













# Feasibility of whole genome and transcriptome profiling in pediatric and young adult cancers

N. Shukla<sup>1,8</sup>, M. F. Levine <sup>1,2,8</sup>, G. Gundem <sup>1,2,8</sup>, D. Domenico<sup>1,2</sup>, B. Spitzer<sup>1</sup>, N. Bouvier<sup>1</sup>, J. E. Arango-Ossa<sup>1,2</sup>, D. Glodzik<sup>2</sup>, J. S. Medina-Martínez<sup>2</sup>, U. Bhanot <sup>3,4</sup>, J. Gutiérrez-Abril<sup>1,2</sup>, Y. Zhou<sup>2</sup>, E. Fiala<sup>1,5</sup>, E. Stockfisch<sup>1</sup>, S. Li<sup>1</sup>, M. I. Rodriguez-Sanchez<sup>1</sup>, T. O'Donohue<sup>1</sup>, C. Cobbs <sup>6</sup>, M. H. A. Roehrl <sup>3,4,7</sup>, J. Benhamida <sup>3</sup>, F. Iglesias Cardenas<sup>1</sup>, M. Ortiz<sup>1</sup>, M. Kinnaman<sup>1</sup>, S. Roberts<sup>1</sup>, M. Ladanyi<sup>3</sup>, S. Modak<sup>1</sup>, S. Farouk-Sait<sup>3</sup>, E. Slotkin<sup>1</sup>, M. A. Karajannis <sup>1</sup>, F. Dela Cruz<sup>1</sup>, J. Glade Bender <sup>1</sup>, A. Zehir <sup>3</sup>, A. Viale<sup>6</sup>, M. F. Walsh <sup>1,5</sup>, A. L. Kung <sup>1,9</sup>✉ & E. Papaemmanuil <sup>1,2,9</sup>✉

The utility of cancer whole genome and transcriptome sequencing (cWGTS) in oncology is increasingly recognized. However, implementation of cWGTS is challenged by the need to deliver results within clinically relevant timeframes, concerns about assay sensitivity, reporting and prioritization of findings. In a prospective research study we develop a workflow that reports comprehensive cWGTS results in 9 days. Comparison of cWGTS to diagnostic panel assays demonstrates the potential of cWGTS to capture all clinically reported mutations with comparable sensitivity in a single workflow. Benchmarking identifies a minimum of 80× as optimal depth for clinical WGS sequencing. Integration of germline, somatic DNA and RNA-seq data enable data-driven variant prioritization and reporting, with oncogenic findings reported in 54% more patients than standard of care. These results establish key technical considerations for the implementation of cWGTS as an integrated test in clinical oncology.

<sup>1</sup>Department of Pediatrics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. <sup>2</sup>Department of Epidemiology & Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. <sup>3</sup>Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. <sup>4</sup>Precision Pathology Biobanking Center, Memorial Sloan Kettering Cancer Center, New York, NY, USA. <sup>5</sup>Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA. <sup>6</sup>Integrated Genomics Operation Core, Center for Molecular Oncology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. <sup>7</sup>Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. <sup>8</sup>These authors contributed equally: Shukla N, Levine MF, Gundem G. <sup>9</sup>These authors jointly supervised this work: Kung AL, Papaemmanuil E. ✉email: [kunga@mskcc.org](mailto:kunga@mskcc.org); [papaemme@mskcc.org](mailto:papaemme@mskcc.org)

Cancer is caused by the accumulation of somatic variants, including point mutations, structural variants (SVs), and copy number alterations (CNAs) that drive oncogenesis, disease progression, and in some cases define therapeutic vulnerabilities. The introduction of next-generation sequencing (NGS)-based targeted gene-panel assays has aided disease diagnosis, guided care, and improved patient outcomes through refinement of treatment options<sup>1–5</sup>. However, targeted panels are optimized to assess clinical biomarkers in common cancers<sup>1,2,4</sup>. Recent studies showed the utility of whole-exome sequencing (WES) in identifying coding mutations in rare cancer genes<sup>6,7</sup>. However, for patients with pediatric or rare cancers that have low mutation burden, distinctive methylation profiles<sup>8,9</sup> are primarily driven by SVs and fusion genes, panel/WES tests fail to identify a clinical biomarker in most cases<sup>1–4,7</sup>. This underscores an unmet need for better diagnostic workflows to guide clinical management.

Cancer whole-genome and transcriptome sequencing (cWGTS) offers the opportunity to assess the full spectrum of germline and somatically acquired mutations, SVs and CNAs, along with quantification of tumor mutation burden (TMB) and genome-wide mutational patterns<sup>10</sup>. The likely clinical utility of cWGTS in pediatric and rare cancers is increasingly evidenced in recent literature<sup>7,11–16</sup>. However, clinical implementation of WGS in oncology is challenged by cost of sequencing, complexity of laboratory, and analytical workflows to process large-scale data within clinically relevant timeframes, concerns about the sensitivity of low-coverage WGS in detecting actionable mutations captured by high-depth panel assays, and the interpretability of cWGTS findings with regard to clinical utility<sup>3</sup>. Here, we demonstrate the feasibility, analytical validity, and resolve critical technical considerations for the implementation of cWGTS in primary cancer care in the context of pediatric, adolescent, and young adult solid tumor patients with rare cancers.

## Results

**Sample processing.** The study cohort included patients presenting with primary diagnostic or relapse/refractory disease. Of 201 patient fresh frozen (FF) tumors nominated for paired cancer/normal whole-genome and transcriptome sequencing (cWGTS), 58 were excluded upon pathology review and 29 did not meet our requirement for >20% tumor purity as assessed by WGS. The majority of the excluded cases were post-therapy neuroblastoma and sarcoma samples with a predominance of necrotic disease. The final cohort included a single sample each from 114 pediatric, adolescent, and young adult patients (median age = 12.6 years, range: 4.5 months to 43.8 years) with solid tumors (Supplementary Tables 1, 2, Supplementary Fig. 1a–d).

**Implementation of a cWGTS workflow for clinical decision support.** To prototype a clinical cWGTS workflow, we developed an end-to-end process (Fig. 1a) that included dedicated: 1. project-management team, 2. lab operators for sample processing, 3. sequencing machines for cWGTS, 4. data-import channel, 5. Biosciences platform for automated deployment of analysis pipelines and API integration with institutional and public databases<sup>17</sup>, 6. reserved computing nodes in a high-performance computing environment, and 7. systematic pipeline for prioritization and reporting of genomic findings (Fig. 1a, Supplementary Fig. 1e). We quantified the end-to-end time from sample acquisition to the generation of an automated report. Time logs were audited starting at the time of surgical biopsy submission to report delivery for review by an interdisciplinary molecular tumor board for samples in our study with audit trails recorded through our biosciences platform (Supplementary Fig. 1e). End-to-end,

this workflow was executed on average in 17 days during the developmental phase (range: 11–29,  $n = 59$  samples), reaching a fully optimized workflow with a final turnaround of 9 days (Fig. 1b,  $n = 16$ ). This is shorter than the standard turnaround time (TAT) for many clinical NGS-panel sequencing tests (2–4 weeks)<sup>1,2,4</sup> markedly faster than the majority of WGS-processing timeframes in literature (3–8 weeks, Fig. 1b)<sup>11–16</sup> and comparable to the TAT achieved by centralized infrastructures of scale such as the Hartwig Foundation<sup>18</sup>. This demonstrates the feasibility of implementing cWGTS profiling to support diagnosis and treatment decisions with a clinically relevant turnaround time within a comprehensive cancer care center.

## Comprehensive genome characterization utilizing cWGTS.

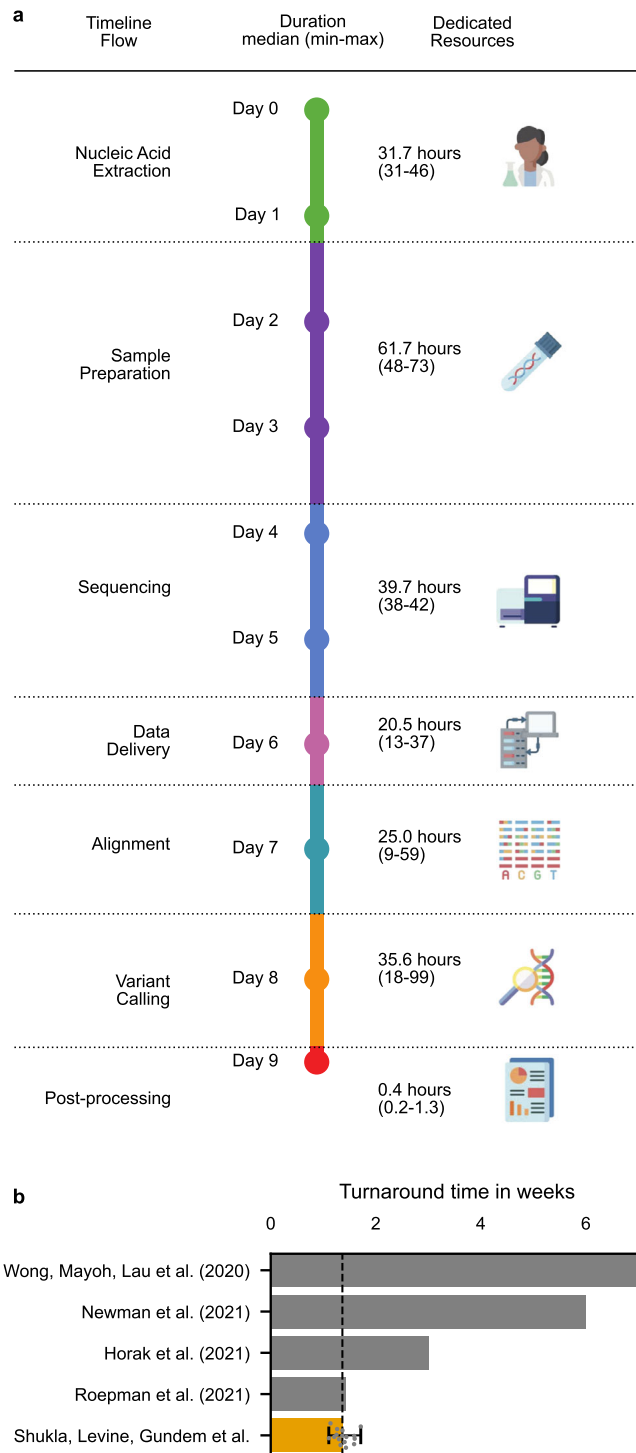
Across all mutational classes, cWGTS identified on average 7353 acquired mutations per sample, including cancer-associated alterations in 99% ( $n = 114$ ) of patients (Supplementary Fig. 2). These include CNAs ( $n = 105$  patients), germline predisposition ( $n = 17$ ), mutations in cancer-associated genes ( $n = 77$ ), translocations/fusion transcripts ( $n = 27$ ), disease-associated SVs ( $n = 75$ ), and outlier TMB or microsatellite-instability (MSI) scores ( $n = 7$ ) (Supplementary Table 3). Further signals of interest included the delineation of mutation signatures<sup>19</sup>, detection of chromothripsis<sup>20</sup> or whole-genome duplication (WGD)<sup>21</sup>, cancer-associated viral sequences (i.e., EBV)<sup>22,23</sup>, estimation of telomere length<sup>24</sup>, and gene expression signatures. SVs, most of which can only be detected using WGS, represented the third most frequent class of genomic alterations.

## Concordance analysis of cWGTS to targeted DNA- and RNA-panel tests.

Within our cohort, targeted DNA profiling of corresponding formalin-fixed paraffin-embedded (FFPE) biopsies by MSK-IMPACT<sup>4</sup> detected actionable biomarkers as defined by OncoKb Levels 1–4<sup>25</sup> in 24% of patients ( $n = 27$ ) (Fig. 2a–c, Supplementary Table 4). Consistent with prior findings demonstrating that patients with rare cancers do not yield clinically relevant biomarkers by panel sequencing<sup>4</sup>, most patients in our cohort (76%,  $n = 87$ ) had no therapy-informing alterations. These results are representative of the expanded pediatric/young-adult patient population at MSK (Supplementary Fig. 3a, b).

We first assessed whether mutations captured by MSK-IMPACT were also detected by WGS. For all discordant samples, we performed MSK-IMPACT on the same DNA aliquots used to generate the WGS libraries. This allowed us to ascertain whether discrepant calls were owing to differences in assay sensitivity (MSK-IMPACT and WGS) or a consequence of intratumor heterogeneity (ITH)<sup>26</sup>.

Of 221 somatic mutations reported by MSK-IMPACT, 174 (79%) were called in WGS (Fig. 2d, e). This includes 68/83 (82%) mutations reported by MSK-IMPACT as oncogenic<sup>25</sup> (Supplementary Fig. 3c). Variants called by both assays ranged from 5% to 97% variant allele frequency (VAF) with high concordance ( $r^2 = 0.75$ ) in VAF estimates (Fig. 2d). The majority of discordant mutations (46/47) were subclonal in MSK-IMPACT (<90% of cancer cell fraction) and 15 were classified as oncogenic (Supplementary Table 4, Supplementary Fig. 3c). Discordant mutations presented with a broad range of VAF (range: 2.2–39%, median = 8.5%) (Fig. 2d) and showed no systematic bias in effective coverage (Supplementary Fig. 3d). The 47 discordant mutations were confined to 26 samples (range 1–7 mutations per patient). Targeted resequencing of the WGS libraries by MSK-IMPACT was performed for 44 discordant variants and none were called, despite a median local depth of sequencing at 469x, supporting ITH as the basis of the discrepancies (Supplementary Fig. 3e). Further corroborating ITH, WGS and targeted



**Fig. 1 End-to-end cWGTs workflow.** **a** Schematic representation of the end-to-end cWGTs workflow, with information on median-time duration (in hours) for each step, as determined by a time trial over four consecutive batches containing  $n = 16$  tumors and representation of dedicated resources necessary to execute the workflow. **b** Comparison of best-reported turnaround times in literature, from sample collection to results ready for tumor board review. For our study, we show an orange bar denoting median time for  $n = 16$  samples with minimum and maximum times denoted with the error bar. These samples were processed post optimization.

resequencing data of the same FF DNA aliquots identified 10 mutations (VAF: 6–31%), which were not reported by MSK-IMPACT from the patient-matched FFPE sample. Three of the 10 additional calls were cancer-associated variants (TP53 L265Yfs\*81, PPM1D S468\*, and HLA-A L102Hfs\*73) (Supplementary Fig. 3e).

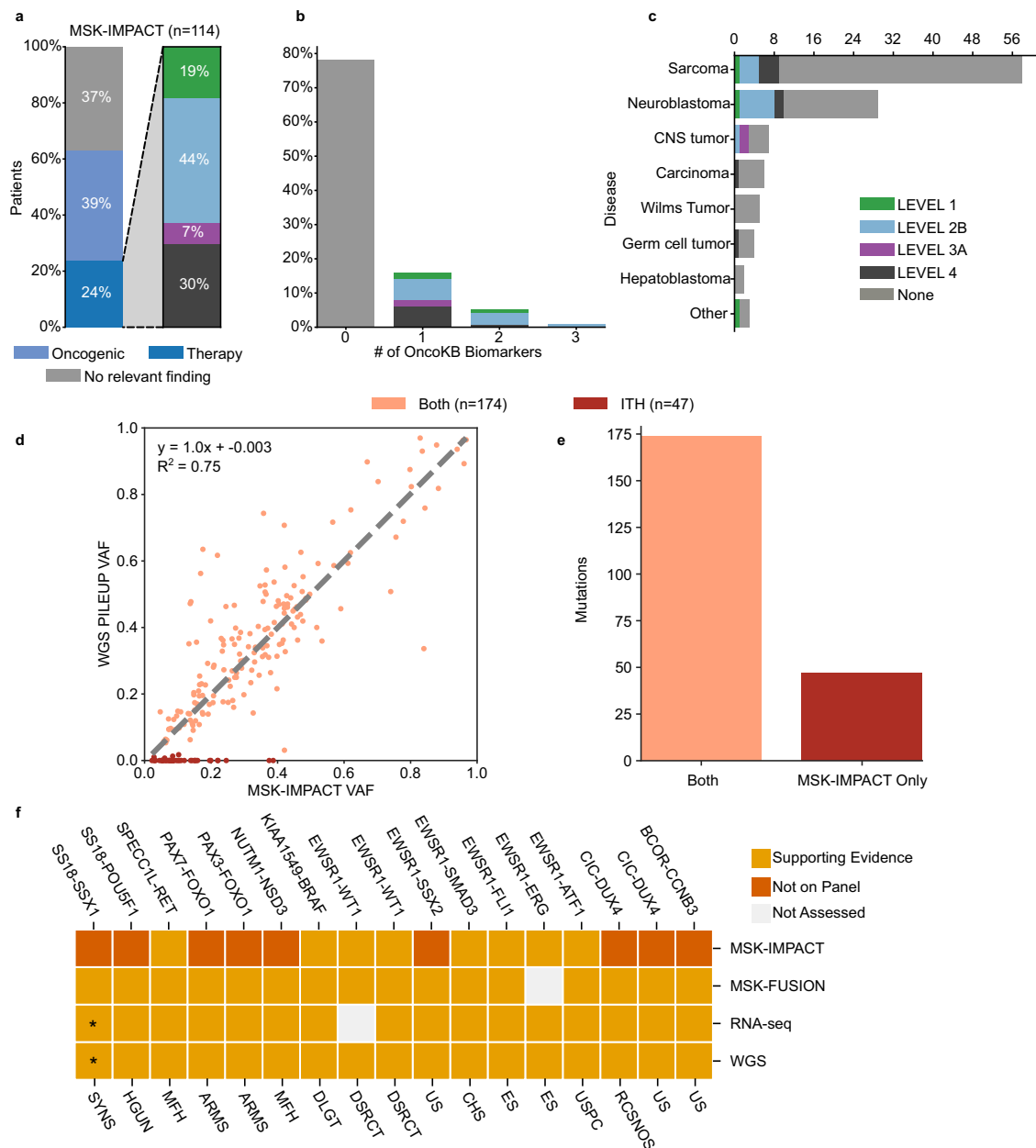
These validation studies demonstrated that discordant calls were due to stochastic sampling of heterogeneous tumors<sup>26</sup> (Supplementary Table 4) and concordance between WGS and MSK-IMPACT is at least 94% (43/47 total discordant) and up to 100% for all 43 mutations in evaluable samples when the same DNA aliquot was used in both assays.

Germline assessment by MSK-IMPACT<sup>27</sup> identified predisposition variants in 13 patients and panel RNA assessment with MSK-Fusion<sup>28</sup> identified oncogenic fusion genes in 18 (Supplementary Tables 5, 6, Fig. 2f). cWGTs captured all 13 germline-predisposition variants and 18 fusions. Importantly, fusion genes were supported by data in both WGS and RNA-seq, which offers the opportunity to orthogonally validate findings within a single workflow (Fig. 2f). These findings demonstrate that cWGTs as an integrative assay allows for the detection of germline, somatic mutations and fusion genes captured by an array of standard-of-care diagnostic tests.

**Technical considerations: optimal depth of coverage for clinical sequencing.** Sensitivity for somatic variant detection is directly dependent on the tumor cellularity of the biopsy and depth of sequencing coverage. The median cWGS depth was 95x (range 67–181) and tumor purity ranged from 21% to 100%, resulting in a median effective coverage of 64x (median depth \* purity estimate). To evaluate optimal depth of sequencing, for each of 97 tumors with WGS coverage  $\geq 60\times$ , we generated 298 derivative subsampled BAM files in the range of 100x, 80x, 60x, and 30–40x (Supplementary Fig. 4a, Supplementary Table 7). De novo variant calling was performed to assess sensitivity of detection for clinically relevant findings by MSK-IMPACT and WGS ( $n = 220$ ), genome-wide mutations across variant classes, and TMB (Fig. 3a, Supplementary Fig. 4b, c). Detection sensitivity correlated with effective coverage and was affected by variant class with slightly less sensitivity for SVs (Fig. 3a). Of the oncogenic findings, >91% were captured at 30–40x and >98% were recalled at 60–100x (Fig. 3a). With lower sequencing coverage, the power to detect subclones is limited (VAF range: 4.3–31%, median = 9.5%) (Fig. 3b, c). Optimal sensitivity for genome-wide mutation calling across variant classes was attained at  $\geq 80\times$  and increased with coverage. Figure 3c provides an overview of variant-detection sensitivity by depth of sequencing coverage, tumor purity, and variant clonal representation.

**Findings of biological and clinical relevance detected by cWGTs only.** The clinical relevance of cWGTs findings that were not identified by clinical panel sequencing (MSK-IMPACT<sup>4</sup>, MSK-fusion<sup>28</sup> and panel testing of 88 cancer-predisposition genes<sup>27</sup>) was determined by a multidisciplinary molecular tumor board. Consistent with recent studies<sup>7,11–16</sup>, cWGTs analyses identified at least one additional cancer-associated oncogenic variant in 54% of patients ( $n = 62$ ). Of these, 33 patients had one or more findings that were of direct clinical relevance, including 7 diagnostic (21%), 15 prognostic (45%), 5 therapy-informing (15%), 5 previously undescribed oncofusions (15%), and 6 germline (18%) biomarkers (Fig. 4a, Supplementary Table 3). Most additional relevant findings were explained by the detection of SVs and fusions and genome-wide mutation signatures (Fig. 4b).

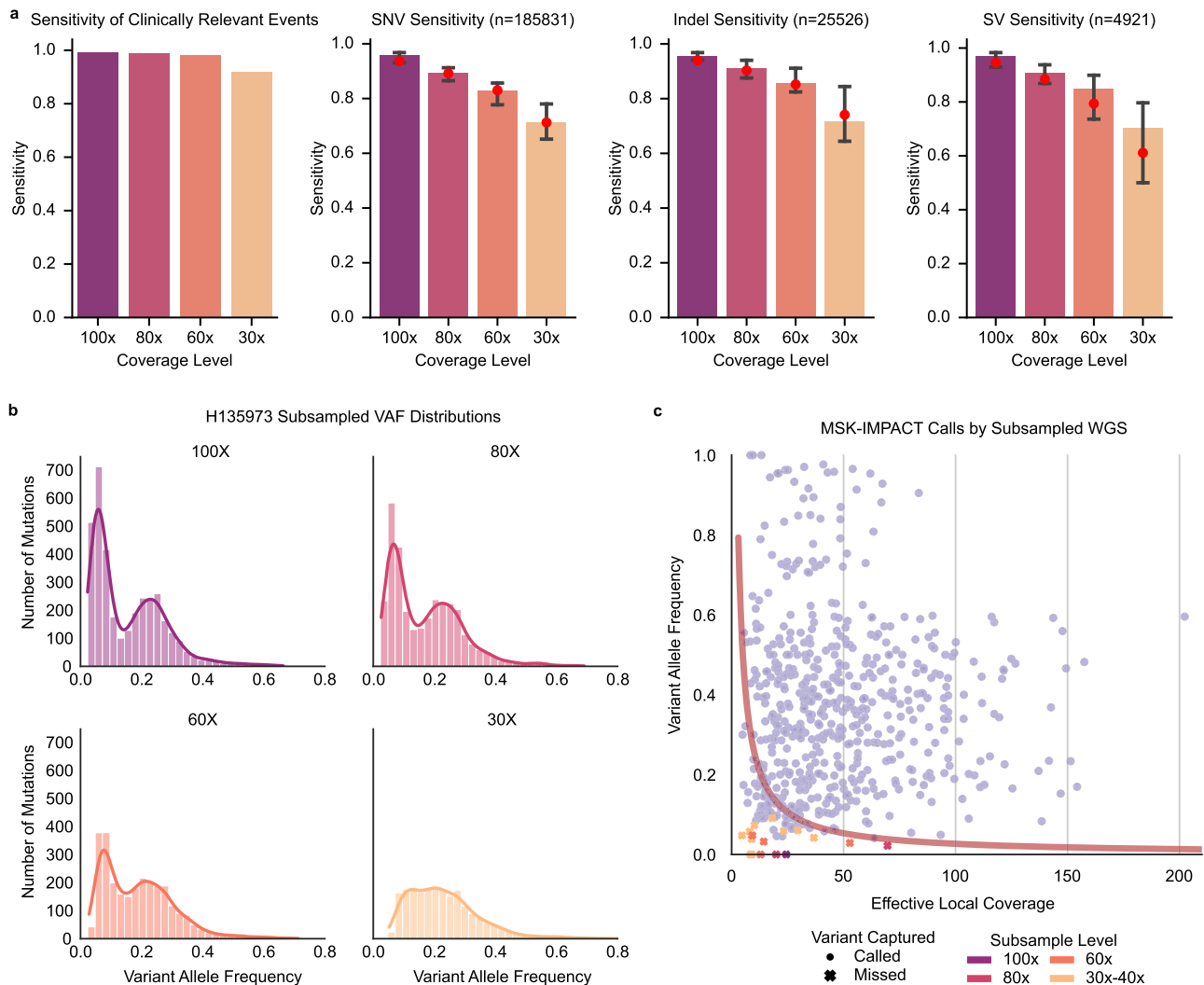
We further inferred the portion of additional findings that would be captured by WES by masking results to the coding regions of genes. Of the 62 patients with incremental findings by



**Fig. 2 Analytical validity of cWGTS for clinical biomarkers.** **a** The left barplot depicts the proportion of patients with therapy-informing, oncogenic, or no relevant findings reported by MSK-IMPACT as defined by OncoKb (Levels 1–4). The right barplot shows the breakdown (0,1,2) of the highest level of OncoKb annotation in the study cohort. **b** Barplot demonstrating breakdown of the highest OncoKb level by the number of informative biomarkers in study cohort. **c** Barplot demonstrating breakdown of the highest OncoKb level by disease class. **d** Scatterplot shows the comparison of variant allele frequency (VAF) of MSK-IMPACT variants as reported by MSK-IMPACT (x axis) and absolute VAF estimates by pileup in WGS data (y axis) (Pearson correlation). Discrepant mutations are observed along the x axis. Mutations are color-coded by call status, where Both is called in both assays and ITH is mutations that were not called in higher- depth resequencing and/or had proportion test  $p$ -value  $< 0.05$ . **e** Barplot demonstrating breakdown of MSK-IMPACT mutations, observed in both WGS and MSK-IMPACT or only MSK-IMPACT (ITH). **f** Validation of oncogenic fusions reported by MSK-IMPACT/MSK-Fusion in cWGTS. The asterisk indicates that the *SS18-SSX1* that was reported by MSK-Fusion was reported as *SS18-SSX2* and supported by spanning reads in WGS. Main oncogene disease code listed underneath for each patient (ARMS alveolar rhabdomyosarcoma, CHS chondrosarcoma, DLGT diffuse leptomeningeal glioneuronal tumor, DSRCT desmoplastic small round-cell tumor, ES Ewing sarcoma, MBL medulloblastoma, MFH undifferentiated pleomorphic sarcoma/malignant fibrous histiocytoma/high-grade spindle-cell sarcoma, RCSNOS round-cell sarcoma, NOS, SYNS synovial sarcoma, US undifferentiated sarcoma, USPC undifferentiated sarcoma of the peritoneal cavity). Source data for panels a–e and f are provided in Supplementary Data 4 and 6.

cWGTS, RNA-seq and WES alone would only capture events in 10 (16%) and 8 ( $n = 13$ ) patients, respectively, or in 17 patients when combined (Fig. 4a, Supplementary Table 3). Thus, only 27% of the findings in cWGTS could be captured by WES and RNA-seq as the majority of additional findings were attributed to SVs.

**Rare variants in established cancer genes.** We identified seven clinically relevant findings targeting known rare cancer genes (Supplementary Tables 5, 8). Of these, three were somatically acquired and included a disease-defining mutation of *KBTD4* (p.R313\_M314insPRR) in a pineal parenchymal tumor of



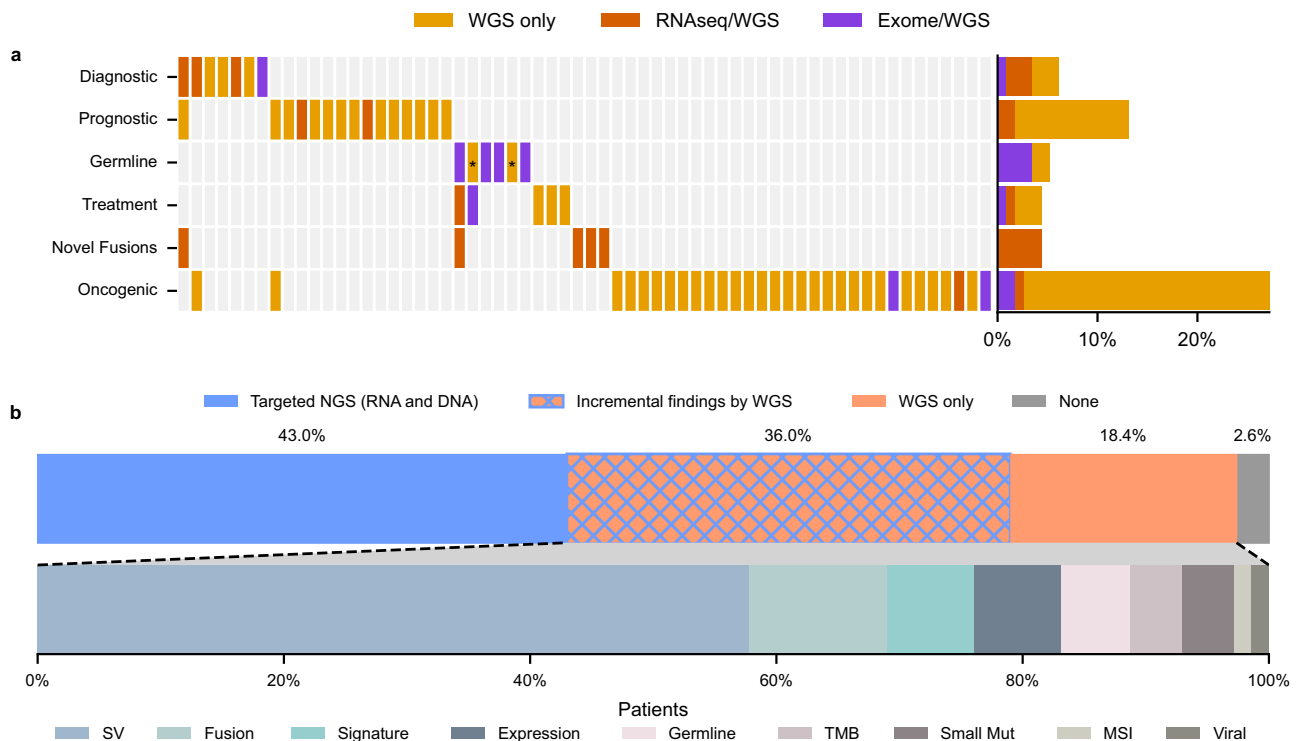
**Fig. 3 Assessment of optimal coverage for WGS.** **a** Barplots demonstrating sensitivity of variant detection and 95% confidence intervals (error bars) by coverage depth (100x, 80x, 60x, and 30–40x) from left to right for: 1. clinically relevant events detected by MSK-IMPACT and WGS ( $n = 220$ ), 2. genome-wide SNVs, 3. genome-wide indels, and 4. genome-wide SVs. Only data from samples with original median coverage  $>100x$  ( $n = 32$ ) are shown. Red dots indicate overall sensitivity of all mutations across all BAMs at the same subsampling level. **b** Histograms of variant allele frequencies for each subsampling level for a representative sample in the study cohort (H135973), showing loss in sensitivity to detect subclonal mutations at lower sequencing depth of coverage. **c** Scatterplot of effective local coverage vs VAF in subsampled BAMs for the clinically relevant calls from MSK-IMPACT. Variants called in subsampled BAMs are shown with circles, while the missed variants are denoted with X's. Trendline shows the cumulative binomial distribution for obtaining at least 2 variant reads, given the effective coverage and variant allele fraction. Source data for panels **a**, **c** are provided at the data repository. Raw data for panel **b** can be accessed at the dbGAP study.

intermediate differentiation<sup>29,30</sup>, a *SETBP1* (p.D868N) mutation in a germ cell tumor, and a *SIX1* mutation (p.Q177R) in a Wilms tumor<sup>31</sup>. Additionally, clinically relevant germline variants were detected in four cancer-associated genes, including an *SBDS* splice-site mutation (c.258 + 2T > C) in a rhabdomyosarcoma, *BARD1* p.E652fs\*69 in a neuroblastoma, *EP300* p.A2259fs\*20, and *EXT2* p.W414\* in two osteosarcoma patients (Supplementary Table 5). These results demonstrate the utility of cWGTs in capturing somatic and germline variants in rare cancer genes not routinely evaluated in targeted panels<sup>2,4</sup>.

**Fusion genes.** Eight in-frame fusion genes were identified from WGS and RNA-seq in patients with no prior findings on clinical testing (Supplementary Fig. 5, Supplementary Table 6), 5 of which were not described before. Of diagnostic relevance, we identify: 1. a t(2;6) (*PAX3-FOXO3*) translocation changing

diagnosis to alveolar rhabdomyosarcoma (ARMS) in a patient who was diagnosed with embryonal rhabdomyosarcoma (ERMS) in the absence of the cardinal ARMS fusions (*PAX3-FOXO1* and *PAX7-FOXO1*)<sup>32</sup> (Fig. 5a, b), 2. a *UACA-LTK* fusion in a metastatic papillary thyroid carcinoma<sup>33</sup>, and 3. a pathognomonic *SH3PXD2A-HTRA1* fusion establishing a diagnosis of schwannoma in a patient evaluated for relapsed stage-IV neuroblastoma<sup>34</sup>.

Of potential therapeutic relevance, we identified an *NTRK3-SLMAP* fusion in a neuroblastoma patient. Activating *NTRK3* fusions are promising therapeutic targets for TRK inhibitors, with activity seen across pediatric and adult cancers<sup>35</sup>. However, screening for *NTRK* fusions is not routinely performed across all disease indications. Additional undescribed fusions included *EPC2-AFF3* and *MAN1A2-ACBD6* identified in two patients with undifferentiated sarcoma, and a *CITED2-MGA* fusion in a round-cell sarcoma not otherwise specified.



**Fig. 4 Additional relevant findings detected by cWGTS as compared with standard of care. a** Heatmap of additional relevant findings by cWGTS colored by what technology (WES, WGS, and RNA-seq) may detect each event. Columns represent patients, while rows are clinical event types. The asterisks for Germline indicate pathogenicity supported by mutational signatures. **b** (top) Stacked-bar breakdown of patients with clinically relevant findings by assay. The blue areas (solid or meshed) represent patients with relevant findings from targeted sequencing (RNA and DNA), while the orange areas (solid or meshed) are for patients with findings from cWGTS. The blue/orange mesh indicates patients that had relevant findings from both targeted sequencing and WGS. (bottom) Stacked-bar breakdown of findings specific to cWGTS from the patients in the orange section (solid or meshed) from top. The relevant findings are colored by event type. SV, structural variant. TMB, tumor mutation burden. MSI, microsatellite instability. Small Mut, small mutations, including substitutions and insertion/deletions. Viral, viral integration. Source data for panels **a**, **b** are provided in Supplementary Data 3.

**Structural variants targeting tumor suppressor genes.** Structural variations of established prognostic relevance<sup>36</sup> were observed in our cohort providing insights of clinical relevance. cWGTS mapped events in *TERT* and *ATR1* in 8 (28%) and 5 (17%) neuroblastoma patients, respectively. Both *TERT* and *ATR1* are increasingly considered as therapy-defining risk-stratification biomarkers for neuroblastoma<sup>37</sup>. The *TERT* SVs could only be identified by cWGTS, and only ¼ of the *ATR1* deletions were reported by MSK-IMPACT<sup>4</sup>.

We also observed recurrent SVs targeting the tumor suppressor gene *DLG2* in 15/29 OS patients<sup>38</sup> and 3/29 neuroblastoma, of which 6 had homozygous deletions (Supplementary Table 9). While *DLG2* has been characterized in osteosarcoma<sup>38</sup>, our findings demonstrate that *DLG2* SVs are also recurrent in neuroblastoma warranting further investigation in future studies.

#### Integration of RNA-seq and WGS for variant annotation.

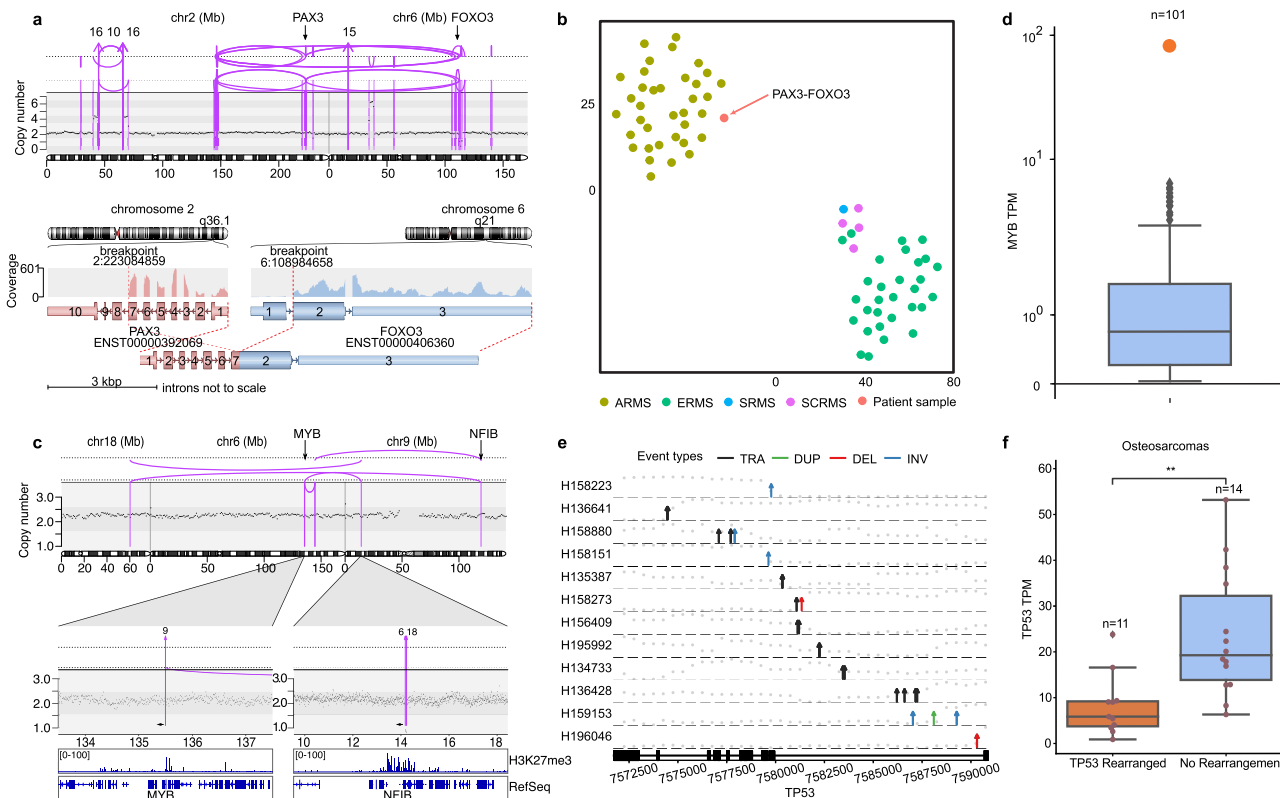
Interpretation of complex SVs in noncoding regions of the genome presents a major challenge for reporting of WGS findings. cWGTS enables the concomitant detection of SVs and assessment of the transcriptomic consequences of the affected loci. For example, a chromoplexy event resulting in overexpression of the *MYB* oncogene<sup>39</sup> through “hijacking” of an *NF1B* enhancer (Fig. 5c, d) was detected in an adenoid cystic carcinoma without informative clinical sequencing findings. While *MYB* overexpression is a cardinal feature of adenoid cystic carcinomas, *MYB* fusions are identified in only 30% of cases using conventional diagnostic assays<sup>39</sup>. Integration of gene expression data was

critical to the annotation and reporting of this complex non-coding SV as the disease defining diagnostic biomarker.

Similarly, among 29 osteosarcoma patients, we identified *TP53* mutations in 12 and mapped noncoding SVs targeting the *TP53* locus in 13. Of these, only 3 were reported by MSK-IMPACT (Fig. 5e). Integration with RNA-seq demonstrated that *TP53* SVs correlated with loss of *TP53* expression, validating their functional relevance (Fig. 5f). Wild-type *TP53* represents an inclusion criterion for p53 pathway modulating drug trials<sup>40</sup>. Here, we show that in the absence of cWGTS, patients with loss of *TP53* by SVs, which have been described in diverse cancers, could be erroneously diagnosed as *TP53* wildtype with implications for assessment of treatment options<sup>40</sup>. We did not identify germline SVs targeting the *TP53* locus.

Taken together, our findings illustrate the necessity to combine RNA and DNA analyses in variant detection, annotation, and prioritization for clinical cWGTS reporting. In our automated workflow, we interrogate DNA mutations for corroborating evidence in the RNA. All recurrent fusion genes reported by panel RNA-seq assays as well as the 8 additional driver fusion genes were orthogonally detected by both WGS and RNA-seq (Supplementary Fig. 5). Furthermore, integration of gene expression data to SV findings resolved functional consequences of SVs targeting noncoding regions on the genome (e.g., *MYB* enhancer hijacking and *TP53* inactivation).

Global gene expression signatures were further used to cluster samples by tumor type, providing further opportunity to resolve a patient’s diagnosis (Supplementary Fig. 6a). Last, in the 101 patients with RNA-seq data, we identified on average 18 gene expression



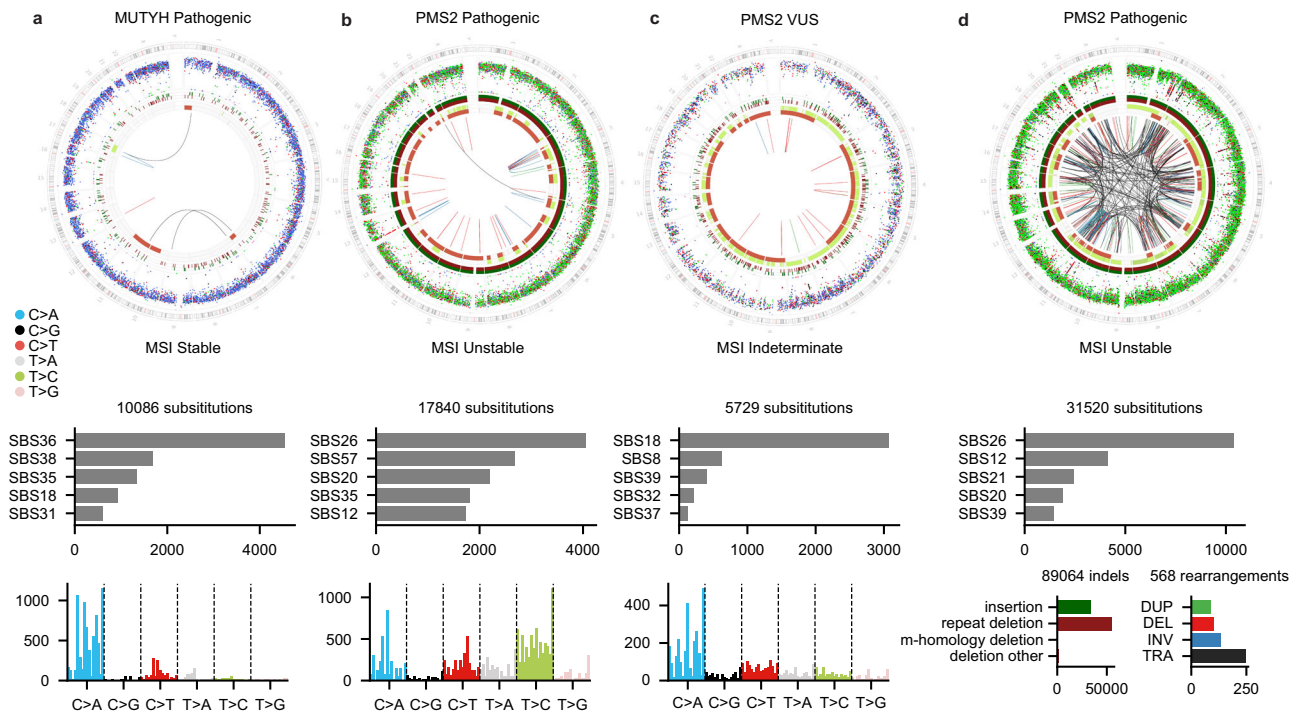
**Fig. 5 Integration of DNA and RNA findings for variant annotation.** **a** Top panel shows absolute copy number on the y axis and the structural variants (SVs) that result in *PAX3-FOXO3* fusion in patient H134768. Lower panel displays RNA fusion product created by the corresponding genomic SVs. **b** tSNE clustering of methylation data from rhabdomyosarcoma samples color-coded by disease subtype (ARMS: alveolar, ERM: embryonal, SCRMS: spindle cell, and SRMS: sclerosing). The patient harboring the *PAX3-FOXO3* fusion clusters with the ARMS samples. **c** Top panel shows the chromoplexy event among chromosomes 6, 9, and 18, resulting in the localization of the *NFIB* enhancer to the *MYB* locus in patient H133676. Lower panel displays H3K27me3 chromatin marks from Drier et al., Nature Genetics 2016. **d** Boxplot shows the *MYB* expression in transcripts per million (TPM) across the cohort. Center line indicates the median and whiskers extend within  $\pm 1.5$  the interquartile range (IQR) from the box. The patient with *MYB-NFIB* event (H133676) is highlighted in orange, demonstrating that the SV event in panels **c** associates with overexpression of *MYB*, validating the SV as an enhancer-hijacking event. **e** Diagram of SV events targeting *TP53* gene body in osteosarcoma patients ( $n = 12$ , the 13th patient's event breakpoints fall outside of the gene body). SVs are shown as arrows with absolute copy number on the y axis (gray dots) overlaid over the exonic structure of *TP53* (TRA: translocation, DUP: duplication, DEL: deletion, INV: inversion). **f** Boxplot shows the comparison of *TP53* expression in RNA between *TP53*-rearranged samples and those without any rearrangement with a center line indicating the median and whiskers extending within  $\pm 1.5$  x the IQR (two-sided Mann-Whitney U test,  $p = 1.645e-03$ ). Raw data for panel **a-c** can be accessed at the dbGAP study. Source data for panel **e** are provided in Supplementary Data 9. Source data for panels **d, f** are provided at the data repository.

biomarkers per sample (range = 1–99)<sup>16</sup> (Supplementary Table 10). However, the evidence with regard to the clinical utility of such expression biomarkers remains to be validated. To this end, we performed a systematic interrogation of genes with aberrant expression with known tissue-specific expression patterns and SVs or fusions detected from cWGTs. Overall, only 8% of the expression biomarkers were associated with an acquired SV or fusion gene, demarcating a subset of high-confidence expression biomarkers (average per patient = 1, range = 0–10). This limits the number of patients in our cohort with an expression biomarker supported by an SV to 54% ( $n = 55$ ) (Supplementary Fig. 6b). Demonstrating the utility of this integrative analysis, we identified a concomitant *KRAS* amplification and overexpression in two patients with no clinically relevant biomarkers (H135462 and H195916) by clinical testing. Of note, in both patients, the amplification was not reported by MSK-IMPACT pointing to the lower sensitivity to detect copy number changes by panel-based assays.

**Integration of germline mutations to somatic mutation signatures for variant annotation.** Annotation of germline variants is restricted to recurrent events in population databases, thus

limiting interpretation for rare founder events. For each genome in our cohort, we quantified the proportion of mutations attributed to each of 73 reference mutation signatures<sup>10</sup>. Two of three patients with a germline mutation in DNA repair genes further harbored mutation profiles suggestive of DNA repair deficiency. Patient H135421 had a pathogenic variant in *MUTYH* and somatic loss of the second allele. About 42% of the mutations were attributed to the *MUTYH* signature SBS36<sup>10</sup> (Fig. 6a). Patient H135466 had a pathogenic variant in *PMS2* (c.538-1G > C) with loss of the wild-type allele by LOH. The tumor was MSI high with hypermutation (TMB = 11.23, indels = 90,246, SNVs = 17,840, and SVs = 44), enrichment of T > C mutations, and repeat-mediated indels characteristic of *PMS2* deficiency<sup>10</sup> (Fig. 6b). In contrast, patient H135073 harbored a variant of unknown significance (VUS) in *PMS2*, a medium MSI score (7.23), and low mutation burden (1.30 Muts/Mb) without evidence of a *PMS2* signature (Fig. 6c). These findings demonstrate the utility of mutation signatures in the assessment of germline mutations in DNA repair genes.

To illustrate this point, in a 12-year-old osteosarcoma patient outside the study cohort, cWGTs characterized a hypermutated genome (TMB = 16.7, indel = 89,588, SNVs = 31,520, and



**Fig. 6** Genome-wide distribution and patterns of somatic mutations for four different patients. **a** Neuroblastoma patient (H135421) harboring a pathogenic germline *MUTYH* variant (c.924 + 3A > C). **b** Immature teratoma patient (H135466) with a pathogenic germline *PMS2* mutation (c.538-1G > C). **c** Malignant peripheral nerve sheath tumor patient (H135073) harboring a germline *PMS2* variant of unknown significance (VUS) (p.W841\*). For each patient, the top panel is a Circos plot showing the different types of somatic mutations along the genome. The outermost ring shows the intermutation distance for all SNVs color-coded by the pyrimidine partner of the mutated base. The middle ring shows small insertions (green) and deletions (red). The innermost ring shows copy number changes, and the arcs show SVs. Middle panel is a barplot showing the absolute number of mutations attributed to the five mutational signatures with the highest exposure in the tumor. Bottom panel is a barplot showing the 96 trinucleotide contexts of SNVs. **d** Genome-wide distribution and patterns of somatic mutations identified in the patient outside the cohort with recurrent osteosarcoma (H201472). WGS results show the sample is hypermutated, with enrichment in SBS26, T > C mutations, repeat-mediated deletions, and MSI unstable. The patient was found to be harboring a pathogenic *PMS2* variant (p.D699H) (repeat deletion: repeat-mediated deletion, m-homology: microhomology-mediated deletion, deletion other: all other deletions, TRA translocation, DUP duplication, DEL deletion, INV inversion). Raw data for this figure can be accessed at the dbGAP study.

SVs = 568) enriched in repeat-mediated deletions consistent with MSI high status (Fig. 6d). This observation prompted consent for germline testing, resulting in identification of a *PMS2* mutation (p.D699H) annotated as likely pathogenic/VUS<sup>41</sup> and a somatic loss of the wild-type allele. MSK-IMPACT reported an indeterminate MSI status (7.5) yet upon testing validated that the germline mutation is pathogenic.

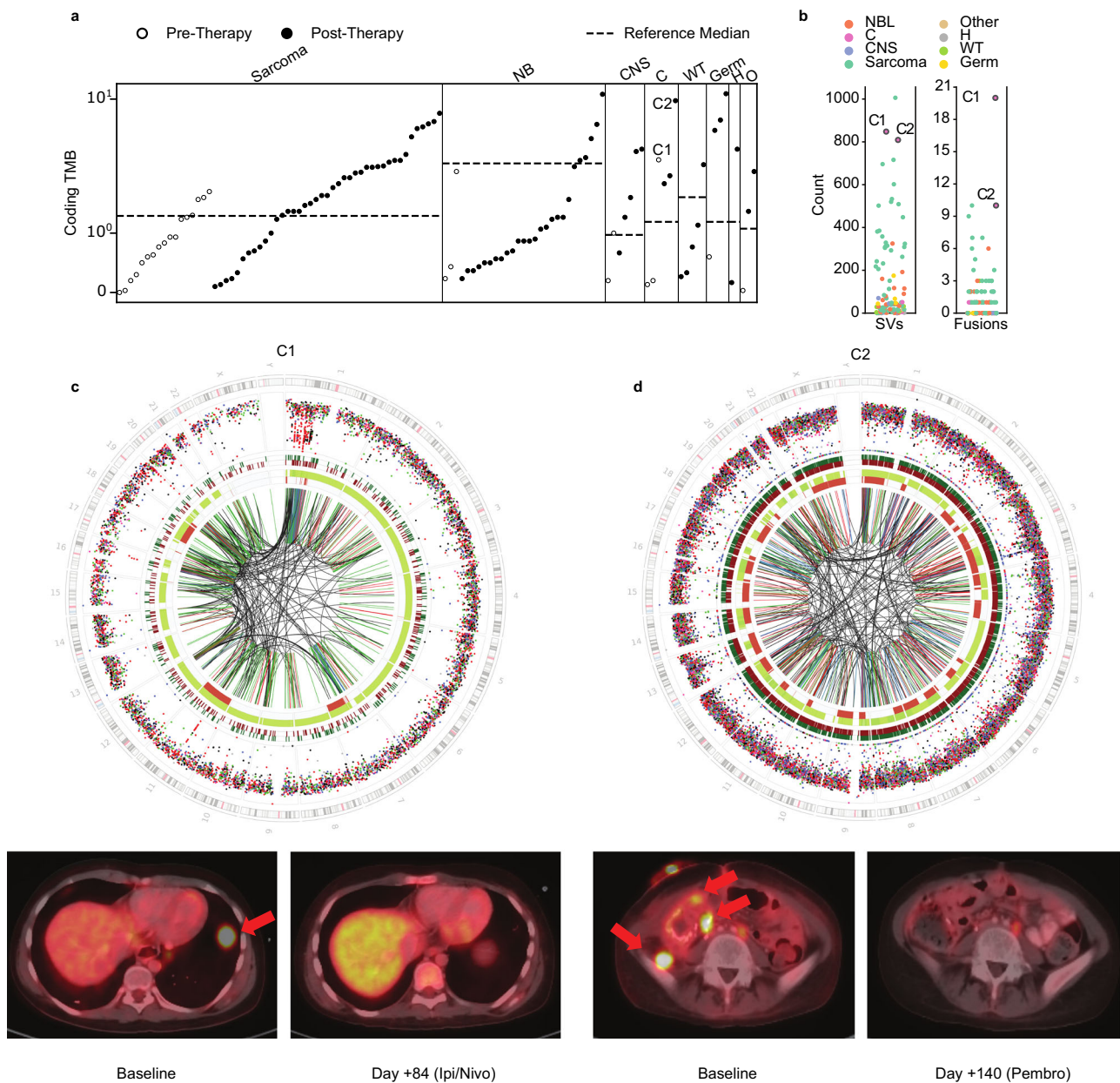
These results demonstrate the utility of integrating composite readouts from cWGTs (germline mutation, allele-specific copy number, and genome-wide TMB) to deliver corroborating evidence for the assessment and reporting of germline-predisposition mutations with implications for family screening, diagnosis, and treatment.

**Genomic alterations of emerging biological and clinical relevance.** Recent studies propose telomere length as prognostic indicators in neuroblastoma among other cancers<sup>42–45</sup>. We recapitulated the established associations between *ATRX* and *TERT* mutations to telomere length<sup>46,47</sup> (Supplementary Fig. 7a–c). *ATRX* mutations were also observed in 8 osteosarcomas, with similar associations to telomere length (Supplementary Fig. 7c). Given the association between adverse risk mutations and telomere length, delineation of the independent prognostic value warrants analyses of data that concomitantly map these mutations, SVs, and telomere length alongside established predictors of outcomes.

We detect chromothripsis<sup>20</sup> in 40% of patients, most recurrently observed in sarcomas (35/58), germ cell tumors (2/4), and less frequently in neuroblastoma (6/29) (Supplementary Fig. 7d). Chromothripsis frequently led to *TP53* loss (10/29)<sup>20,48</sup>, amplification of *MYC*, *VEGFA*, and *MDM2*. Additionally, in 2 patients, chromothripsis resulted in oncogenic fusions (*MAN1A2-ACBD6* and *PAX3-FOXO3*) (Supplementary Fig. 7e). Previous studies have proposed an association between whole-genome duplication (WGD) and poor outcomes in cancer<sup>21</sup>. WGD was seen in 42/114 patients with an enrichment in sarcoma (24/54), carcinomas (3/7), and neuroblastoma (11/30) (Supplementary Fig. 7f).

**Biological and clinical implications of tumor mutation burden across variant classes.** Panel-based approaches derive estimates of TMB, MSI scores, and mutation signatures<sup>49</sup>, whereas WGS directly quantifies genome-wide mutation burden across all variant classes (Fig. 7a, b and Supplementary Fig. 8a). We observed higher overall (8.3-fold) TMB estimates in our cohort, relative to reports in pediatric cancer (Fig. 7a)<sup>10,48,50</sup>. TMB was higher in therapy-exposed compared with treatment-naive patient samples (0.1–11.2 in treated vs 0–2.7 treatment-naive, Mann–Whitney test,  $p = 1.892e-04$ ) and correlated with evidence of treatment-related signatures (i.e., temozolomide, platinum) (Supplementary Fig. 8b)<sup>10,51</sup>, observed in 45/114 patients pointing to persistence of clones that were exposed to and survived cancer therapy<sup>52</sup>.





**Fig. 7 Genome-wide mutational burden in the context of immunotherapy.** **a** Distribution of coding tumor mutational burden (TMB) as assessed by WGS across the cohort ( $n = 114$ ), colored by treatment status of the patient at the time of sampling. Dotted line indicates median-coding TMB (SNVs and indels) as previously reported by the Zero Childhood Cancer study. Patients are grouped by disease category (NB: neuroblastoma, CNS: central nervous system, C: carcinoma, WT: Wilms tumor, Germ: germ cell tumor, H: hepatoblastoma, O: other). Carcinoma patients C1 and C2 who responded to immunotherapy are labeled. **b** Distribution of structural variant (SV) (right) and gene fusion (left) burden across the samples with both WGS and RNA-seq available ( $n = 101$ ). Patient C2 had a poor-quality RNA sample, so clonal fusions from another time point from the same patient are shown. **c** (top) Genome-wide distribution and patterns of somatic mutations for tumor C1 (H135022), patient with metastatic adrenocortical carcinoma, depicting high SV burden. Circos plots are shown as described in Fig. 6. PET imaging shows resolution of a large pulmonary metastatic lesion (red arrow) following treatment with nivolumab and ipilimumab. **d** Genome-wide distribution and patterns of somatic mutations for H135462, a 14-year-old with relapsed refractory poorly differentiated clear-cell carcinoma with high TMB and SV burden. Circos plots are shown as described in Fig. 5. PET imaging shows resolution of multiple metastatic lesions (red arrows) following treatment with pembrolizumab. Source data for panels a and b are provided at the data repository. Raw data for panel c, d can be accessed at the dbGAP study.

Patients H135022 (adrenocortical carcinoma) and H135462 (clear-cell carcinoma) had progressive on-treatment metastatic disease and in the absence of therapy-informing biomarkers by clinical testing were at the end of their therapeutic options. cWGTS analyses revealed a profoundly rearranged genome scoring these two patients as the highest in fusion burden and SV burden in the cohort (Fig. 7b–d). H135022 was treated with checkpoint blockade

(nivolumab/ipilimumab), resulting in complete response after three cycles of therapy, and is disease-free 26 months after therapy cessation (Fig. 7c), whereas patient H135462 was treated with pembrolizumab, achieved a complete response after 6 cycles, and remains disease-free 10 months after therapy (Fig. 7d).

These findings demonstrate the value of cWGTS to fully assess the level of genomic instability across variant classes and

highlight the need to further evaluate SV and fusion gene burden as biomarkers of response to immune checkpoint blockade therapies.

### Derivation of comprehensive WGS profiling in cell-free DNA.

Our study evaluated key technical considerations of cWGTS in FF biopsies, as an optimal source of tumor DNA. However, limited biopsies may restrict access to cWGTS for all patients. Cell-free DNA (cfDNA) from blood plasma represents an alternative source of DNA for tumor profiling<sup>53</sup>. Recently, cfDNA NGS profiling including WGS has been used to detect tumor-specific CNAs and fragmentation patterns in pediatric tumors<sup>54–56</sup>. However, the potential of high-depth WGS in tissue-naïve identification of the genetic changes (SNVs, indels, and SVs) in a patient's cancer genome from cfDNA has been largely unexplored.

In an exploratory analysis, we performed WGS from matched FF samples and cfDNA from seven patients collected at the time of tumor biopsy (Supplementary Table 11). cfDNA genome-wide coverage ranged from 94 to 102× (Fig. 8a) with a wide range of tumor content in cfDNA ~10–83% (Fig. 8b). To assess the suitability of cfDNA for unbiased genome-wide mutation detection, we performed de novo mutation calling in cfDNA for all variant classes (SNVs, indels, SVs, and CNA) and compared the results to FF analyses.

WGS from cfDNA did not present technical limitations in data generation or false-positive variant calling across mutation classes. Derivation of high-quality variant calls was contingent upon the quantity of circulating tumor DNA (ctDNA). In four patients with ctDNA content sufficient for CNA detection, we establish good concordance between the FF and cfDNA CNA profiles (Fig. 8c, Supplementary Fig. 9a–g). Strikingly, for patients with high ctDNA content (i.e., IH158182), we derived a near-complete picture of the genome-wide mutation patterns demonstrating that cancer genomic landscape can be fully recapitulated by cfDNA WGS (Fig. 8d). Importantly, for patient H135967, we showcase that even with an estimated ctDNA content of 20%, the same threshold used for analyses of FF material, we can detect all the known oncogenic events in the FF sample across variant classes, which include a *TP53* substitution, *MYC* and *CCNE1* amplifications, and SVs targeting *ARID1A* and *ATRX* (Supplementary Fig. 9c).

We further demonstrate the potential of cfDNA to capture a comprehensive representation of different types of variants across the tumor phylogeny compared with solid biopsies (Supplementary Table 11) through the detection of cfDNA-specific subclones (Fig. 8e, Supplementary Fig. 9a–g). These results provide the proof-of-concept for the feasibility of deriving tumor-agnostic comprehensive WGS profiling from a liquid biopsy.

### Discussion

We present a comprehensive technical assessment for cWGTS implementation in clinical care practice in oncology. We demonstrate that using a single integrated workflow, cWGTS captures the full spectrum of cancer-associated genomic alterations that are assessed using a diversity of standard-of-care diagnostic assays. With implementation of best laboratory and computational practices, we execute an end-to-end sample-to-report turnaround time within 9 days, which is aligned to clinical needs for diagnosis and care decisions, and is comparable to infrastructures of scale<sup>18</sup>. Despite 5–10-fold lower sequencing coverage compared with panel-based assays, we demonstrate that in matched biopsies, cWGTS recovered all clinically reported variants by high-depth targeted profiling assays. We establish >80× coverage and tumor purity of at least 20% to attain this sensitivity. However, this sets a stringent quality threshold on

fresh frozen tumor specimens that are not as broadly available as FFPE. However, a major limitation of WGS in FFPE is a high error rate in genome-wide calls<sup>57</sup>. To this end, we provide proof-of-concept feasibility data demonstrating that comprehensive WGS profiling can also be leveraged in patients who have high cfDNA content in circulation at the time of diagnosis or relapse. Our findings pave the way for future studies focused on analytical validation and optimizations of comprehensive tumor-agnostic WGS profiling from cfDNA for diagnostic purposes.

To support cWGTS variant annotation and prioritization, we implemented an analytical workflow that learns from variant annotation databases and integrates signals from germline mutations, somatic DNA, and RNA-seq findings. This allows us to annotate, validate, and prioritize SVs of diagnostic (e.g., *MYB* enhancer hijacking), prognostic (e.g., *ATRX/TERT*), and therapeutic relevance (e.g., *TP53* loss-of-function SVs). Consistent with recent literature for pediatric and rare cancers<sup>7,15,16,58</sup>, >50% of patients had additional findings of established biological or clinical significance. The majority of these findings were SVs in cancer genes, fusion genes, and genome-wide mutation signatures that targeted panels are not optimally designed to identify. Importantly, we demonstrate that only a minority of such additional findings would be captured by WES and RNAseq alone or in combination. Larger cohort studies are warranted to determine the incidence and prevalence of clinically relevant biomarkers captured by cWGTS.

The clinical relevance of cWGTS extends beyond that of rare cancers<sup>59</sup>. We show that by cWGTS, we detect the full spectrum of cancer-associated mutations in 99% of patients. The vision of patient-tailored medicine warrants the delivery of clinical decisions that extend beyond a single druggable biomarker and rather consider the composite readouts from a patient's cancer genome that inform on a patient's a priori risk of developing cancer, diagnosis, likelihood of treatment response, risk of progression, and therapeutic vulnerabilities. With increasing implementation of cWGTS on well-annotated clinical specimens<sup>15,16,58</sup>, our ability to interpret cWGTS findings will improve, and by extension, the clinical utility of cWGTS will expand. As the economic barriers to cWGTS are mitigated in time, a single comprehensive assessment of the cancer genome is positioned to replace multiple targeted diagnostic tests in prospective clinical sequencing<sup>59</sup>.

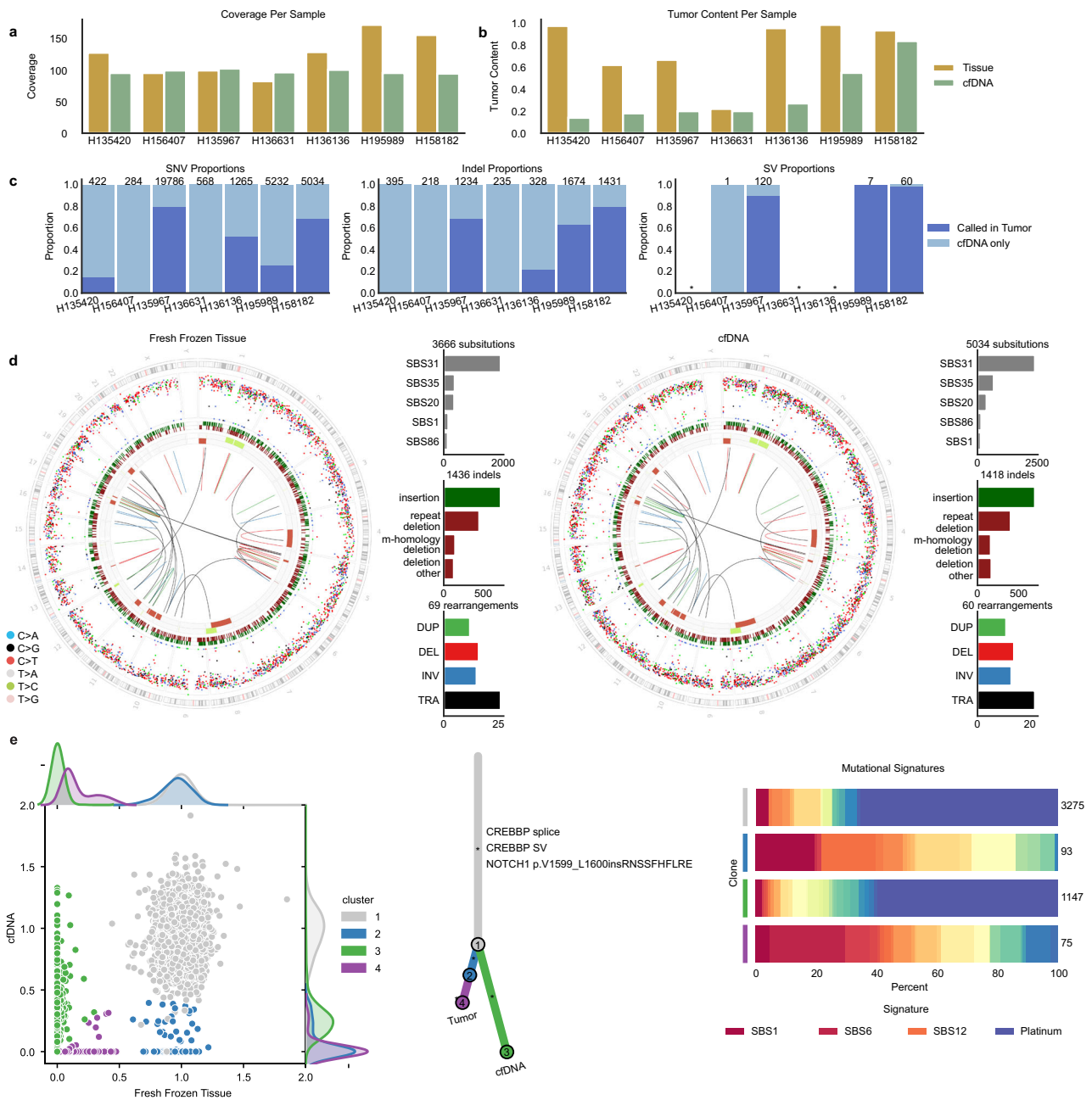
### Methods

**Study participants.** Patients who were seen within the Department of Pediatrics at Memorial Sloan Kettering Cancer Center with presumed or established solid tumor malignancies (including CNS tumors) were eligible to enroll on an institutional prospective tumor/germline-sequencing protocol (ClinicalTrials.gov number, NCT01775072) with informed consent from the patients or their guardians. This study was approved by the MSKCC Institutional Review Board/Privacy Board. Patients with newly diagnosed as well as relapsed/refractory disease were eligible. Adults with pediatric-type malignancies or rare cancers up to the age of 39 were also eligible to enroll.

**Clinical profiling.** DNA extracted from formalin-fixed paraffin-embedded (FFPE) tumor and blood samples (as a matched normal) was sequenced using MSK-IMPACT, an FDA-approved targeted panel used to sequence patients' tumors at MSKCC. MSK-IMPACT captures protein-coding exons of 468 cancer-associated genes, introns of frequently rearranged genes, and genome-wide copy number (CN) probes<sup>4</sup>. Tumor and normal samples were sequenced at 800× and 600×, respectively. Established pipelines followed by manual review were used to characterize germline and somatic mutations, CN variants, and if targeted, genomic rearrangements as previously described<sup>1</sup>. Germline data for alterations in cancer-predisposition genes were analyzed in 88 genes as previously described<sup>2</sup>. For select tumor indication MSK-Fusion<sup>3</sup>, a New York State-approved RNA-capture assay that targets common RNA fusion genes in solid tumors was also performed. Clinically relevant findings were annotated using OncoKb tiers 1–4<sup>4</sup>.

### Research-sequencing approaches

**DNA extraction.** For 114 subjects enrolled in the study, tumor DNA was extracted from fresh frozen (FF) or OCT tissue biopsies and matched normal DNA from



**Fig. 8 Comparison of WGS data from matched fresh frozen tumor tissue and cfDNA.** **a** Coverage values ordered by estimated tumor context in cfDNA. **b** Estimates of tumor content. **c** Barplots showing the proportion of de novo mutation calls in cfDNA that are present in the matched fresh frozen tumor broken down by variant type. cfDNA samples with no high-confidence SVs denoted with an asterisk. **d** Genome-wide distribution and mutation patterns of matched fresh frozen (left) and cfDNA (right) samples for H158182. Circos plots are shown as described in Fig. 6. **e** Individual-level clonality analysis for H158182. (left) Scatterplot of cancer cell fraction (CCF) values for all substitutions color-coded by the estimated cluster. (middle) Phylogenetic tree representation of clusters annotated with clinically relevant variants. (right) Clone-level mutational signature analysis showing the proportion of mutations attributed to each mutational signature with total numbers of mutations in each cluster shown on the right. Whereas drivers associated with these clones could not be determined, cfDNA-specific SNV calls recapitulated mutation signatures in the FF sample, and were enriched for platinum-associated mutational signatures pointing to the existence of therapy-exposed tumor subclones in circulation. (repeat deletion: repeat-mediated deletion, m-homology: microhomology-mediated deletion, deletion other: all other deletions, TRA: translocation, DUP: duplication, DEL: deletion, INV: inversion). Source data for panels **a**, **b** are provided in Supplementary Data 11. Source data for panel **c** are provided at the data repository. Raw data for panels **d**, **e** can be accessed at the dbGAP study.

buffy coat using the DNeasy Blood & Tissue Kit (Qiagen catalog # 69504) according to the manufacturer’s protocol. FFPE tissue was deparaffinized using heat treatment (90 °C for 10’ in 480 µL PBS and 20 µL 10% Tween-20), centrifugation (10,000 × g for 15’), and ice chill. Paraffin and supernatant were removed, and the pellet was washed with 1 mL of 100% EtOH followed by an incubation overnight in 400 µl of 1 M NaSCN for rehydration and impurity removal. Tissues were subsequently digested with 40 µl of Proteinase K (600 mAU/

ml) in 360 µl of Buffer ATL at 55 °C. DNA isolation proceeded with the DNeasy Blood & Tissue Kit (QIAGEN catalog # 69504) according to the manufacturer’s protocol modified by replacing AW2 buffer with 80% ethanol. All DNA was eluted in 0.5X Buffer AE.

*Whole-genome sequencing.* After PicoGreen quantification and quality control by Agilent BioAnalyzer, 500 ng of genomic DNA were sheared using a LE220-plus

Focused-ultrasonicator (Covaris catalog # 500569) and sequencing libraries were prepared using the KAPA Hyper Prep Kit (Kapa Biosystems KK8504) with modifications. Briefly, libraries were subjected to a 0.5X size select using AMPure XP beads (Beckman Coulter catalog # A63882) after post-ligation cleanup.

PCR-free libraries were pooled equivolume for sequencing. Samples were run on a NovaSeq 6000 in a 150-bp/150-bp paired-end run, using the NovaSeq 6000 SP, S1, S2, or S4 Reagent Kit (300 cycles) (Illumina). Tumors were covered to an average of 95X (range = 67–181) and normals at 50X (range = 32–159).

**RNA extraction.** Tumor tissue from FF biopsies was homogenized in 1 mL TRIzol Reagent (ThermoFisher catalog # 15596018) followed by phase separation with 200  $\mu$ L chloroform. RNA was extracted from the aqueous phase using the miR-Neasy Micro Kit (Qiagen catalog # 217084) on the QIAcube Connect (Qiagen) according to the manufacturer's protocol with 350  $\mu$ L input. Samples were eluted in 15  $\mu$ L of RNase-free water.

**Whole-transcriptome RNA sequencing.** After RiboGreen quantification and quality control by Agilent BioAnalyzer, 18ng–1 $\mu$ g of total RNA with an RNA integrity number varying from 1 to 9.9 underwent ribosomal depletion and library preparation using the TruSeq Stranded Total RNA LT Kit (Illumina catalog # RS-122-1202) according to instructions provided by the manufacturer with 8 cycles of PCR. Samples were barcoded and run on a HiSeq 2500 in Rapid Mode or HiSeq4000 at PE100 or on a NovaSeq 6000 at PE150, using the HiSeq Rapid SBS Kit v2, HiSeq 3000/4000 SBS Kit, or NovaSeq 6000 SP, S1, S2, or S4 Reagent Kit (300 cycles) (Illumina). Sequencing was performed to achieve a median of 83 million paired reads per sample.

**cfDNA extraction and whole-genome sequencing.** Cell-free DNA (cfDNA) was extracted from plasma using MagMAX cfDNA isolation kit. After PicoGreen quantification, 47–500 ng of cfDNA were used to make sequencing libraries using the KAPA Hyper Prep Kit (Kapa Biosystems KK8504) with 4 cycles of PCR and pooled equimolar. One sample with sufficient input was prepared PCR-free. Samples were run on a NovaSeq 6000 in a PE150 run, using the NovaSeq 6000 SBS v1 Kit and an S4 flow cell (Illumina). The average coverage per sample was 91X.

**Workflow optimization.** In order to achieve stable turnaround times of 9 days, dedicated resources and optimizations were needed, such as to minimize human steps in the process. In the sequencing core, lab technicians along with sequencers were needed to process and quality-control the incoming samples. A high-throughput connection was used to transfer sequencing data to the bioinformatics core with automatic notifications. An ETL cron job was developed to synchronize relevant deidentified metadata regularly from clinical systems. The data and bioinformatics analyses were tracked and automated using the Isabl platform<sup>5</sup>. In order to achieve stable algorithm turnaround times, parallelization was often split by the estimated amount of work (i.e., number of reads; [https://github.com/papaemmelab/split\\_bed\\_by\\_reads](https://github.com/papaemmelab/split_bed_by_reads)) rather than genomic length. Processing was performed within a heavily shared internal high-performance computing (HPC) cluster with around 4000 cores. The results were automatically curated and prioritized using both cached databases and live APIs in order to reduce interpretation time.

**Bioinformatic analysis.** Analysis of cWGTS data was executed using Isabl platform<sup>5</sup> and included: 1. data QC; 2. ensemble variant calling for germline and somatically acquired mutations from at least two out of three algorithms run for each variant class; 3. signature extraction (i.e., mutation signatures, microsatellite-instability score, and gene expression); 4. variant classification; and, 5. the generation of a clinical prototype summary report. Briefly, upon completion of each sequencing run, Isabl imports paired tumor-normal FASTQ files, executes alignment, quality-control algorithms, and generates tumor purity and ploidy estimates. For samples with sufficient coverage (>60 $\times$ ) and tumor purity (>20%), ensemble variant calling for each variant class (substitutions, insertions and deletions, and structural variations) is performed. High-confidence somatic mutations were classified with regard to their putative role in cancer pathogenesis and statistical post-processing enables the derivation of microsatellite-instability scores and mutation signatures<sup>6</sup>. RNA-seq data were independently analyzed for acquired fusions and gene expression metrics in a subset ( $n = 101$ ). For a subset of patients with consent ( $n = 100$ ) for germline analyses, the normal genome was also independently analyzed.

Clinical relevance of mutations in common cancer genes was annotated using OncoKb, COSMIC, Ensembl Variant Effect Predictor, VAGrENT, gnomAD, and ClinVar databases<sup>725</sup> (refs. 60–63). Additionally, integration of signals across data modalities (germline, somatic mutations, somatic signatures, CN segments, and gene expression profiles) was executed to further determine the significance of observed events. Population filtering, database comparison, and somatic data integration were performed using methods in accordance with the American College of Medical Genetics and ClinGen Somatic/Germline Data Integration subcommittee<sup>64–66</sup>. Last, the findings were automatically embedded into a single-page summary (.html) report containing high-level clinical data, quality-control metrics, genetic findings, and relevant data-visualization plots (i.e., CIRCOS plots, mutation signatures, and gene expression clustering by tSNE). Putative findings of

clinical relevance identified by WGS and RNA-seq were reviewed by an interdisciplinary team of clinical oncologists, molecular pathologists, and cancer genomics experts. Typically 8–10 cases were reviewed in an hour-long tumor board meeting that was held biweekly. The findings were categorized with regard to their relevance in clinical practice as 1. diagnostic, 2. risk predisposition for germline variants, 3. prognostic, 4. therapy-informing, 5. pathogenic, 6. likely pathogenic, or 7. variant of unknown significance (VUS).

## Pipeline overview

**Whole-genome/transcriptome alignment and quality control.** Whole-genome paired-end reads were aligned to human reference genome (GRCh37d5) using BWA-mem (v0.7.17) as a part of the pcap-core v2.18.2 wrapper (<https://github.com/cancerit/PCAP-core>)<sup>67</sup>. The wrapper includes marking of duplicates using Picard. Whole-transcriptome sequencing reads were aligned using Spliced Transcripts Alignment to a Reference, STAR (v2.5.4b, <https://github.com/alexdobin/STAR>) with Ensembl 75 for transcript information<sup>68</sup>. Upon alignment, BAM files for tumor/normal WGS and tumor RNA-seq data for each individual were compared using Compair<sup>69</sup> in order to detect potential sample swaps and cross-individual contamination. Genome-wide median coverage was calculated using Mosdepth<sup>70</sup> with minimum mapping quality of 20. Tumor purity and ploidy was estimated using Battenberg (<https://github.com/cancerit/cgpBattenberg>) and somatic substitution calls. Additionally, a quality-control report is generated per sample using MultiQC (v1.9) (<https://github.com/ewels/MultiQC>) to aggregate alignment and read statistics from FastQC (v0.11.5) (<https://github.com/s-andrews/FastQC>), Picard (v2.25.6) (<https://github.com/broadinstitute/picard>), and RNA-SeQC (v1.1.8.1) (<https://software.broadinstitute.org/cancer/cga/rna-seq>)<sup>70,20</sup>.

**Identification of somatic mutations in whole-genome sequences.** Somatic alterations were detected comparing the tumor against the matched normal for each variant type. All bioinformatic tools were launched using an in-house wrapper. Allele-specific subclonal CN changes were detected using Battenberg (cgpBattenberg v1.4.0) (<https://github.com/cancerit/cgpBattenberg>)<sup>71</sup>. Single-nucleotide variants (SNVs) were identified using Strelka2 (v2.9.1 with manta v1.3.1), (<https://github.com/Illumina/strelka>), MuTect2 (gatk:v4.0.1.2), (<https://github.com/broadinstitute/gatk>), and CaVEMan (cgpCavemanWrapper v1.7.5) (<https://github.com/cancerit/cgpCaVEManWrapper>)<sup>72–74</sup>. Variant post-processing was done using default flags for Strelka2 and MuTect2, while for CaVEMan, cgpCavemanPostprocessing (v1.5.2) was used filtering for sequencing artifacts with >=3 mutant alleles in at least 1% of samples within a panel of 100 unmatched blood normal (<https://github.com/cancerit/cgpCaVEManPostProcessing>). Small insertions and deletions (indels) were detected using Strelka2, MuTect2, and Pindel (cgpPindel v1.5.4) (<https://github.com/cancerit/cgpPindel>) and filtered against a panel of 100 unmatched normals<sup>75</sup>. Structural genomic variants (SVs) were identified using SvABA (~v1.0.0 commit 47c7a88) (<https://github.com/walaj/svaba>), GRIDSS (v2.2.2) (<https://github.com/PapenfussLab/gridss>), and BRASS (v4.0.5 with GRASS v1.1.6) (<https://github.com/cancerit/BRASS>) using a panel of 100 in-house unmatched normals<sup>76,77</sup>.

Finally, microsatellite-instability status was assessed using MSISensor (v0.5) (<https://github.com/ding-lab/msisensor>)<sup>78</sup>.

**Variant consolidation and annotation.** VCF files for SNVs and indels were merged with an in-house wrapper using chromosome, position, reference allele, and alternative allele. The merged VCFs were annotated with VAGrENT (v3.3.0, <https://github.com/cancerit/VAGrENT>) and VEP (v92, <https://github.com/Ensembl/ensembl-vep>)<sup>63,79</sup>. VCF files for SVs were merged using MergeSVvcs (v1.0.2, <https://github.com/papaemmelab/mergeSVvcs>). High-confidence mutations were designated as those that were passed by at least 2 callers.

**Designation of putative oncogenic mutations.** We define a variant as oncogenic if it represents an established “driver” mutation on the basis of prior literature and recurrence in cancer genome. For SNVs, indels, and fusion genes, these annotations are derived from OncoKb. For SVs, we annotate events that target known oncogenes and tumor suppressor genes and use prior literature as reference (e.g., for TERT, ATRX, and TP53 SVs).

**Identification of germline mutations.** Germline single-nucleotide polymorphisms (SNPs) and indels were detected using Strelka2 and Freebayes (v1.2.0, <https://github.com/ekg/freebayes>) with an in-house wrapper. VCF files were merged and annotated using the same procedure used for the somatic variants<sup>80</sup>. Germline variants called by both callers were considered high-confidence. Germline variants were prioritized for review by filtering for recurrence in the current cohort, frequency in any population of 1000 genomes/Gnomad and ClinVar.

**Characterization of gene fusions.** Gene fusions were identified using three different callers: FusionCatcher (v1.0.0, <https://github.com/ndaniel/fusioncatcher>), STAR-Fusion (v1.3.1, <https://github.com/STAR-Fusion/STAR-Fusion>), and FuSeq (v1.1.1, <https://github.com/ngiavtr/FuSeq>)<sup>81–83</sup>. Calls were merged by gene pair and annotated using FusionCatcher's databases. Fusions were considered confident if

called by at least 2 callers. Events were visualized with the plotting functionality by Arriba (<https://github.com/suhrig/arriba>)<sup>84</sup>.

**Gene expression analysis.** Gene expression profiles were ascertained in transcripts per million (TPM) using SALMON (v0.10.0, <https://github.com/COMBINE-lab/salmon>)<sup>85</sup>. tSNE was performed on RNA-seq data from 101 tumors from the cohort and an in-house reference consisting of 155 pediatric tumors using python scikit-learn (v0.21.1, <https://scikit-learn.org/>) and visualized interactively using python bokeh (v1.2.0, <https://docs.bokeh.org/>)<sup>86</sup>.

**Identification of gene expression biomarkers.** Expression biomarkers were assessed using the methodology outlined by Horak et al.<sup>16</sup>. Only actionable genes outlined in the publication supplementary as assessed by severe overexpression, overexpression, severe underexpression, or underexpression, were evaluated. An internal reference cohort of 274 tumor RNA samples was used as a baseline. A gene was considered over-/underexpressed if expression was in the top or bottom ten percent of the reference cohort, while severe over-/underexpression was categorized as expression in the top or bottom five percent.

**Calculation of tumor mutational burden.** Tumor mutational burden (TMB) was calculated using high-confidence, somatic substitutions and indels that fall within coding regions. The totals for these variant classes were combined and then converted to coding TMB using a divisor of 30 to approximate the length of the human exome in Mb. Values greater than 2 coding mutations per Mb were considered pediatric high and values greater than 10 coding mutations per Mb were considered hypermutators, thresholds set by the study in Grobner et al.<sup>16</sup>.

**Identification of mutation signatures of point mutations.** Mutational signature analysis was performed with the MutationalPatterns package (v1.6.1, <https://bioconductor.org/packages/release/bioc/html/MutationalPatterns.html>) and using the COSMIC Mutational Signatures (v3.1) with the addition of Temozolomide signature from Kucab et al.<sup>87,88</sup>.

**Assessment of ITH between matched MSK-IMPACT and WGS samples.** ITH between the matched FFPE and FF samples that underwent MSK-IMPACT and WGS sequencing was assessed by comparing CN changes and substitutions/indels falling within 468 genes included in MSK-IMPACT. For substitutions/indels, the clonal representation in the two assays was compared by performing a proportion test comparing the VAFs reported and adjusting for assay-specific local depth and purity. Mutations with a p-value <0.05 have a statistically significant difference in clonal presentation suggesting ITH. CN profiles from MSK-IMPACT generated using FACETS<sup>89</sup> were compared with the Battenberg output from WGS. In patients where DNA was available, resequencing of discordant mutations was performed using the MSK-IMPACT panel on the same DNA that underwent WGS sequencing at a median depth of 438X<sup>90</sup>. Tumor purity of both assays was taken into account to mitigate the effects of technical issues on mutation calling.

**Inference of clonal structure.** Clonal structure was analyzed using high-confidence SNVs called in each biopsy or the union of SNVs whenever multiple biopsies were available for a patient. DPCLust (v0.2.2, <https://github.com/Wedge-Oxford/dpclus>) was used for calculation of cancer cell fraction corrected for purity and local CN, as well as clustering and assignment of mutations across samples with the exception of the Gibbs Sampling Dirichlet Process step that was optimized internally<sup>71</sup>. Clonal ordering was deduced using clonevol (v0.99.11, <https://github.com/hdng/clonevol>)<sup>91</sup>. Mutational signatures were computed in each cluster independently. Figures were generated with matplotlib (v3.1.0, <https://matplotlib.org/>).

**Estimates of telomere length.** The ratio of telomere length in tumor vs normal was estimated using Telseq (v0.0.2, <https://github.com/zd1/telseq>)<sup>92</sup>.

**Derivation of subsampled BAM files and sensitivity assessment.** A total of 298 subsampled BAMs were generated using samtools (v1.11, <https://github.com/samtools/samtools>) view command with the subsampling option<sup>93</sup>. Median coverages were calculated for the original BAMs using Mosdepth (v0.2.5, <https://github.com/brentp/mosdepth>) with mapping quality >20 and then used to calculate fractions to downsample to approximately 100×, 80×, 60×, and 30×–40× where original coverage was allowed<sup>70</sup>. Mosdepth was used to verify that the median coverage of the subsampled BAM fell within +/-5× of the desired coverage. De novo variant calling and annotation was then performed independently on the subsampled BAM files using the same procedure as cWGTS as described above.

**cfDNA tumor content and variant comparison.** Tumor content in cfDNA specimens was estimated using Battenberg and manual inspection with the help of SNV VAF density plots. De novo variant calling was performed independently using the methods described for identification of somatic mutations in WGS. Further analysis was done to compare specific clinical variants identified by MSK-IMPACT using hileup (v1.0.0, <https://github.com/brentp/hileup>). Clonal structure across tissue and cfDNA samples was inferred using the same methods as described before followed by analysis of clone-specific mutational signatures. All mutations across

the clonal structures were then piled up across corresponding FF WGS data to look for evidence of mutations in both specimens.

**Variant curation and characterization of incremental findings from cWGTS.** Variant curation of targeted NGS assays was performed as previously described<sup>4,28</sup>. To assess variants identified by cWGTS, a multidisciplinary team of disease experts, clinical geneticists, molecular pathologists, and genomics experts assembled regularly to classify molecular alterations (somatic and germline), mutation signatures, and gene expression data. Incremental findings of cWGTS were defined as established oncogenic alterations or signatures not identified by matched MSK-IMPACT somatic or germline NGS (DNA) or ArcherDx targeted NGS (RNA). Incremental findings were further classified as clinically relevant if they met one of the following criteria: (1) diagnostic finding—defined as an alteration that provided justification or an alteration in cancer diagnosis or cancer-subtype diagnosis, (2) established prognostic finding—defined as an alteration with established prognostic relevance with robust support from scientific literature, (3) likely pathogenic or known pathogenic germline- predisposition event, (4) treatment-informing finding—defined as an alteration that provides direct justification of a therapeutic modality, or (5) a driver oncogenic fusion.

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

## Data availability

The raw data for WGS and RNA-seq data generated in this study have been deposited in the dbGAP database under accession code [phs002620.v1.p](https://dbgap.ncbi.nlm.nih.gov/oa/GET.cgi?acc=phs002620.v1.p). These data are available under restricted access due to individual privacy concerns. Permanent employees of an institution at a level equivalent to a tenure-track professor or senior scientist with laboratory administration and oversight responsibilities may request access through dbGAP. The requests, which are managed by NCI's Data Access Committee, take less than 2 days for approval and access is permitted for 12 months. The processed MSK-IMPACT data are available in a study-specific dataset at cbioPortal [[https://www.cbioportal.org/study/summary?id=mixed\\_kunga\\_msk\\_2022](https://www.cbioportal.org/study/summary?id=mixed_kunga_msk_2022)]. Summary and processed data for the figures are available in the source data file as well as the data repository at [https://github.com/papaemmelab/Shukla\\_Levine\\_Gundem](https://github.com/papaemmelab/Shukla_Levine_Gundem). Annotation databases included public resources such as Cancer Gene Census, OncoKb, ClinVar, 1000genomes, gnomAD, and Ensembl Variant Effect Predictor (VEP) databases. The remaining data are available within the article and Supplementary Information file.

## Code availability

Scripts for generating the figures are provided at [https://github.com/papaemmelab/Shukla\\_Levine\\_Gundem](https://github.com/papaemmelab/Shukla_Levine_Gundem).

Received: 17 September 2021; Accepted: 21 April 2022;

Published online: 18 May 2022

## References

1. Beaubier, N. et al. Integrated genomic profiling expands clinical options for patients with cancer. *Nat. Biotechnol.* **37**, 1351–1360 (2019).
2. Frampton, G. M. et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat. Biotechnol.* **31**, 1023–1031 (2013).
3. Chakravarty, D. & Solit, D. B. Clinical cancer genomic profiling. *Nat. Rev. Genet.* <https://doi.org/10.1038/s41576-021-00338-8> (2021).
4. Zehir, A. et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat. Med.* **23**, 703–713 (2017).
5. Gong, J., Pan, K., Fakhri, M., Pal, S. & Salgia, R. Value-based genomics. *Oncotarget* **9**, 15792–15815 (2018).
6. Parsons, D. W. et al. Diagnostic yield of clinical tumor and germline whole-exome sequencing for children with solid tumors. *JAMA Oncol.* **2**, 616–624 (2016).
7. Rusch, M. et al. Clinical cancer genomic profiling by three-platform sequencing of whole genome, whole exome and transcriptome. *Nat. Commun.* **9**, 3962 (2018).
8. Koelsche, C. et al. Sarcoma classification by DNA methylation profiling. *Nat. Commun.* **12**, 498 (2021).
9. Capper, D. et al. DNA methylation-based classification of central nervous system tumours. *Nature* **555**, 469–474 (2018).
10. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
11. Davies, H. et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.* **23**, 517–525 (2017).

12. Wong, M. et al. Whole genome, transcriptome and methylome profiling enhances actionable target discovery in high-risk pediatric cancer. *Nat. Med.* **26**, 1742–1753 (2020).
13. Wrzeszczynski, K. O. et al. Analytical validation of clinical whole-genome and transcriptome sequencing of patient-derived tumors for reporting targetable variants in cancer. *J. Mol. Diagn.* **20**, 822–835 (2018).
14. Duncavage, E. J. et al. Genome sequencing as an alternative to cytogenetic analysis in Myeloid cancers. *N. Engl. J. Med.* **384**, 924–935 (2021).
15. Newman, S. et al. Genomes for kids: the scope of pathogenic mutations in pediatric cancer revealed by comprehensive DNA and RNA sequencing. *Cancer Discov.* <https://doi.org/10.1158/2159-8290.CD-20-1631> (2021).
16. Horak, P. et al. Comprehensive genomic and transcriptomic analysis for guiding therapeutic decisions in patients with rare cancers. *Cancer Discov.* <https://doi.org/10.1158/2159-8290.CD-21-0126> (2021).
17. Medina-Martínez, J. S. et al. Isabi Platform, a digital biobank for processing multimodal patient data. *BMC Bioinformatics* **21**, 549 (2020).
18. Roepman, P. et al. Clinical validation of whole genome sequencing for cancer diagnostics. *J. Mol. Diagn.* **23**, 816–833 (2021).
19. Marabelle, A. et al. Association of tumour mutational burden with outcomes in patients with advanced solid tumours treated with pembrolizumab: prospective biomarker analysis of the multicohort, open-label, phase 2 KEYNOTE-158 study. *Lancet Oncol.* **21**, 1353–1365 (2020).
20. Stephens, P. J. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
21. Bielski, C. M. et al. Genome doubling shapes the evolution and prognosis of advanced cancers. *Nat. Genet.* **50**, 1189–1195 (2018).
22. Jonigk, D. et al. Molecular and clinicopathological analysis of Epstein-Barr virus-associated posttransplant smooth muscle tumors. *Am. J. Transpl.* **12**, 1908–1917 (2012).
23. Shannon-Lowe, C. & Rickinson, A. The global landscape of EBV-associated tumors. *Front. Oncol.* **9**, 713 (2019).
24. Lindrose, A. R. et al. Method comparison studies of telomere length measurement using qPCR approaches: a critical appraisal of the literature. *PLoS ONE* **16**, e0245582 (2021).
25. Chakravarty, D. et al. OncoKB: annotation of the oncogenic effect and treatment implications of somatic mutations in cancer. *J. Clin. Oncol.* **34**, 11583–11583 (2016).
26. Gerlinger, M. et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).
27. Fiala, E. M. et al. Prospective pan-cancer germline testing using MSK-IMPACT informs clinical translation in 751 patients with pediatric solid tumors. *Nat. Cancer* **2**, 357–365 (2021).
28. Benayed, R. et al. High yield of RNA sequencing for targetable kinase fusions in lung adenocarcinomas with no mitogenic driver alteration detected by DNA sequencing and low tumor mutation burden. *Clin. Cancer Res.* **25**, 4712–4722 (2019).
29. Stevens, T. M. et al. NUTM1-rearranged neoplasia: a multi-institution experience yields novel fusion partners and expands the histologic spectrum. *Mod. Pathol.* **32**, 764–773 (2019).
30. Lee, J. C. et al. Recurrent KBTBD4 small in-frame insertions and absence of DROSHA deletion or DICER1 mutation differentiate pineal parenchymal tumor of intermediate differentiation (PPTID) from pineoblastoma. *Acta Neuropathol.* **137**, 851–854 (2019).
31. Wegert, J. et al. Mutations in the SIX1/2 pathway and the DROSHA/DGCR8 miRNA microprocessor complex underlie high-risk blastemal type Wilms tumors. *Cancer Cell* **27**, 298–311 (2015).
32. Skapek, S. X. et al. Rhabdomyosarcoma. *Nat. Rev. Dis. Prim.* **5**, 1 (2019).
33. Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. *Cell* **159**, 676–690 (2014).
34. Agnihotri, S. et al. The genomic landscape of schwannoma. *Nat. Genet.* **48**, 1339–1348 (2016).
35. Drilon, A. et al. Efficacy of larotrectinib in TRK fusion-positive cancers in adults and children. *N. Engl. J. Med.* **378**, 731–739 (2018).
36. Ackermann, S. et al. A mechanistic classification of clinical phenotypes in neuroblastoma. *Science* **362**, 1165–1170 (2018).
37. Valentijn, L. J. et al. TERT rearrangements are frequent in neuroblastoma and identify aggressive tumors. *Nat. Genet.* **47**, 1411–1414 (2015).
38. Chen, X. et al. Recurrent somatic structural variations contribute to tumorigenesis in pediatric osteosarcoma. *Cell Rep.* **7**, 104–112 (2014).
39. Drier, Y. et al. An oncogenic MYB feedback loop drives alternate cell fates in adenoid cystic carcinoma. *Nat. Genet.* **48**, 265–272 (2016).
40. Duffy, M. J. et al. p53 as a target for the treatment of cancer. *Cancer Treat. Rev.* **40**, 1153–1160 (2014).
41. Arora, S. et al. Functional analysis of rare variants in mismatch repair proteins augments results from computation-based predictive methods. *Cancer Biol. Ther.* **18**, 519–533 (2017).
42. Lawlor, R. T. et al. Alternative lengthening of telomeres (ALT) influences survival in soft tissue sarcomas: a systematic review with meta-analysis. *BMC Cancer* **19**, 232 (2019).
43. Pezzolo, A. et al. Intratumoral diversity of telomere length in individual neuroblastoma tumors. *Oncotarget* **6**, 7493–7503 (2015).
44. Ohali, A. et al. Telomere length is a prognostic factor in neuroblastoma. *Cancer* **107**, 1391–1399 (2006).
45. Koneru, B. et al. Telomere maintenance mechanisms define clinical outcome in high-risk neuroblastoma. *Cancer Res.* **80**, 2663–2675 (2020).
46. Hartlieb, S. A. et al. Alternative lengthening of telomeres in childhood neuroblastoma from genome to proteome. *Nat. Commun.* **12**, 1269 (2021).
47. Sieverling, L. et al. Genomic footprints of activated telomere maintenance mechanisms in cancer. *Nat. Commun.* **11**, 733 (2020).
48. Gröbner, S. N. et al. The landscape of genomic alterations across childhood cancers. *Nature* **555**, 321–327 (2018).
49. Campbell, B. B. et al. Comprehensive analysis of hypermutation in human cancer. *Cell* **171**, 1042–1056.e10 (2017).
50. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
51. Pich, O. et al. The mutational footprints of cancer therapies. *Nat. Genet.* **51**, 1732–1740 (2019).
52. Landau, H. J. et al. Accelerated single cell seeding in relapsed multiple myeloma. *Nat. Commun.* **11**, 3617 (2020).
53. Bettgowda, C. et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci. Transl. Med.* **6**, 224ra24 (2014).
54. Klega, K. et al. Detection of somatic structural variants enables quantification and characterization of circulating tumor DNA in children with solid tumors. *JCO Precis. Oncol.* **2018**, PO.17.00285 (2018).
55. Peneder, P. et al. Multimodal analysis of cell-free DNA whole-genome sequencing for pediatric cancers with low mutational burden. *Nat. Commun.* **12**, 3230 (2021).
56. Andersson, D., Fagman, H., Dalin, M. G. & Ståhlberg, A. Circulating cell-free tumor DNA analysis in pediatric cancers. *Mol. Asp. Med.* **72**, 100819 (2020).
57. Robbe, P. et al. Clinical whole-genome sequencing from routine formalin-fixed, paraffin-embedded specimens: pilot study for the 100,000 Genomes Project. *Genet. Med.* **20**, 1196–1205 (2018).
58. Wong, M. et al. Whole genome, transcriptome and methylome profiling enhances actionable target discovery in high-risk pediatric cancer. *Nat. Med.* **26**, 1742–1753 (2020).
59. Nangalia, J. & Campbell, P. J. Genome sequencing during a patient’s journey through cancer. *N. Engl. J. Med.* **381**, 2145–2156 (2019).
60. Landrum, M. J. et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868 (2016).
61. Forbes, S. A. et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
62. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
63. McLaren, W. et al. The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
64. Richards, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
65. Alberts, D. & Hess, L. M. *Fundamentals of Cancer Prevention* (Springer Science & Business Media, 2008).
66. Walsh, M. F. et al. Integrating somatic variant data and biomarkers for germline variant classification in cancer predisposition genes. *Hum. Mutat.* **39**, 1542–1552 (2018).
67. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
68. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
69. Bergmann, E. A., Chen, B.-J., Arora, K., Vacic, V. & Zody, M. C. Conpair: concordance and contamination estimator for matched tumor-normal pairs. *Bioinformatics* **32**, 3196–3198 (2016).
70. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2017).
71. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
72. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
73. Van der Auwera, G. A. & O’Connor, B. D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra* (O’Reilly Media, Incorporated, 2020).
74. Jones, D. et al. cgpaVEManWrapper: simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. *Curr. Protoc. Bioinforma.* **56**, 15.10.1–15.10.18 (2016).

75. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
76. Wala, J. A. et al. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* **28**, 581–591 (2018).
77. Cameron, D. L. et al. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Res.* **27**, 2050–2060 (2017).
78. Niu, B. et al. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics* **30**, 1015–1016 (2014).
79. Menzies, A. et al. VAGrENT: variation annotation generator. *Curr. Protoc. Bioinforma.* **52**, 15.8.1–15.8.11 (2015).
80. Erik Garrison, G. M. Haplotype-based variant detection from short-read sequencing. *arXiv* <https://doi.org/10.48550/arXiv.1207.3907> (2012).
81. Nicorici, D. et al. FusionCatcher - a tool for finding somatic fusion genes in paired-end RNA-sequencing data. <https://doi.org/10.1101/011650> (2014).
82. Haas, B. J. et al. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol.* **20**, 213 (2019).
83. Vu, T. N. et al. A fast detection of fusion genes from paired-end RNA-seq data. *BMC Genomics* **19**, 786 (2018).
84. Uhrig, S. et al. Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Res.* **31**, 448–460 (2021).
85. Srivastava, A. et al. Alignment and mapping methodology influence transcript abundance estimation. *Genome Biol.* **21**, 239 (2020).
86. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
87. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med.* **10**, 33 (2018).
88. Kucab, J. E. et al. A compendium of mutational signatures of environmental agents. *Cell* **177**, 821–836.e16 (2019).
89. Shen, R. & Seshan, V. E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.* **44**, e131 (2016).
90. 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 (2015).
91. Dang, H. X. et al. ClonEvol: clonal ordering and visualization in cancer sequencing. *Ann. Oncol.* **28**, 3076–3082 (2017).
92. Ding, Z. et al. Estimating telomere length from whole genome sequence data. *Nucleic Acids Res.* **42**, e75 (2014).
93. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).

## Acknowledgements

The authors would like to acknowledge Drs T. Heaton, M. LaQuaglia M, and J. Gerstle for executing surgical resections, Drs N. Schultz, D. Chakravarty, and M. Nissan for assistance with OncoKb annotations, Drs. D. Solit and A. Kentsis for support and interesting discussions, and Drs. N.K. Cheung, B. Kushner, E. Basu, P. Meyers, L. Wexler, P. Kothari, I. Dunkel, S. Gilheeny, Y. Khakoo, K. Kramer, S. Prockop, and T. Trippett for participation and clinical contributions to the study. E.P. is a Josie Robertson Investigator and is supported by the European Hematology Association, American Society of Hematology, Gabrielle's Angels Foundation, V Foundation, and The Geoffrey Beene Foundation, and a Damon-Runyon Rachleff Innovator Award recipient. Funding

for this study was supported by the Olayan Fund for Precision Pediatric Cancer Medicine.

## Author contributions

E.P., A.L.K., and N.S. designed the study. M.F.L., G.G., D.D., J.G.A., J.S.M.M., Y.Z., and J.E.A.O. developed algorithmic infrastructure and M.F.L., G.G., D.D., D.G., J.G.A., and J.E.O. performed bioinformatic analysis. N.S., N.B., E.S., L.S., I.S.R., and T.O. oversaw coordination of patient consent, sample processing, and clinical data acquisition. U.B., M.H.A.R., and S.S.F. oversaw biospecimen banking and pathology review. N.S., M.F.L., G.G., B.S., D.G., E.F., J.B., T.O., F.I.C., M.O., M.K., S.R., S.M., E.S., M.K., M.L., F.D.C., J.G.B., A.Z., M.F.W., A.L.K., and E.P. oversaw variant annotation and data interpretation. C.C. and A.V. executed laboratory processing of biospecimens and sequencing. M.F.L., G.G., D.D., and E.P. prepared figures and tables. N.S., G.G., A.L.K., and E.P. wrote the paper with input from M.F.L., D.D., and M.F.W. All authors reviewed and approved the paper for submission.

## Competing interests

E.P., A.L.K. and J.S.M.M. are founders, equity holders and hold fiduciary roles in Isabl Inc. E.P., A.L.K., J.S.M.M., M.F.L., G.G., J.Z., J.E.A.O., J.G.A., N.S. are inventors on intellectual property related to a software platform for genomic data analytics, a non-provisional patent application has been filed 17/629292 titled "SYSTEMS AND METHODS FOR CANCER WHOLE GENOME AND TRANSCRIPTOME SEQUENCING (CWGTS)". G.G. is a consultant in Isabl Inc. D.G. is a consultant and a shareholder of Repare Therapeutics. D.G. is a consultant and stock-option holder in MNM Diagnostics.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-022-30233-7>.

**Correspondence** and requests for materials should be addressed to A. L. Kung or E. Papaemmanuil.

**Peer review information** *Nature Communications* thanks Edwin Cuppen and the other anonymous reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022