

Enabling whole genome sequencing analysis from FFPE specimens in clinical oncology

Received: 26 February 2024

Accepted: 16 October 2025

Published online: 27 November 2025

 Check for updates

Dylan Domenico^{1,2}, Gunes Gundem^{1,2}, Max F. Levine^{1,2},
Juan Esteban Arango-Ossa^{1,2}, Pauline Robbe³, Georgios Asimomitis^{1,2},
Cassidy Cobbs^{1,2}, Emily Stockfisch¹, Tara O'Donohue¹, Charlotte Brierley⁵,
Janine Senz⁶, Dawn Cochrane⁶, Neeman Mohibullah⁴, Umesh Bhanot^{1,2},
Joachim Silber⁷, Neerav Shukla¹, Sohrab P. Shah^{1,2}, Britta Weigelt^{1,2},
Oliver Zivanovic⁸, Andrew McPherson^{1,2}, Anna Schuh^{1,2}, Andrew L. Kung¹ &
Elli Papaemmanuil^{1,2} 

The adoption of whole genome sequencing (WGS) in clinical oncology is challenged by low data quality and increased artifacts in standard-of-care formalin-fixed paraffin-embedded (FFPE) samples. Analysis of 56 fresh frozen (FF) and FFPE matched pairs demonstrates that FFPE processing results in a median 20-fold enrichment in artifactual calls across mutation classes and impairs detection of clinically relevant biomarkers such as homologous recombination deficiency (HRD). We demonstrate that implementation of consensus calling reduces artifactual structural variant (SV) calls by 98% but is not sufficient in mitigating artifactual calls for single nucleotide variants (SNVs) and indels as compared to FF data. We develop FFPErase, a machine learning framework that filters SNV/indel artifacts and delivers clinical grade variant reporting allowing accurate quantification of clinically relevant biomarkers. Comparison of FFPErase WGS calls to clinical reporting by FDA-approved panel tests demonstrates 99% sensitivity and enables reporting of 24% more clinically relevant findings.

Cancer genome studies have accelerated the discovery of novel diagnostic, prognostic and therapeutic biomarkers. Prospective sequencing of cancer biopsies is now routine in clinical oncology and represents a pivotal step in treatment planning. To this end, most clinical testing assays leverage targeted panel sequencing, a cost-effective, scalable solution that is amenable to the short fragment sizes and low DNA quality associated with FFPE.

However, panel tests are designed to support the detection of small mutations in the genes most frequently mutated in common

cancers. These targeted approaches do not yield clinically relevant findings in up to 60% of cancer patients, especially those with less common cancer diagnoses^{1,2}. WGS is the only DNA profiling assay that enables comprehensive characterization of a cancer genome through the identification of the full spectrum of clinically relevant mutations across variant classes, such as SNVs, indels, SVs, copy number variants (CNVs) and complex mutation signatures, such as tumor mutational burden (TMB), microsatellite instability (MSI) and HRD³.

¹Department of Pediatrics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. ²Halvorsen Center for Computational Oncology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. ³RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. ⁴Integrated Genomics Operation Core, Center for Molecular Oncology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. ⁵Medical Research Council Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK. ⁶Department of Molecular Oncology, BC Cancer, Vancouver, BC, Canada. ⁷Department of Pathology and Laboratory Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA. ⁸Gynecology Service, Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, NY, USA. ⁹Oxford Molecular Diagnostics Centre, Radcliffe Department of Medicine, University of Oxford, Oxford, UK.

 e-mail: papaemme@mskcc.org

In recent years, seminal studies have demonstrated the diagnostic utility of WGS, enabling detection of clinically relevant biomarkers in 68–90% of patients with rare cancers, pediatric tumors, hematological neoplasms, and advanced metastatic disease^{4–9}. With decreasing sequencing costs, WGS is becoming increasingly accessible. However, a major limitation of WGS applications in oncology is the reliance on FF tissue as a source of tumor DNA due to FFPE preparations resulting in hyper-fragmented DNA (225–300 bps) that is smaller than the optimal WGS range (360–480 bps), and generating low coverage data with an enrichment in artifactual variant calls^{10–14}. Development of end-to-end laboratory and analytical workflows to generate high-fidelity WGS data from FFPE material would open up molecularly guided clinical decision support for patients who do not benefit from existing diagnostic approaches. Importantly, routine WGS profiling of FFPE would advance biomarker discovery through the analysis of ongoing and retrospective clinical trial cohorts^{15,16}.

Recent efforts to address the challenges of WGS in FFPE material have focused on optimization of DNA extraction and library preparation protocols^{17–23}, or computational tools to post-process mutation calls^{24–26}. Of these, FFPESig and FFPolish filter artifactual SNVs but not indels and FFPESig specifically is trained on targeted data rather than WGS. More recently, FFPImpact was developed as a scoring method to quantify the level of FFPE damage within a sample; however, this tool does not output a set of high-quality filtered calls²⁷. Importantly, none of these methodologies has been quality-controlled against validated clinical sequencing assays. As such, there is still no computational tool that effectively assesses the impact of FFPE damage across the spectrum of variant classes in a cancer genome, including SNVs, indels as well as SVs, and clinically relevant mutation signatures (TMB, HRD, MSI, etc.)^{28,29}. As laboratory workflows are established for the derivation of high-quality sequencing data, there is a pressing need to develop generalizable tools for analysis and variant reporting in FFPE-derived WGS data.

Here, we analyze 56 trios of matched FF, FFPE and normal control WGS data to identify the genome-wide consequences of FFPE damage across all variant classes using a consensus calling approach³⁰. We developed FFPEraser, a random forest classifier that improves FFPE artifact classification and concordance between matched FF/FFPE datasets and delivers clinical-grade reporting across all variant classes and clinically relevant mutation signatures, thereby unlocking the potential of WGS in clinical oncology.

Results

Cohort and data characteristics

We acquired matched FF/FFPE WGS data from 168 samples across three centers, comprising 56 patient trios (FF, FFPE, germline control) and 16 cancer types (Fig. 1a, Supplementary Data 1). The dataset includes 40 published trios from Oxford University Hospitals, 10 in-house trios from Memorial Sloan Kettering Cancer Center (MSKCC), and six trios from BC Cancer (BCC).

Formalin fixation, DNA extraction, and WGS were performed independently in each center, resulting in median genome-wide coverages of 93x for FF and 51x for FFPE samples (Supplementary Fig. 1a). The target FFPE coverage for MSKCC, Oxford and BCC was 80x, 70x and 40x, respectively, and the average drop in coverage from FF to FFPE was 36x. Purity and ploidy estimates remained consistent across FF and FFPE. Although protocols differed between institutions, FFPE libraries across all centers had shorter average insert sizes (166–358 bp) than FF (356–503 bp) and increased GC bias (Supplementary Fig. 1b, c). These results demonstrate lower effective coverage and library quality in FFPE-derived WGS.

Comparison of genome-wide mutation calls between FFPE and FF

To evaluate the effect of FFPE processing in variant calling from WGS, we first compared the output of single algorithm variant calling

approaches to consensus calling using three individual variant callers per mutation class. We evaluated FF/FFPE mutations considering 1. the union of all passed calls and 2. consensus calls (i.e., variants supported by at least two callers).

Our data demonstrated that consensus variant calling significantly decreases the proportion of FFPE specific SVs (92–12%). However, this improvement is not seen in SNV and indel calls, where the median proportion of FFPE-specific mutations remained high (62 and 73%, respectively). This highlights the need to implement a consensus variant calling methodology but also to incorporate additional approaches for artifact filtration across variant classes (Fig. 1b, c). Retaining a consensus calling approach, we called 2,346,008 SNVs and 387,851 indels in FF samples and 3,946,938 SNVs and 466,232 indels in FFPE, signifying a median 2.0x and 2.4x increase in SNVs and indels, respectively, with variability across centers (Supplementary Data 2)^{10,18}. Across the dataset, there were up to 152x more SNVs and 42x more indels called in FFPE-derived WGS data, lowering the precision of small mutation calling to 50% for SNVs and 62% for Indels. SV calling precision was less affected amongst the samples (median fold-change 0.76 (0.19–1.42)) (Fig. 1d, Supplementary Fig. 1d).

Overall, 85% of SNVs (1,992,218) and 75% of indels (289,231) in FF were also called in FFPE (Supplementary Data 3). While SV consensus calling maintained high precision (80%, 5094/6364), sensitivity was lower (57%, 5094/8880) due to reduced coverage on SV detection amongst other factors such as shorter read fragments and lower mapping quality (Supplementary Fig. 2a, b). Coverage impact was evident as FF-specific SV loci ($n = 3776$) were on average 15x lower in FFPE compared to shared FF/FFPE calls ($n = 5102$). With 52% of FF-specific loci having at least 2 supporting reads in FFPE, increased coverage could recover these SVs. CNV profiles in FFPE showed higher noise levels resulting in hyper-segmentation and less reliable determination of copy number state (Supplementary Fig. 2c–f)^{17,18}.

Impact of FFPE damage on the ability to identify clinically relevant molecular biomarkers

A major motivation for expanding WGS application to FFPE biospecimens is the reporting of clinically relevant biomarkers. We therefore evaluated the sensitivity and specificity of FFPE WGS data as compared to matched FF WGS in calling established molecular biomarkers. Annotations with OncoKB³¹ and COSMIC³² identified 117 small mutations (27 indels, 90 SNVs), 46 SVs, and 32 amplifications in FF derived WGS data. Sensitivity for the same driver detection in FFPE was 89% (174/195). An additional set of 13 FFPE-specific findings were reported likely due to intratumor heterogeneity⁹. (Supplementary Fig. 3, Supplementary Data 4–5, Supplementary Information)

Next, the implications of FFPE damage were assessed on TMB and cancer-associated mutation signatures. Derived from our consensus call set, genome-wide TMB was elevated in FFPE (median: 10.28, range: 1.42–536.38) compared to FF (median: 3.45, range: 0.04–561.56), yet coding TMB was unaffected (Fig. 2a–c), suggesting elevated artifact representation in non-coding regions. Of note, without deploying consensus calling coding TMB was elevated an average 7-fold (Supplementary Fig. 4a).

Samples with a high artifact rate showed enrichment in specific mutation signatures. For example, an increased contribution of SBS37 was observed in 45/56 FFPE samples (median proportion: 23.4%) compared to the corresponding FF (12/56, median proportion: 3.6%) (Fig. 2d–e, Supplementary Fig. 4b, c, Supplementary Data 6). Additionally, FFPE-derived data exhibited a 2.8x increase in both insertions and repeat-mediated deletions (Supplementary Fig. 4d). In contrast, SV profiles remained largely unaffected (median cosine similarity: 0.97, range: 0.02–1.00) (Supplementary Data 7)³³.

We hypothesized that the rise in genome-wide mutation burden and corresponding artifact signatures could impair the detection of composite mutation signatures like HRD. Utilizing HRDdetect³⁴ and

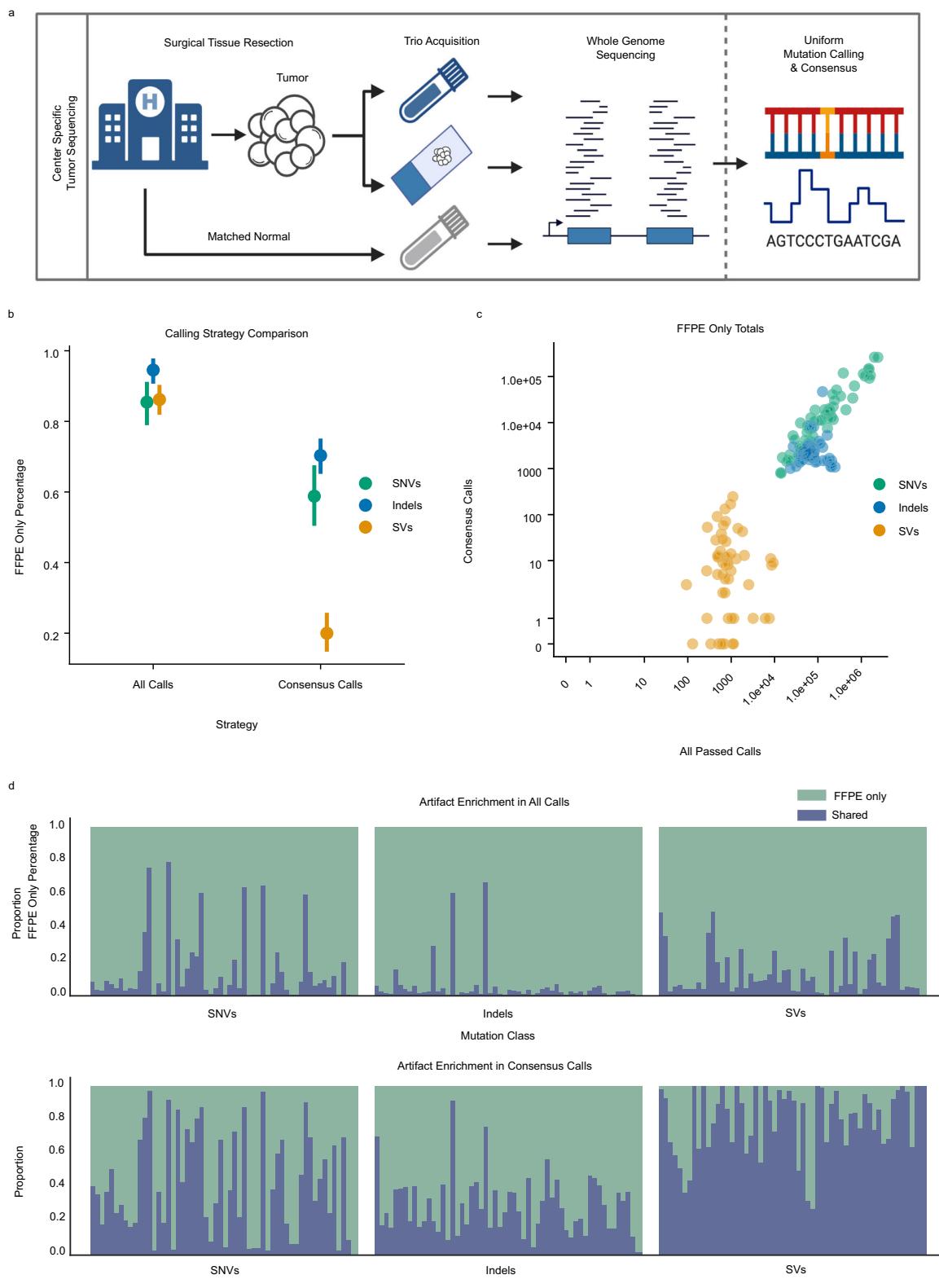
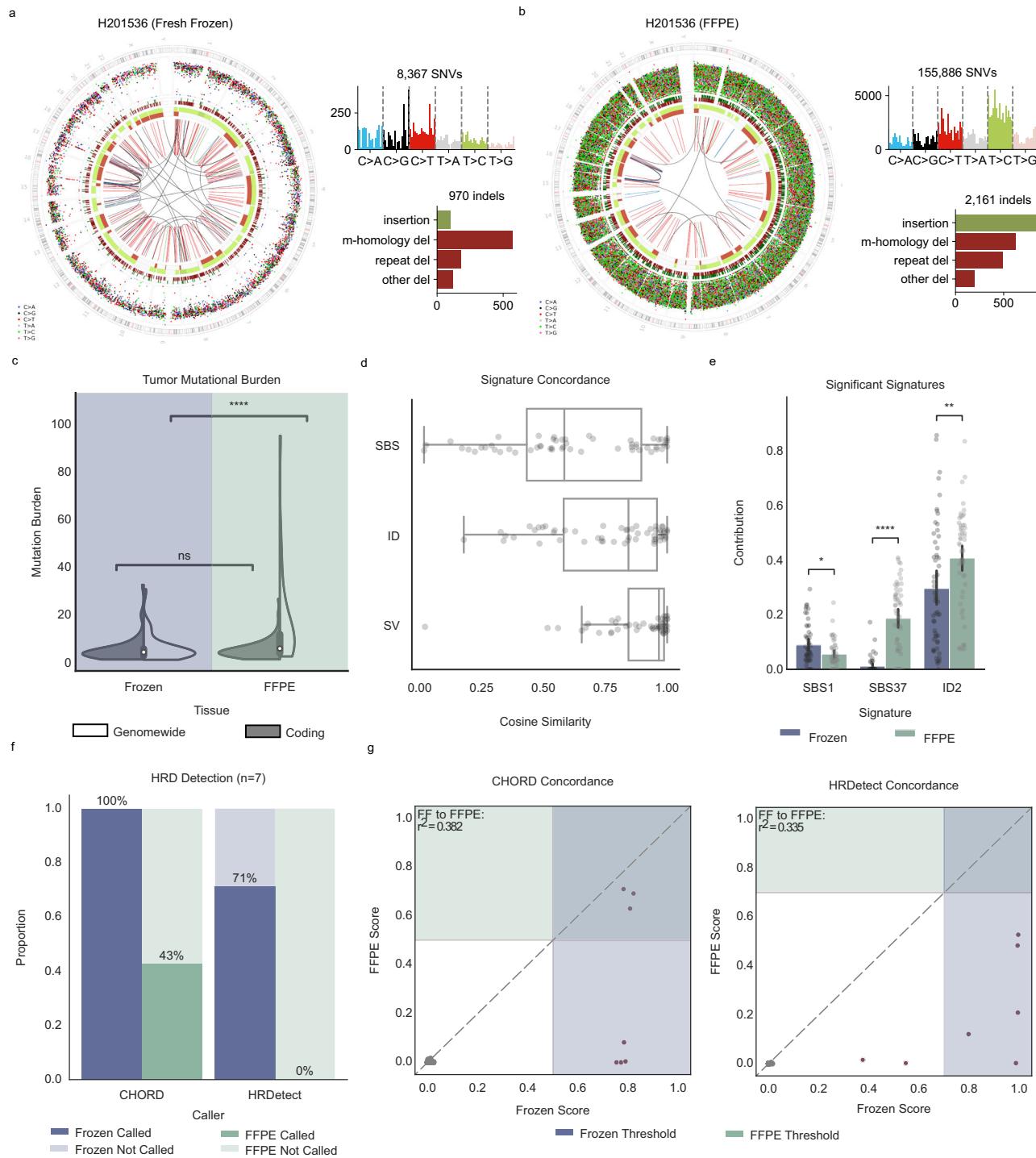


Fig. 1 | Cohort Characteristics. **a** Graphical representation of data accrualment, demonstrating acquisition of trios and sequencing at multiple centers. **b** Pointplot showing distribution across 56 FF/FFPE pairs in percentage of calls unique to FFPE sample in all passed calls versus consensus calls for each variant class. Center represents mean percentage and error bars represent 95% confidence intervals. **c** Scatterplot showing call totals for all passed calls versus consensus calls for each variant class. **d** Stacked barplots demonstrating proportion of detected mutation

calls within FFPE samples that are shared with matched FF in each mutation class. Top plots showing artifact enrichment in all passed calls and bottom blots showing artifact enrichment specifically in high confidence consensus calls. FF, Fresh Frozen; FFPE, Formalin-fixed, paraffin-embedded; SNV, single nucleotide variant; indel, small insertion or deletion; SV, structural variant. Source data are provided as a Source Data file. Panel a created in BioRender. Papaemmanuil, E. (2025) <https://BioRender.com/qlm8gy2>.



CHORD²⁹, 7 samples were flagged as likely HRD by both callers in the FF data. However, HRD scores in the corresponding FFPE data were below the detection cutoff for 7/7 cases by HRDetect, and 4/7 by CHORD, resulting in incorrect HRD classification (Fig. 2f, g, Supplementary Data 8). These discrepancies demonstrate that increased SNV/indel artifacts in FFPE affects two key predictor features of HRD: SBS3 and microhomology-mediated indels, leading to false HRD classification (Supplementary Fig. 4e).

Evaluation of MSI status by TMB, MSIsensor, and MIMcall scores in FF identified 2 MSI-positive samples (TMBs: 99 and 562) (Supplementary Data 9). Both MSI positive samples exhibited increased contribution of MMR-associated signatures (SBS15, SBS21)^{28,35-38}. Contrary to HRD, FFPE artifacts did not affect presentation or detection of MMR-

related signatures. There was one false positive MSI sensor call in FFPE (H158880; FFPE: 11.09, FF: 1.68) and none in MIMcall. However, this sample had low TMB (FFPE: 1.71, FF: 1.43) and no presentation of MMR-associated signatures in FF, demonstrating the necessity of integrating multiple genome-wide features in MSI assessment.

Taken together, this analysis demonstrates that whilst common cancer-associated mutations can be captured in FFPE-derived WGS data, without correction, clinically relevant signatures can be missed (e.g., HRD) or misclassified (e.g., MSI). Importantly, comprehensive assessment of FFPE-derived data across biomarker classes, inclusive of small mutations, structural variants and complex mutation signatures, is warranted to ensure accurate interpretation of reporting of these signatures.

Fig. 2 | Genomewide Signal Impact. **a** Example of a WGS FF tumor resection (H201536) through 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. **b** Circos plot as described for matched FFPE tumor resection. **c** Violin plot showing TMB distribution across cohort ($n = 56$) for FF and FFPE samples excluding hypermutator cases H203508 and H203526 for interpretability. Both coding and and genomewide TMB calculations are shown (two-sided Mann-Whitney U test, Coding: $p = 3.489e-01$, Genomewide: $p = 2.381e-06$). In the middle of the split violin is a box and whisker plot for the all data points (coding and non-coding) with the median as a white dot, a thick bar for the interquartile range (25–75th percentile) and a thin bar that goes to the minimum and maximum of the data. **d** Boxplots showing cosine similarity in COSMIC SBS and ID signature contribution matrices (median SBS: 0.58, ID: 0.84) as well as cosine similarity in SV context matrices (median: 0.97) between 56 matched FF/FFPE

samples. The line in the middle of each box represents the median, while the upper and lower ends of each box represent the interquartile range (IQR) (25th–75th percentile) and the whiskers represent 1.5x the IQR. **e** Barplot showing significant signature differences between FF and FFPE for 56 matched samples (two-sided Mann-Whitney U test, SBS1 $p = 3.641e-02$; SBS37 $p = 1.211e-12$; ID2 $p = 8.420e-03$). Bars at mean contribution of signature of all 56 samples and error bands showing 95% confidence intervals. **f** Barplot showing HRD detection status for callers HRDetect and CHORD in patients with HRD ($n = 7$). **g** Concordance in HRDetect and CHORD scores for the seven HRD positive patients identified with scores above the recommended threshold for each tool. Red dotted lines show threshold of detection per tool (0.5 for CHORD and 0.7 for HRDetect) for FFPE sample score. FF each mutation class. FF, Fresh Frozen; FFPE, Formalin-fixed, paraffin-embedded; SNV, single nucleotide variant; indel, small insertion or deletion; SV, structural variant; HRD, homologous recombination deficiency. Raw data for (a,b) can be accessed at the EGA study. Source data for (c–g) are provided in Supplementary Data 2, 6, 8 and the data repository. Source data are also provided as a Source Data file.

Rationale for the development of an FFPE-specific artifact classifier

The distinct mutation patterns observed in FFPE, such as the enrichment of SBS37 signature and repeat-mediated deletions, present an opportunity to develop informatics solutions that learn and correct for FFPE-specific artifacts. To this end, we evaluated 33 SNV and 29 indel-associated features (Supplementary Data 10) and noted significant importance in features related to VAF, read mapping quality, and insert size (Supplementary Fig. 5a-b)³⁹.

We designed a machine learning classifier tailored to discern FFPE-specific artifacts. We hypothesized that certain patterns of FFPE-associated artifacts are likely shared across preparation protocols, while some may be protocol or laboratory-specific. To test this hypothesis, we evaluated three classification models: 1. Consider mutations derived from samples from a single center for training/testing; 2. Consider mutations from samples derived from two centers and testing on data from a third; 3. A combined approach, training a model per sample with mutations across all centers but excluding mutations from the test sample. All models utilized a balanced random forest architecture, with the 1- and 3-center approaches employing a leave-one-out cross-validation and the 2-center approach representing validation on a fully independent dataset. (Fig. 3a-c, Supplementary Data 11, Supplementary Information).

The classifier's performance across all approaches showcased a 93–94% median accuracy for SNVs and 89–91% for indels. Importantly, models where the test center data were incorporated during training (1- and 3-center) marginally outperformed the 2-center approach (Supplementary Fig. 5c) suggesting that there are both shared, as well as center-specific effects of FFPE damage that can only be captured if these are represented in the training dataset. Using the 3-center approach, precision for SNV/indel calling increased from 52 to 92% for FFPE mutations relative to the FF calls, while maintaining an 80% sensitivity. Notably, the cross center approach was the only to not filter high-confidence driver mutations shared by FF and FFPE, demonstrating that it extracts the most robust FFPE artifact profile (Fig. 3d, e, Supplementary Data 12).

Post-filtering results in high concordance in drivers and global signatures

Next, we used the 3-center approach to filter artifacts in each hold-out sample and assessed the impact of FFPE filtration on reporting of clinically relevant drivers and global mutational patterns. Concordance for TMB, COSMIC signatures, and HRD detection drastically improved.

Upon filtration, median genome-wide TMB decreased from 10.3 (1.4–536.4) to 2.9 (0.3–497.4), much closer to FF TMB estimates (median: 3.5) (Fig. 4a, Supplementary Data 13). The cosine similarity between FF and FFPE for SNV signatures improved to 0.94 (0.40–1.00)

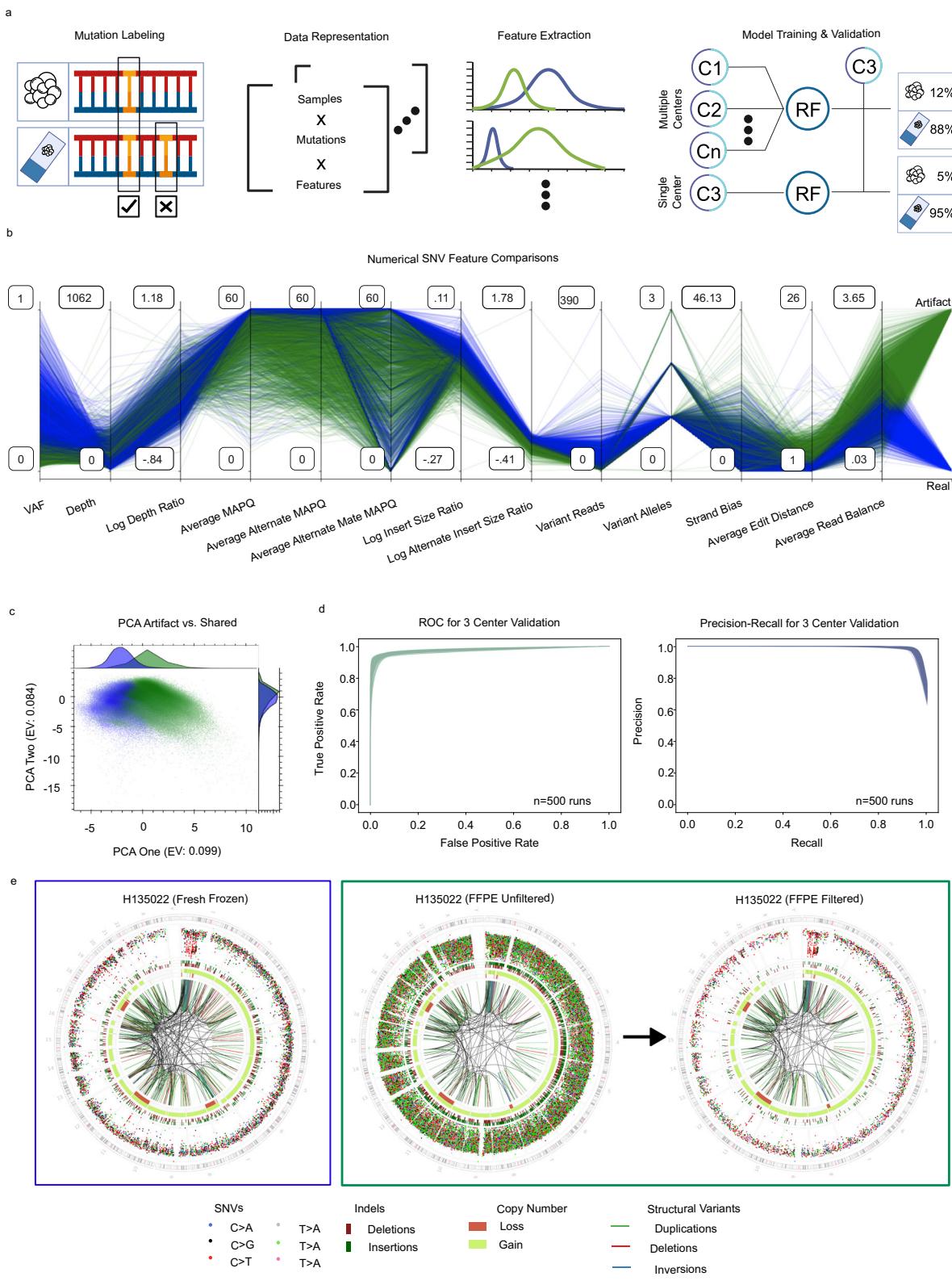
(Fig. 4b, Supplementary Fig. 6a-b, Supplementary Data 14), and the contribution of FFPE-specific signatures (e.g., SBS37) was notably decreased. Similar results were obtained for indels (Fig. 4c). To verify that signature concordance after filtering was independent of the signature fitting method, we calculated signature exposures using both MutationalPatterns⁴⁰ and MuSiCal⁴¹ and observed consistent top signatures and concordance between matched FF and filtered FFPE samples (Supplementary Fig. 6c).

Post-filtration, 7/7 HRD+ tumors were correctly classified in FFPE (Fig. 4d). For the 2 cases identified as MSI+, artifact filtration reduced mutation burden by 7–8% without removing contribution of MMR-associated signatures. These results demonstrate that FFPErase removes artifacts without losing sensitivity, thus enabling accurate detection of biological signals related to HRD and MMR. Importantly, FFPErase accurately classified all drivers shared between FF/FFPE ($n = 103$). Of the 11 FFPE-unique driver calls, 4 (1 SNV, 3 indels) were classified as artifacts. 2 of these were indels located in the polymorphic HLA-A locus, a region that is highly variable, suggesting that classifier performance in loci with decreased mapping quality may require more training data.

Performance of FFPErase as compared to existing approaches

We compared the FFPErase results from the 3-center approach to two publicly available artifact removal strategies, FFPolish and FFPESig. Additionally, we adapted the scoring tool FFPEImpact to output filtered mutational contexts (Supplementary Information)^{24,25}. FFPolish leverages a feature extraction approach like FFPErase for SNVs but not indels. When applied to our data, FFPolish had a lower accuracy compared to FFPErase (63 vs. 95%) and critically removed 42% of SNV drivers shared by matched FF/FFPE tissue (Fig. 5a, Supplementary Data 15). FFPESig does not provide an artifact detection strategy but rather corrects trinucleotide contexts using FFPE signatures learned from targeted sequencing data. The adapted version of FFPEImpact outputs a similarly filtered trinucleotide context but does not provide per-mutation classification. To assess the performance of all tools, we quantified cosine similarity between matched FF and post-filtered FFPE contexts and show that the similarity index was higher using FFPErase compared to all other methods (FFPErase: 0.99, FFPolish: 0.91, FFPEImpact: 0.90, FFPESig: 0.83) (Fig. 5b-d, Supplementary Data 16). These findings highlight that FFPErase is the only tool to deliver a per-mutation classification, filter putative artifacts and result in near FF variant calls. Furthermore, we applied our FFPErase filtration to a representative dataset from the FFPolish study²⁴ consisting of 31 Burkitt lymphoma and 38 cervical cancer cases and achieved a median accuracy of 97% for SNVs and 90% for indels across samples (Supplementary Fig. 7).

Next, we evaluated FFPErase performance against clinical findings in a cohort of 34 cases for which we had FFPE-derived WGS data and



matched clinical sequencing by FDA-approved MSK-IMPACT¹ from the same biopsy. A total of 92 clinically relevant events were reported by MSK-IMPACT. 99% were retained in FFPErase-processed WGS data. Importantly, analysis of WGS from the FFPE specimen revealed additional, clinically relevant events in 41% (14/34) of patients increasing the number of findings by 24% (Fig. 6a, Table 1, Supplementary Data 17). These findings were not captured by panel tests and were mostly represented by SVs.

Clinical use case: artifactual clone from FFPE removed from ovarian cancer with PARP inhibitor resistance

Beyond driver mutations and global signatures, FFPE artifact filtration can directly impact the analysis of tumor clonal structure, with direct implications for surveillance studies during disease progression and treatment response.

We present a case study that highlights the clinical utility of FFPErase in an ovarian cancer patient with a germline BRCA2 mutation

Fig. 3 | Artifact Classifier. **a** Graphical representation of training and validation process from mutation labeling to validation stratification on the basis of centers. **b** Parallel axis plot showing feature distributions for random sampling of 6704 SNVs (50% shared, 50% artifact). Final axis categorizes FFPE unique mutations as 1 and mutations shared with FF as 0 for the ARTIFACT class. **c** PCA reducing feature space to two principal components and shading mutations as shared with FF (blue) or FFPE unique (green) for SNVs from all samples excluding hypermutator cases (1823304 artifacts and 560,618 real). Axis legends adjusted for extreme outliers (greater than 0.00001 percentile and less than 99.999 percentile). **d** AUC for ROC and Precision-

Recall for 500 random training samplings of 750 (50% shared, 50% artifacts) mutations from 3 samples per center (nine total) and tested on all other sample mutations. **e** Timepoints for an example tumor (H135022) derived from FF and FFPE pre and post filtration are shown in the form of a Circos plot as described in Fig. 2. FFPE, formalin-fixed, paraffin-embedded; PCA, principal component analysis; FF, fresh frozen; SNV, single nucleotide variant; AUC, area under curve; ROC, receiver operating characteristic. Raw data for panels b-e can be accessed at the dbGAP, EGA, and 100,000 genomes studies. Source data can be generated using the tool repository. Panel a created in BioRender. Papaemmanuil, E. (2025) <https://BioRender.com/nzk6z38>.

and a heterogeneous response to platinum and PARP inhibitors (PARPi). WGS data were generated for the two responsive pre-treatment FF specimens and one resistant post-therapy FFPE sample with a BRCA2 reversion mutation (Fig. 6b). Both pre-treatment FF tumors had TMBs of 4.09 and 4.40 with high HRDetect scores (97 and 96%). In contrast, prior to filtering the FFPE sample had a high TMB (20.19) and low HRDetect score (37%). FFPErase designated 80% of SNVs and 43% of indels as artifacts refining TMB estimates to 4.44 and HRDetect score to 99% in the processed data.

Prior to filtering, clonal reconstruction across three samples identified 8 clones. HRD-related signature SBS3 was detected in all clones except for two that were unique to FFPE (5 and 7). Filtration with FFPErase demonstrates that clone 7 was primarily defined by artifactual mutations (CCF: 20.2%, 92% artifacts) and corrects the mutational profile of clone 5 (CCF: 77.8%, 17% artifacts). Notably mutation signature analysis of clone 5 post FFPErase filtering showed no evidence of SBS3 suggesting HR proficiency⁴². This is consistent with the BRCA2 reversion. Clone 5 also had high contributions of cisplatin signatures (SBS31/35) (Fig. 6c, d, Supplementary Fig. 8)⁴³. Taken together, cleanup with FFPErase enables accurate genomic characterization of a clone associated with treatment resistance that would have been missed without correction.

Discussion

Challenges in performing WGS in FFPE biospecimens pose a critical barrier to adoption of WGS in clinical oncology. In this study, we deliver a comprehensive meta-analysis of three independent FFPE WGS datasets with matched FF and germline control WGS data ($n=56$ trios)¹⁸. We study the genome-wide presentation of FFPE-related artifacts across mutation classes including SNVs/indels, SVs and global mutational signatures inclusive of HRD, MSI and TMB.

Despite variation in FFPE preparation and sequencing protocols, we identified common artifact patterns across the three centers, suggesting that these are associated with core components of FFPE processing. Consistent with prior studies, SNV/Indel calling is confounded by an enrichment of FFPE-associated artifacts in 95% of tumors. Artifact mutations in FFPE are enriched for SBS37, insertions and repeat-mediated deletions. While FFPE-associated artifacts do not impair driver mutation detection and prioritization in our dataset, they do affect detection accuracy of clinically relevant mutation signatures such as HRD. For example utilizing FFPE WGS data as input, HRDetect fails to accurately classify 7/7 patients with HRD+ status in FF data, whilst CHORD only captures 3/7 HRD+ cases. Whilst genome-wide TMB is increased in FFPE data, coding TMB is not, which is consistent with clinical testing data⁴⁴. SV calling was mostly consistent between FF and FFPE when utilizing consensus calling, requiring two out of three callers to call an event. SV sensitivity was predominantly affected by variation in coverage, which may result in a loss of detection and reporting of specific SVs; however, genome-wide SV profiles (i.e., HRD) are preserved.

The enrichment of artifactual calls highlights a clear need to differentiate artifacts from real mutations in FFPE-derived WGS data. For SVs, employing consensus calling greatly increases precision, but not for small mutations. To this end, we leverage the data in this analysis and develop FFPErase, a machine learning model that is informed by the analysis of SNVs/indels from matched FF/FFPE WGS data from three

centers. FFPErase achieves high accuracy in artifact classification increasing the precision of mutation calling by 40% without filtering out driver events. Benchmarking showed improved model performance when in-house and external samples were both included in training. This observation indicates that FFPE-associated artifacts are generalizable across FFPE processing workflows, yet center-specific artifacts are also present. FFPErase delivers a center-agnostic FFPE filtering workflow. Most importantly, this tool enables the characterization of clinically relevant mutation signatures like HRD without reducing TMB associated with biologically relevant signals such as MSI.

Existing methods like FFPESig and FFPolish have limited capabilities in artifact removal, focusing solely on SNVs, training on targeted data or lacking a per-mutation artifact prediction. FFPEimpact is designed to score the extent of FFPE damage and does not remove artifacts from the data. Our comparative analysis to these tools shows that FFPErase achieves the highest accuracy in artifact detection (95%) and cosine similarity between matched FF and FFPE contexts (0.99), while critically preserving 95% of shared SNVs. To ensure generalizability, the tool was further tested against a fourth, external dataset of 69 tumor samples from 2 never seen cancer types achieving 97% accuracy. Notably, FFPErase is the only tool benchmarked against an FDA-approved targeted panel assay. FFPErase not only retains 99% of clinically relevant events reported by MSK-IMPACT, but also uncovers additional clinically significant events in 41% of patients, primarily SVs.

Sequential sampling is increasingly implemented in clinical practice, as a means to understand disease progression and treatment response. We present a case study that shows how the FFPE-related artifacts affect resolution and interpretation of tumor clonal phylogenies. Sequential WGS profiling of an ovarian cancer patient with heterogeneous response to PARP inhibition shows the emergence of a treatment-resistant clone in a post-therapy FFPE resection. Deconvolution of the subclonal structure and HRD status is challenged by the presence of artifacts in the FFPE resection. Upon filtering with FFPErase, we deliver a clean representation of the clonal architecture of the tumor and an accurate clone-specific assessment of HRD status that informs disease presentation and treatment resistance.

A limitation of our study is that FFPErase does not address hypersegmentation in CN from FFPE-derived WGS data. Perhaps adaptation of methods such as dryclean or ACE can support this, but would need significantly larger datasets^{45,46}. Additionally, although we demonstrate that high depth coverage can be achieved in FFPE, optimization of laboratory protocols to improve library quality and sequencing depth is warranted but beyond the scope of this study. We note that applying FFPErase to targeted/exome sequencing data rather than WGS would require retraining the model using relevant data (e.g., exome or panel) to account for variations in sequencing depth and coverage patterns between these assays.

Our work provides a clear path forward for the implementation of FFPE-derived WGS datasets in oncology through the combination of consensus variant calling and artifact filtration methods such as FFPErase. Applications such as FFPErase enable the utilization of FFPE biospecimens and ultimately help democratize the development of clinical WGS workflows in a wide variety of settings (academic centers, community hospitals, etc.) and accelerate correlative studies for biomarker discovery.

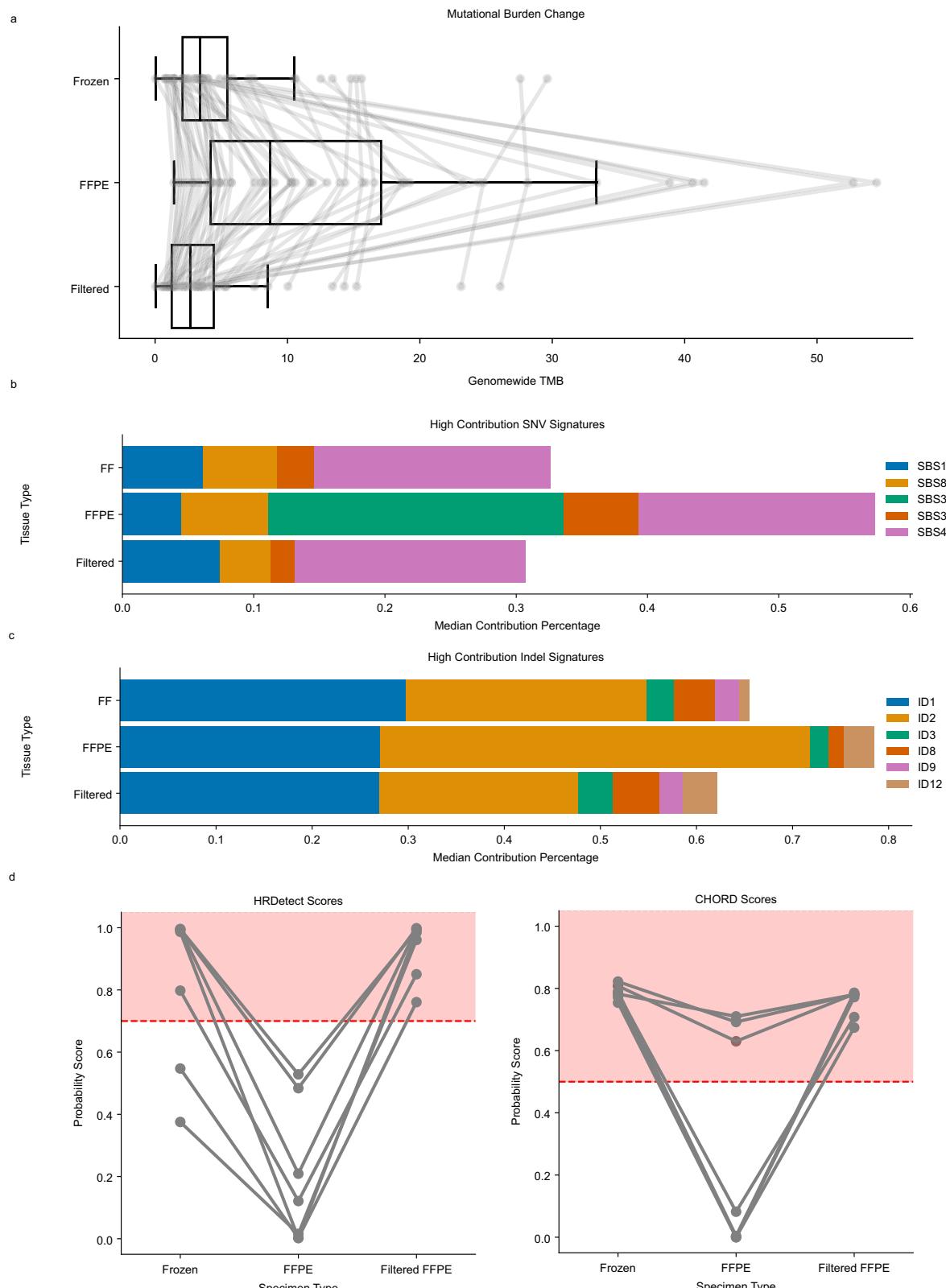
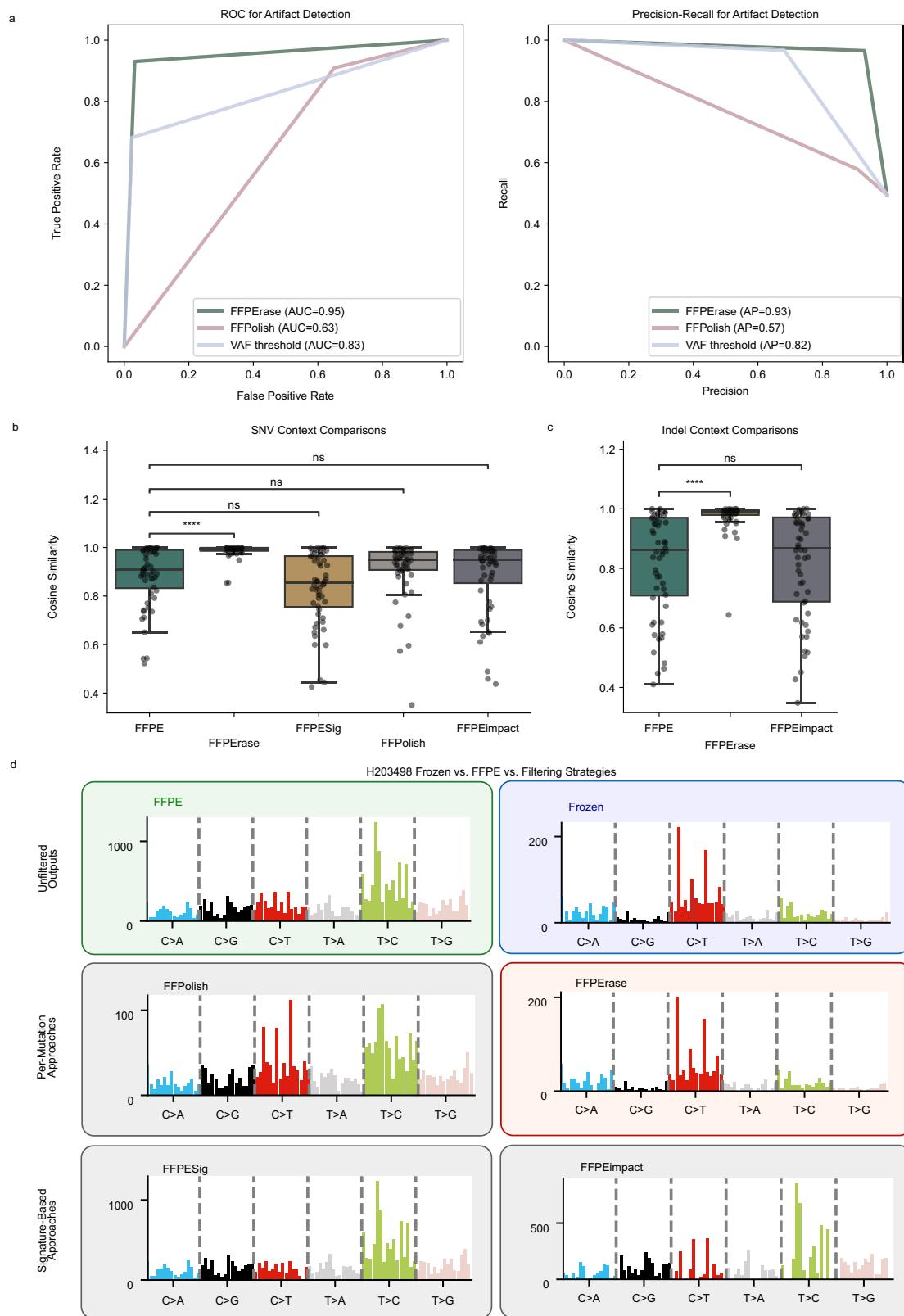


Fig. 4 | Filtration Impact on Global Signatures. **a** Boxplots showing distribution of genomewide TMB in both FF and FFPE samples prior to filtration with FFPFerse compared to FFPE samples post filtration for all 56 matched FF/FFPE pairs. The line in the middle of each box represents the median, while the upper and lower ends of each box represent the interquartile range (IQR) (25–75th percentile) and the whiskers represent 1.5x the IQR. **b** Stacked bar plot showing median contribution percentage of highest contribution signatures in SNVs for FF, FFPE, and filtered FFPE. **c** Same as (b) for indel signatures. **d** Distribution comparisons for FF vs. unfiltered FFPE vs. filtered FFPE of HRDetect (Frozen vs. FFPE $p = 9.876e-03$, Frozen

vs. Filtered $p = 1$) and CHORD (Frozen vs. FFPE $p = 4.281e-03$, Frozen vs. Filtered $p = 2.479e-01$) scores for 7 HRD positive cases by FF scoring with shaded regions for each tool's HRD threshold. (two-sided Mann-Whitney U test) TMB, tumor mutational burden; FF, fresh frozen; FFPE, formalin-fixed, paraffin, embedded; COSMIC, Catalog Of Somatic Mutations In Cancer; SBS, single base substitution; ID, insertion/deletion; HRD, homologous recombination deficiency. Source data for all panels are provided in Supplementary Data 6, 8, 13 and the data repository. Source data are also provided as a Source Data file.



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. The rest of the data was sourced from public or request-for-access repositories as described^{18,24}.

Fig. 5 | Comparison to Existing Methods. **a** ROC and PR AUC plots comparing FFPErase, FFPolish, and a VAF thresholding method (setting a filter on every mutation with VAF < 0.2*sample purity) results on all samples within the cohort. **b** Boxplot showing the distribution of mutational context cosine similarities as compared to FF data for unfiltered and filtered FFPE by use of FFPERase, FFPSig, FFPolish and FFPEimpact with statistical annotations between unfiltered and filtered data by each tool for all 56 matched pairs. (two-sided Mann–Whitney U test, FFPErase p = 6.024e-07, FFPSig p = 9.951e-02, FFPolish p = 1; FFPEimpact p = 1; ns = 5e-02 < p < =1, * = 1e-02 < p < = 5e-02, ** = 1e-03 < p < = 1e-02, *** = 1e-04 < p < = 1e-03, **** = p < 1e-04) The line in the middle of each box represents the median, while the upper and lower ends of each box represent the interquartile range (IQR) (25th–

75th percentile) and the whiskers represent 1.5x the IQR. **c** Same as **(b)** for indels for all 56 matched pairs (two-sided Mann–Whitney U test, FFPErase p = 1.838e-10, FFPEimpact p = 1). **d** Example 96-mutational context plots for substitutions from H203498 FF/FFPE alongside the filtered FFPE profiles using FFPErase, FFPSig, FFPolish and FFPEimpact. SBS, single base substitution; FF, fresh frozen; FFPE, formalin-fixed, paraffin-embedded; ROC, receiver operating characteristic; PR, precision recall; AUC, area under curve. Raw data for panels **a** and **d** can be accessed at the dbGAP, EGA, and 100,000 genomes studies. Source data for **(b, c)** are provided in Supplementary Data 16 and the data repository. Source data are also provided as a Source Data file.

Study design and WGS analysis

Raw data for 56 tumor specimens were collected across three centers, representing independent workflows for FFPE sample preparation and sequencing (Supplementary Data 1). For each tumor, trios of samples for each tumor, including two resections, one prepared as FF and the other as FFPE, as well as a matched normal sample from peripheral blood, were sourced. An additional 35 FFPE-only tumor specimens with matched normal peripheral blood were collected and processed from MSKCC. Patient specimens spanned 16 cancer indications and age at resection ranged from 5 to 89 years. Both FF and FFPE tumors were assessed using the same analytical pipelines based on consensus calling across all variant classes^{9,30}.

Sample Collection

Patient samples from three independent centers were collected, processed and sequenced separately. A trio of samples was collected for each patient, two tumor samples from the same resection that were prepared as FF and FFPE samples, and one matched normal sample for matched analysis. Ten trios were collected from MSKCC, six from BCC, and forty from Oxford University Hospitals.

Tissue fixation, processing, and embedding

For the 45 subjects from MSKCC, surgical samples were placed into standard tissue cassette(s) and completely immersed in fresh 10% Neutral Buffered Formalin (NBF) and loaded into the Peloris tissue processor. Surgical routine samples were processed for about 7 h in multiple steps per reagent consisting of formalin (90 min), alcohol (125 min), xylene (90 min) and paraffin (90 min). Cassettes were then removed and embedded in paraffin blocks using the Tissue-Tek AutoTEC a120.

DNA extraction

For the 45 subjects from MSKCC, tumor DNA was extracted from fresh frozen (FF) OCT embedded tissue biopsies and matched normal from peripheral blood using the DNeasy Blood & Tissue Kit (Qiagen catalog # 69504) according to the manufacturer's protocol for all but one sample where the AllPrep DNA/RNA Mini Kit (Qiagen catalog # 80204) was used. FFPE tissue was deparaffinized using heat treatment (90 °C for 10' in 480 µL PBS and 20 µL 10% Tween 20), centrifugation (10,000 x 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 Buffer ATL at 55 °C. DNA isolation proceeded with the Mag-Bind Blood & Tissue DNA HDQ 96 Kit (Chemagic SKU M6399-01) according to the manufacturer's protocol modified by replacing AW2 buffer with 80% ethanol. All DNA was eluted in 0.5X Buffer AE.

For the six subjects from BCC, tumor DNA was extracted from FF tissue biopsies and matched normal. FFPE tissue was deparaffinized using xylene and then extracted using the QIAamp DNA FFPE Tissue Kit (Qiagen catalog # 56404) according to the manufacturer's protocol.

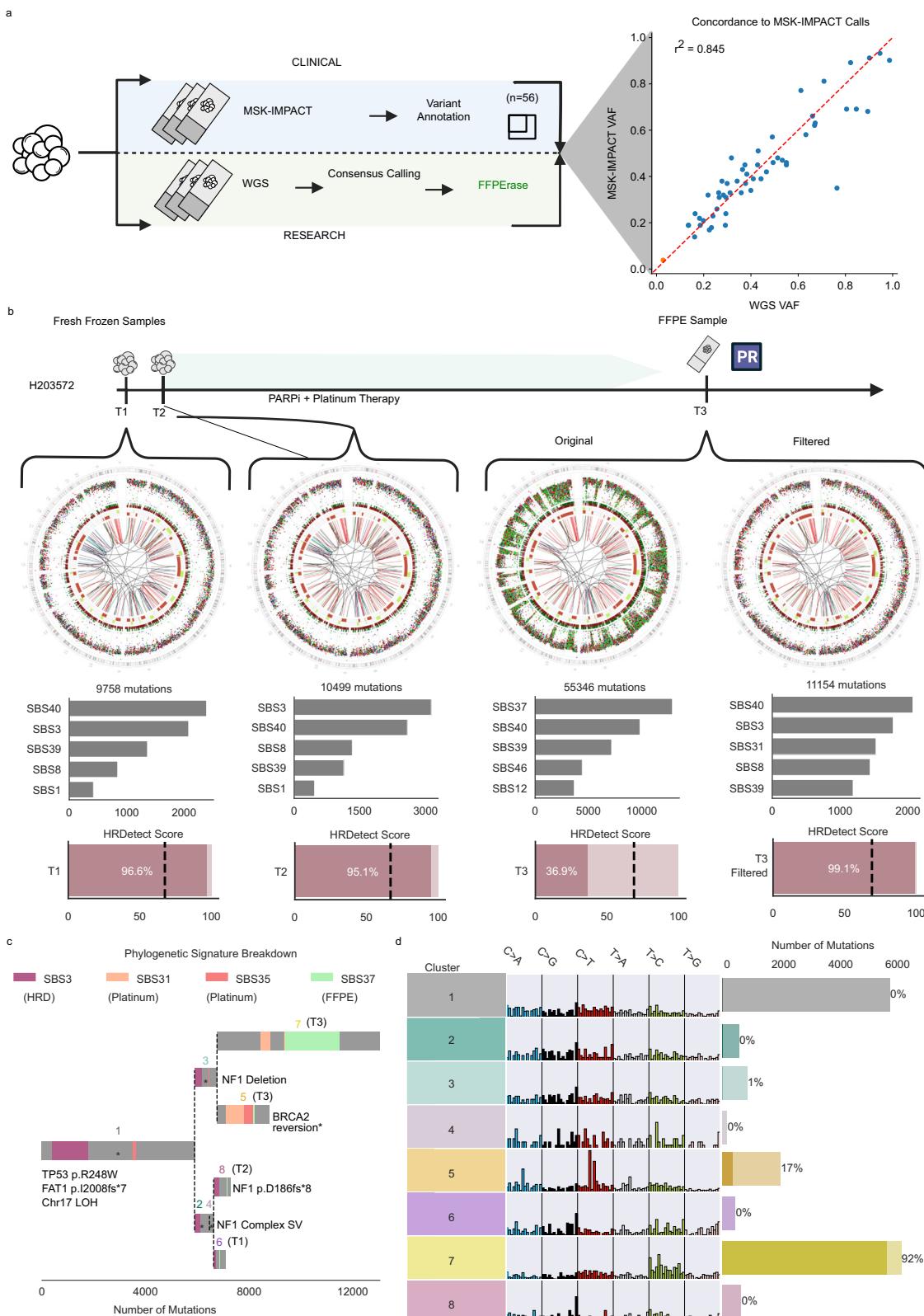
DNA extraction for the 40 samples from Oxford University Hospitals is detailed in the associated publication¹⁸.

Whole genome sequencing

For the 45 subjects from MSKCC, after PicoGreen quantification and quality control by Agilent BioAnalyzer, 500 ng of genomic DNA were sheared from all FF samples and where possible from all but 5 FFPE samples (minimum gDNA: 140.9 ng) using a LE220-plus Focused-ultrasonic (Covaris catalog # 500569). Sequencing libraries were prepared using the KAPA Hyper Prep Kit (Kapa Biosystems KK8504) and were subjected to a 0.5X size select using aMPure XP beads (Beckman Colter catalog # A63882) after post-ligation cleanup. For the FF samples, PCR-free libraries were pooled equivolume for sequencing and for the FFPE samples five cycles of PCR were performed. Samples were then 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).

For the six subjects from BCC, FF samples were prepared as follows: One microgram of genomic DNA was arrayed in a 96-well microtitre plate and subjected to shearing by sonication (Covaris). Sheared DNA was end-repaired and size selected using AMPure XP beads targeting a 300–400 bp fraction. After 3' A-tailing, full length TruSeq adaptors were ligated. Libraries were purified using AMPure XP beads. Library fragment sizes were assessed using an aliquot of PCR amplified library DNA on the Agilent 2100 Bioanalyzer DNA1000 chip, or Caliper GX DNA1000 chip. The PCR-free library concentration was quantified using a qPCR Library Quantification kit (KAPA, KK4824). Four cycles of PCR were applied to rescue low-yield libraries prior to sequencing with paired-end 100 base reads on the Illumina HiSeq platform according to manufacturer recommendations. FFPE was prepared as follows: gDNA was normalized to 250 ng in a volume of 62 µL elution buffer (Qiagen) and transferred into a microTUBE plate for shearing on an LE220 (Covaris) acoustic sonicator using the conditions: Duty Factor - 20%, Peak Incident Power - 450 W, Cycle per burst - 200, Duration - 2 × 60 s with an intervening spin. FFPE protocol has a dominant DNA peak in the size range between 300 and 400 bp. To improve the library quality of FFPE-derived DNA, solid-phase reversible immobilization (SPRI) bead-based size selection was performed before library construction to remove smaller DNA fragments from highly degraded FFPE DNAs. If not removed early in the library construction process, these smaller fragments would otherwise dominate the final amplified library. FFPE DNA damage and end-repair and phosphorylation were combined in a single reaction using an enzymatic premix (NEB), then bead purified using a 0.8:1 (bead: sample) ratio to remove small FFPE fragments. Repaired DNA fragments were next A-tailed for ligation to paired-end, partial Illumina sequencing adapters then purified twice with PCR Clean DX beads (0.9:1 ratio). Full-length adapter products were achieved by performing 8 cycles PCR with primers introducing fault-tolerant hexamer “barcodes” allowing multiplexing of libraries. Indexed PCR products were double purified with 0.9:1 beads. Concentration of final libraries was determined using size profiles obtained from a high sensitivity Caliper LabChip GX together with Quant-iT (Invitrogen) quantification. Libraries were sequenced with paired-end 150 base reads on a HiSeqX instrument running v2.5 chemistry (Illumina).

WGS for the 40 samples from Oxford University Hospitals is detailed in the associated publication¹⁸.



Bioinformatic analysis

Sequencing data was gathered and analyzed analogously using the Isabl platform 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, MSI score, homologous recombination deficiency); 4. Variant classification; and, 5. The generation of a clinical prototype

summary report⁹. 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. 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>)⁴⁷. The wrapper includes marking of duplicates using Picard. For tumor samples

Fig. 6 | Detailed Use-Case for PARP-inhibitor treated Ovarian Case H203572. **a** Graphical representation of concordance test against FDA-approved MSK-IMPACT panel calls with breakout scatterplot of called mutations by VAF in WGS vs. VAF in MSK-IMPACT. **b** Timeline for ovarian cancer patient H203572 in respect to WGS samples with corresponding circos plots described in Fig. 3 for: pre-therapy right ovary resection from FF tissue, pre-therapy infracolic omentum resection from FF tissue, post- therapy distal pancreas resection from FFPE tissue both pre and post filtration with FFPErase. Additional panels below consisting barplots showing the absolute number of mutations attributed to the five SBS signatures with the highest exposure in the tumor, below that the 96 trinucleotide contexts of SNVs, and lastly the global HRD probability score from HRDetect for each sample. **c** Phylogenetic tree for pre-filtration clonality analysis with each bar representing

the number of mutations per abstracted clone and colored by proportions of study relevant SBS signatures with gray denoting other signature contributions. Drivers are annotated on the clone they are associated with and are matched by using mean CCF per clone as compared to the mutation CCF. **d** 96 trinucleotide contexts for each SNV clone pre-filtration along with bars on the right that are length adjusted for number of mutations within the clone and show percentage of variants filtered by FFPErase. VAF, variant allele frequency; FF, fresh frozen; SBS, single base substitution; FFPE, formalin-fixed, paraffin-embedded. Source data for panels a is provided in Supplementary Data 17 and the data repository. Source data are also provided as a Source Data file. Raw data for (b-d) can be accessed at the dbGAP study. Panel 6b created in BioRender. Papaemmanuil, E. (2025) <https://BioRender.com/cdp997a>.

ensemble variant calling for each variant class (substitutions, insertions and deletions and structural variations) was performed. High confidence somatic mutations are classified with regards to their putative role in cancer pathogenesis and statistical post-processing enables the derivation of MSI scores, mutation signatures and HRD scores. Clinical relevance of mutations in common cancer genes was annotated using OncoKb, COSMIC, Ensembl Variant Effect Predictor, VAGrENT, gnomAD and ClinVar databases. Details of the variant calling and annotation can be found in the Supplementary Information.

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>)⁴⁸. 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>)⁴⁸⁻⁵⁰. 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 normals⁵¹. 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^{33,52,53}. Finally, microsatellite-instability status was assessed using MSI Sensor (v0.5) (<https://github.com/ding-lab/msisensor>)³⁵ and MIMcall (<https://github.com/afujimoto/MIMcall>)³⁸.

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>)^{54,55}. VCF files for SVs were merged using MergeSVvcfs (v1.0.2, <https://github.com/papaemmelab/mergeSVvcf>). High-confidence mutations were designated as those that were passed by at least 2 callers and >600 bp in length for SV calls.

Calculation of TMB

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 and 2900 to approximate length of the human genome in Mb. Values

greater than ten mutations per Mb were considered hypermutators, thresholds set by the study in Grobner et al.⁸.

Identification of mutation signatures for substitutions and indels

Mutational signature analysis was performed with the MutationalPatterns package (v3.4.1, <https://bioconductor.org/packages/release/bioc/html/MutationalPatterns.html>) for both substitutions and indels using signatures for the SBS and ID variant classes from COSMIC Mutational Signatures (v3.1) with the addition of Temozolomide signature from Kucab et al. for substitution signatures.

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/dpclust>) 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 which was optimized internally⁴⁸. Clonal ordering was deduced using clonevol (v0.99.11, <https://github.com/hdng/clonevol>)⁵⁶. Mutational signatures were computed in each cluster independently. Figures were generated with matplotlib (v3.1.0, <https://matplotlib.org/>).

Development of artifact classification model

The artifact classifier was designed using scikit-learn (v0.24.1, <https://scikit-learn.org/0.24/>) and imbalanced-learn (v0.8.0, <https://imbalanced-learn.org/stable/>). Custom mutational feature extraction was performed on the high confidence call set produced from Isabl for both somatic substitutions and indels to format for training and test. Optimization of classification was done using nextflow (<https://github.com/nextflow-io/nextflow>) in order to parallelize artifact detection and achieve computation time of under 2 hours. The final artifact classifier was developed to run given a VCF and a BAM as input and was trained using a balanced Random Forest model. It leverages 33 features per single nucleotide variant and 29 features per indel that are either at the sample or mutation level and has a depth of 100 trees. Test strategies involving training and test data from the same center employed a leave-one-out cross validation strategy where a model was produced for each test sample using equal amounts of randomly sourced mutations from all other available samples for training. This was done for the 1-center and 3-center approaches. Machine learning architectures were explored and compared including a convolutional neural network approach leveraging pileup images⁵⁷. These are detailed in Supplementary Information. We implemented the code in such a way that the model can be retrained with new data utilizing the preprocessing and training modules⁵⁸.

Feature extraction

Features are extracted from tumor BAMs using a VCF to designate mutation loci and leveraging a combination of custom scripts and

Table 1 | Driver Annotations for FFPE WGS to MSK-IMPACT Validation

Sample ID	Disease	IMPACT Drivers (Called in FFPE: 0/1)	Additional Findings by WGS (Tiled by IMPACT: 0/1)	Relevance
H136375_T06	Osteosarcoma	TP53 p.X261_splice (1) RAC1 AMP (1)	TERT SV (0) PRKACA DEL (0)	Oncogenic Oncogenic
H156416_T02	Leiomyosarcoma			
H158219_T01	Sarcomatoid Malignant Neoplasm NOS	ATRX p.Q973* (1) TP53 p.P278S (1) PIK3CA p.H1047R (1) FBXW7 LOSS (1)	DLG2 DEL (0) ZFHX3 DEL (1)	Oncogenic Oncogenic
H196094_T02	Leiomyosarcoma			
H196388_T02	Glioblastoma Multiforme	TP53 p.V157F (1) CDKN2A/B LOSS (1) BRAF-KIAA1549 SV (1) JAK1 p.P861fs*4 (1) 17p LOSS (1) 10 LOSS (1)	ATRX DUP (1)	Oncogenic
H196388_T03	Glioblastoma Multiforme	TP53 p.V157F (1) CDKN2A/B LOSS (1) BRAF-KIAA1549 SV (1) JAK1 p.P861fs*4 (1) 17p LOSS (1) BRAF AMP (1) 10 LOSS (1)	ATRX DUP (1) 17p CNLOH (1)	Oncogenic Prognostic
H200646_T02	Embryonal Rhabdomyosarcoma	MGA (1) MDM2 AMP (1)		
H201688_T01	Diffuse Leptomeningeal Glioneuronal Tumor			
H202874_T02	Colon Adenocarcinoma	PIK3CA p.H1047R (1) ATM p.M2384fs*19 (1) NF1 p.T2196fs*5 (1) NF1 p.D946fs*1 (1) ARID1A p.D1850fs*34 (1) High TMB (1)	MSI High (MSISensor + Signatures)	Therapy Informing
H203305_T01	Adrenocortical Carcinoma	CTNNB1 p.G34E (1) RB1 DEL (1)	PRKAR1A p.*338E (1) KDM6A DEL (1)	Oncogenic Oncogenic
H206102_T01	Undifferentiated Sarcoma	NF1 INV (1) EGFR AMP (1)	TP53 DEL (1)	Oncogenic
H209568_T01	Rhabdomyosarcoma	TP53 p.Y126C (1) NF1 DEL (1)		
H209570_T02	Embryonal Rhabdomyosarcoma	DICER1 p.E1705K (1) TP53 p.E258* (1) 17p LOSS (1)		
H209578_T01	Astrocytoma	KRAS p.Q61H (1)		
H210036_T01	Undifferentiated Sarcoma	TERT SV (1) CXXC5-MN1 (1)		
H210038_T01	Pilocytic Astrocytoma	NF1 DEL (1)	NF1 DEL (Overlapping) (1)	Oncogenic
H210039_T01	Schwannoma	NF2 p.Q410* (1) CDKN2A/B/MTAP LOSS (1) 22 CNLOH (1)	ATM splice SNV (1)	Oncogenic
H210178_T01	Low-Grade Neuroepithelial Tumor	FGFR1 p.N577K (1)		
H210491_T01	Anaplastic Astrocytoma	TP53 p.R273C (1) TP53 p.G226D (1) PIK3CA p.H1065Y (1) PTEN p.K267fs*9 (1) KRAS p.G13D (1) IDH1 p.R132C (1) STK11 p.A241T (1) SMARCA4 p.W1178*4 (1) TCF7L2 c.875+1 G > A (1) MGA c.2092+1 G > A (1) SMAD3 p.W30* (1) REST p.P893fs*5 (1) MSH6 p.F1088fs*2 (1) BCOR p.R1053fs*2 (1) MLH1 LOSS (1) High TMB (1) ATRX p.V1347fs*2 (1) PIK3R1 p.I442del (1)	SBS11	Therapy Informing
H210492_T01	Myoepithelial Carcnioma	EWSR1-KLF1 (1)		

Table 1 (continued) | Driver Annotations for FFPE WGS to MSK-IMPACT Validation

Sample ID	Disease	IMPACT Drivers (Called in FFPE: 0/1)	Additional Findings by WGS (Tiled by IMPACT: 0/1)	Relevance
H210494_T02	Alveolar Rhabdomyosarcoma	KMT2D p.R755Pfs*3 (1) PAX3-FOXO1 (1)		
H210579_T01	Neuroblastoma	NF1 DUP (1) MYCN AMP (1)	TERT SV (1) 17 CNLOH (1)	Prognostic Prognostic
H210584_T01	Neuroblastoma		DLG2-EDC4 (+ 11q loss) (0)	Prognostic
H210588_T01	Undifferentiated Sarcoma	EWSR1-NF2 (1)	BCORL1 DEL (0)	Oncogenic
H210589_T01	Glioblastoma	TP53 p.D281Y (1) TP53 p.R175H (1) TERT promoter SNV (1) PTPN11 p.E76K (1) CDKN2A/B/MTAP DEL (1) CDKN2C DEL (1) PTEN DEL (1)	RPL5 p.E133* (0) MGA p.L1376fs*20 (1) FAT1 SV (1)	Oncogenic Oncogenic Oncogenic
H210624_T01	Gastric Remnant Adenocarcinoma	ARID1A p.Q487fs*132 (1) KRAS p.G12C (1) TGFB2 DEL (1)		
H210686_T01	Medulloblastoma	PTCH1 p.Y1181* (1) PTCH1 p.A1088fs*59 (1) MAX p.R60Q (1) PIK3R1 p.C659fs*3 (1) TERT promoter SNV (1) KDM6A p.W1193* (1)		
H210687_T02	Embryonal Rhabdomyosarcoma			
H210818_T01	Sertoli-Leydig Cell Tumor	AKT1 p.R76_C77insWRPRPNTFIIR (1)		
H210819_T01	Uterine Adenosarcoma	DICER1 p.D1810F (1) TP53 p.D259V (1) DIS3 c.580+2 T > G (1)		
H210830_T01	Dysembryoplastic Neuroepithelial Tumor	FGFR1 DUP (1)		
H210834_T01	Neuroblastoma	SMARCA4 p.E1579* (1)		
H210835_T01	Spindle Cell Neoplasm	CTNNB1 p.T41I (1)		
H210844_T01	Schwannoma		SH3PXD2A-HTRA1 (0)	Diagnostic
H210862_T01	Pineal Parenchymal Tumor of Intermediate Differentiation	JAZF1-ETV1 (1)		

Data for driver mutations discovered by MSK-IMPACT and whether or not they were also discovered in WGS from FFPE. Additionally any findings that were found only by WGS and whether or not they were tiled in the panel test. (SV Structural Variant, DEL Deletion, DUP Duplication, CNLOH Copy Neutral Loss of Heterozygosity, AMP Amplification).

Picard's CollectSequencingArtifactMetrics (v2.25.6, <https://github.com/broadinstitute/picard>). Features were assessed and selected using permutation feature importance and random forest mean decrease in impurity. Feature correlation was also assessed using Spearman's correlation to determine relationships between all features.

Reporting summary

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

Data availability

The raw data from MSKCC have been deposited in the dbGAP database under accession code phs002620.v1.p1 [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002620.v1.p1] for the fresh frozen and normals of the 10 trio samples used for training and the matched FFPE have been deposited under accession code phs004176.v1.p1 [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs004176.v1.p1]. The remaining MSKCC raw data from FFPE tumors and matched normals used for validation have also been deposited under the dbGAP accession code phs004176.v1.p1 [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs004176.v1.p1]. 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 raw data from BCC generated in this study has been deposited in the EGA database under accession code EGAD00001011331. The raw data from the Oxford dataset was accessed upon request from the original study authors. Primary data from the 100,000 Genomes Project, which are held in a secure research environment, are available to registered members of the Research Network. Membership of the Research Network is open to all individuals, students, or staff affiliated with UK academic research institutions, NHS trusts, relevant charitable organizations, foreign universities and research institutions, governmental departments, and foreign healthcare organizations involved in significant research activity. The datasets for the validation on the National Cancer Institute (NCI) Cancer Genome Characterization Initiative (CGCI) Burkitt Lymphoma Genome Sequencing Project (BLGSP) and HIV+ Tumor Molecular Characterization Project - Cervical Cancer (HTMCP-CC) is available by request through <https://portal.gdc.cancer.gov/> under accession codes phs000527.v1.p1 [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000527.v1.p1], phs000528.v1.p1 [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000528.v1.p1]. Summary and processed data for the figures are available in the data repository at https://github.com/papaemmelab/Domenico_NC_FFPE. Source data are provided with this paper.

Code availability

FFPErase has been made available online at <https://github.com/papaemmelab/nf-ffperase> (<https://doi.org/10.5281/zenodo.16326434>). Scripts for generating the figures are provided where possible at https://github.com/papaemmelab/Domenico_NC_FFPE.

References

1. Zehir, A. et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat. Med.* **23**, 703–713 (2017).
2. Beaubier, N. et al. Clinical validation of the tempus xT next-generation targeted oncology sequencing assay. *Oncotarget* **10**, 2384–2396 (2019).
3. Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).
4. 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).
5. Horak, P. et al. Comprehensive genomic and transcriptomic analysis for guiding therapeutic decisions in patients with rare cancers. *Cancer Discov.* **11**, 2780–2795 (2021).
6. 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 (2021).
7. Roepman, P. et al. Clinical validation of whole genome sequencing for cancer diagnostics. *J. Mol. Diagn.* **23**, 816–833 (2021).
8. Gröbner, S. N. et al. The landscape of genomic alterations across childhood cancers. *Nature* **555**, 321–327 (2018).
9. Shukla, N. et al. Feasibility of whole genome and transcriptome profiling in pediatric and young adult cancers. *Nat. Commun.* **13**, 2485 (2022).
10. Do, H. & Dobrovic, A. Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. *Clin. Chem.* **61**, 64–71 (2015).
11. Srinivasan, M., Sedmak, D. & Jewell, S. Effect of fixatives and tissue processing on the content and integrity of nucleic acids. *Am. J. Pathol.* **161**, 1961–1971 (2002).
12. Haile, S. et al. Sources of erroneous sequences and artifact chimeric reads in next generation sequencing of genomic DNA from formalin-fixed paraffin-embedded samples. *Nucleic Acids Res.* **47**, e12 (2019).
13. Zhao, Y. et al. Whole genome and exome sequencing reference datasets from a multi-center and cross-platform benchmark study. *Sci. Data* **8**, 296 (2021).
14. Oh, E. et al. Comparison of accuracy of whole-exome sequencing with formalin-fixed paraffin-embedded and fresh frozen tissue samples. *PLoS One* **10**, e0144162 (2015).
15. Sah, S. et al. Functional DNA quantification guides accurate next-generation sequencing mutation detection in formalin-fixed, paraffin-embedded tumor biopsies. *Genome Med.* **5**, 77 (2013).
16. Mathieson, W. & Thomas, G. A. Why Formalin-fixed, paraffin-embedded biospecimens must be used in genomic medicine: an evidence-based review and conclusion. *J. Histochem. Cytochem.* **68**, 543–552 (2020).
17. Schweiger, M. R. et al. Genome-wide massively parallel sequencing of formaldehyde fixed-paraffin embedded (FFPE) tumor tissues for copy-number- and mutation-analysis. *PLoS One* **4**, e5548 (2009).
18. 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).
19. de Schaetzen van Brien, L. et al. Comparative analysis of somatic variant calling on matched FF and FFPE WGS samples. *BMC Med. Genomics* **13**, 94 (2020).
20. Munchel, S. et al. Targeted or whole genome sequencing of formalin fixed tissue samples: potential applications in cancer genomics. *Oncotarget* **6**, 25943–25961 (2015).
21. Bailey, S. T. et al. High-quality whole-genome sequencing of FFPE samples. *J. Clin. Oncol.* **36**, e13500 (2018).
22. Oba, U. et al. An efficient procedure for the recovery of DNA from formalin-fixed paraffin-embedded tissue sections. *Biol. Methods Protoc.* **7**, bpac014 (2022).
23. Xiao, W. et al. Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing. *Nat. Biotechnol.* **39**, 1141–1150 (2021).
24. Dodani, D. D., Nguyen, M. H., Morin, R. D., Marra, M. A. & Corbett, R. D. Combinatorial and machine learning approaches for improved somatic variant calling from formalin-fixed paraffin-embedded genome sequence data. *Front. Genet.* **13**, 834764 (2022).
25. Guo, Q. et al. The mutational signatures of formalin fixation on the human genome. *Nat. Commun.* **13**, 4487 (2022).
26. Domenico, D. et al. High-resolution interpretable classification of artifacts versus real variants in whole genome sequencing data from archived tissue. In *Proceedings of the 2023 ICML Workshop on Computational Biology (ICML-CB 2023)*. (Honolulu, Hawai'i, USA 2023). Preprint at https://icml-compbio.github.io/2023/papers/WCBICML2023_paper116.pdf.
27. Basyuni, S. et al. Large-scale analysis of whole genome sequencing data from formalin-fixed paraffin-embedded cancer specimens demonstrates preservation of clinical utility. *Nat. Commun.* **15**, 7731 (2024).
28. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
29. Nguyen, L. et al. Pan-cancer landscape of homologous recombination deficiency. *Nat. Commun.* **11**, 5584 (2020).
30. Medina-Martínez, J. S. et al. Isabl Platform, a digital biobank for processing multimodal patient data. *BMC Bioinforma.* **21**, 549 (2020).
31. Chakravarty, D. et al. OncoKB: a precision oncology knowledge base. *JCO Precis. Oncol.* **2017**, 11 (2017).
32. Forbes, S. A. et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
33. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
34. Davies, H. et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.* **23**, 517–525 (2017).
35. Niu, B. et al. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics* **30**, 1015–1016 (2014).
36. Middha, S. et al. Reliable Pan-cancer microsatellite instability assessment by using targeted next-generation sequencing data. *JCO Precis. Oncol.* **2017**, 84 (2017).
37. Meier, B. et al. Mutational signatures of DNA mismatch repair deficiency in and human cancers. *Genome Res.* **28**, 666–675 (2018).
38. Fujimoto, A. et al. Comprehensive analysis of indels in whole-genome microsatellite regions and microsatellite instability across 21 cancer types. *Genome Res.* **30**, 334–346 (2020).
39. McGraw, K. O. & Wong, S. P. A common language effect size statistic. *Psychol. Bull.* **111**, 361–365 (1992).
40. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. Mutational patterns: comprehensive genome-wide analysis of mutational processes. *Genome Med.* **10**, 33 (2018).
41. Jin, H. et al. Accurate and sensitive mutational signature analysis with MuSiCal. *Nat. Genet.* **56**, 541–552 (2024).
42. Murciano-Goroff, Y. R. et al. Reversion mutations in germline BRCA1/2-mutant tumors reveal a BRCA-mediated phenotype in non-canonical histologies. *Nat. Commun.* **13**, 7182 (2022).

43. Boot, A. et al. In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors. *Genome Res.* **28**, 654–665 (2018).

44. Campbell, B. B. et al. Comprehensive analysis of hypermutation in human cancer. *Cell* **171**, 1042–1056.e10 (2017).

45. Deshpande, A., Walradt, T., Hu, Y., Koren, A. & Imlielinski, M. Robust foreground detection in somatic copy number data. *bioRxiv* 847681 <https://doi.org/10.1101/847681> (2019).

46. Poell, J. B. et al. ACE: absolute copy number estimation from low-coverage whole-genome sequencing data. *Bioinformatics* **35**, 2847–2849 (2019).

47. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

48. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).

49. Van der Auwera, G. A. & O'Connor, B. D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. (O'Reilly Media, 2020).

50. Jones, D. et al. cgpCaVManWrapper: 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).

51. 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).

52. Wala, J. A. et al. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* **28**, 581–591 (2018).

53. Cameron, D. L. et al. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Res.* **27**, 2050–2060 (2017).

54. McLaren, W. et al. The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).

55. Menzies, A. et al. VAGrENT: variation annotation generator. *Curr. Protoc. Bioinforma.* **52**, 15.8.1–15.8.11 (2015).

56. Dang, H. X. et al. ClonEvol: clonal ordering and visualization in cancer sequencing. *Ann. Oncol.* **28**, 3076–3082 (2017).

57. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).

58. Domenico & Arango Ossa, Enabling whole genome sequencing analysis from FFPE specimens in clinical oncology, <https://github.com/papaemmelab/nf-ffperase>, (2025).

Acknowledgements

We acknowledge the MSK Biobank and Pathology Core Facility for tissue procurement and processing. This research was funded in part through the NIH/NCI Cancer Center Support Grant P30 CA008748. This work utilized resources from the High Performance Computing Group at Memorial Sloan Kettering Cancer Center. 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. D.D., M.F.L., G.G., J.E.A.O., and G.A. developed algorithmic infrastructure and performed bioinformatic analysis. N.S., E.S., and T.O. oversaw coordination of patient consent, sample processing, and clinical data acquisition for MSKCC samples. U.B. and J.S. oversaw biospecimen banking and pathology review for MSKCC samples. D.D., M.F.L., G.G., N.S., and E.P. oversaw variant

annotation and data interpretation. C.C. and N.M. executed laboratory processing of biospecimens and sequencing for MSKCC samples. P.R., A.S., and C.B. oversaw data acquisition and interpretation for Oxford dataset. J.S. and D.C. oversaw data acquisition and interpretation for the BCC dataset. A.M., B.W., S.P.S. and O.Z. oversaw data acquisition and interpretation of PARP inhibitor case report and O.Z. performed surgical resection and acquisition for case report. D.D., G.G., and E.P. prepared figures and tables. D.D., G.G., and E.P. wrote the paper with input from A.L.K. and J.E.A.O. All authors reviewed and approved the paper for submission.

Competing interests

G.G. is a consultant in Isabl Inc. B. W. reports a research grant from REPARSE Therapeutics paid to the institution, and has an immediate family member who is employed by AstraZeneca, outside the submitted work. A.L.K. is on the Scientific Advisory Board of Emendo Biotherapeutics, Karyopharm Therapeutics, Imago BioSciences, and Darwin-Health; is co-Founder and on the Scientific Advisory Board of Isabl; has equity interest in Imago BioSciences, Emendo Biotherapeutics and Isabl; is an inventor of MSK intellectual property licensed to Isabl; and receives royalty income from Labcorp. E.P. is a founder, equity holder and holds fiduciary roles in Isabl Inc. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-65654-7>.

Correspondence and requests for materials should be addressed to Elli Papaemmanuil.

Peer review information *Nature Communications* thanks Hidewaki Nakagawa, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions 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-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence 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 licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025