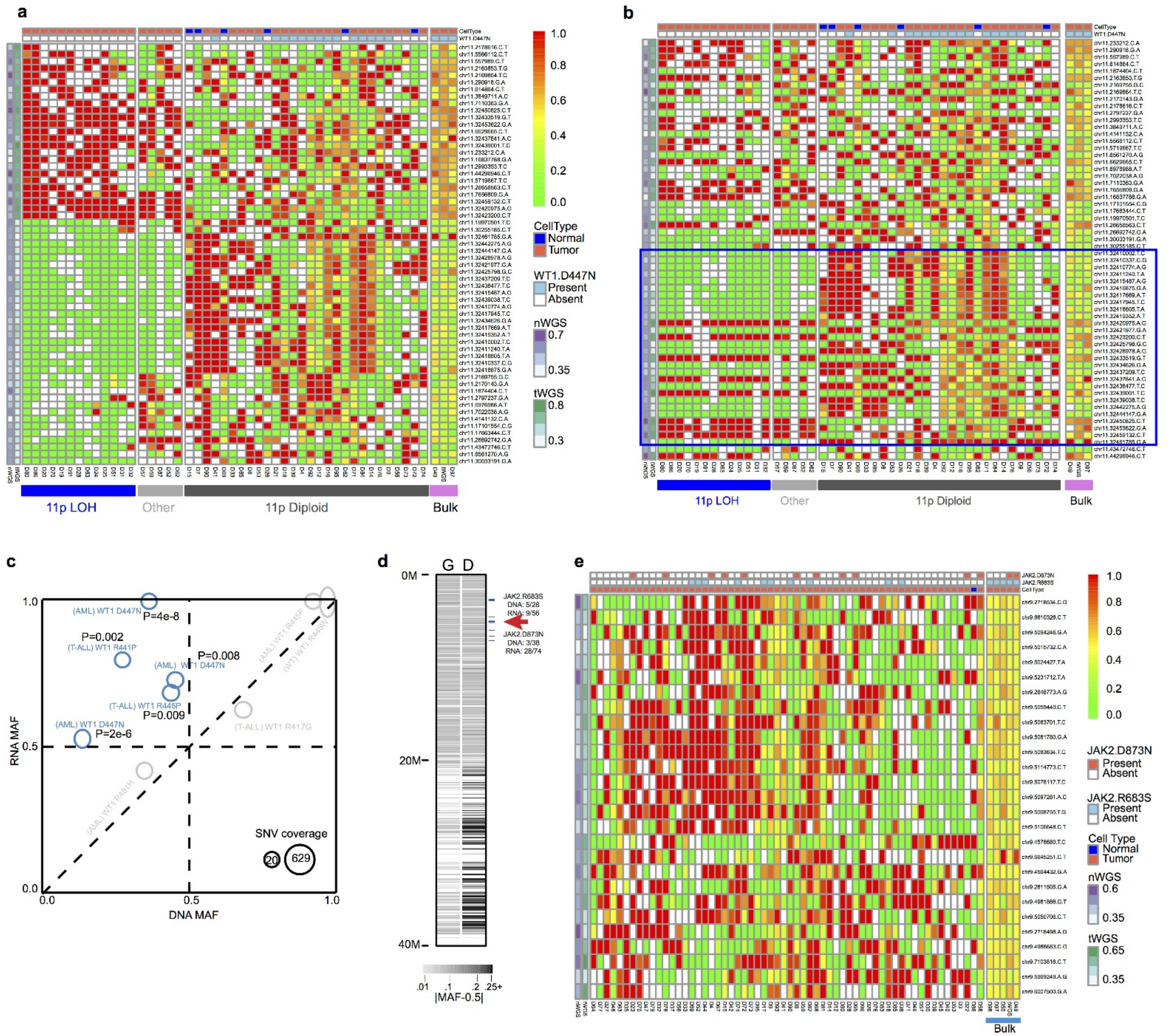


**Extended Data Figure 7 | Clustering analysis of tumour RNA-seq data and immune cell infiltration analysis. a**, Clustering analysis was carried out for 739 primary tumours with RNA-seq data available. The top 1,000 most variable expressed genes were clustered using Ward's minimum variance method. Each disease is annotated as shown in the first row with colour indicated in the key. **b**, **c**, Immune cell infiltration in osteosarcomas and NBL. Macrophage M0 and M2 were the dominant immune cell populations observed in osteosarcoma tumours (**b**). T and B cells, followed by macrophages, were the major immune cell types observed in NBL tumours (**c**).

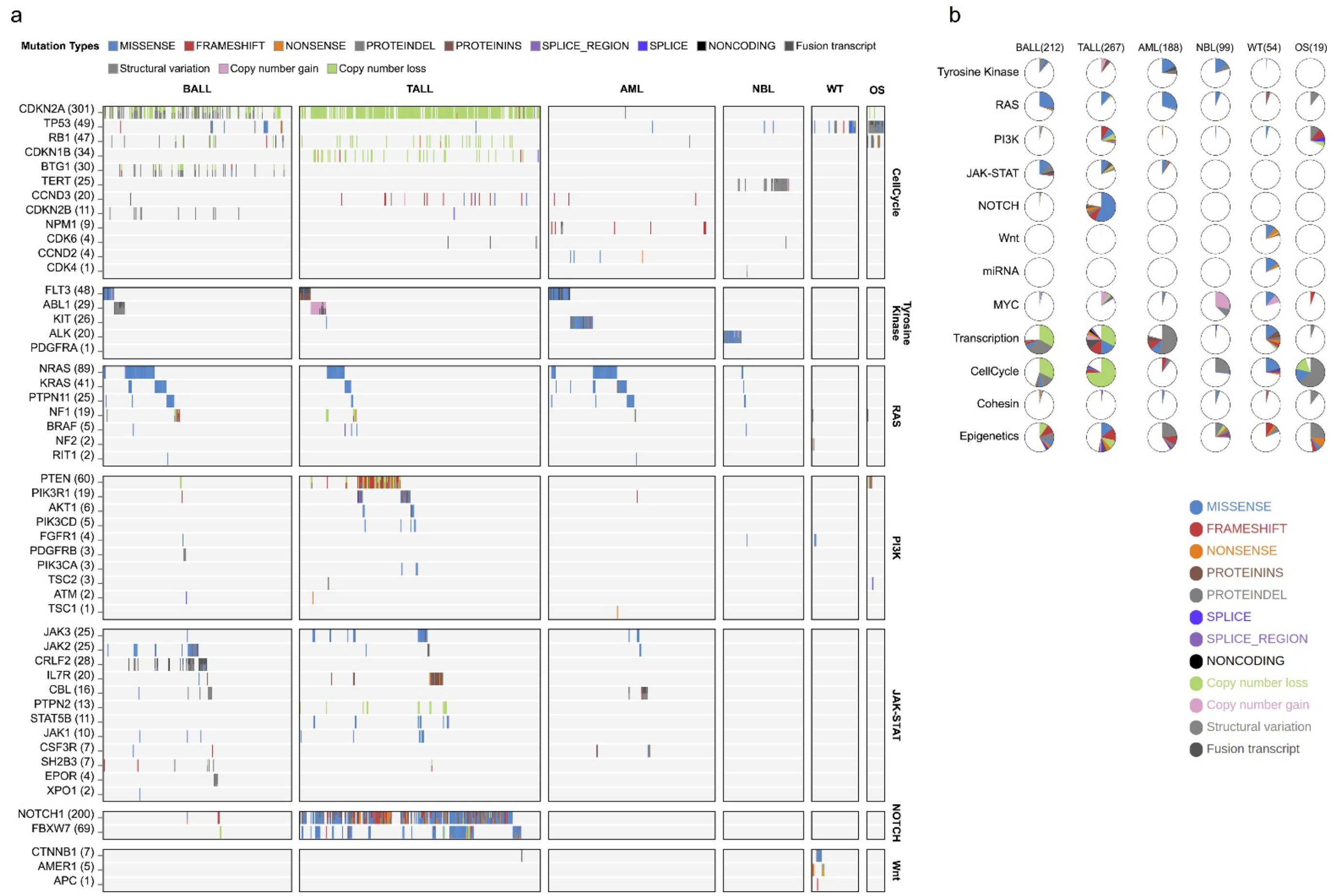


| SMC6.D1069N | Mutant | Total | MAF | Fisher P value |
|---|---|---|---|---|
| #DNA reads | 622 | 640 | 0.97 | $<10^{-100}$ |
| #RNA reads | 0 | 329 | 0 | |

**Extended Data Figure 8 | Analysis of allele-specific expression.** **a**, Mutant allele and total read count for *SMC6* D1069N in DNA and RNA of NBL case PAPZYP. This is to illustrate variants with suppressed mutant allele expression despite high DNA MAF and a high level of gene expression in RNA-seq. *P* value was calculated using two-sided Fisher's exact test. DNA coverage of the *MYCN* and *SMC6* region indicating multiple segments with high amplification (estimated at 26 copies). Details of the last three exons (E26, E27 and E28) of *SMC6* are shown with DNA structural variants highlighted by vertical red bars. The mutation *SMC6* D1069N is present in a region disrupted by structural variants, which dissociate the last three exons from the rest of *SMC6*. The high DNA MAF was therefore within a gene fragment that could not be transcribed and the expressed reference allele was from the intact gene. **b**, Non-expressed truncating (black) and non-truncating (blue) mutations showed a similar ($P = 0.52$, two-sided Wilcoxon rank-sum test) median MAF (horizontal black lines). The number of SNVs in each category is shown in parentheses. **c**, Hot spot mutations exhibited elevated mutant allele expression. Each mutation is shown as an oval positioned by its DNA MAF (*x* axis) and RNA MAF (*y* axis). The read count in DNA and RNA is depicted by the radius along the *x*-axis and *y*-axis direction, respectively. Mutations on chromosome X are shown as dotted ovals. Read counts from CGI and WES were combined whenever possible. **d**, Within-sample analysis to evaluate the effect of normal cell contamination on ASE. Shown are two samples with hotspot SNVs (red dots in cases PAPEWB and PATPBS) and two samples with truncating mutations (red circles in cases PAJNJJ and PARBFJ), which had a sufficient number of expressed coding mutations. The purity of each tumour is indicated. Dots represent SNVs and circles represent indels. Smaller symbols indicate the presence of CNA or LOH. An asterisk indicates a significant difference in MAFs between DNA (*x* axis) and RNA (*y* axis), which requires a minimum MAF difference of 0.2 (dashed lines) and a two-sided Fisher's exact test $P < 0.01$ (exact *P* values indicated in each panel). A dot in case PAJNJJ with DNA MAF of 0.5 and RNA MAF of 1.0 is not significant owing to low coverage (2×) in RNA-seq. In all four cases, within-sample concordance of DNA and RNA MAF for all except the ASE mutation suggest that normal cell contamination has a negligible effect on ASE.



**Extended Data Figure 9 | Allele-specific expression of *WT1* and *JAK2*.** **a**, **b**, Hierarchical clustering of single-cell sequencing data for AML case PAPWIU, in which rows were ordered by clustering (**a**) or by position (**b**). Each row represents one germline SNP and each column is a single cell. Three clusters (11p LOH, Other, and 11p diploid) were detected according to variant allele frequency, ranging from 0.0 (green) to 1.0 (red). The top two rows indicate the cell type (tumour or normal) and *WT1* D447N mutation status. **b**, Variants within the *WT1* locus are highlighted with a blue box. The cluster 'Other' matches the 11p LOH cluster within the *WT1* locus as the samples in this cluster had mono-allelic genotypes at *WT1*, probably caused by a focal deletion. The cluster 'Other' could also be caused by chimeric cells. However, as all cells in this cluster have the same pattern matching, the 11p LOH clusters within the *WT1* gene (the blue box in **b** represents the genomic location of chr11:32,410,002-32,461,785 and *WT1* is located at chr11:32,409,322-32,457,081). A *WT1* focal deletion better explains the profile in 'Other'. **c**, All nine missense *WT1* mutations with DNA and RNA data. The lowest RNA coverage is 16 for *WT1* R445P in AML case PABLDZ. Five mutations exhibiting allele-specific expression mutations (two-sided Fisher's exact test $P < 0.01$; exact $P$ values also listed for each mutation) are highlighted in blue (grey for $P \geq 0.01$). AML case PABLDZ (*WT1* R445P) had LOH at the *WT1* locus; LOH was present in the predominant clone at the diagnosis and may mask the presence of ASE in a subclone. **d**, **e**, Two *JAK2* mutations (R683S and D873N) were detected in B-ALL case PAPEWB, in which D873N showed ASE (DNA MAF is 3/38, RNA MAF is 28/74, Fisher's exact test $P < 0.01$). A single-cell sequencing experiment was designed to investigate whether the ASE could be attributed to subclonal CNA undetectable in the bulk tumour. **d**, The 27 germline SNPs in *JAK2* locus were selected along with the two somatic *JAK2* mutations and the other 46 somatic variants. **e**, Heat map of genotype clusters generated from the 64 assays (4 bulk and 60 single cells) passing single-cell sequencing quality control and the original CGI genotype data. The absence of a cluster of mono-allelic genotypes indicates the absence of 9p LOH, which in turn confirms ASE of D873N.



**Extended Data Figure 10 | Pathway-centric overview of mutational landscape in paediatric cancers. a**, Heat map of somatic mutations in selected pathways across six histotypes. **b**, Pie chart of mutation frequency in selected pathways. The number of samples in the calculation is indicated for each histotype. An interactive version of the data is available at the ProteinPaint portal (https://pecan.stjude.org/proteinpaint/study/pan-target).

nature research

Corresponding author(s): Jinghui Zhang

# Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study. For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

## ▸ Experimental design

1. Sample size

Describe how sample size was determined.

We analyzed all genome and transcriptome data collected from diagnosis tumor samples available under the accession number cited in the Online Methods section, except two cases: one is a duplicated neuroblastoma case, and the other is an osteosarcoma case which was collected from a patient age > 40 yrs.

2. Data exclusions

Describe any data exclusions.

Besides the two cases excluded from the analysis as described above, WES data from another 23 osteosarcoma samples were included only for determining driver mutation prevalence but not other analyses (see section "Whole exome data analysis" of the Online Methods for details).

3. Replication

Describe the measures taken to verify the reproducibility of the experimental findings.

The western blot for KRAS novel isoforms shown in Extended Data Figure 6g and 6h were done in duplicates, with similar observations.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Not relevant. Patient samples were grouped by disease types in this analysis.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

The investigators were not blinded to group allocation during the analysis. Patient samples were grouped by disease types. Detailed comparison was carried out among different type of tumors.

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

| n/a | Confirmed | |
|---|---|---|
| □ | ☒ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.) |
| □ | ☒ | A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| □ | ☒ | A statement indicating how many times each experiment was replicated |
| □ | ☒ | The statistical test(s) used and whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| □ | ☒ | A description of any assumptions or corrections, such as an adjustment for multiple comparisons |
| □ | ☒ | Test values indicating whether an effect is present<br>*Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.* |
| □ | ☒ | A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range) |
| ☒ | □ | Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation) |

*See the web collection on statistics for biologists for further resources and guidance.*

## Software

Policy information about availability of computer code

7. Software

| | |
|---|---|
| Describe the software used to analyze the data in this study. | Public available softwares: bwa (v0.7.12) for WES data alignment; Bambino for SNV/Indel analysis from WES; GRIN, MuSiC (v0.1), MutSigCV (v1.4) for driver mutation analysis; HTSeq for expression analysis from RNA-seq; R packages (v3.3.2) for statistical analysis; ESTIMATE (v1.0.13) and CIBERSORT for immune infiltration estimation; HotNet2 for pathway analysis; Circos (v0.69) for circular genome visualization; ProteinPaint portal for genomic data visualization.<br>In house tools: STRONGARM for RNA-seq data alignment; CICERO for fusion detection from RNA-seq; Medal Ceremony for mutation pathogenicity analysis; ad hoc perl scripts were used in filtering genomic variants from CGI dataset. |

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

## Materials and reagents

Policy information about availability of materials

8. Materials availability

| | |
|---|---|
| Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party. | No unique materials were used. |

9. Antibodies

| | |
|---|---|
| Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species). | Monoclonal antibody against human KRAS-N terminus (catalogue # H00003845-M02 Novusbio Littleton, CO), anti-beta-actin (catalogue # 4967, Cell Signaling Tech Danvers, MA), and anti-flag antibody (catalogue # TA50011, Origene Rockville, MD) were used for western blot. |

10. Eukaryotic cell lines

| | |
|---|---|
| a. State the source of each eukaryotic cell line used. | 293T cells purchased from ATCC (catalogue CRL-3216). |
| b. Describe the method of cell line authentication used. | The 293T cell line was purchased from ATCC (CRL-3216) and authenticated by ATCC. |
| c. Report whether the cell lines were tested for mycoplasma contamination. | We tested the mycoplasma contamination on 293T cells using MycoAlert kit (Lonza) and the result is negative. |
| d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use. | No commonly misidentified cell lines were used. |

## Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

### 11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.

No animals were used.

Policy information about studies involving human research participants

### 12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

Patients were recruited through collaborations with the Children's Oncology Group (COG) and Therapeutically Applicable Research to Generate Effective Treatments (TARGET) project. Diagnosis for these patients include B-ALL (N=689, age information missing for 6 cases, median age: 5.4 yr), T-ALL (N=267, age information missing for 3 cases, median age: 9.3 yr), AML (N=210, median age: 9.4 yr), NBL (N=316, median age: 3.1 yr), WT (N=128, median age: 4.2 yr) and OS (N=89, median age: 14.4 yr). Gender (N=8, female N=3, male N=5) and race (N=6, Hispanic N=1, Caucasion N=4, Asian N=1) information were included for the patients with UV signature in Extended Data Figure 2. Other clinical information was not used for this study.