doi:10.1038/nature25795

SUPPLEMENTARY NOTE For

Pan-cancer genome and transcriptome analyses of 1,699 pediatric leukemias and solid tumors

Supplementary Note 1. Verification of somatic point mutations

We first tested the process on 14 diagnosis-remission-relapse trio samples analyzed by both CGI and WES [1]. Of the 661 resulting CGI SNVs, 580 (88%) were verified by WES, while the indel verification rate was 67% (48/72). Notably, all 53 variants (45 SNVs and 8 indels) on driver genes identified in this study were cross-validated by WES.

We next used 104 primary tumor samples analyzed by both WES and CGI WGS to cross-validate CGI mutations. This analysis did not include the 14 tumor samples used earlier in evaluating our filtering algorithm. Among the 2,582 coding SNV/indels detected from CGI WGS data, 2,104 (82%) mutations were detected in the corresponding WES data. The WGS SNVs included 12 pairs of SNVs located 3 or 4bp apart in the same sample with none verified by WES, suggesting this a potential sequencing artifact. The overall specificity is slightly lower than the previously reported 88% specificity of CGI somatic SNV/indel detection [2]. In addition, 215 SNV/indels detected in WES were absent in CGI WGS resulting in a sensitivity of 92.3% (2,582 out of 2,797), higher than the previously reported sensitivity of 85%[2].

2,179 somatic point mutations (2,009 SNVs and 170 indels) identified in 190 B-ALL tumors were subjected to custom capture followed by sequencing on an Illumina HiSeq 2000, as previously described [1]. The average read-depth for both the tumor and the normal samples was 68X, and 85% of the mutations were verified.

For B-ALL case CAAABF, 90.4% SNVs were validated by Illumina WGS (SJHYPO123; same DNA specimen), with a high MAF concordance between platforms (**Extended Data Fig. 2e**).

Supplementary Note 2. Verification of structural rearrangements

Of the 1,011,810 putative CGI SVs, 3,265 passed these filters. Experimental verification using 14 CGI diagnosis-remission-relapse trio samples from a previous publication [1] showed a validation rate of 78%, with 79 out of the 101 SVs experimentally verified by targeted capture sequencing. We also compared our SV detection with translocations annotated in the clinical data file downloaded from the TARGET data matrix. For AML, the oncogenic fusion information was available for t(6;9), t(8;21), t(3;5)(q25;q34), t(6;11)(q27;q23), t(9;11)(p22;q23), t(10;11)(p11.2;q23), t(11:19)(q23:p13.1), and inv(16) for 98 cases. We correctly identified the corresponding fusions for 96 cases (98%). For B-ALL, the oncogenic fusion information was available for ETV6-RUNX1, MLL rearrangement, TCF3-PBX1, BCR-ABL1 and TCF3-HLF for 50 cases. We identified fusions for 46 cases (92%). Combining the two datasets resulted in an overall detection rate of 96%.

Supplementary Note 3. Verification of copy number alterations.

We compared the *MYCN* amplification status derived from CONSERTING with that of the original CGI analysis to evaluate the accuracy of the recalled CNAs. A subset of 32 NBL tumors carried a clinically-validated high-amplitude amplification of *MYCN*, a known oncogenic driver in pediatric neuroblastoma [3]. CGI's HMM CNA model only reported *MYCN* amplifications in 15 of these 32 tumors, while CONSERTING identified high-amplitude amplifications in 31 tumors. For the NBL with a negative CONSERTING finding (case PASJZC), a review of the initial diagnosis data indicated that this discrepancy could be explained by tumor heterogeneity and

tumor material sampling bias. Moreover, two additional subclonal *MYCN* amplification events were predicted in the remaining tumor samples (PARACM, PATHVK), demonstrating that CONSERTING achieved higher sensitivity than the original CGI analysis.

Supplementary Note 4. Selection of a candidate gene in a somatic alteration that affect multiple genes

We removed genes co-occurring with structural alterations affecting known oncogene or tumor suppressor genes (e.g. MTAP deletions were removed, as those are passengers of the CDKN2A deletion; STK11 deletions were removed, as those co-occur with the TCF3-PBX1 fusion), genes lacking support from gene expression (e.g., unexpressed genes disrupted by structural alterations), or genes in regions with frequent rearrangements due to the cell lineage (TCR/IGK/IGH/IGL regions in B-ALL). For recurrent gene fusions, a partner gene was not counted unless there was evidence for other types of somatic alterations (e.g., IGH locus, PBX1). To avoid double-counting, we only kept one gene to represent the following oncogenic fusions: *TCF3* for TCF3-PBX1 and TCF3-HLF; *CRLF2* for IGH-CRLF2 and P2RY8-CRLF2; *GLIS2* for CBFA2T3-GLIS2; *ZNF384* for TCF3-ZNF384 and EP300-ZNF384; *MEF2D* for MEF2D-BCL9; *NUP214* for DEK-NUP214; *ABL1* for BCR-ABL1; *KMT2A* for KMT2A rearrangements; *MLLT10* for PICALM-MLLT10; *CBFB* for CBFB-MYH11 and *RUNX1* for RUNX1-RUNX1T1. Input from disease experts within the TARGET team was also taken into account when finalizing the driver status of a candidate (e.g., *AUTS2* deletion which occurred in 9 NBL and 5 B-ALL samples was not considered a driver lesion).

Supplementary Note 5. *KRAS* novel isoform cloning and Transfection

To demonstrate the protein expression of two novel KRAS isoforms identified in leukemia (KRAS isoform 154 a.a. and 150 a.a.), clones were generated on a wild-type KRAS cDNA construct (Origene, Rockville, MD) by site-direct mutagenesis (Agilent, Santa Clara, CA). The following primer pairs were used 5′- gaaggcatcatcaacaccttacttccacgtgtactgtcttgtctttgctga -3′ and 5′- tcagcaaagacaagacagtacacgtggaagtaaggtgttgatgatgccttc -3′ for isoform 154 a.a., and 5′- aaggcatcatcaacacctcactgtcttgtctttgctg -3′ and 5′- cagcaaagacaagacagtgaggtgttgatgatgcctt -3′ for isoform 150 a.a.. All constructs were sequenced for verification. In all, 5×10^5 293T cells (ATCC, catalogue # CRL-3216, Manassas, VA; authenticated by ATCC, and test negative for mycoplasma contamination using MycoAlert kit (Lonza)) per well of a six-well plate were cultured in DMEM (Lonza, Walkersville, MD) with 10% of FBS (Sigma, Atlanta, GA). Two microgram of plasmid DNA was transfected into cells with X-tremeGENE HP DNA transfection reagent (Roche, Indianapolis, IN).

Supplementary Note 6. *KRAS* Western blot

293T cells were collected 72h after transfection and total protein extracted with RIPA lysis buffer. Frozen primary tumor samples were recovered and protein extracted with Triton-X lysis buffer. 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.

Supplementary Note 7. RNA-seq clustering

Cluster analysis was performed for 739 primary tumors analyzed by RNA-seq. After applying quantile normalization on the FPKM matrix, we excluded genes with low expression (FPKM<1)

in >70% samples. Median absolute deviation (MAD) was calculated for each gene across the cohort, and the top 1,000 genes were selected for cluster analysis using Ward's minimum variance method. Consequently, the samples were predominantly clustered by cell of origin (**Extended Data Fig. 7a**), consistent with previous observations [4].

Supplementary Note 8. RNA-seq analysis of immune cell infiltration

Expression-based infiltration analysis was applied to the three histotypes of solid tumors, NBL (n=90), WT (n=79) and OS (n=19). We first applied ESTIMATE[5] to evaluate if an immune cell infiltration signature (ImmuneScore) was present in each sample. For tumors with a positive ImmuneScore (42 NBL and 13 OS samples), we applied CIBERSORT[6] to derive the relative immune cell component. Notably, different immune cell representations were observed between NBL and OS (**Extended Data Fig. 7b**). In NBL, T and B cells were the major immune cell types, followed by macrophages, while in OS, macrophages M0 and M2 were the dominant immune cell populations. This indicates tumor-immune interactions may play different roles on different histotypes.

Supplementary Note 9. Within-sample comparison of DNA MAF and RNA MAF

Low tumor purity may confound ASE identification when comparing DNA and RNA MAF for a single variant. If ASE were an artifact of normal-in-tumor contamination, we would expect that all variants within the same sample to be classified as ASE variants. By contrast, bona-fide ASE is not expected to be a global profile as only a limited number of variants are expected to have ASE. To clarify these two possibilities, we performed within-sample comparison of DNA MAF and RNA MAF for samples having multiple expressed coding variants. Representative samples (**Extended Data Fig. 8d**) demonstrated that within the same sample, most coding variants have

highly concordant DNA and RNA MAF, indicating that the ASE variants in these samples are not an artifact of normal-in-tumor contamination.

Supplementary Note 10. Imputation of ethnicity for two leukemia patients

To genetically impute the unknown ethnicity of two patients (PANXDR and CAAABF) with mutation signature indicating UV-light exposure, we extracted genotypes for our internally curated ~240,000 SNP data set from WGS data available for 654 patients in the TARGET cohort. Principal component analysis was then carried out using the GLU software package (https://github.com/bioinformed/glu-genetics). Among patients with known ethnicity, the first eigenvector (EV1) separated Caucasian from African American populations. Both samples co-clustered with Caucasian patients (**Extended Data Fig. 2b**).

Supplementary Note 11. ALL incidence analysis

Age-specific rates of ALL by race/ethnicity were obtained from the most recent registry (1973-2014) of Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2014), National Cancer Institute, DCCPS, Surveillance Research Program, released April 2017, based on the November 2016 submission.

1 Ma, X. *et al.* Rise and fall of subclones from diagnosis to relapse in pediatric B-acute lymphoblastic leukaemia. *Nature communications* **6**, 6604, doi:10.1038/ncomms7604 (2015).
2 Molenaar, J. J. *et al.* Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature* **483**, 589-593, doi:10.1038/nature10910 (2012).
3 Brodeur, G. M., Seeger, R. C., Schwab, M., Varmus, H. E. & Bishop, J. M. Amplification of N-myc in untreated human neuroblastomas correlates with advanced disease stage. *Science* **224**, 1121-1124 (1984).
4 Hoadley, K. A. *et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929-944, doi:10.1016/j.cell.2014.06.049 (2014).

5 Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature communications* **4**, 2612, doi:10.1038/ncomms3612 (2013).
6 Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nature methods* **12**, 453-457, doi:10.1038/nmeth.3337 (2015).