# SJPedPanel: A pan-cancer gene panel for childhood malignancies

Pandurang Kolekar[1]*, Vidya Balagopal[2]*, Li Dong[1]*, Yanling Liu[1], Scott Foy[1], Quang Tran[1], Heather Mulder[1], Anna LW Huskey[2], Emily Plyler[1], Zhikai Liang[1], Jingqun Ma[2], Joy Nakitandwe[3], Jiali Gu[2], Maria Namwanje[2], Jamie Maciaszek[2], Debbie Payne-Turner[2], Saradhi Mallampati[2], Lu Wang[2], John Easton[1#], Jeffery M. Klco[2#], Xiaotu Ma[1#]

[1]: Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN, USA
[2]: Department of Pathology, St. Jude Children's Research Hospital, Memphis, TN, USA
[3]: Department of Pathology and Laboratory Medicine, Diagnostics Institute, Cleveland Clinic, Cleveland, OH, USA

*: Contributed equally

[#]: correspondence authors
XM (Xiaotu.Ma@stjude.org); JMK (Jeffery.Klco@stjude.org); JE (John.Easton@stjude.org)

## Abstract

### Background

Large scale genomics projects have identified driver alterations for most childhood cancers that provide reliable biomarkers for clinical diagnosis and disease monitoring using targeted sequencing. However, there is lack of a comprehensive panel that matches the list of known driver genes. Here we fill this gap by developing SJPedPanel for childhood cancers.

### Results

SJPedPanel covers 5,275 coding exons of 357 driver genes, 297 introns frequently involved in rearrangements that generate fusion oncoproteins, commonly amplified/deleted regions (e.g., *MYCN* for neuroblastoma, *CDKN2A* and *PAX5* for B-/T-ALL, and *SMARCB1* for AT/RT), and 7,590 polymorphism sites for interrogating tumors with aneuploidy, such as hyperdiploid and hypodiploid B-ALL or 17q gain neuroblastoma. We used driver alterations reported from an established real-time clinical genomics cohort (n=253) to validate this gene panel. Among the 485 pathogenic variants reported, our panel covered 417 variants (86%). For 90 rearrangements responsible for oncogenic fusions, our panel covered 74 events (82%). We re-sequenced 113 previously characterized clinical specimens at an average depth of 2,500X using SJPedPanel and recovered 354 (91%) of the 389 reported pathogenic variants. We then investigated the power of this panel in detecting mutations from specimens with low tumor purity (as low as 0.1%) using cell line-based dilution experiments and discovered that this gene panel enabled us to detect ~80% variants with allele fraction of 0.2%, while the detection rate decreases to ~50% when the allele fraction is 0.1%. We finally demonstrate its utility in disease monitoring on clinical specimens collected from AML patients in morphologic remission.

### Conclusions

SJPedPanel enables the detection of clinically relevant genetic alterations including rearrangements responsible for subtype-defining fusions for childhood cancers by

targeted sequencing of ~0.15% of human genome. It will enhance the analysis of specimens with low tumor burdens for cancer monitoring and early detection.

**Keywords**

Pediatric Cancer, Gene Panel, Next Generation DNA sequencing, Genetic Testing, Molecular Diagnostics, Cancer Early Detection and Monitoring

## Background

Extensive insights on the genetic underpinnings (i.e., driver alterations) of childhood cancers [1-3] have been uncovered in the past decade using next generation sequencing. To date, diagnostic sequencing has become part of clinical service in some institutions[4-6]. Although whole genome, and to lesser extent whole exome sequencing, is preferrable to maximize the detection of cancer-associated variants in the clinical setting, there are significant resource and infrastructure requirements for these modalities that are not amenable to the majority of clinical labs. Further, the broad coverage of whole genome and whole exome sequencing renders it challenging to achieve ultra-deep sequencing that is essential for the analysis of specimens with low tumor purity such as for detecting minimal residual disease and for disease monitoring. Targeted gene panel-based sequencing holds the promise to address these challenges.

Although there have been multiple gene panels designed for adult cancers, such as MSK-IMPACT[6], currently there is no comprehensive gene panel for pediatric cancers. This is important considering the recent pan-cancer study of 1,699 childhood cancers that demonstrated a dramatic difference between adult and childhood cancers, where 55% of the 142 driver genes in pediatric cancers are not found in adult pan-cancer studies [3]. In this study, 62% of driver alterations in childhood cancers are copy number alterations (CNVs) or structural variations (SVs) whose boundaries typically do not fall into protein coding regions. Indeed, our recent study of oncogenic fusions[7] indicated that 55.7%, 22.5%, and 18.5% of pediatric leukemia, brain, and solid tumors have subtypes defined by oncogenic fusions, for which the DNA breakpoints typically fall into intronic regions. These facts render base pair level ascertainment of driver alterations in childhood cancers challenging by using conventional capture sequencing kits such as exome sequencing and call for a dedicated gene panel for pediatric cancers that includes both coding and non-coding targets to maximize the detection of key alterations in pediatric cancers.

Here, we highlight the prominent features of our pan-cancer gene panel (termed SJPedPanel) for childhood cancers by comparison with five existing cancer gene panels. We validate its superior coverage of genes relevant to childhood cancers

using a well-described real-time clinical cohort via *in silico* analysis, followed by re-sequencing a subset of these cases for experimental validation. We also demonstrate the power of our gene panel in detecting rare variants using ultra-deep sequencing via serial dilution experiments, as well as disease monitoring in remission samples from acute myeloid leukemia (AML) patients.

## Results

### Panel design

We designed our SJPedPanel by integrating findings from 44 published tumor-normal paired genomics studies of childhood cancers that spans leukemia, brain, and solid tumor [2-4,8-48]. SJPedPanel includes 1.069 million exonic base pairs from 5,275 coding exons for detecting protein coding mutations in 357 known driver genes for childhood cancers [**Fig. 1a**; see Additional file 1: Table ST1, and Additional file 2 for plots of chromosomal distribution of genes generated using an R/Bioconductor package chromPlot [49]]. To account for the structural variations (SVs) that result in subtype-defining oncogenic fusions for which DNA breakpoints typically fall in intronic regions [7], 1.438 million base pairs from 297 introns of 94 genes [Additional file 1: Table ST1] were included. Moreover, 0.209 million bases from promoter regions were targeted for detecting promoter alterations including rearrangements and point mutations such as *TAL1* in T-ALL[50] [**Fig. 1b**]. Highly recurrent oncogenes (*MYCN* in neuroblastoma[3]) and tumor suppressor genes (such as *CDKN2A*, *PAX5*, and *SMARCB1*) were targeted by probes tiling the entire gene region for detecting CNVs. To account for the fact that breakpoints of structural alterations can fall outside gene regions, we extended the target regions to frequent DNA breakpoints by using patient data from ProteinPaint [51] and GenomePaint [52]. Collectively, 2.82 million base pairs were designed for potential SNV, Indel, SV and CNV/LOH driver alterations. Notably, a few known childhood cancer drivers are intentionally excluded due to genomic space considerations. For example, *MECOM* [53] and *GFI1B* [54] are known to be involved in enhancer-hijacking alterations and were excluded due to the large space needed to cover the many possible breakpoints.

In addition, 7,590 SNPs were selected for detecting copy number variations and loss of heterozygosity (CNV/LOH) across the genome [see Additional file 1: Table ST2].

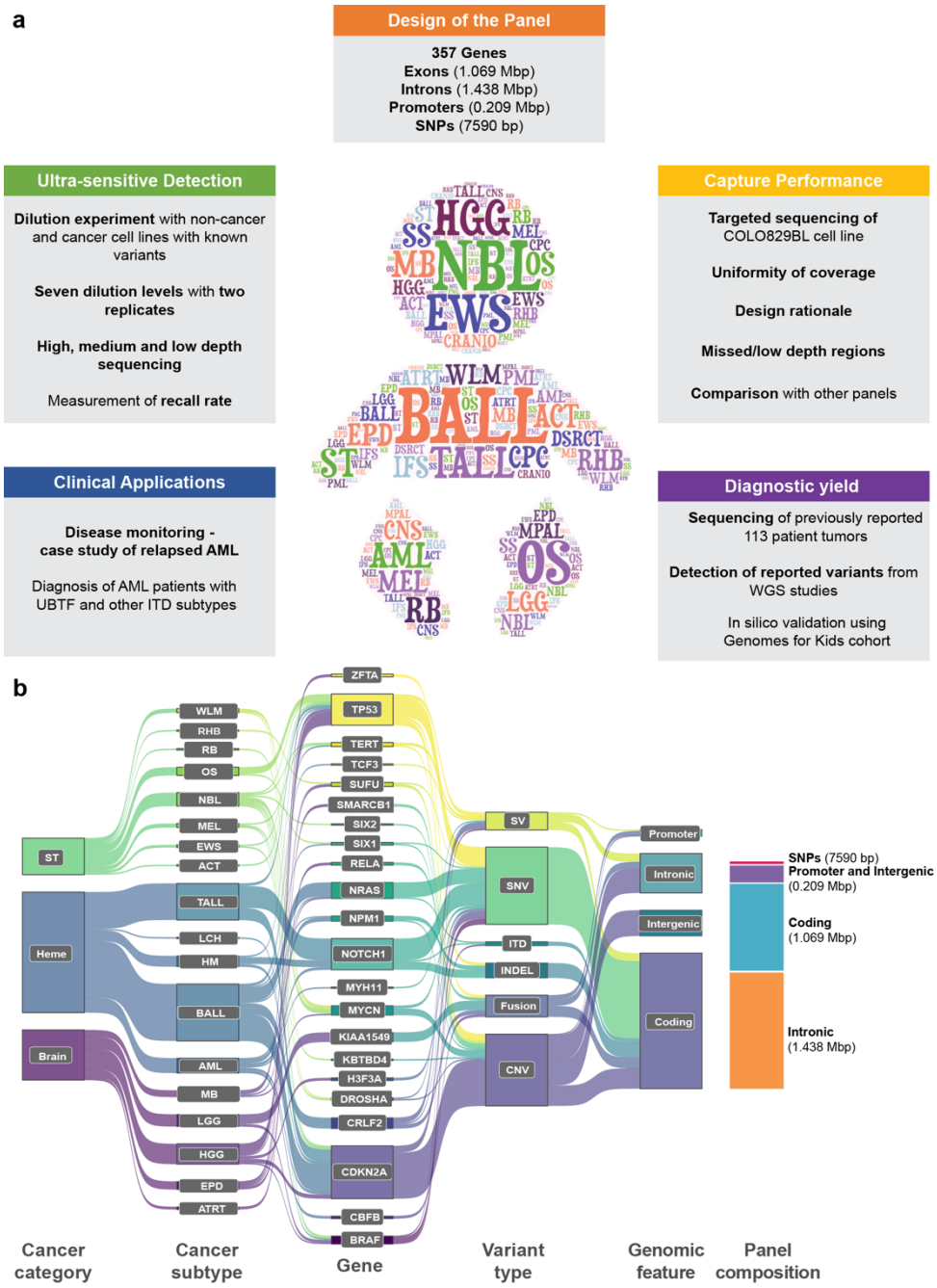


**Fig. 1. Design of pediatric cancer gene panel.** (a) This study includes panel content, investigation of ultra-sensitive detection, capture performance, diagnostic yield, and clinical applications. (b) A Sankey diagram [55] showing spectrum of childhood cancers (Heme: hematological malignancies; ST: solid tumors; Brain: brain tumors), cancer subtypes, genes, variant types, and genomic features targeted by SJPedPanel. Stacked bar plot at the right end shows space distribution of different genomic features covered by SJPedPanel.

The median distance between these SNPs is 332 Kb, with 25th and 75th quantile being 60 Kb and 593 Kb, respectively [see Additional file 3: Fig. S1a]. Notably, >80% of these SNPs have population frequency between 40% and 60% [see Additional file 3: Fig. S1b], which ensures that that nearly 50% of patients are heterozygous at each SNP site. Therefore, around 3,000 (=7590×0.5×0.8) heterozygous SNPs are expected for each patient, which leads to a theoretical resolution of ~1 Mb for CNV/LOH detection. The number of SNPs chosen per chromosome is roughly proportional to the lengths of chromosomes [see Additional file 3: Figs. S1c-d, and Additional file 4 for plots of chromosomal distribution of SNPs generated using an R/Bioconductor package IdeoViz [56]]. Considering the read length and the insert length (for target capture and sequencing), these 7,590 SNPs actually occupy ~250*7590=1.8975 million base pairs. Thus, our panel consisted of ~4.7 million base pairs, or ~0.15% of the human genome. The compact size of our panel enables us to reach 30,000X at the cost of a standard WGS (~30X) per sample, thus enabling cost-effective cancer monitoring and/or early detection [**Fig. 1a**]. The gene panel was manufactured by Twist Bioscience.

**Comparison of gene content with other panels**

We first compared the gene content between our panel and five other commonly used commercial panels for childhood cancers, including FoundationOne Heme, FoundationOne CDx [57,58], MSK-IMPACT [6,59], OncoKids [5], and Oncomine Comprehensive assay v3 (OCAv3) [60] [Additional file 1: Tables ST3a and ST3b]. We used the list of 183 driver genes reported in two recent childhood pan-cancer studies [2,3] involving 2,578 cases. As seen in **Table 1**, SJPedPanel covers 159 (87%) genes whereas all other panels covered <60% of the reported pediatric cancer driver genes [**Fig. 2a** and Additional file 1: Table ST4].

A comparison of gene names among the panels indicated that SJPedPanel has unique coverage of 110 genes [**Fig. 2b**, Additional file 1: Table ST3b] when compared to the other panels combined, such as *DGCR8* and *SIX1* for Wilms tumor [61], *SHH* for medulloblastoma [62], *ZFTA* for ependymoma [22,63], *UBTF*-TD [64] and *PICALM* [64,65] for AML. Among all the panels, SJPedPanel provides the largest

intronic regions (297 introns from 94 genes; **Table 1**) responsible for rearrangements that generate fusion oncoproteins.

**Table 1.** Overview of the panels selected for comparison.

| Panel name | Focus area | No. of genes | Genes screened for | | | SNPs | %Pediatric cancer driver genes covered out of reported 183 genes† | Publication |
|---|---|---|---|---|---|---|---|---|
| | | | Coding exons | Selected introns | Promoter regions | | | |
| SJPedPanel | Pediatric cancers | 357 | 357 | 94 | 10 | 7590 | 87% | This study |
| FoundationOne® Heme | Adult blood cancers | 418 | 408 | 31 | | | 57% | He et al., 2016 [66] |
| MSK-IMPACT® | Adult solid tumors | 468¶ | 468 | 13 | 1 | 862 | 52% | Cheng et al., 2015 [6] |
| OncoKids® | Pediatric cancers | 181 | 137 | 44 | 1 | | 49% | Hiemenz et al., 2018 [5] |
| FoundationOne® CDx | Adult solid tumors | 324 | 309 | 34 | 1 | | 44% | Whitepaper by Company§ [58] |
| Oncomine™ Comprehensive Assay v3 | Adult cancers | 161 | 146 | 15 | 1 | | 29% | Hovelson et al., 2015 [67] |

†: 183 pediatric cancer genes are reported in two pediatric pan-cancer studies, Ma et al., 2018 [3] and Grobner et al., 2018 [2]
¶: MSK-IMPACT panel has been reported to consist of 468 genes considering 2 different transcript isoforms for CDKN2A gene, however based on unique gene names MSK-IMPACT panel consists of 467 genes.
§: Sources of content for all the panels are available in Additional file 1: Table ST3a

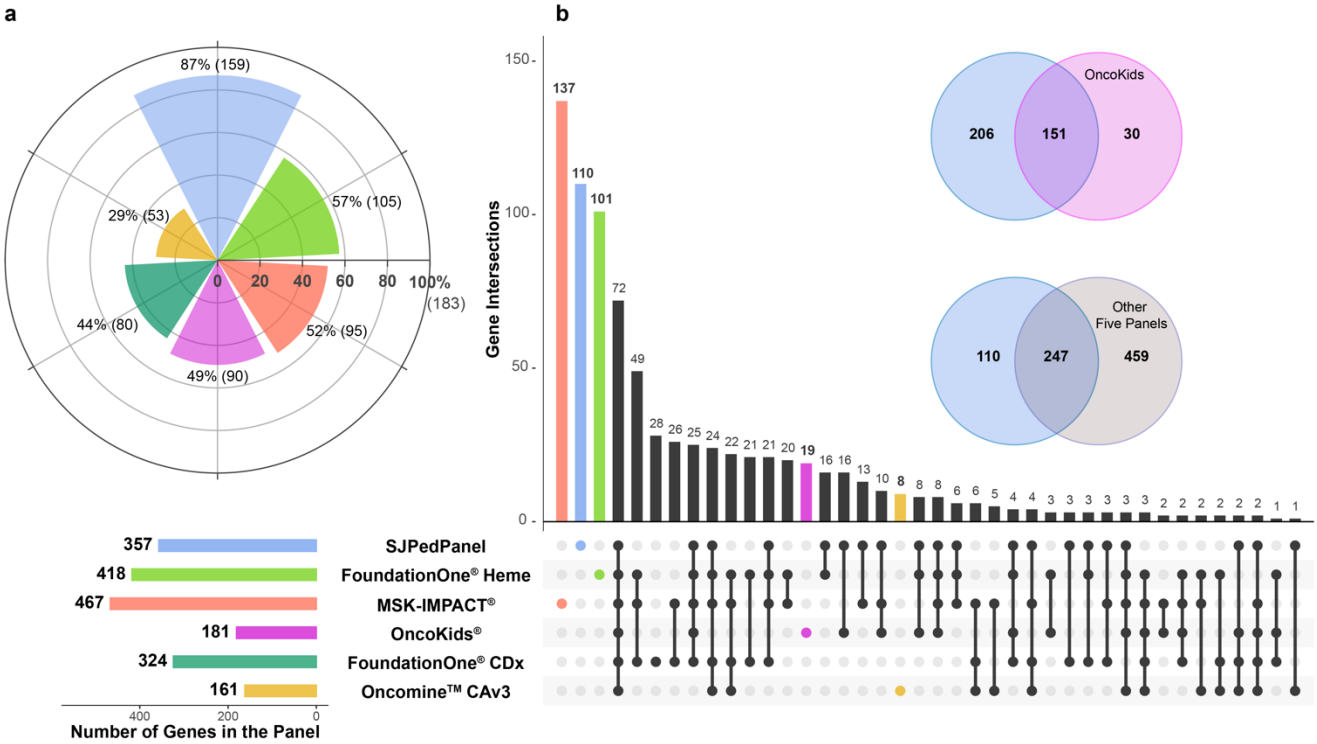


**Fig. 2. Comparison of gene content between SJPedPanel with other panels**. **(a)** Pediatric cancer relevance of 6 panels based on coverage of 183 driver genes from childhood pan-cancer studies [2,3]. The horizontal bars at the bottom indicate numbers of genes designed in each panel coded by corresponding color. **(b)** Analysis of common and unique genes among 6 panels using UpSet plot [68]. Venn diagrams indicate intersection of SJPedPanel genes with OncoKids, and other fiver panels combined.

On the other hand, among the 459 genes specific to other panels, only three genes were reported in two recent pediatric pan-cancer studies with low patient frequencies (*ZNF217: 0.59%, PCBP1: 0.31%, and CARD11: 0.21%*) [2,3]. MSK-IMPACT panel has the maximum number of genes (467), of which 137 are exclusive from other panels [6,59]. Most of these genes are relevant to adult cancers with the highest concentration in adult solid tumors [6,59]. Similarly, FoundationOne Heme panel consists of 418 genes with a focus on adult hematological malignancies [57,66]. OncoKids is the only other pediatric cancer panel under comparison and covers 30 genes that are not included in our panel [see **Fig. 2b** and Additional file 1: Table ST3b]. *MECOM* is excluded due to large space needed for the diverse promoter hijacking events [53]. *CALR, RARA* and *SS18* were not included due to an overall paucity of alterations in these genes in current pediatric genome cohorts. Collectively, SJPedPanel offers by far the most comprehensive coverage of genetic alterations for the study of childhood malignancies.

**Capture performance of the panel**

A critical consideration in genomic sequencing (especially in panel sequencing) is the coverage uniformity. To study this question, we sequenced four targeted sequencing libraries (C1- C4) prepared using COLO829BL (ATCC #CRL-1980), a non-cancer cell line that has been extensively used in literature for clinical proficiency testing or benchmarking [69,70]. To ensure reproducibility, technical replicates were generated to achieve high (C1, C2, ~2000X) and low depth (C3, C4, ~200X) of sequencing. Libraries in each set were sequenced with either the Illumina NovaSeq 6000 (C1, C3) or NextSeq 500 (C2, C4).

As expected, the average depth was highly correlated ($r^2$: 0.98) with the number of raw reads sequenced [Additional file 3: Supplementary Fig. S2a]. With this data, we investigated the capture uniformity at base pair level [**Fig. 3a-b**] and at region level [**Fig. 3c-d**]. Because highly uniform capture data would ensure most bases/regions to have similar depth (therefore a histogram with very small standard deviation), we choose to use coefficient of variation ($C_V$, defined by $\sigma / \mu$ of the histogram) to measure sequencing uniformity. Here $\sigma$ and $\mu$ are the estimate of standard deviation and mean, respectively, by trimming 2.5% of extreme values from both ends of the

histograms [**Fig. 3**]. At base pair level, we observed that $C_V$ is close to 0.35 for libraries sequenced by NovaSeq, and to be between 0.37 and 0.38 for libraries sequenced by NextSeq. At region level, NovaSeq data has $C_V$ close to 0.25, while NextSeq data has *Cv* range from 0.22 to 0.28.

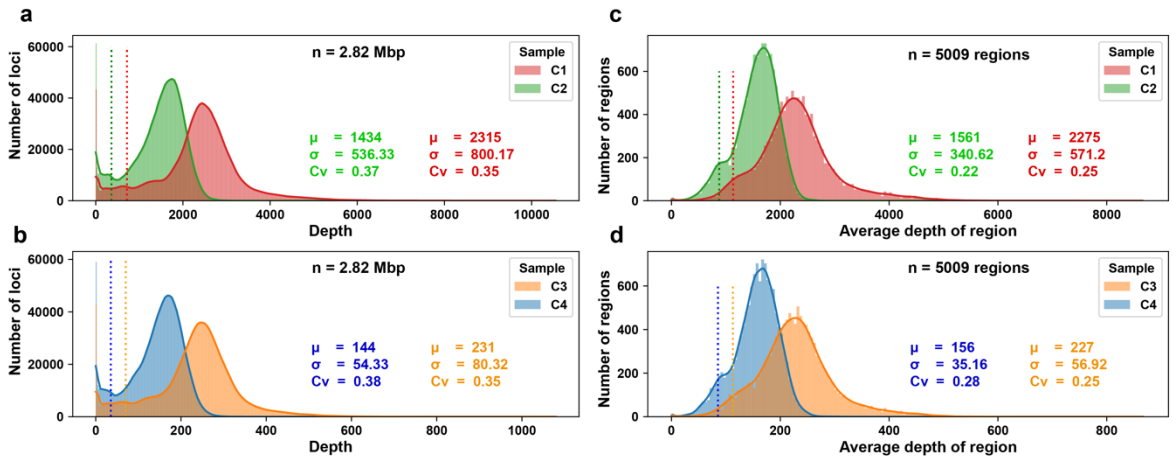


**Fig. 3. Capture uniformity per base (a, b) and per region (c, d) of the panel.** Uniformity of coverage across in the SJPedPanel for high depth samples sequenced on Illumina NovaSeq 6000 (C1) and NextSeq 500 (C2) respectively; and low depth samples sequenced on Illumina NovaSeq 6000 (C3) and NextSeq 500 (C4) respectively. The histograms are made at base pair level (2.82 Mbp) **(a, b)** and region level (n=5009) **(c, d)**. The vertical dotted lines indicate ($\mu$ -2$\sigma$) of the respective distributions. The statistical parameters ($\mu$: average depth, $\sigma$: standard deviation, $C_v$: coefficient of variance) were calculated by trimming observations in upper and lower 2.5 percentiles. All the sample and region level QC parameters are available in the Additional file 1: Tables ST5 and ST6.

Overall, the standard deviation is less than or around 1/3 of the mean, which ensures that most of the target bases/regions are sufficiently covered. Using the two-sigma rule (that approximates the 95% confidence interval), we also measured the percentage of bases/regions with depth higher than ($\mu$ - 2$\sigma$). We found that 97% and 95% of bases have depth higher than this threshold for NovaSeq data and NextSeq data, respectively [Additional file 1: Table ST5 and Additional file 3: Fig. S2b]. Similar trends were observed from the region level analyses [**Figs. 3c-d** and Additional file 1: Table ST6]. These data indicated that SJPedPanel has satisfactory capture efficiency that is reproducible over different sequencing platforms.

We next analyzed regions that are consistently poorly covered (i.e., less than $\mu$ - $3\sigma$) in the COLO829 data. We identified 27 regions (26 regions are small exons), of which 10 regions consistently showed no coverage across all the four samples [see column “Remark” in Additional file 1: Table ST6]. These 27 regions occupy 8091 bp (~0.3%) of the panel and more than half of these bases (4972 bp) belong to only two regions, *NUTM2A* (3152 bp including intron 1 with 2050 bp) and *DUX4* (1820 bp including exon 1) [see Additional file 1: Table ST7 and Additional file 3: Supplementary Fig. S3]. For *STAG2* the coverage is slightly below the pre-defined cut-offs for 3 regions (351 bp) (e.g. ~ 500X for C1 sample, Additional file 1: Table ST7). While looking for the potential reasons for poor coverage of these regions, we observed that flanking regions (+/- 50 bp) of most of these poorly covered regions comprise either high GC content, such as exon 1 of *MLLT1* (94% GC), or homopolymer runs, such as T-runs around three regions of *STAG2* gene [see Additional file 1: Table ST7], which can be informative for future optimization.

Out of the five panels compared, only MSK-IMPACT panel was reported to have 31 consistently poorly covered regions [59], which all happened to be targeted by SJPedPanel as well. Interestingly, SJPedPanel demonstrated sufficient depth of coverage ($>$ ($\mu$ - $2\sigma$) of respective COLO829BL sample level cut-offs) for 29 out of the 31 regions [Additional file 1: Table ST8]. The remaining two regions, exon 2 of *NOTCH2* and exon 15 of *PMS2,* consistently showed poor coverage as in MSK-IMPACT panel and were also part of the 27 poorly covered regions of SJPedPanel discussed above [Additional file 1: Table ST7].

Similarly, we analyzed the depth of coverage at designed SNPs. Notably, 99.5% of all the 7590 SNPs have depth more than ($\mu$ - $2\sigma$) of the respective sample level cut-offs [Additional file 1: Table ST9; see Additional file 3: Supplementary Fig. S4a for median and minimum depth of coverages for SNPs]. As expected, the variant allele frequencies (VAF) of all the SNPs in control COLO829BL samples were clustered around either 0, 0.5 or 1 [Additional file 3: Supplementary Fig. S4b]. A total of 3300 heterozygous SNPs ($0.3 \leq$ VAF $\leq 0.7$) are observed in COLO829BL, supporting the informativeness of our designed SNPs as mentioned above (~3000 heterozygous SNPs expected from any donors).

### *In silico* comparison of SJPedPanel and WES in a real-time clinical genomics cohort

We next compared the coverage of reported pediatric cancer alterations using whole exome sequencing (WES), because WES is effectively a capture-sequencing kit that targets coding exons of all genes rather than a panel of genes (thus an upper bound of all coding-region-based gene panels). We first asked whether our panel could offer comparable coverage of driver alterations (with a focus on coding SNVs and Indels) in pediatric cancers to WES. The recently published "Genome for Kids" (G4K) study [71] reported pathogenic and likely pathogenic (P/LP) variants (called driver alterations hereafter) using three platform sequencing (WGS, WES and RNAseq) from 253 pediatric cases that encompassed 20 cancer subtypes in a real-time clinical genomics setting, thus enabling us to assess the potential of SJPedPanel to cover driver alterations from diverse childhood cancer types. Here, we performed *in silico* analysis of regions targeted by SJPedPanel and WES using the curated positions of 485 driver alterations (including SNV/Indel/SV/ITD; Additional file 3: Supplementary Fig. S5a) from the G4K study [see Additional file 1: Table ST10]. SJPedPanel covered 86% of the 485 reported driver alterations as compared to 76% by WES [**Fig. 4a,** last pair of bars for "All" variants with gray background]. We next classified the variants into SNV, Indel, Fusion/SV (structural rearrangements that result in fusion oncoproteins), Other SV (structural rearrangements that do not result in fusion oncoproteins but affect cancer driver genes such as disrupting tumor suppressor genes), and ITD, by using the class labels in the G4K study [71]. As expected, WES only covered 12% of the Fusions/SV that have either of the breakpoints in exonic regions, while SJPedPanel covered 82% of these events. On the other hand, while WES covered all the reported driver SNVs and Indels, SJPedPanel did not cover 7% and 5% of SNVs and Indels, respectively. Interestingly, the host genes of these uncovered SNVs and Indels are rarely mutated (<0.1%) childhood cancers [2,3] [**Fig. 4a**, see Additional file 1: Table ST10, "Comment" column]. Of note, SJPedPanel covered 100% of the ITDs, while WES doesn't cover 13% of these. In fact, the ITDs missed by WES have DNA breakpoints that fall in introns and resulted in duplication of involved exons, such as tandem duplications in *PAX5* [72] and *KMT2A* [73], for which selected intronic regions were designed in

SJPedPanel. Of note, SJPedPanel successfully captured [Additional file 3: Supplementary Fig. S6] the recently described *UBTF* exonic tandem duplications[64].

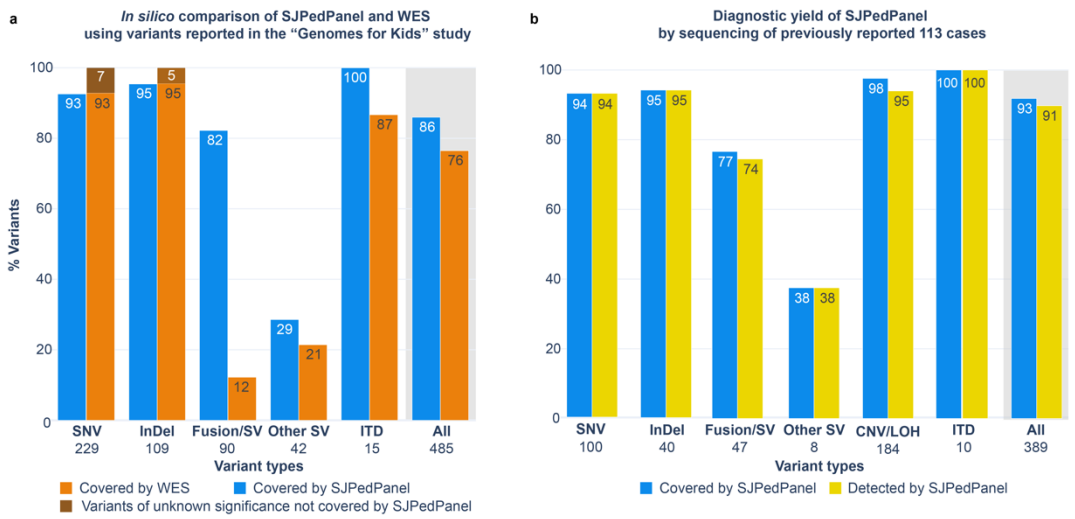


**Fig. 4. (a) *In silico* coverage comparison between SJPedPanel and WES** using percent coverage of variants (SNVs, Indels, Fusion, SV and ITDs) reported in the "Genomes for Kids" study [71]. The last pair of bars with gray background shows combined percent coverage over "All" 485 variants. **(b) Diagnostic yield of SJPedPanel by sequencing of previously reported 113 cases.** Y-axis shows percentage of covered and detected variants by SJPedPanel over each variant type. The last pair of bars with gray background for "All" variants show combined detection rate. Numbers of reported driver alterations are indicated at the bottom of bars for corresponding variant types.

We also asked what percentage of patients could benefit from SJPedPanel vs WES. As it turned out, at least one variant per case would have been covered by SJPedPanel in 93% of 210 cases (median: 1, range: 1 to 13), in contrast to 75% of the cases using WES (median: 2, range: 1 to 13) [Additional file 3: Supplementary Fig. S5b]. Most of this gain is due to the capture of intronic breakpoints that result in oncogenic fusions [**Fig. 4a**]. This data demonstrated that SJPedPanel has superior potential for detection and reporting of driver alterations in pediatric tumors than WES.

**Comparison of diagnostic yield between SJPedPanel and WES**

To validate above *in silico* findings, we re-sequenced 113 clinical cases with available specimen from previous clinical studies [4,71] using SJPedPanel. These samples reflect the broad tumor types and subtypes common in childhood cancer, including 27 hematological malignancies, 43 solid tumors and 43 brain tumors [see Additional file 1: Table ST11 for cohort description]. Common subtypes, such as AML (n=14), ALL (n=10), rhabdomyosarcoma (n=5), neuroblastoma (n=5), osteosarcoma (n=3), Wilms tumor (n=3), high-grade glioma (n=8) and medulloblastoma (n=14) are represented in addition to rare entities, such as melanoma (n=1) and desmoplastic small round cell tumor (n=2) (see Additional file 3: Fig. S7). Among these cases, 389 driver alterations (Methods) are reported via three-platform (WGS, WES, RNAseq) sequencing. These include 100 SNVs, 40 Indels, 55 SVs, 184 CNV/LOHs and 10 ITDs [Additional file 1: Tables ST12-ST16]. Of these, 361 (92.8%) variants were targeted by the panel, including 94 SNVs (94%), 38 Indels (95%), 36 Fusion/SVs (76.59%), 3 other SVs (37.5%), 180 CNV/LOHs (97.82%), and 10 ITDs (100%) [**Fig. 4b**, Additional file 1, Table ST17]. In total 28 P/LP variants were not covered by our panel [Additional file 1, Table ST17]. Of these, 6 were SNVs, 2 were Indels, 11 were Fusion/SVs (not designed), 5 were other SVs and 4 were focal CNVs. The uncovered variants belonged to 26 genes, which are not mutated in published pediatric pan-cancer cohorts [2,3] except for *COL1A1* that has a low mutation frequency of 0.2% [see Additional file 1: Table ST4, Additional file 3: Supplementary Fig. S8 and Additional file 1: Table ST17], further supporting their omission from our panel design.

We achieved a mean depth of ~2500X for all the samples tested [Additional file 1: Table ST18a], which ensures 95% confidence of detection of variants with ≥1% allele fraction (AF) [74]. As expected, we found an average of 3300 heterozygous SNPs (median: 3405, range: 2227 to 3844) in germline samples (n=30). These samples were pooled as a reference to interrogate CNV/LOH in tumor samples [**Method**; Additional file 1: Table ST18b]. By using a “rotation control” method [75] coupled with a recently developed Indel/SV genotyping tool [**Method**], we detected 98% (354 of the 361 covered variants; Additional file 1: Table ST17) of reported driver alterations. SNVs, Indels, ITDs showed detection rate of 100% [**Fig. 4b**]. The

Fusion/SVs showed an overall recall rate of 97% (35 out of 36) of the covered alterations. One fusion/SV marker (*RUNX1::RUNX1T1* from the case SJCBF100) that was covered in the panel design with single breakpoint in *RUNX1* was missed due to insufficient depth (21X) of coverage [Additional file 1: Tables ST13 and ST18a], therefore poor capture efficiency in certain genomic regions warrants a future study. By contrast, we could detect 38% (3 out of 8) of other SVs, and the rest of the missed SVs had their breakpoints either in intergenic or non-covered regions, which is consistent with a much larger genomic space for breakpoints in tumor suppressor genes. On the other hand, SJPedPanel detected 94.6% of the reported CNV/LOHs [Additional file 1: Table ST17]. Collectively, SJPedPanel detected 91% of the 389 reported driver alterations from these 113 cases [**Fig. 4b**, pair of last bars with gray background]. Of note, at least one variant was detected for 96.5% cases (n=113), with a median of 3 variants per case (Range = 0-16; Additional file 1: Table ST19). Consistent with the *in-silico* analysis [**Fig. 4a**], a comparison of SNV, Indel, SV and ITD variants (n=205 out of 389) discovered using three platform sequencing[4,71] (WGS, WES, RNASeq) indicated that SJPedPanel covers 88% of these variants whereas WES covers 78% [Additional file 3: Supplementary Fig. S9]. These data further confirm the superior performance achieved using SJPedPanel for childhood cancers with a panel size approximately 10% the size of WES.

We also highlight the successes and challenges in panel design by using structural variants as examples. First, due to the large genomic size of intronic regions that can be involved in oncogenic fusions, inclusion/exclusion of intronic regions involves a difficult balance between panel size and effective coverage of patient population. In our 2023 study of fusion gene pairs involved in 5,190 childhood cancers [7], 72 representative genes are selected for 274 fusion gene pairs and SJPedPanel included 53 of these 72 genes. The maximum mutation frequency of the 19 genes not included [Additional file 1: Table ST3c] in our panel is 0.1% [7]. Furthermore, inclusion of all relevant introns of all genes involved in oncogenic fusions would need ~8 Mbp (described in Source data file of Fig, 2k-l mentioned in Liu et al., 2023 [7]). With the observation that some genes can have multiple fusion partners (e.g., 40 of the 53 included representative genes have between 2 to 32 partner genes), we intentionally left out some partner genes by relying on common representative genes, which reduced the space from ~8 Mbp (280 introns) to ~1Mbp (156 introns).

For example, including only the intronic regions of *KMT2A* was sufficient to detect *KMT2A::MLLT1* fusion in pediatric T-ALL patient (SJMLL002) even though we did not include *MLLT1* introns in the panel content [Additional file 3: Fig. S10]. Despite this success, SVs can be challenging to detect. For example, SJPedPanel missed *RUNX1::RUNX1T1* in SJCBF100 because *RUNX1T1* introns were not included (~100K base pairs needed). As expected, panel also missed an inversion involving *RB1* gene in case SJRB0051 [Additional file 3: Fig. S11] because the DNA breakpoints fall into an intronic region of *RB1* and *RB1* introns are not covered (~180 kb are needed to cover all *RB1* introns).

Apart from SV/Fusions, SJPedPanel covers well known ITDs such as *FLT3*, *NOTCH1*, *BRAF* etc. [Additional file 1: Table ST16]. SJPedPanel also provides exclusive coverage of *UBTF* gene that was recently described in pediatric AML [64]. Although in literature *UBTF* ITDs are typically mistakenly detected as small Indels [3,64,76,77], our panel successfully detected *UBTF* TDs in two pediatric AML cases (SJAML015373 and SJAML016569; see Additional file 3: Supplementary Fig. S6) [64].

**Determining limit of detection**

One of the important applications of panel sequencing is disease monitoring, where the tumor burden is typically less than 1% and thus variants are rare and challenging to detect. To investigate the applicability of our panel, we first performed dilution experiments (with 7 tumor concentrations of 10%, 5%, 2.5%, 1%, 0.5%, 0.2%, 0.1%, in addition to pure normal of 0% and pure cancer of 100%) using 6 pediatric cancer cell lines (697, EW-8, K562, ME-1, MOLM-13, and Rh30) and 1 non-cancer cell line (GM12878) as normal control. These 6 lines collectively contain 26 unique P/LP variants (14 SNVs, 4 Indels, and 8 SVs) [78,79] [Additional file 1: Table ST20]. The lack of shared driver alterations allowed for pooling of dilutions to reduce the experimental complexity while keeping the diversity of mutation types. For example, at ladder of 0.5%, we mixed 5:5:5:5:5:5:970 cell equivalents from the 6 cancer lines and the normal line, respectively. To ensure sufficient power of detecting variants with low allele fractions, we achieved an average depth of 5,000X for 0.5%, >7,000X for 0.2% and 0.1% dilution concentrations [**Method**, Additional file 1: Table ST21].

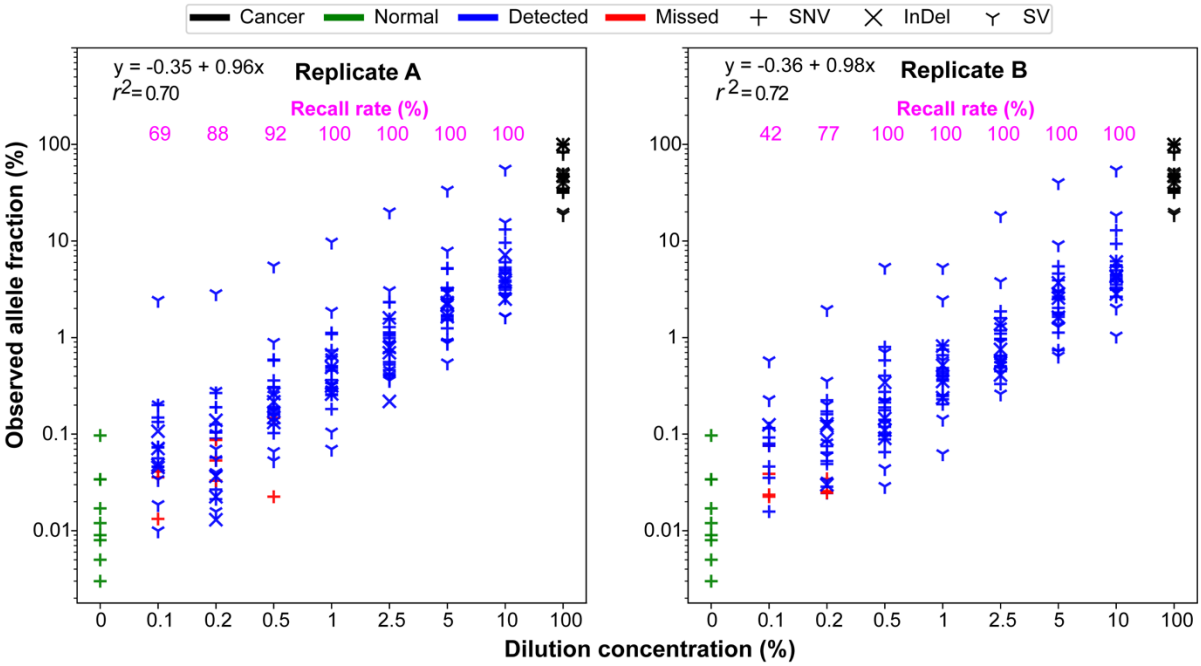


**Fig. 5. Determining the limit of detection with SJPedPanel.** The observed allele fractions of 26 driver alterations are shown on Y-axis as a function of corresponding dilution concentration shown in X-axis. The detection rate for each dilution concentration is shown on top as magenta text. The observed allele fractions of variants from normal and pure tumor cell lines are also shown using green and black points, respectively. The variants detected ($Q < 0.05$) are shown in blue, whereas those missed ($Q > 0.05$) are shown in red. Shown are results from replicate A (a) and replicate B (b).

We used SequencErr [80] and a newly developed SVIndelGenotyper [81] (manuscript under review) to perform allele counting followed by variant calling using binomial models with false discovery rate control (**Method**), where the pure normal of 0% was used to estimate background error rates as no cancer-driving somatic alterations are expected in non-cancer cell line GM12878. As seen in **Fig. 5**, the observed allele fractions closely represent corresponding dilution ladders, with R-squared values of 0.7 and 0.72, for biological replicates A and B, respectively [Additional file 1: Table ST22]. Notably, although we achieved >90% detection rate when the dilution concentration is above 0.5%, it diminishes quickly at lower dilutions. At a dilution concentration of 0.1%, recall rate was 69% and 42% in replicates A and B, respectively.

To further investigate the effect of sequencing depth on detection rate, we performed *in silico* down-sampling experiment [Method; Additional file 1: Table ST23]. For all the dilutions with concentrations ≥ 1%, which were initially sequenced at 5000X and 2500X [Additional file 3: Fig. S12a and S12b], recall rate is found to be close to 100% even after down sampling their depths to 1000X. However, recall rate declined with down sampled depths for dilution concentrations < 1%. For samples with dilution concentration of 0.5% [Additional file 3: Fig. S12b top panel], recall rate dropped from 97% at 3000X to 75% at 1000X. Thus, markers with allele fraction of 0.5% can be reliably detected at 2500X~3000X, which is concordant with our theoretical binomial calculation of 2,100X [**Method**, **Fig. 5**]. However, for markers with allele fraction of 0.2% and 0.1% the recall rate was <50% even at initial 10,000X data [Additional file 3: Fig. S12c]. This finding suggests that the current LoD is between 0.1% and 0.5% and is consistent with a recent report using cfDNA data [82].

The above data suggests a depth of 2,100X will ensures 95% chance of detecting >=5 mutant reads if the true allele fraction is >0.5%, However, considering the sequencing uniformity parameter, where 95% of targeted regions are covered at $\mu$-$2\sigma = 0.33\mu$, we would recommend 3 × 2,100 = 6,300X so that >95% of targeted regions will be covered >2,100X and ensure the 95% chance of detecting variants with allele fraction of 0.5%. Since the total space of this panel is ~0.15% of a human genome, 6,300X corresponds to whole genome sequencing at 9X coverage.

**Case study: Real time tracking of relapsed AML using deep sequencing**

To test the suitability of the panel for disease monitoring in a real-world setting, we chose two AML cases (SJAML016582 and SJAML016551) that had remission samples with low tumor purity. These cases have multiple pathogenic and likely pathogenic variants (P/LP) reported by clinical sequencing of diagnosis and relapse samples [Additional file 1: Table ST24] and are ideal for panel sequencing. The diagnosis, remission, and relapse samples were sequenced to an average depth of 5,000X using SJPedPanel per above power calculations.

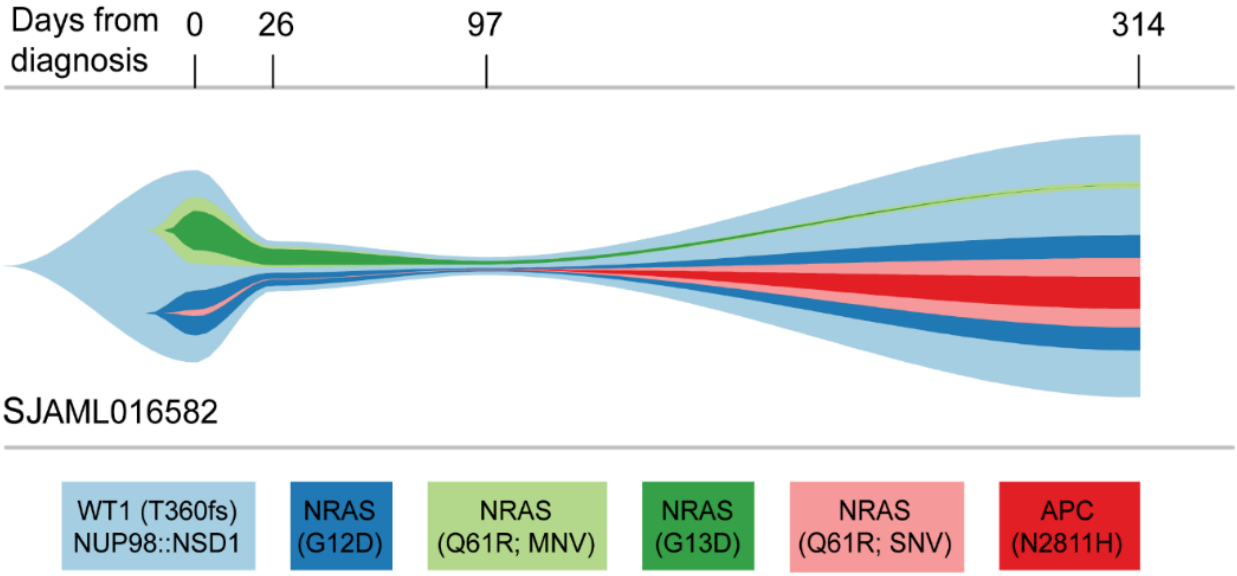


**Fig. 6. Real time tumor tracking in case SJAML016582.** The estimated cellular fractions of subclones at four timepoints from diagnosis (day 0) to relapse (day 314) are shown as a river-plot. Subclones with very low cellularity (e.g., NRAS Q61R (SNV) at diagnosis) were adjusted for visualization purposes using an R package fishplot [83]. Actual values are available in the Additional file 1: Table ST24-a.

In case SJAML016582, apart from subtype defining *NUP98*::*NSD1* fusion, 4 pathogenic variants were detected at diagnosis (day 0) and 6 pathogenic variants were detected at relapse (day 315), with 2 variants shared between diagnosis and relapse. With the ultra-deep panel sequencing data, we recovered all variants initially detected by whole genome sequencing, including 4 at diagnosis and 6 at relapse. Interestingly, panel sequencing detected the SNV that encodes *NRAS* Q61R in diagnosis tumor with allelic fraction (AF) 0.12%, and the MNV that encodes *NRAS* Q61R in relapse tumor with AF 0.07%, both of which are beyond the detection power of whole genome sequencing [Additional file 1: Table ST24-a; **Fig. 6**]. Further, in the day 26 remission data, we detected a high tumor burden (6.56%, Method). Notably, the tumor burden continued to decrease down to AF of 2.88% at day 97 as reflected by the SV responsible for *NUP98::NSD1* fusion.

In addition to disease monitoring, we also applied ultra-deep sequencing for detecting measurable residual disease (MRD) to investigate patient response to chemotherapy. Generally flow cytometry is a method of choice for such applications [84] in addition to real time PCR [85] and droplet digital PCR [86]. However, these

approaches are not as scalable as ultra-deep sequencing. In order to evaluate the efficacy of SJPedPanel to detect MRD, we sequenced three samples from diagnosis (day 0), MRD (day 23) and relapse (day 344) for AML patient SJAML016551. Here the diagnosis and relapse samples are used to define ancestral genetic alterations that are obligated to be present in the MRD sample. For this case, five pathogenic variants, including a *KMT2A::MLLT10* structural variant, were identified to be shared between diagnosis and relapse and are expected to be detected in MRD sample. Although flow-based MRD detection was negative for this case, our deep sequencing detected all 5 pathogenic variants with allele fraction range between 0.7-1% [Additional file 1: Table ST24-b; Additional file 3: Supplementary Fig. S13]. Together, our data demonstrate the potential ability of SJPedPanel in measuring MRD and monitoring disease progression that could aid in early detection of relapse.

## Discussion

We developed SJPedPanel, a hybridization capture-based assay targeting 357 pediatric relevant genes, including specific oncogenes and tumor suppressors implicated in 44 pediatric cancer genomic studies[2-4,7,26,33,35,61,71], many through the extensive efforts of the Pediatric Cancer Genome Project [1] and NCI TARGET project [3], as well as real-time clinical sequencing efforts at St. Jude Children's Research Hospital employing WGS, WES and RNA-Seq platforms [4,71]. SJPedPanel includes 297 introns that contain structural variants in 94 genes, accounting for 1.438 Mbp of genomic space, as well as 0.209 Mbp for detecting promoter-hijacking SVs. Furthermore, 7,590 common SNPs are covered, allowing for the detection of copy number changes and loss of heterozygosity (LOH) at a resolution of ~1Mb. Due the large differences in the genomic landscape of cancers in children and adults[3], SJPedPanel has unique coverage of 110 genes frequently implicated in pediatric cancers when compared to other commonly used panels (FoundationOne Heme, FoundationOne CDx, MSK-IMPACT, OncoKids and Oncomine Comprehensive Assay v3). One of the limitations of the current study is that the SJPedPanel was compared with other panels using in silico methods due to limited availability of specimens, which can be addressed in future prospective studies.

We used *in silico* and experimental strategies to evaluate the performance of SJPedPanel for detecting clinically relevant somatic mutations. Using the previously published G4K study, SJPedPanel was found to cover 86% of the reported somatic markers, including SNV, Indel and SV. Similar findings were obtained by using real-time clinical sequencing samples [4,71] in which 91% of the reported clinically relevant variants were detected, including all SNVs, Indels and ITDs. At patient level, at least one variant was detected in 96% cases (median 3). These findings establish the ability of SJPedPanel to detect clinically relevant somatic mutations in a wide range of samples from a real-time clinical genomics setting for childhood cancers.

While tumor-normal paired whole genome sequencing remains the gold standard for cancer diagnostics, the overall cost and required bioinformatic pipelines and infrastructure currently limits its broad application. On the other hand, gene-panel based genomics testing can enable many centers to perform NGS-based cancer diagnostics at overall lower cost and faster turn-around time. The content of SJPedPanel allows for more comprehensive detection of the alterations common in pediatric cancer compared to both WES and other panels. An inherent limitation of DNA sequencing panels is the lack of coverage at all critical loci or newly discovered recurrent alterations; however, panel content can readily be updated. For example, the current version of SJPedPanel lacks sufficient coverage to identify the recurrent *ASPSCR1::TFE3* fusion characteristic of alveolar soft part sarcoma or *SSX1*/*SSX2*::*SS18* in synovial sarcoma. Such genes will be incorporated in future versions. To maximize the utility of this panel, the targeted genomic locations are included in Tables ST1 and ST2 [Additional file 1] and future updates will be made readily available to public.

As a proof-of-principle, we demonstrate the application of using SJPedPanel for minimal residual disease (MRD) detection and post-treatment disease monitoring. It would be interesting to see the clinical value of this application using large patient cohorts in the future. Because this is a pan-cancer gene panel for childhood malignancies, it will be relatively straightforward to develop sub panels dedicated to certain cancer subtypes to further reduce the size of the panel, and in turn enabling much higher depth with a similar cost. We expect SJPedPanel to significantly impact the management and diagnosis of children with cancer.

## Material and Methods

### Panel design

Based on extensive research and literature review of pan-cancer genome profiling studies a list of exonic and/or intronic regions (n=5,009 regions, 2.82 Mbp) from 357 genes that are frequently implicated in pediatric cancers was compiled to detect single nucleotide variants (SNVs), small insertions & deletions (Indels), gene fusions, structural variants (SVs), and internal tandem duplications (ITDs). We also curated a list of 7,590 single nucleotide polymorphic markers (SNPs) evenly spread across human chromosomes to detect large genomic structural rearrangements such as copy number variations (CNVs) and loss of heterozygosity (LOH). The details of all the genomics regions and SNPs used to assemble the pediatric pan cancer, termed as SJPedPanel, are available in Additional file 1: Tables ST1 and ST2.

### Capture efficiency of the SJPedPanel

We generated one high-depth (~2000X) and one low-depth (~200X) targeted sequencing libraries (two replicates) using SJPedPanel using COLO829BL cell line (a gold-standard cell line for clinical validation and other established studies [69,70]). These libraries were sequenced on Illumina NovaSeq and NextSeq platforms, respectively. The data generated was used to evaluate the capture performance of the probes and uniformity of coverage across regions and loci of the SJPedPanel [**Fig. 3** and Additional file 3: Fig. S2].

### Dilution experiment

A dilution experiment using 6 cancer cell lines (ME-1, 697, Rh30, EW-8, K562, Molm13, courtesy of Elizabeth Stewart, Department of Oncology, St. Jude Children's Research Hospital) and a non-cancer cell line (GM12878) was designed to achieve seven tumor concentrations with two replicates each. The seven dilutions were divided in three groups a) ultra-low (0.1%, 0.2%), b) low (0.5%, 1%), and c) medium (2.5%, 5% and 10%), which were sequenced at depths of 10,000X, 5000X and 2500X respectively. The cell lines were also sequenced independently in undiluted forms at 25,000X to estimate the original allele fractions of 26 cell-line-specific markers (14 SNVs, 4 Indels, 8 SVs; all these markers are confirmed to be

exclusively detected from one of the six cell lines) given in Additional file 1: Table ST20. Recall rate of these known markers across different dilutions was used to assess the limit of detection of the SJPedPanel.

The limit of detection (LoD) is determined by two critical factors: 1) the sequencing depth (also known as power) and 2) the noise level. For example, if the true allele fraction is 1%, a sequencing depth of 913X will ensure 95% chance of detecting this variant with >=5 mutant alleles [74]. In consideration of the high range of dilution concentrations, we aimed for 2,500X depth for dilution ladders >1%, 5,000X for ladders 0.5% and 1%, and 10,000X for ladders 0.1% and 0.2%. On the other hand, the noise level is typically regarded as background error rate. Mutations with higher background error rates are more difficult to detect because the true signal can be overwhelmed by the background noises. We previously developed computational error suppression methods to achieve error rate of ~$10^{-6}$-$10^{-4}$ for substitutions [87] and similar methods and results have been achieved for Indels and SVs (manuscript under review).

### *In silico* down sampling experiment

We performed *in silico* down sampling of data from a set of cancer cell line dilution samples to find out the trade-off between recall rate, down sampled depth of sequencing and associated cost estimates. The samples originally sequenced at 2,500X were further down sampled to simulate depths of sequencing at 1,000X, 1,500X and 2,000X, whereas the samples which were sequenced at 5,000X and 10,000X were down sampled to simulate depths of sequencing at 1,000X, 1,500X, 2,000X, 3,000X. For each of the desired down sampled depth, 10 samples were simulated, each consisting of randomly sampled reads at loci of 14 SNVs. These simulated samples were used to determine the trade-off between recall rate and depth of sequencing.

### Investigating diagnostic yield using clinical sample re-sequencing

Based on sample availability we selected 113 specimens from previously sequenced pediatric cancer cases treated at St. Jude Children's Research Hospital to represent a wide range of cancer subtypes common in pediatrics. Samples were chosen primarily from the pilot study cohort (n=40) [4] and G4K studies (n=73) [71] and had

previously reported clinically relevant markers identified by triple platform approach of whole genome, whole exome, and transcriptome sequencing. The recall rate of clinically relevant markers from these cases was used to establish the diagnostic yield of the SJPedPanel. Here seven markers from 3 hypermutator cases (SJHGG030335, SJHGG030336 and SJST030211) were downgraded to VUS (variant of unknown significance) for this analysis per communication with corresponding author of the G4K study [see “Comment” column in the Additional file 1: Table ST12], resulting in 140 of SNV/Indel, 55 of fusion/SVs, 184 CNV/LOH and 10 ITDs (total = 389). A list of cases and their cancer subtypes used for these purposes is provided in the [Additional file 1: Table ST11].

**Library preparation, capture, and sequencing**

DNA samples were obtained and subjected to DNA-seq library preparation and target enrichment followed by sequencing in the Clinical genomics laboratory as described below. An input of 100ng of DNA was used to construct libraries using the Twist Library Preparation Enzymatic Fragmentation (EF) Kit 2.0 (Twist Biosciences, CA) following the manufacturer’s instructions. Capture oligos were designed to detect putative SNVs, Indels, SVs, ITDs and CNVs in 357 genes of clinical interest. SJPedPanel was synthesized at Twist Biosciences and is described in detail in the section on panel design. Eight libraries were pooled at a time and target enrichment for the SJPedPanel baits was carried out using Twist hybrid capture protocol following manufacturer’s instruction. Paired-end 150-cycle sequencing was performed on NovaSeq or NextSeq instruments (Illumina Inc, CA) as appropriate. Where necessary, additional sequencing (“top off”) was performed to ensure that a sequencing depth of at least 1000X was achieved in all cases.

**Early detection of relapsed AML cases**

To test the panel’s capability for disease monitoring, two pediatric AML cases (SJAML016582 and SJAML016551) with material available at diagnosis, relapse, and remission timepoints were analyzed. Both samples provide multiple trackable somatic markers, including structural variants and SNVs. Samples were subjected to deeper sequencing depths of >5000X after targeted capture to ensure detection of low-level variants at <1%. Average of variant allele fractions (VAF) of detected somatic variants were used to estimate tumor burden at corresponding time points.

**Coverage comparison between SJPedPanel and Whole exome sequencing**

The content of SJPedPanel was compared with that of whole exome sequencing (WES) manifest to highlight the differences in coverage of hg19 genomic regions. An Illumina Exome 2.0 Plus hg19 BED file [88] padded with 10 bp was used for region intersection analyses. We utilized recently reported somatic variants from the Genomes for Kids (G4K) [71], a real-time three-platform sequencing study of 309 pediatric cancer patients to benchmark the coverage of reported pathogenic and likely pathogenic variants between SJPedPanel and WES.

**Comparison of SJPedPanel with other panels**

We compared the content of the SJPedPanel with DNA content of five other commercially available panels including FoundationOne Heme [57], FoundationOne CDx [58] by Foundation Medicine Inc., MSK-IMPACT [6,59] by Memorial Sloan Kettering Cancer Center, OncoKids [5] by Children's Hospital Los Angeles and Oncomine Comprehensive Assay v3 by Thermo Fisher Scientific Inc. [60]. These panels collectively represent the breadth and diversity of clinical gene panels. Since most of the providers do not provide the exact coordinates of the regions in the panel, we compared the content of these panels using standardized gene names with the help of official gene symbols and synonyms from the NCBI gene database [89].

**Bioinformatics analyses**

The adapter trimmed paired end FASTQ files generated on the Illumina NovaSeq/NextSeq platforms were assessed for sequence and instrument quality using FastQC v0.11.9 [90] and SequencErr v2.0.9 [80]. The reads were mapped against GRCh37 build using BWA aln 0.7.12-r1039 [91]. The utility commands in SAMtools v1.7 [92] and BEDTools v2.25.0 [93] were used perform simple operations using BAM and BED files. The count files obtained from BAM files using SequencErr were further passed as an input to DeepSeqCoverageQC v0.3.1 [94] to compute depth of coverage QC metrics of the sequenced samples over loci/regions of the SJPedPanel. The genotyping of SVs and Indels to compute the allele fractions was carried out using SVIndelGenotyper [81]. The CNVs were detected using CNVkit v0.9.10 [95]. The BAM files of 30 germline samples were used to create a pooled reference of per-bin copy number estimates. The segment & bin-level call files along with CNV diagrams generated by CNVkit batch command were used to review the

CNV calls in tumor samples. To determine the LOH in sequenced samples, the minor allele frequencies of 7590 SNPs were computed using count files generated by SequencErr and subsequently used to generate allelic imbalance plots over chromosomes. The output files and diagrams generated by CNVkit v0.9.10 and the allelic imbalance figures used to review CNV and LOH events are available from Zenodo repository [96].

A previously developed "rotation control" method [75] was used to obtain the background count of the variants for binomial testing and Q-values [97] were used to assess the statistical significance of detection based on binomial testing. Briefly, frequency of any variant from the original mutation-positive sample was considered as foreground ($V_f$) and frequency of the same variant derived from the aggregated counts from rest of the samples, which were expected to be wild type in the original analysis, was considered as background or control ($V_b$) for the binomial testing. Thus, the background counts were rotated w.r.t. change in the foreground variant and sample of interest. A sample is called mutation-positive if the Q-value was found to be < 0.05. All the statistical analyses were performed using R v4.0.3 [98].

## References


1 Downing, J. R. *et al.* The Pediatric Cancer Genome Project. *Nat Genet* **44**, 619-622, doi:10.1038/ng.2287 (2012).

2 Grobner, S. N. *et al.* The landscape of genomic alterations across childhood cancers. *Nature* **555**, 321-327, doi:10.1038/nature25480 (2018).

3 Ma, X. *et al.* Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. *Nature* **555**, 371-376, doi:10.1038/nature25795 (2018).

4 Rusch, M. *et al.* Clinical cancer genomic profiling by three-platform sequencing of whole genome, whole exome and transcriptome. *Nat Commun* **9**, 3962, doi:10.1038/s41467-018-06485-7 (2018).

5 Hiemenz, M. C. *et al.* OncoKids: A Comprehensive Next-Generation Sequencing Panel for Pediatric Malignancies. *J Mol Diagn* **20**, 765-776, doi:10.1016/j.jmoldx.2018.06.009 (2018).

6 Cheng, D. T. *et al.* Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. *J Mol Diagn* **17**, 251-264, doi:10.1016/j.jmoldx.2014.12.006 (2015).

7 Liu, Y. *et al.* Etiology of oncogenic fusions in 5,190 childhood cancers and its clinical and therapeutic implication. *Nature communications* **14**, 1739, doi:10.1038/s41467-023-37438-4 (2023).

8 Cheung, N. K. *et al.* Association of age at diagnosis and genetic mutations in patients with neuroblastoma. *JAMA* **307**, 1062-1071, doi:10.1001/jama.2012.228 (2012).

9 Gruber, T. A. *et al.* An Inv(16)(p13.3q24.3)-encoded CBFA2T3-GLIS2 fusion protein defines an aggressive subtype of pediatric acute megakaryoblastic leukemia. *Cancer Cell* **22**, 683-697, doi:10.1016/j.ccr.2012.10.007 (2012).

10 Roberts, K. G. *et al.* Genetic alterations activating kinase and cytokine receptor signaling in high-risk acute lymphoblastic leukemia. *Cancer Cell* **22**, 153-166, doi:10.1016/j.ccr.2012.06.005 (2012).
11 Robinson, G. *et al.* Novel mutations target distinct subgroups of medulloblastoma. *Nature* **488**, 43-48, doi:10.1038/nature11213 (2012).
12 Wu, G. *et al.* Somatic histone H3 alterations in pediatric diffuse intrinsic pontine gliomas and non-brainstem glioblastomas. *Nat Genet* **44**, 251-253, doi:10.1038/ng.1102 (2012).
13 Zhang, J. *et al.* A novel retinoblastoma therapy from genomic and epigenetic analyses. *Nature* **481**, 329-334, doi:10.1038/nature10733 (2012).
14 Zhang, J. *et al.* The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature* **481**, 157-163, doi:10.1038/nature10725 (2012).
15 Chen, X. *et al.* Targeting oxidative stress in embryonal rhabdomyosarcoma. *Cancer Cell* **24**, 710-724, doi:10.1016/j.ccr.2013.11.002 (2013).
16 Holmfeldt, L. *et al.* The genomic landscape of hypodiploid acute lymphoblastic leukemia. *Nat Genet* **45**, 242-252, doi:10.1038/ng.2532 (2013).
17 Jaffe, J. D. *et al.* Global chromatin profiling reveals NSD2 mutations in pediatric acute lymphoblastic leukemia. *Nat Genet* **45**, 1386-1391, doi:10.1038/ng.2777 (2013).
18 Paugh, B. S. *et al.* Novel oncogenic PDGFRA mutations in pediatric high-grade gliomas. *Cancer Res* **73**, 6219-6229, doi:10.1158/0008-5472.CAN-13-1491 (2013).
19 Shah, S. *et al.* A recurrent germline PAX5 mutation confers susceptibility to pre-B cell acute lymphoblastic leukemia. *Nat Genet* **45**, 1226-1231, doi:10.1038/ng.2754 (2013).
20 Zhang, J. *et al.* Whole-genome sequencing identifies genetic alterations in pediatric low-grade gliomas. *Nat Genet* **45**, 602-612, doi:10.1038/ng.2611 (2013).
21 Chen, X. *et al.* Recurrent somatic structural variations contribute to tumorigenesis in pediatric osteosarcoma. *Cell Rep* **7**, 104-112, doi:10.1016/j.celrep.2014.03.003 (2014).
22 Parker, M. *et al.* C11orf95-RELA fusions drive oncogenic NF-kappaB signalling in ependymoma. *Nature* **506**, 451-455, doi:10.1038/nature13109 (2014).
23 Roberts, K. G. *et al.* Targetable kinase-activating lesions in Ph-like acute lymphoblastic leukemia. *N Engl J Med* **371**, 1005-1015, doi:10.1056/NEJMoa1403088 (2014).
24 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, doi:10.1158/2159-8290.CD-14-0622 (2014).
25 Wu, G. *et al.* The genomic landscape of diffuse intrinsic pontine glioma and pediatric non-brainstem high-grade glioma. *Nat Genet* **46**, 444-450, doi:10.1038/ng.2938 (2014).
26 Andersson, A. K. *et al.* The landscape of somatic mutations in infant MLL-rearranged acute lymphoblastic leukemias. *Nat Genet* **47**, 330-337, doi:10.1038/ng.3230 (2015).
27 Li, B. *et al.* Negative feedback-defective PRPS1 mutants drive thiopurine resistance in relapsed childhood ALL. *Nat Med* **21**, 563-571, doi:10.1038/nm.3840 (2015).
28 Lu, C. *et al.* The genomic landscape of childhood and adolescent melanoma. *J Invest Dermatol* **135**, 816-823, doi:10.1038/jid.2014.425 (2015).
29 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).
30 Pinto, E. M. *et al.* Genomic landscape of paediatric adrenocortical tumours. *Nat Commun* **6**, 6302, doi:10.1038/ncomms7302 (2015).
31 Tong, Y. *et al.* Cross-Species Genomics Identifies TAF12, NFYC, and RAD54L as Choroid Plexus Carcinoma Oncogenes. *Cancer Cell* **27**, 712-727, doi:10.1016/j.ccell.2015.04.005 (2015).
32 Zhang, J. *et al.* Germline Mutations in Predisposition Genes in Pediatric Cancer. *N Engl J Med* **373**, 2336-2346, doi:10.1056/NEJMoa1508054 (2015).
33 Faber, Z. J. *et al.* The genomic landscape of core-binding factor acute myeloid leukemias. *Nat Genet* **48**, 1551-1556, doi:10.1038/ng.3709 (2016).
34 Iacobucci, I. *et al.* Truncating Erythropoietin Receptor Rearrangements in Acute Lymphoblastic Leukemia. *Cancer Cell* **29**, 186-200, doi:10.1016/j.ccell.2015.12.013 (2016).
35 Liu, Y. F. *et al.* Genomic Profiling of Adult and Pediatric B-cell Acute Lymphoblastic Leukemia. *EBioMedicine* **8**, 173-183, doi:10.1016/j.ebiom.2016.04.038 (2016).
36 Zhang, J. *et al.* Deregulation of DUX4 and ERG in acute lymphoblastic leukemia. *Nat Genet* **48**, 1481-1489, doi:10.1038/ng.3691 (2016).
37 de Rooij, J. D. *et al.* Pediatric non-Down syndrome acute megakaryoblastic leukemia is characterized by distinct genomic subsets with varying outcomes. *Nat Genet* **49**, 451-456, doi:10.1038/ng.3772 (2017).

38 Liu, Y. *et al.* The genomic landscape of pediatric and young adult T-lineage acute lymphoblastic leukemia. *Nat Genet* **49**, 1211-1218, doi:10.1038/ng.3909 (2017).
39 Northcott, P. A. *et al.* The whole-genome landscape of medulloblastoma subtypes. *Nature* **547**, 311-317, doi:10.1038/nature22973 (2017).
40 Alexander, T. B. *et al.* The genetic basis and cell of origin of mixed phenotype acute leukaemia. *Nature* **562**, 373-379, doi:10.1038/s41586-018-0436-0 (2018).
41 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, doi:10.1038/nm.4439 (2018).
42 Pajtler, K. W. *et al.* Molecular heterogeneity and CXorf67 alterations in posterior fossa group A (PFA) ependymomas. *Acta Neuropathol* **136**, 211-226, doi:10.1007/s00401-018-1877-0 (2018).
43 Stewart, E. *et al.* Identification of Therapeutic Targets in Rhabdomyosarcoma through Integrated Genomic, Epigenomic, and Proteomic Analyses. *Cancer Cell* **34**, 411-426 e419, doi:10.1016/j.ccell.2018.07.012 (2018).
44 Brady, S. W. *et al.* The Clonal Evolution of Metastatic Osteosarcoma as Shaped by Cisplatin Treatment. *Mol Cancer Res* **17**, 895-906, doi:10.1158/1541-7786.MCR-18-0620 (2019).
45 Newman, S. *et al.* Clinical genome sequencing uncovers potentially targetable truncations and fusions of MAP3K8 in spitzoid and other melanomas. *Nat Med* **25**, 597-602, doi:10.1038/s41591-019-0373-y (2019).
46 Brady, S. W. *et al.* Pan-neuroblastoma analysis reveals age- and signature-associated driver alterations. *Nature communications* **11**, 5183, doi:10.1038/s41467-020-18987-4 (2020).
47 Brady, S. W. *et al.* Therapy-induced mutagenesis in relapsed ALL is supported by mutational signature analysis. *Blood* **136**, 2235-2237, doi:10.1182/blood.2020008107 (2020).
48 Waanders, E. *et al.* Mutational landscape and patterns of clonal evolution in relapsed pediatric acute lymphoblastic leukemia. *Blood Cancer Discov* **1**, 96-111, doi:10.1158/0008-5472.BCD-19-0041 (2020).
49 Orostica, K. Y. & Verdugo, R. A. chromPlot: visualization of genomic data in chromosomal context. *Bioinformatics* **32**, 2366-2368, doi:10.1093/bioinformatics/btw137 (2016).
50 Mansour, M. R. *et al.* Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346**, 1373-1377, doi:10.1126/science.1259037 (2014).
51 Zhou, X. *et al.* Exploring genomic alteration in pediatric cancer using ProteinPaint. *Nat Genet* **48**, 4-6, doi:10.1038/ng.3466 (2016).
52 Zhou, X. *et al.* Exploration of Coding and Non-coding Variants in Cancer Using GenomePaint. *Cancer Cell* **39**, 83-95 e84, doi:10.1016/j.ccell.2020.12.011 (2021).
53 Ottema, S. *et al.* The leukemic oncogene EVI1 hijacks a MYC super-enhancer by CTCF-facilitated loops. *Nat Commun* **12**, 5679, doi:10.1038/s41467-021-25862-3 (2021).
54 Northcott, P. A. *et al.* Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511**, 428-434, doi:10.1038/nature13379 (2014).
55 Sjoberg, D. *ggsankey: An R package*, <https://github.com/davidsjoberg/ggsankey> (2021).
56 Pai S, R. J. *IdeoViz: Plots data (continuous/discrete) along chromosomal ideogram. doi:10.18129/B9.bioc.IdeoViz, R package version 1.37.0*, <https://bioconductor.org/packages/IdeoViz.> (2023).
57 *Foundation One Heme*, <https://assets.ctfassets.net/w98cd481qyp0/42r1cTE8VR4137CaHrsaen/baf91080cb3d78a52ada10c6358fa130/FoundationOne_Heme_Technical_Specifications.pdf> (2019).
58 *Foundation One CDX*, <https://assets.ctfassets.net/w98cd481qyp0/YqqKHaqQmFeqc5ueQk48w/d12f19680205941ea3fee417f08e9524/F1CDx_Technical_Specifications.pdf> (2022).
59 *MSK-IMPACT Panel*, <https://www.accessdata.fda.gov/cdrh_docs/reviews/den170058.pdf> (2017).
60 *Thermo Fisher Oncomine Comprehensive Assay v3*, <https://assets.thermofisher.com/TFS-Assets/LSG/brochures/oncomine-comprehensive-assay-v3-flyer.pdf> (2022).
61 Gadd, S. *et al.* A Children's Oncology Group and TARGET initiative exploring the genetic landscape of Wilms tumor. *Nat Genet* **49**, 1487-1494, doi:10.1038/ng.3940 (2017).
62 DeSouza, R. M., Jones, B. R., Lowis, S. P. & Kurian, K. M. Pediatric medulloblastoma - update on molecular classification driving targeted therapies. *Front Oncol* **4**, 176, doi:10.3389/fonc.2014.00176 (2014).

63 Arabzade, A. *et al.* ZFTA-RELA Dictates Oncogenic Transcriptional Programs to Drive Aggressive Supratentorial Ependymoma. *Cancer Discov* **11**, 2200-2215, doi:10.1158/2159-8290.CD-20-1066 (2021).
64 Umeda, M. *et al.* Integrated Genomic Analysis Identifies UBTF Tandem Duplications as a Recurrent Lesion in Pediatric Acute Myeloid Leukemia. *Blood Cancer Discov* **3**, 194-207, doi:10.1158/2643-3230.BCD-21-0160 (2022).
65 Jeha, S. *et al.* Clinical significance of novel subtypes of acute lymphoblastic leukemia in the context of minimal residual disease-directed therapy. *Blood Cancer Discov* **2**, 326-337, doi:10.1158/2643-3230.BCD-20-0229 (2021).
66 He, J. *et al.* Integrated genomic DNA/RNA profiling of hematologic malignancies in the clinical setting. *Blood* **127**, 3004-3014, doi:10.1182/blood-2015-08-664649 (2016).
67 Hovelson, D. H. *et al.* Development and validation of a scalable next-generation sequencing system for assessing relevant somatic variants in solid tumors. *Neoplasia* **17**, 385-399, doi:10.1016/j.neo.2015.03.004 (2015).
68 Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938-2940, doi:10.1093/bioinformatics/btx364 (2017).
69 Craig, D. W. *et al.* A somatic reference standard for cancer genome sequencing. *Sci Rep* **6**, 24607, doi:10.1038/srep24607 (2016).
70 Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191-196, doi:10.1038/nature08658 (2010).
71 Newman, S. *et al.* Genomes for Kids: The Scope of Pathogenic Mutations in Pediatric Cancer Revealed by Comprehensive DNA and RNA Sequencing. *Cancer Discov* **11**, 3008-3027, doi:10.1158/2159-8290.CD-20-1631 (2021).
72 Gu, Z. *et al.* PAX5-driven subtypes of B-progenitor acute lymphoblastic leukemia. *Nat Genet* **51**, 296-307, doi:10.1038/s41588-018-0315-5 (2019).
73 Choi, S. M., Dewar, R., Burke, P. W. & Shao, L. Partial tandem duplication of KMT2A (MLL) may predict a subset of myelodysplastic syndrome with unique characteristics and poor outcome. *Haematologica* **103**, e131-e134, doi:10.3324/haematol.2017.185249 (2018).
74 Ma, X., Arunachalam, S. & Liu, Y. Applications of probability and statistics in cancer. *Quantitative Biology* **8**, 15-108, doi:10.1007/s40484-020-0203-8 (2020).
75 Li, B. *et al.* Therapy-induced mutations drive the genomic landscape of relapsed acute lymphoblastic leukemia. *Blood* **135**, 41-55, doi:10.1182/blood.2019002220 (2020).
76 Patel, J. P. *et al.* Prognostic relevance of integrated genetic profiling in acute myeloid leukemia. *N Engl J Med* **366**, 1079-1089, doi:10.1056/NEJMoa1112304 (2012).
77 Tyner, J. W. *et al.* Functional genomic landscape of acute myeloid leukaemia. *Nature* **562**, 526-531, doi:10.1038/s41586-018-0623-z (2018).
78 Bairoch, A. The Cellosaurus, a Cell-Line Knowledge Resource. *J Biomol Tech* **29**, 25-38, doi:10.7171/jbt.18-2902-002 (2018).
79 Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564-576 e516, doi:10.1016/j.cell.2017.06.010 (2017).
80 Davis, E. M. *et al.* SequencErr: measuring and suppressing sequencer errors in next-generation sequencing data. *Genome Biol* **22**, 37, doi:10.1186/s13059-020-02254-2 (2021).
81 SVIndelGenotyper. <https://github.com/stjude/SVIndelGenotyper> (2023).
82 Deveson, I. W. *et al.* Evaluating the analytical validity of circulating tumor DNA sequencing assays for precision oncology. *Nat Biotechnol* **39**, 1115-1128, doi:10.1038/s41587-021-00857-z (2021).
83 Miller, C. A. *et al.* Visualizing tumor evolution with the fishplot package for R. *BMC Genomics* **17**, 880, doi:10.1186/s12864-016-3195-z (2016).
84 Loken, M. R. *et al.* Residual disease detected by multidimensional flow cytometry signifies high relapse risk in patients with de novo acute myeloid leukemia: a report from Children's Oncology Group. *Blood* **120**, 1581-1588, doi:10.1182/blood-2012-02-408336 (2012).
85 Gabert, J. *et al.* Standardization and quality control studies of 'real-time' quantitative reverse transcriptase polymerase chain reaction of fusion gene transcripts for residual disease detection in leukemia - a Europe Against Cancer program. *Leukemia* **17**, 2318-2357, doi:10.1038/sj.leu.2403135 (2003).
86 Mencia-Trinchant, N. *et al.* Minimal Residual Disease Monitoring of Acute Myeloid Leukemia by Massively Multiplex Digital PCR in Patients with NPM1 Mutations. *J Mol Diagn* **19**, 537-548, doi:10.1016/j.jmoldx.2017.03.005 (2017).

87 Ma, X. *et al.* Analysis of error profiles in deep next-generation sequencing data. *Genome Biol* **20**, 50, doi:10.1186/s13059-019-1659-6 (2019).
88 *Illumina DNA Prep with Exome 2.0 BED Files*, <https://support.illumina.com/downloads/Illumina-dna-prep-exome-20-bed-files.html> (2023).
89 Brown, G. R. *et al.* Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res* **43**, D36-42, doi:10.1093/nar/gku1055 (2015).
90 FastQC. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (2019).
91 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595, doi:10.1093/bioinformatics/btp698 (2010).
92 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
93 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).
94 DeepSeqCoverageQC. <https://github.com/pandurang-kolekar/DeepSeqCoverageQC> (2023).
95 Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol* **12**, e1004873, doi:10.1371/journal.pcbi.1004873 (2016).
96 Kolekar, P. *SJPedPanel: Supplementary Data - Output files and diagrams generated by CNVkit v0.9.10 and the allelic imbalance figures used to review CNV and LOH events.*, <https://zenodo.org/doi/10.5281/zenodo.8173838> (2023).
97 Storey JD, B. A., Dabney A, Robinson D. *qvalue: Q-value estimation for false discovery rate control. doi:10.18129/B9.bioc.qvalue, R package version 2.34.0*, <https://bioconductor.org/packages/qvalue.> (2023).
98 Team, R. C. *R: A Language and Environment for Statistical Computing*, <https://www.R-project.org/> (2023).

## Declarations

### Availability of data and materials

The cell line data generated for this study have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB64356 (https://www.ebi.ac.uk/ena/browser/view/PRJEB64356). The accession numbers of the samples are listed in Additional File 1: Tables ST5 and ST21.

### Funding

This work was supported in part by the National Cancer Institute of the National Institutes of Health under Award Number R01CA273326 (to X.M.), the Fund for Innovation in Cancer Informatics (www.the-ici-fund.org, to X.M. and J.M.K.), Cancer Center Support Grant P30CA021765 (Developmental Fund to J.M.K. and X.M.) from the National Institutes of Health, and the American Lebanese Syrian Associated Charities (ALSAC). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or other funding agencies.

### Authors' contributions

X.M. and J.M.K. conceived the research. P.K. collected data, performed data analyses, generated figures and implemented the *DeepSeqCoverageQC* algorithm. V.B. and L.D. performed sequencing experiments. J.M.K. coordinated clinical specimen acquisition. Y.L. and Q.T. implemented *SVInDelGenotyper* algorithm. P.K., V.B., L.D.,Y.L.,S.F.,Q.T.,H.M.,A.L.H,E.P.,Z.L.,J.M.,J.N,J.G,J.M.,D.P.,S.M.,L.W., E.S.,J.E.,J.M.K. and X.M. contributed to analyses. All authors read and approved the final manuscript.

### Authors' information

Pandurang Kolekar, Vidya Balagopal and Li Dong contributed equally to this work.

### Ethics approval and consent to participate

This study was approved by the St. Jude institutional review board (IRB) and informed consent was obtained for samples collection from the patient, parents or

guardians. Subjects were not compensated for participation. All patient samples are de-identified.

**Competing interests**

The authors declare no competing financial interests.

**Acknowledgements**

X.M and J.M.K thank Elizabeth Stewart for providing cell lines for our dilution experiments.