npj | precision oncology

Article

Published in partnership with The Hormel Institute, University of Minnesota

https://doi.org/10.1038/s41698-026-01390-5

EXaCT-2: an augmented and customizable oncology-focused whole exome sequencing platform

Peter Waltman[1,2,3], Pooja Chandra[1,2,3], Ken W. Eng[2,3,4], David C. Wilkes[2,5], Hyeon Park[5], Carlos Pabon[6], Princesca Delpe[1,2,3], Bhavneet Bhinder[1,2,3], Jyothi Manohar[2], Troy Kane[2], Evan Fernandez[5], Kathryn Gorski[1,2,3], Noah Greco[1,2], Manuele Simi[1,2,3], Jeffrey M. Tang[1,2,3], Pantelis Zisimopoulos[1,2,3], Abigail King[1,2], Majd Al Assaad[2,5], Theresa Ten Eyck[6], Douglas Roberts[6], Jorge Monge[7], Francesca Demichelis[8], Wayne Tam[9], Madhu M. Ouseph[5], Alexandros Sigaras[1,2,3], Himisha Beltran[10], Hannah Rennert[5], Neal Lindeman[5], Wei Song[5], James Solomon[5], Juan Miguel Mosquera[2,5], Rob Kim[2], Jeffrey Catalano[2,5], Duane C. Hassane[2,3,11], Michael Sigouros[1,2], Olivier Elemento[1,2,3,12], Alicia Alonso[2,11,12,13] & Andrea Sboner[2,3,5,12,13] ✉

We developed and benchmarked Exome Cancer Test v.2.0 (EXaCT-2), a novel whole-exome sequencing (WES) assay based on Agilent's SureSelect hybrid-capture technology and expanded with custom probes targeting cancer-informative genomic regions. EXaCT-2 provides ~1,400 cancer genes with the depth of coverage typical of targeted panels, while achieving the genomic breadth to detect somatic copy number alterations (SCNAs), common cancer-related rearrangements, oncogenic viruses and B-cell receptor (BCR) clonotypes. Evaluated with a cancer patient cohort of 244 matched tumor/normal pairs and compared with clinically-validated results, EXaCT-2 achieved a mean sequencing depth of ~400× for critical cancer genes and ~100× for the remainder of the exome, with SCNA characterization showing improved boundary detection and overall segmentation. The assay demonstrated enhanced sensitivity for detecting sub-clonal, low-allele-frequency mutations missed by standard exome assays, such as mutations in GC-rich genes like KRAS. Analysis is performed by a modular, bespoke pipeline that leverages a workflow manager (Nextflow), in combination with containerized open-source tools. In addition to mutations and SCNAs, the pipeline reports common cancer rearrangements, hematologic oncogenic viruses, BCR clonotypes, and global molecular metrics, such as tumor mutational burden (TMB) and microsatellite instability (MSI). Collectively, these results establish EXaCT-2 as a comprehensive platform for integrated cancer genome profiling.

Next-generation sequencing (NGS) provides a comprehensive evaluation of genomic alterations that can inform diagnosis, prognosis, and suggest therapeutic avenues. It is now routinely available at many healthcare facilities and molecular laboratories. However, most oncology tests are based on a limited set of genes and their variants (targeted panels), which leads to the following problems: (1) updates when new cancer genes are discovered: our current knowledge of the disease is biased by using datasets enriched with people of European descent[1–3]; (2) overestimation of important features

[1]Department of System and Computational Biology, Weill Cornell Medicine, New York, NY, USA. [2]Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY, USA. [3]Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA. [4]Field Applications Department, Illumina, Ann Arbor, MI, USA. [5]Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY, USA. [6]Diagnostics and Genomics Group, Agilent Technologies, Santa Clara, CA, USA. [7]WCM Myeloma Center, Weill Cornell Medicine, New York, NY, USA. [8]Department of Cellular, Computational, and Integrative Biology, Interdepartmental Center of Medical Sciences (CISMed), University of Trento, Trento, Italy. [9]Department of Anatomic Pathology, Hematopathology, Molecular Genetic Pathology, Northwell Health, New York, NY, USA. [10]Department of Medicine, Dana-Farber Cancer Institute, Boston, MA, USA. [11]Department of Medicine, Weill Cornell Medicine, New York, NY, USA. [12]Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA. [13]These authors jointly supervised this work: Alicia Alonso, Andrea Sboner.
✉e-mail: ans2077@med.cornell.edu

such as tumor mutational burden (TMB)[4]; (3) limitation in evaluating mutational signatures, a genomic feature increasingly relevant to the understanding cancer development, progression, and therapeutic intervention[5]; (4) impaired evaluation of somatic copy number alterations (SCNAs)[6–9]; (5) limited identification of gene rearrangements. All these issues are due to the limited genomic territory they interrogate. Using broader NGS assays, such as WES and Whole Genome Sequencing (WGS) can overcome some of these limitations[10]. However, these assays are unable to achieve the depth of coverage that targeted panels can provide[10,11]. Lower coverage can impair detection of important events in tumors with high normal cell contamination or sub-clonal events.

Our team previously developed one of the first clinically approved whole-exome sequencing (WES) tests for advanced cancer care based on Agilent HaloPlex target enrichment: EXaCT-1[12,13]. Building on this experience and stemming from our commitment to further improve precision cancer care, we present EXaCT-2, a novel sequencing platform designed for comprehensive genomic profiling in a single workflow that significantly expands and improves the repertoire of clinically relevant somatic alterations. The EXaCT-2 assay, based on Agilent SureSelect hybridization capture, combines the *depth* of targeted panels with the *breadth* of a WES assay, while adding features of both WGS and Array CGH. This is accomplished by incorporating additional capture probes for increased coverage of cancer-relevant genes, for accurate estimation of both tumor purity and the boundaries of SCNAs, for detection of rearrangement breakpoints, assessment of immunoglobulin clonotypes, and for hematologic oncogenic viral genomes characterization. This integrated approach allows concurrent detection of single-nucleotide variants, small insertions and deletions, copy-number alterations, structural variants, mutational signatures, and oncoviruses from the same library. EXaCT-2 provides a unified platform for both precision oncology research and clinical translation.

Results

Key features of EXaCT-2

EXaCT-2, based on the Agilent SureSelect V6 + COSMIC assay (Santa Clara, CA)[14], combines broad exome coverage, features of whole-genome sequencing, and deep targeted coverage of cancer-critical regions, capabilities that we classify as *Breadth of Coverage* and *Depth of Coverage* (see schematic in Fig. 1A). Briefly, breadth of coverage improvements are achieved by incorporating probes for intronic regions (to allow for rearrangement detection), intergenic regions (for improved somatic copy-number alteration detection), BCR regions (for comprehensive rearrangement profiling), and known blood cancer-associated viruses. Depth of coverage improvements include adding probes to boost the coverage of selected clinically relevant cancer genes to reach levels comparable to targeted assays (Table S1).

To highlight the expanded breadth of coverage provided by the EXaCT-2 assay, we compared its increased genomic coverage to standard

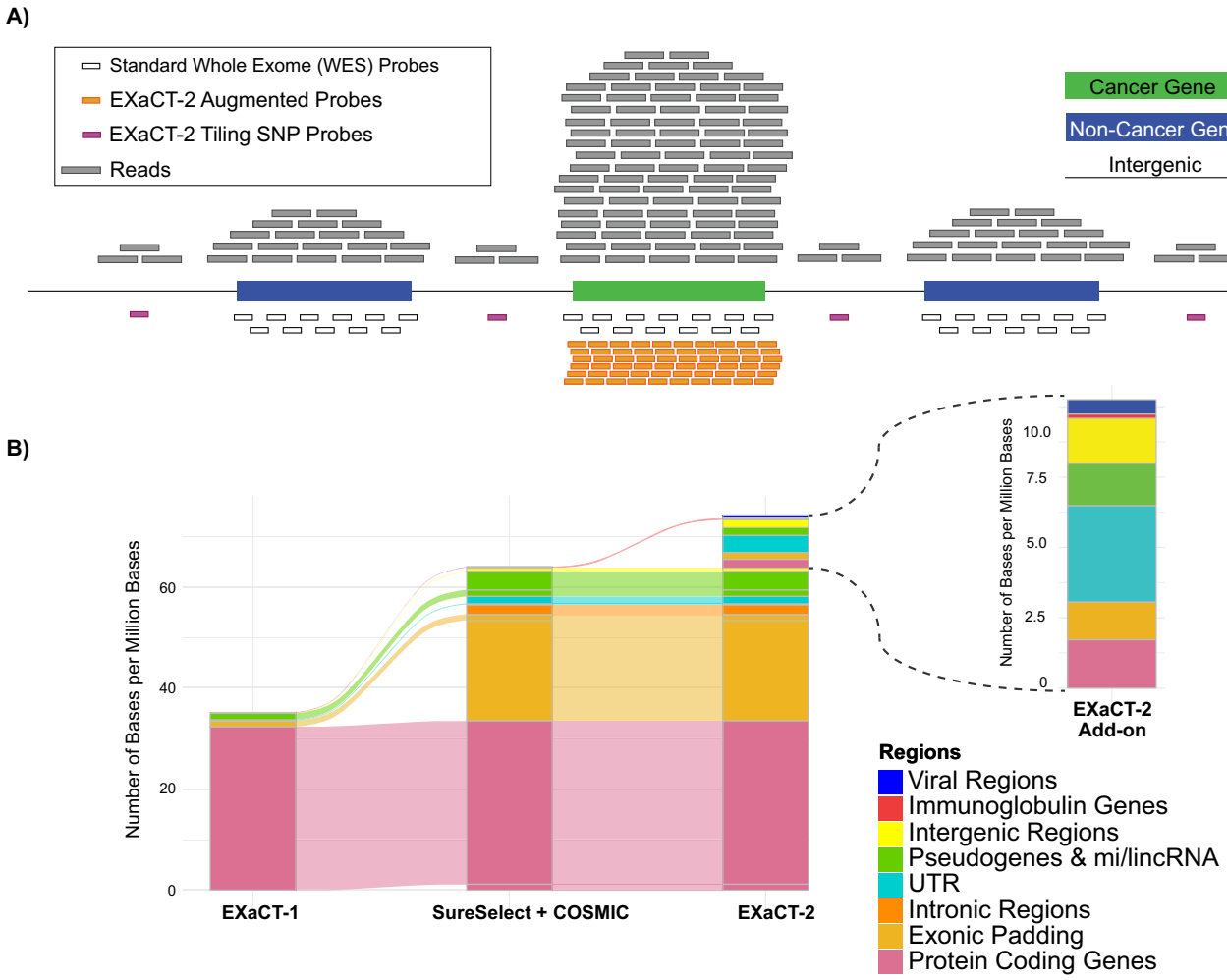

Fig. 1 | EXaCT-2 Assay Design. A Schematic of the probe design strategy. **B** Representation of the bases covered by EXaCT-2. The plot shows comparison with the whole exome assays EXaCT-1 and the standard SureSelect V6 + COSMIC. The inset shows EXaCT-2 additional bases in specific regions such as intergenic and intronic areas, aiming to improve molecular profiling of specimens with enhanced resolution.

whole-exome platforms such as EXaCT-1 and Agilent SureSelect V6 + COSMIC (Fig. 1B; Table S2). EXaCT-2 uniquely targets approximately 10Mbp of additional sequence (Fig. 1B – inset), encompassing regions with established or emerging relevance for cancer diagnosis and management: 1)~3.1Mbp of intergenic targets for 32,700 well-established non-coding SNPs to refine copy number boundaries resulting in an expansion of overlapping non-coding RNAs; 2)~0.62Mbp for 340 intronic hotspots of DNA breakpoints and MSI regions; 3)~0.15Mbp to achieve near-complete coverage of immunoglobulin variable regions; 4)~0.5Mbp covering 4 viral agents, commonly involved in hematological or soft-tissue tumors; and 5)~3.4Mbp for expanded coverage of the non-coding untranslated regions. Across protein-coding regions, the three assays remain relatively unchanged (~21,000 genes – 35 Mbp). EXaCT-2 captures an 8.9% increase in protein-coding bases compared with EXaCT-1, comprising ~3.6% (1.1 Mbp) contributed by SureSelect and ~5.3% (1.7 Mbp) unique to EXaCT-2. Owing to the hybrid-capture design, exonic padding regions, defined as the 50 non-coding bases flanking an exon, account for about 22Mbp of the assay compared to EXaCT-1 (Fig. 1B).

To provide the higher depth of coverage across the genomic regions most relevant to cancer profiling, we augmented the coverage for nearly 1,400 cancer genes identified through expert curation, including regions with high-GC content and repetitive sequences, typically under-represented (Table S1, Fig. 1A). This was achieved while maintaining a coverage for the remaining exome that is typical for other whole-exome sequencing (WES) assays (80-120x). To achieve this augmented coverage, a booster probe set was designed to improve coverage in known cancer regions with <50x coverage, while an augmented probe set was designed to comprehensively cover a subset of these genes (by selectively adding probes in a 1:5 ratio - see Methods). Given that tumor samples often contain varying proportions of normal cells, diminishing the signal of somatic mutations and sub-clonal events, this augmented design maximizes sensitivity for detecting clinically relevant mutations in heterogenous tumor contexts.

Sequencing depth was empirically optimized to achieve a target augmented coverage of about 380–400x for cancer genes, enabling the detection of variants at 5% allele frequency with a probability of 90%[15]. This effective 400x coverage is comparable to those provided by targeted panels such as MSK-IMPACT[11]. Libraries were sequenced to generate between 100 million and 250 million paired-end reads (mean: 210.8 million, SD: 34.2 million - see section "Depth of Coverage in Cancer Relevant Genes"). Under these conditions, we simultaneously achieved ~100x coverage across the non-augmented regions.

In the following sections, we use these two categorizations to organize the discussion of our evaluation of these improvements on a Discovery and a Clinically Validated cohort with 118 and 84 patients, respectively (see Table S14, Fig. S1 and Methods).

Increased breadth of coverage improves resolution of somatic copy number alteration (SCNA) boundaries

We illustrate in Fig. 2 the effect of the EXaCT-2 extended genomic territory on the assessment of the boundaries of somatic copy number alterations (SCNAs). For this analysis, we selected 4 cases from the Discovery Cohort that were also profiled in the Clinical Cohort using EXaCT-1. SCNAs were identified by their respective analytical pipelines (CNVkit[16] and CNVSeeqer[13]) from their matched tumor/normal specimens (see Methods). The systematic comparison of the 4 cases demonstrated that while the segmentations that were generated from the two assays were highly concordant in terms of gene level alterations (99.9% of altered genes detected in both analyses - Chi-square p-value < $2.2e^{-16}$ Table S3), the inclusion of probes covering intergenic SNPs allows the segmentation from the EXaCT-2 assay to be much less under- and over-segmented than with an exome-based assay like EXaCT-1 (Fig. 2A; Figs. S2, S3, Tables S4, S5).

Figure 2B, using Case_01, illustrates how the expanded probe design of EXaCT-2 enhances detection of copy number alterations by refining breakpoints beyond coding regions. Both assays detected a deletion involving the Desmocollin genes (log2 ratio < −0.5: Table S4); but boundaries are different. EXaCT-1 placed the deletion within the second exon of Cadherin 2 (CDH2; 18:25,727,685), whereas EXaCT-2, leveraging intergenic SNPs (Fig. 2C), localized the boundary to the intergenic region (18:27646011).

Detection of rearrangements in common cancer driver genes

The EXaCT-2 assay evaluated 66 genes often involved in clinically-relevant gene fusions (Table S6) and covers a broad spectrum of cancer types (Fig. S4)[17]. As this is a DNA-based assay, two assumptions guided the design: 1. genomic rearrangements are the basis for the generation of gene fusions, and 2. rearrangement breakpoints frequently occur within specific introns. For this reason, we designed the EXaCT-2 assay to include probes that cover the introns that harbor some of the most common clinically relevant breakpoints (Table S7). Remarkably, this focused approach proved to be highly effective in our evaluation with the 12 specimens from our Clinically Validation Cohort which harbor known rearrangements identified using traditional RNA-seq-based methods (Table S8; Fig. 3A for an example: KIAA1549-BRAF, a clinical marker in Pediatric Low-Grade Astrocytoma[18]). EXaCT-2, using Delly[19], identified all the rearrangements targeted by the assay, i.e., those that have tiled breakpoints ($n = 10$). Another fusion with a breakpoint not specifically targeted by the assay was also correctly identified as its breakpoint occurred in the exon padding region (Tables S8 and S9, Methods).

Improved B-Cell Receptor (BCR) rearrangement detection

B cell development involves the coordinated rearrangement of Immunoglobulin (Ig) genes to produce functional antibodies, generating an estimate of 10^{18} unique clonotypes[20]. In Monoclonal Gammopathies[21], clonal expansion of B cells expressing a unique and distinct BCR leads to abnormal monoclonal Ig production. In these cases, identifying the sequence of the BCR is useful in diagnosis, risk stratification, and monitoring disease progression.

To improve the detection of BCR clonotypes, the EXaCT-2 assay expanded the Ig variable gene segments (V) to include about 98% of the known repertoire, incorporating diversity (D) and joining (J) intronic regions (Fig. 1). Utilizing the bioinformatics tool MiXCR[22], EXaCT-2 demonstrated effective clonal BCR detection (Fig. 3B) on a set of known positive multiple myeloma samples from the Clinical Validation Cohort assessed with Lymphotrack (Invivoscribe, Inc., USA). For the lymphoma cell line HBL-1HBL-1[23,24] MiXCR correctly identified the heavy chain rearrangement at a frequency of 17.6% and various light chain rearrangements (top row). Negative controls (DLBCL-1, -2, -3 and Urothelial Carcinoma, obtained from the Discovery Cohort) showed no predominant clonotype and displayed a dispersion of BCR clonotypes at a very low frequency of 2.8% (bottom rows). For the two patients with multiple myeloma, clonal immunoglobulin heavy chain rearrangements were detected at frequencies of 64.4% and 60.1%, respectively. The EXaCT-2 assay demonstrated high analytical sensitivity, detecting these rearrangements with as little as 1 ng of input DNA (Fig. S6). It should be noted, however, that the assay was performed using high quality DNA isolated from bone marrow aspirates and cell lines, and the performance has not been evaluated on decalcified or formalin-fixed, paraffin-embedded (FFPE) bone marrow core biopsy specimens, and BCR detection may differ for degraded or lower-quality DNA.

Identification of viruses common in hematological and soft-tissue malignancies

We assessed the assay's viral-detection probes by quantifying reads aligned to the viral genome. We evaluated Epstein-Barr Virus (types 1 & 2 – EBV1/2), Kaposi's sarcoma-associated herpesvirus (KSHV), Human T-cell lymphotropic virus (HTLV), and Merkel cell polyomavirus (MCPyV), using 14 samples from the Clinical Validation Cohort. This set included 13 virus-positive tumors: 4 KSHV-positive samples, and 3 each positive for EBV, HTLV and MCPyV along with one negative control (Table S10). All samples were confirmed using qualitative PCR, except for the MCPyV specimens, which were clinically diagnosed as Merkel cell carcinomas. In all viral

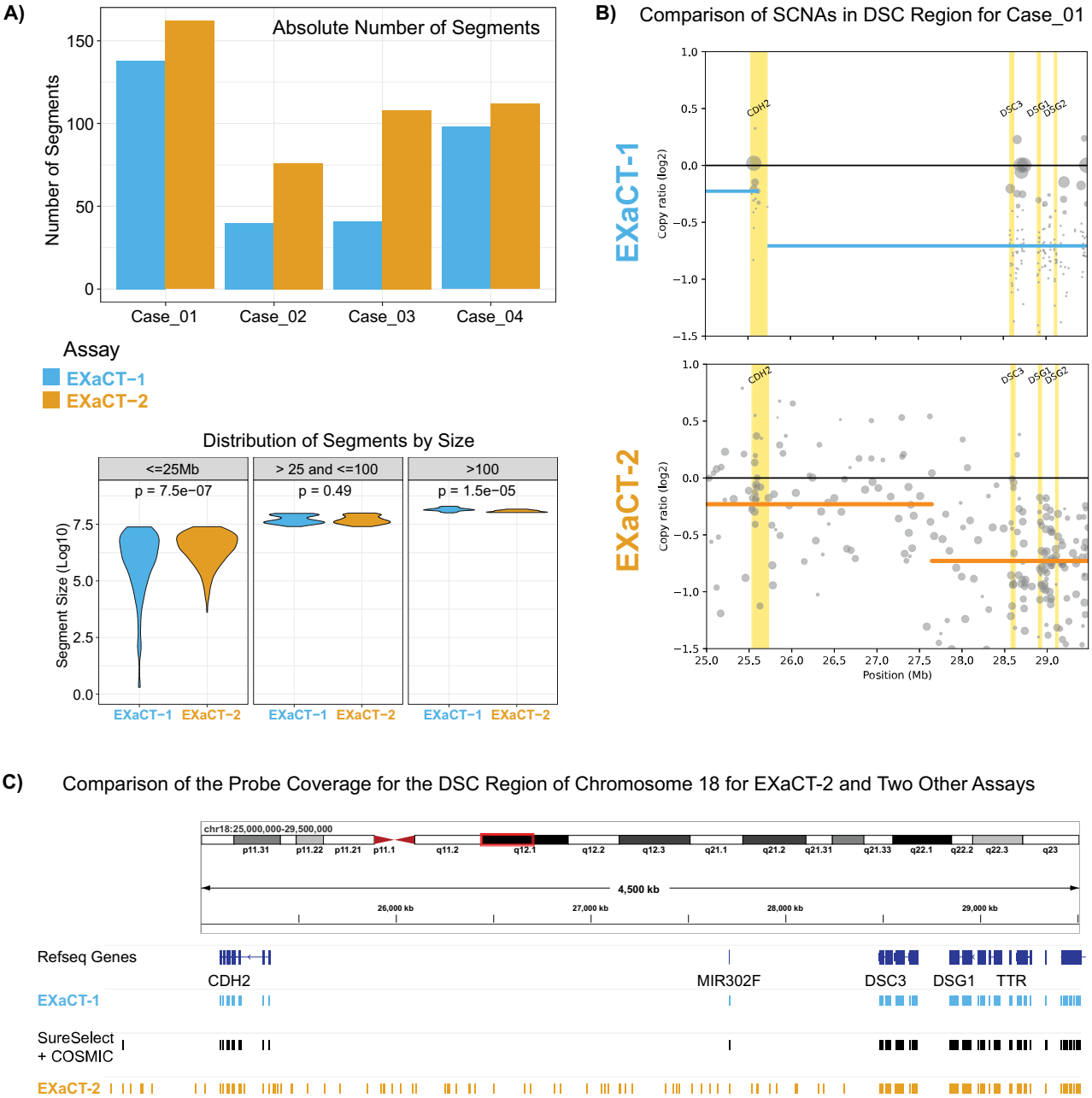

Fig. 2 | Somatic Copy Number Alteration (SCNAs) comparison between 4 cases assayed by both EXaCT-1 and EXaCT-2. A – top) Absolute Number of segments in EXaCT-1 and EXaCT-2 for the 4 cases. **A** – bottom) Comparison by assay of segment sizes within the 3 categories of small (< = 25 Mb), medium (> 25 and < = 100 Mb), and large (> 100 Mb), with statistical difference assessed by *t*-test. **B** The segmented copy-ratio from the analyses of the same sample by EXaCT-1 (top) and EXaCT-2 (bottom), with both indicating a deletion (log2 ratio < −0.5). Note, the size of the data point is proportional to the weight of that point in the segmentation performed by CNVkit (see Methods). **C** Illustration of the probe coverage in the DSC region of chromosome 18 for the SCNAs shown in panel B for three different WES assays: EXaCT-1, SureSelect + COSMIC, and EXaCT-2. Note, all SCNAs are limited to the four cases.

positive cases, the EXaCT-2 assay was able to detect the presence of each virus (Table S10, see Methods for further discussion).

Depth of Coverage in Cancer Relevant Genes

We evaluated the normal samples from the Discovery Cohort (Table S14) to demonstrate that the augmented regions achieve significantly higher coverage than the remainder of the exome. Analyses were limited to normal samples to avoid potential confounding effects from somatic copy number alterations, and coverage values were calculated after removal of PCR duplicates. Across libraries sequenced with 100–250 million paired-end reads, coverage in the augmented regions consistently exceeded the target augmented coverage of ~400× (mean = 409×, SD = 98.6×; Fig. 4A), while the standard regions retain ~100x coverage (mean = 90, SD = 21). Based on these data, we established a minimum sequencing depth of 100 million reads for downstream analyses and adopted a standard sequencing cutoff of 125–150 million reads per sample to balance coverage consistency across experiments and sequencing cost.

Within this cohort, DNA quality had limited impact on average coverage across DIN values of 3–8 (Fig. S7). In contrast, FFPE samples from the Clinical Validation Cohort with DIN < 4 exhibited reduced coverage under standard input conditions. To achieve coverage near the target augmented depth, these samples underwent DNA repair and were prepared using a fourfold increase in DNA input, resulting in coverage sufficient for downstream analyses (Fig. S8; Methods).

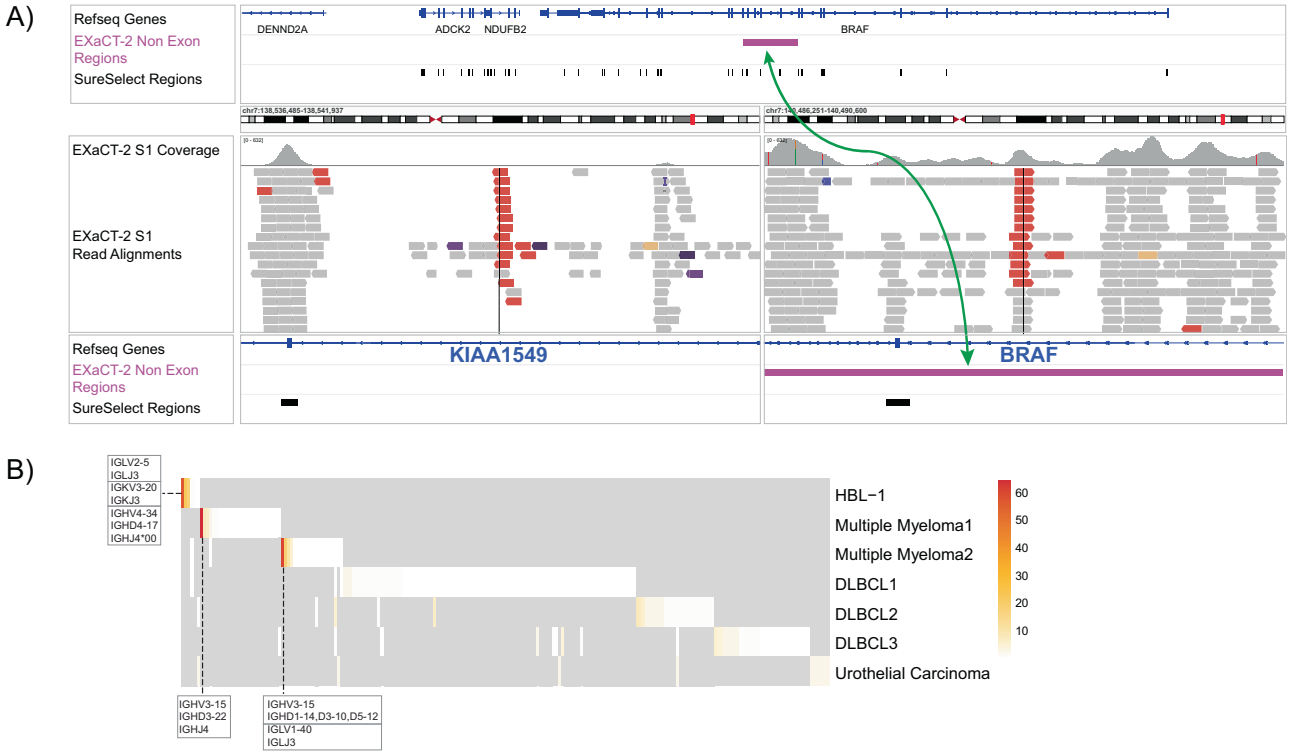

Fig. 3 | Types of rearrangements captured by EXaCT-2. A BRAF rearrangement. In pink, a region known to be a fusion hotspot for BRAF is captured by EXaCT-2. Below that is an IGV representation of BRAF-KIAA1549 rearrangement for one of the positive cases in the Clinical Validation Cohort: Reads involved in the rearrangements are in red. Note, that the reads in the KIAA1549 locus align to an untargeted intronic region and are only recovered due the BRAF intronic region targeted by EXaCT-2. **B** BCR rearrangement. Heatmap indicating the relative read frequency of each BCR clonal rearrangement within the corresponding sample from the Clinical Validation Cohort. Gray areas represent a read frequency of 0, indicating the absence of that clonotype in the sample. Additionally, all the highlighted rearrangements in the box have a frequency of > = 15%. (DLBCLs and the urothelial carcinoma are part of the Discovery Cohort as negative controls).

In Fig. 4B we demonstrate the effectiveness of the depth augmentation on enhancing the coverage of notoriously difficult to assay genes, such as CEBPA, KRAS, or commonly altered genes such as IDH1 and EGFR (Fig. 4B, S9, 10). On average, the EXaCT-2 assay provides 15.8x greater coverage for these genes when compared to EXaCT-1. For this comparison, a representative set of 28 normal samples profiled by the EXaCT-1 platform were used.

To further illustrate this improvement, we evaluated in the Discovery Cohort the exon coverage of the KRAS protein, one of the most common oncogenic drivers[25–27], heavily researched over the past decade[28]. Indeed, we observed significantly higher coverage for exons 2-5 (Fig. S10A). The most significant improvement occurring for exon 2, the location of the well-known G12/G13 mutations that are known oncogenic drivers in various carcinomas such as colorectal and endometrial[28–31] (Fig. 5A, S10A). However, the occurrence of this mutation in a highly GC-rich region has made the detection of it with standard NGS panels to be difficult[11]. As shown in Fig. 5B, the depth augmentation of the EXaCT-2 assay achieves a coverage nearly equivalent to that of a targeted panel.

Moreover, the EXaCT-2 evaluation of 28 validation samples from the Clinical Cohort harboring 9 clinically relevant cancer variants achieved perfect recovery of these mutations with a high degree of variant allele frequencies (VAF) concordance (Fig. S15, Table S16; Pearson correlation: 0.69; p: 5.6e-5).

Novel clinical insights revealed by higher coverage on cancer genes

Given this higher coverage in KRAS, we sought to determine if this translated into the EXaCT-2 assay having a higher sensitivity to detect this clinically important mutation. Using the TCGA Pan-Cancer project (whole exome) as a benchmark, we first compared the number of samples that harbored this mutation, relative to the rest of their respective cohorts, limiting the comparison to cancers with a known association with the KRAS G12* mutation (colon, lung and pancreatic adenocarcinoma). This evaluation determined that the EXaCT-2 assay had a significantly greater proportion of samples harboring the KRAS G12* mutation than the TCGA (Chi Squared test; p-value = 0.0006, Table S11). In a subsequent examination, we compared the number of reads carrying the alternative allele (alt reads, henceforth) for mutations with VAF < = 0.2, reasoning that mutations with more alt reads are more likely to be genuine rather than artifacts. To do this evaluation, we stratified the range of variant allele frequencies (VAFs) from 0-0.15 by increments of 0.05, and determined that in every comparison, the EXaCT-2 assay provides significantly higher alt read coverage than TCGA WES (Fig. 6A).

We next sought to determine whether this increased alt read coverage was generalizable to all the cancer genes whose coverage was augmented in the EXaCT-2 assay. To make the comparison more stringent, we further limited the augmented regions to those that were classified as being problematic by the Genome in a Bottle consortium[32], as well as those that possess a GC-enrichment greater than 80% by the UCSC Genome Browser[33] (see Methods). Limiting our analysis to all mutations that were supported by 15 or more alt reads, the EXaCT-2 assay detects a significantly higher number of mutations than the TCGA Pan-cancer (WES) cohort (7.30% vs 4.49% - Table S12, Chi-square p-value < $2.2e^{-16}$). This difference can be illustrated by comparing the distribution of the mutational VAFs that are provided by the EXaCT-2 assay with those generated from standard sequencing platforms like the ones used by the TCGA Pan-Cancer project[34] (Fig. S11, see Figs. S12–S14 for different thresholds). In this comparison, the critical difference occurs in the lower tails of these distributions which demonstrates that EXaCT-2 assay has the sensitivity to capture more sub-

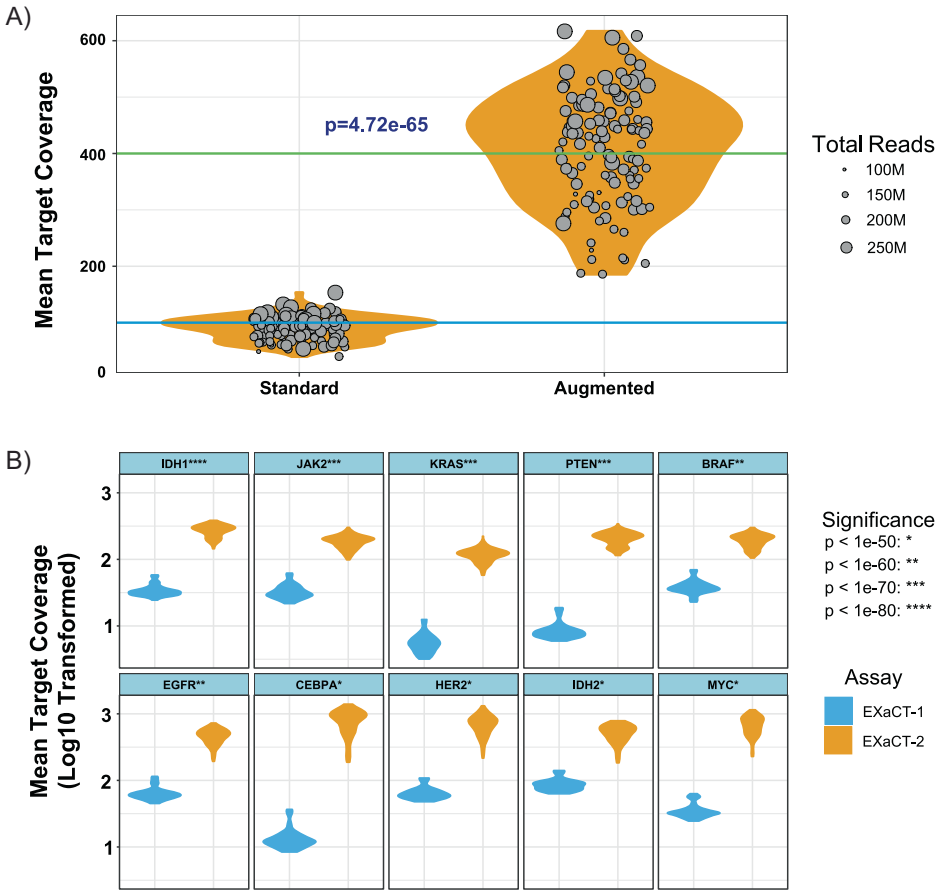

Fig. 4 | Coverage Analyses. A Mean target coverage between standard (non-augmented) and augmented regions. Each dot represents a sample where size is proportional to the Total no. of reads per million. Note, normal samples only from the Discovery Cohort were used for these plots to prevent somatic copy number events in the tumor samples from altering the coverage calculations. **B** Comparison of mean target coverage on selected cancer genes. Each plot reports the distribution of log10-transformed mean target coverage on the indicated genes for EXaCT-2 and 28 EXaCT-1 samples from the Clinical Validation Cohort.

clonal mutations than is possible with standard sequencing assays, and with far greater confidence.

The next two cases illustrate the clinical implications of the high sensitivity of EXaCT-2 in cancer relevant regions. We identified 10 samples in our Discovery cohort that possessed a KRAS G12* mutation with allele frequencies less than or equal to 0.1. In the first case, 5 samples belong to a 69-year-old, Caucasian female patient with pancreatic ductal adenocarcinoma (PDAC) which were collected over the span of more than 6 years[35]. As shown in Fig. 6B, EXaCT-2 detected the G12V mutation in all samples[36–38], in contrast with the previous analysis[35]. In that study, using a pyrosequencing assay specifically designed for KRAS[39], only the two later samples were determined to possess the KRAS G12V mutation. EXaCT-2 identified this mutation with the allele frequency progressively increasing from 0.03 to 0.34 over this time period and treatment protocol, demonstrating higher sensitivity and suggesting and expansion of a sub-clonal population harboring the mutation.

In the second case, we analyzed paired samples from a patient with endometrial cancer, including a tumor specimen and adjacent histologically normal tissue (Fig. 6C). The KRAS G12A mutation was detected in the tumor at an allele frequency of 0.31 and was also present in the adjacent benign tissue at a lower allele frequency (0.017), suggesting the presence of cells harboring this mutation outside the primary tumor. This finding is consistent with prior reports of microscopic tumor cell infiltration into surrounding tissue, which remains a challenge in the field, and may have implications for treatment strategy[40–43].

TMB and MSI

EXaCT-2, which evaluates both coding and non-coding regions, can further explore correlations between TMB and MSI (Figs. S17, S18), confirming that higher TMB does not necessarily correlate with higher MSI[44,45].

Bioinformatics Pipeline

We leveraged the use of cloud-based computing for scalability and the use of modern workflow managers (Nextflow) with containerized solution to run the bioinformatics pipeline and ensure its reproducibility and portability, as suggested by the Association for Molecular Pathology (AMP)[46–48]. Importantly, the pipeline has been designed in a modular fashion so that replacement of specific tools can be easily adopted and validated. Moreover, the use of Nextflow and containers facilitate the deployment of the pipeline in different computational environments with *no changes* to its logic (see Methods)[46,49,50].

Costs

Having established the biological insights yielded by EXaCT-2, we also evaluated the effort and costs involved in running the assay. EXaCT-2 requires 3 days for library preparation and a sequencing depth of 125 million reads per sample, comparable with targeted panels and only ~15% more expensive than TSO500 and ~24% cheaper than EXaCT-1 (see Table S13 for a full comparison). The bioinformatics pipeline for EXaCT-2 is computationally intensive, but with optimized computing and parallel processing, the total turnaround time for a 12-sample run is approximately 10–12 days.

Discussion

In this study, we described EXaCT-2, an enhanced whole-exome sequencing assay whose aim is to achieve a comprehensive cancer profiling that may improve patient management. We employed the well-established SureSelect target-enrichment technology, which consistently deliver uniform coverage and reproducibility[51]. Building on this foundation, EXaCT-2 expands the interrogated genomic territory by an additional 10 Mb compared with current WES or targeted platforms. It also boosts coverage of clinically important cancer genes to levels comparable to targeted panels

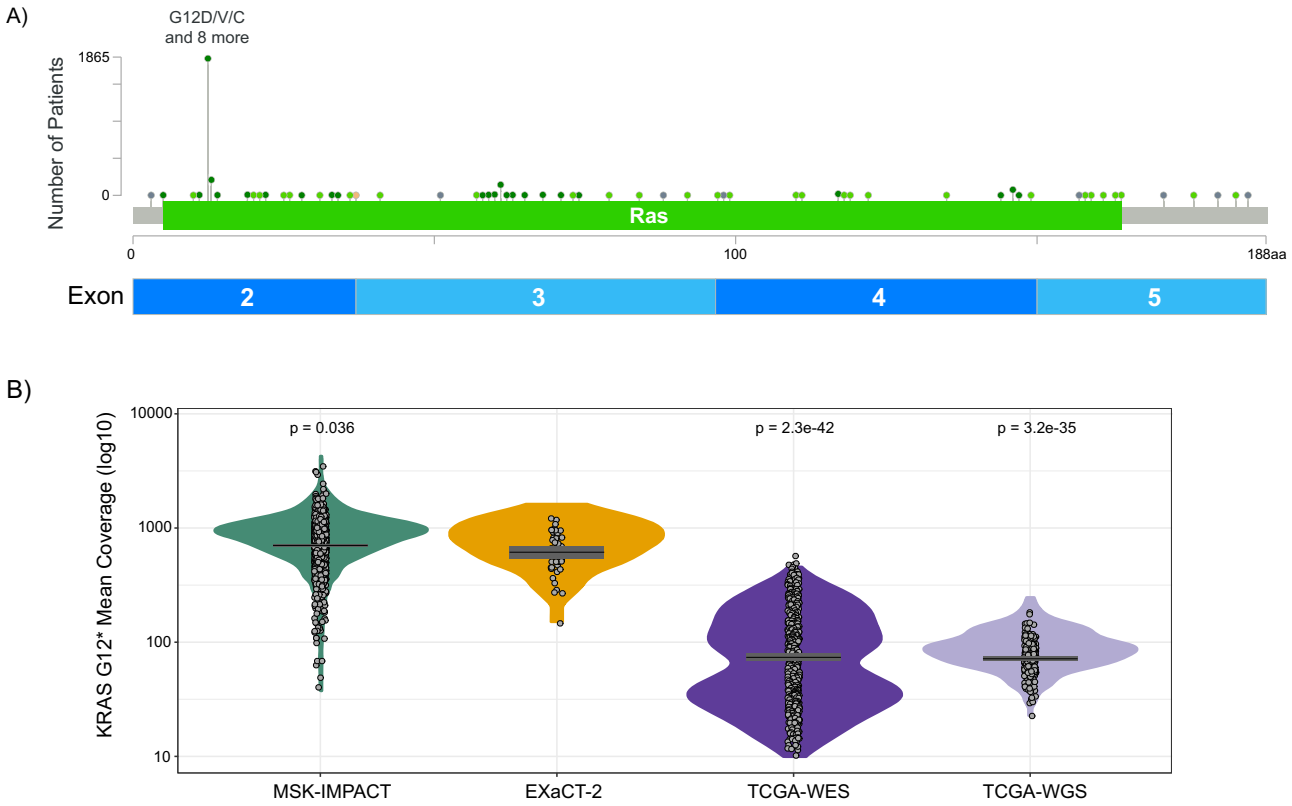

Fig. 5 | KRAS coverage across protein coding exons. A Schematic of the locations of common KRAS mutations from cBioPortal. The plot shows the frequency of the $KRAS^{G12*}$ mutations. Image generated using a combined cohort of patients with colorectal, pancreatic and non-small-cell lung cancers (see Methods). **B** Comparison of the read coverages for $KRAS^{G12D}$ mutations detected by different assays. The plot shows KRAS positive samples assayed by MSK-IMPACT[11] ($n = 1285$), EXaCT-2 ($n = 44$ from the Discovery Cohort), and TCGA Pan-cancer (WES[34] and WGS[105]; $n = 561$ and 227, respectively). Non-EXaCT-2 samples were retrieved from cBioPortal.

(~400x); while maintaining standard coverage (~100x) across the remainder of the exome.

The increased *breadth* enables the detection of common actionable rearrangements (such as BRAF, ALK, NTRK1) and improves copy number boundary determination, a feature that is becoming an important biomarker for certain therapeutic options (Figs. 2, 3)[52,53]. Additionally, it detects oncogenic viruses and expands the evaluation of the BCR repertoire (Fig. 3). The detection on the same assay of rearrangements (and oncogenic viruses) can reduce the need for multiple tests. This assay is also customizable, allowing rapid adaptation to new discoveries on understudied populations and rare tumors. Indeed, large scale efforts to sequence a more diverse population already pointed to new genomic findings[54–56]. We already demonstrated its utility by one recent study on thymoma, a rare cancer of the thymus gland (Fig. S16)[57]. Two other recent publications have used EXaCT-2 to extensively characterize tumor cohorts including organoids and lung cancer samples[58,59].

The increased *depth* of coverage of known cancer genes enables the detection of subclonal mutations at early timepoints. This enhanced coverage also enables reliable genomic profiling in samples with low tumor purity, where normal cell contamination would otherwise obscure critical variants (Figs. 4, 5). As demonstrated in Fig. 6, the greater *depth* of coverage provided by EXaCT-2 enables it with the sensitivity to detect critical mutations at emergent levels, thereby providing clinicians more informed treatment choices at earlier stages of disease onset.

While the higher depth in EXaCT-2 mimics conventional targeted panels by providing high sensitivity for well-characterized driver mutations, its broader interrogated territory captures the extended spectrum of variants, facilitating accurate estimation of global genomic metrics—including tumor mutational burden (TMB) and microsatellite instability (MSI)—and mitigating biases associated with ancestry-specific knowledge (Fig. S17, S18)[4,54–56,60–72].

One caveat is that the assay requires matched tumor/normal specimens, which increase costs and logistical complexity; however, this aspect is common to other assays[11,13]. Additionally, EXaCT-2 is designed to detect a defined set of selected rearrangements ($n = 66$ genes) comparable to the RNA component of the TSO500 assay ($n = 55$ genes)[73]. However, as a DNA-based assay, it cannot determine if rearrangements are expressed, so complementing it with a transcriptome-based assay can provide a more comprehensive molecular picture[74].

In summary, EXaCT-2 is a comprehensive sequencing platform that strengthens precision oncology approaches[54,75–82]. It combines the features of a clinical-grade assay with the capacity for novel discovery, by balancing the strengths of targeted panels and WGS assays and offering a more comprehensive view of tumor alterations—particularly important given the broader genomic diversity observed in non-European populations, which remains underrepresented in current targeted panels[3,83–85].

Methods

Cohorts Used for This Study

We collected 2 cohorts, described below and summarized in Table S14.

For the Discovery Cohort, tumor and matched normal samples from 118 recently diagnosed cancer patients were collected by WCM Englander Institute of Precision Medicine (EIPM) under approved research protocols (IRB1305013903, IRB0107004999, IRB1008011221, and IRB1011011386). In 19 cases, multiple tumor samples were available from the same patient, either from recurrences or metastatic biopsies. In addition, 107 tumor-derived organoids were established from these tumors. Matched normal samples were reused for analysis when multiple tumor or organoids

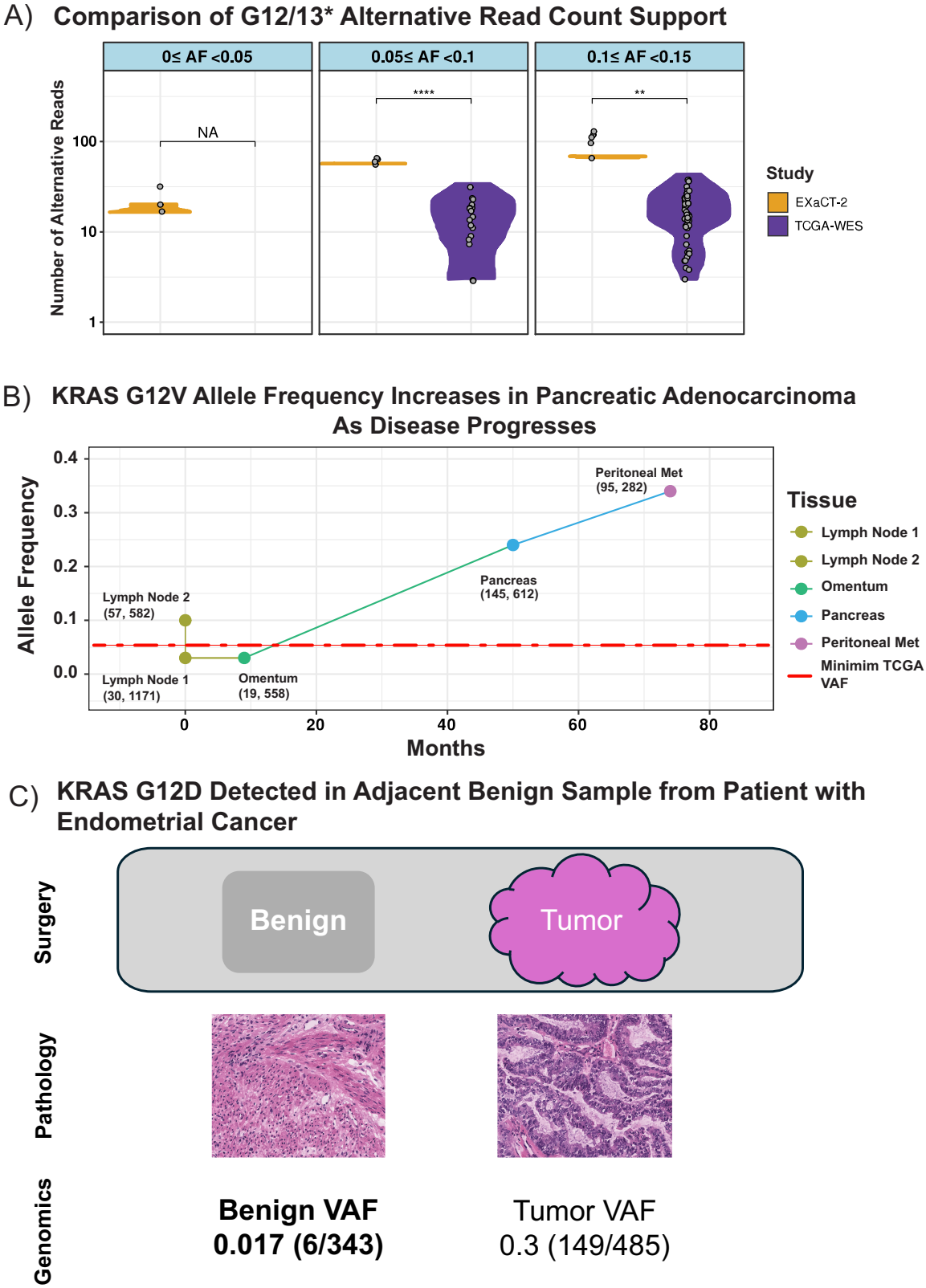

Fig. 6 | Analysis of KRAS mutations in the EIPM Discovery Cohort. A Number of alternative reads per allele frequency. EXaCT-2 shows higher number of alternative reads at the same level of variant allele frequencies compared to TCGA WES data; (**B**) KRAS G12V mutations in a longitudinal series of 5 PDAC samples. The early timepoints showed the presence of a sub-clonal population of cell harboring the mutation. **C** KRAS G12D detected in adjacent benign tissue. EXaCT-2 detects 6 alt reads out of 343x in the benign tissue.

originated from a single patient, resulting in a total of 244 matched tumor-normal pairs available for analysis (Table S14 and Fig. S1).

The Clinical Validation Cohort consisted of samples with previously characterized genomic alterations from 84 patients analyzed by the WCM CLIA-certified clinical laboratory or by our research team using the previous WES assay (EXaCT-1). It included, 32 matched tumor/normal pairs and 28 tumor-only samples encompassing 14 samples used for viral analysis, and 12 samples used for fusion/rearrangement analysis. In addition, two confirmed multiple myeloma samples (IRB10004608) served as positive control for BCR rearrangement analysis. For germline SNV sensitivity assessment, two well-characterized HapMap cell lines, NA12878 and GM24385 were included, and cell line HBL-1 was used as a positive control for BCR rearrangement detection (Table S14).

Sample Collection and DNA Extraction

H&E-stained slides of tumor samples were reviewed and annotated by board-certified pathologists, and tumor DNA was extracted from OCT-embedded fresh-frozen tissues or FFPE. Germline DNA for matched normal was obtained from blood, buccal swabs, or benign tissue. DNA from Multiple Myeloma DNA was derived from CD8+ cells from bone marrow biopsies. Organoids were collected after 5 passages in culture[59]. DNA from the two well-characterized HapMap cell lines, NA12878 and GM24385, was obtained from the Coriell Institute (Camden, NJ) for germline SNV sensitivity analysis. DNA from HBL-1, used as a positive control for BCR rearrangement analysis, was kindly provided by Dr. Wayne Tam.

DNA from OCT-embedded tissue was extracted using the Maxwell 16 Tissue DNA Purification Kit (Promega, Cat# AS1030), and DNA from FFPE tissue was extracted using the Maxwell 16 FFPE Plus DNA Kit (Promega, Cat# AS1135). DNA from buffy coat fractions was isolated with the Maxwell 16 LEV Blood DNA Kit (Promega, Cat# AS1290), while organoid DNA was extracted using the DNeasy Blood & Tissue Kit (Qiagen, Cat# 69506). Saliva samples were processed using the Oragene collection kit (DNA Genotek, Cat# ORG-500).

DNA quantity and quality were assessed using the Qubit dsDNA HS Assay (Thermo Fisher Scientific) and the Agilent TapeStation System, respectively.

EXaCT-2 Assay Formulation and Probe Blending

Probe design, testing, and blending were performed iteratively. A booster probe set was designed to improve coverage in known cancer regions with <50X coverage, while an augmented probe set was designed to comprehensively cover a subset of these genes. Probes for the augmented regions, booster sets, and challenging genomic targets were designed using high-density tiling (5X). Probes are 120 bp in length, roughly corresponding to the median exon size of human genes[86]. The final assay formulation (Agilent, Cat G9496D) consisted of: SureSelect V6 + COSMIC (1X), Non-Coding Regions (1X), Augmented Targets (5X), Booster Probe Set (5X), CEBPA (300X), and Viral Genomes (1/30X).

EXaCT-2 Library Preparation and Sequencing

Libraries were prepared with 50 ng (DIN > −4) or 200 ng (DIN < 4) of DNA using SureSelect Enzymatic Fragmentation and XT Low Input kits following manufacturer instructions (Agilent, Cat 5191 and G9703A). Indexed libraries (~1500 ng) were hybridized with 5 µl of the EXaCT-2 Target Enrichment probes (Agilent, Cat G9496D). Libraries from normal and tumor samples were normalized to 10 nM, pooled 1:1, and sequenced for 150 cycles on a paired-end flow cell (Illumina NovaSeq 6000) to a target an optimal depth of 125–150 million reads per sample. Tumor-only and cell line libraries were prepared and sequenced similarly.

Based on prior experience with degraded FFPE samples prepared elsewhere, the fourteen viral FFPE DNA samples obtained from the Clinical Cohort (DIN < 4) were repaired using NEBNext FFPE DNA Repair Mix (M6630; New England BioLabs, Ipswich, MA) prior to library preparation.

Bioinformatics Processing

Bioinformatics pipeline overview. The pipeline starts with the alignment of the FASTQ files to generate BAM files. Once aligned, the pipeline verifies that the tumor/normal samples originated from the same patient using SPIA (the SNP Panel Identification Assay)[87]. Subsequent analyses are only performed if the samples are verified by SPIA. If they are verified, the pipeline for somatic analysis employs a collection of tools to detect the following: SNV and indels, SCNAs, tumor purity and ploidy, rearrangements, tumor mutational burden, microsatellite instability, and viruses. Briefly, tools used on each matched tumor-normal sample pair are: SPIA to guarantee that the matched tumor/normal controls belong to the same patient[87], an updated version of the MC3 algorithm (SNV & Indel Identification)[88], CNVkit (SCNA Identification)[16], CLONET (tumor purity/ploidy)[89], and MSISensor-Pro (Microsatellite instability)[90,91]. Tools used on the tumor-only samples are: Delly (Rearrangement Detection)[19], an in-house bespoke tool for Viral Detection; MiXCR (Immuno-globulin (Ig) rearrangements)[22]. SNV and indels are annotated by Nirvana[92], which provides clinical grade annotations for these mutations, including impact assessment and population frequency. There is a filtration step that eliminates low quality calls and common artifacts. Finally, in-house calculation for Tumor Mutation Burden (TMB)[93,94] is determined. All versioning information is provided in Table S19. Finally, for the SCNA comparisons, the SCNAs that were identified with the EXaCT-1 assay were generated using a bespoke tool called CNVSeeqer, as described previously[13].

As mentioned previously, we used Nextflow, an open-source bioinformatics workflow manager, to manage all the different components and steps of the analysis. This framework also allows the logic and the flow of the pipeline to be decoupled from the computational environment where the pipeline is executed[47]. For this study, all analyses were performed using the Azure Batch Service on Microsoft Azure Cloud, which was used to run the EXaCT-2 pipeline. This enabled us to use parallelism in performing each step when possible.

In the following sections we provide more details on these steps.

Read alignment and preprocessing. The raw output of the Illumina NovaSeq 6000 was processed using Real-Time Analysis (RTA) v1.18.54 and bcl2fastq2 Conversion Software v2.20, (Illumina Inc., 2019) to generate demultiplexed FASTQ files. Adapter sequences were removed using fastp[95], and the trimmed sequencing reads were aligned to the human reference genome (human_g1k_b37+viral sequences) using the Burrows–Wheeler Alignment tool (bwa v0.7.17, using 'mem' option). After alignment, PCR duplicates were marked using Sambamba (v0.6.7), followed by realignment around the indels, fixing mates and base quality score recalibration using Genome Analysis Toolkit (GATK, v4.2.0.0), and all unmapped reads were removed (all read with MAPQ equal to 0). Additional quality control was performed on the resulting BAM files, and matched tumor-normal samples were analyzed with SPIA[87] to confirm that they originated from the same individual.

Somatic Variant Calling. Somatic variants were identified using a consensus mutation calling approach which integrated seven different mutation callers, this approach is a modified version of the previously published pipeline by the TCGA's Multi-Center Mutation Calling in Multiple Cancers (MC3) working group[88]. Our pipeline is comprised of 6 mutation callers (VarScan2[96], SomaticSniper[97], MuTect2[98], Strelka2[99], RADIA[100], and MuSE[101]) to identify single-nucleotide variants (SNVs) and 4 mutation callers (VarScan2, MuTect2, Strelka2, Pindel[102]) to identify insertion/deletions (indels). Each caller was run individually on paired tumor and matched normal samples, with the caller-specific set of variants filtered using its own caller-specific criteria, using default settings.

These caller-specific sets of variants were merged into a single VCF file, and read depths reported for the same variant called by different tools was averaged. This merged VCF included only those variants that passed the respective caller-specific filtering criteria, and variants were restricted to the

genomic coordinates of the capture kit, plus flanking regions equal to 215 bp. Note, when averaging the read depths of variants that were called by 2 or more callers, if a variant was called by a given method, but didn't pass its internal filtering criteria, the associated read-depth information was excluded from the average.

Finally, this merged set of variants was filtered further based on the following criteria: within-sample allele strand bias, frequency (0.1), total read depth in the tumor (30 total reads, including those with and without the variant), and population frequency (0.01 in either TOPmed (Trans-Omics for Precision Medicine)[103] gnomAD (Genome Aggregation Database)[104]. Frequently mutated, but biologically unrelated genes such as TTN and the MUC genes were also excluded from the final result (Table S18).

Somatic Copy Number Alterations. Somatic Copy Number Alterations (SCNAs) are structural modifications occurring in specific genomic areas, resulting in alterations to the copy number of those regions. These changes are interpreted as either the duplication or deletion of genomic segments. Based on the algorithm from CNVkit[16], each genomic segment is assigned a Log2 Ratio, which is obtained by taking the base-2 logarithm of the ratio between the tumor sample's normalized depth and the matching normal sample's normalized depth. To categorize these genomic segments as Copy number Gain or Loss, and Amplified or Deleted, we set thresholds on the Log2 ratio as:

- Amplification : > = 1
- Gain : [0.5,1)
- Loss : (−1,−0.5]
- Deletion : < = −1.

Segments with Log2 Ratio < 0.5 or > −0.5, are considered Neutral.

B-Cell Receptor Rearrangements. To identify all clonal rearrangements within a sample, the MiXCR tool (version 4.6.0)[22] was employed. The dataset consisted of whole-exome paired-end samples, and the following command line was utilized to extract clonotype information:

mixcr analyze exome-seq --species hsa --no-reports --no-json-reports Fastq1 Fastq2 sampleID.

mixcr exportClones --dont-split-files sampleID.contigs.clns sampleID.tsv. In the above command, 'Fastq1' and 'Fastq2' contain reads from the forward and reverse strands, respectively, of the DNA fragments. 'sampleID' represents the sample name used for the output files. The generated output files include 'sampleID.contigs.clns' and 'sampleID.tsv', with the latter containing read fraction, nucleotide, and amino acid clonotype sequences, as well as Phred quality information for each clonotype present in the sample.

The predominant clonotype in the sample is typically identified based on the highest read fraction, indicating its abundance. Further analysis involves computing the Read frequency, calculated as *read frequency = read fraction * 100*, which represents the percentage of all reads in the sample assigned.

Total Mutation Burden (TMB) estimation. The calculation of TMB involves dividing the total count of mutation events by the number of bases per million in the specified coding regions of EXaCT-2. This process specifically considers certain SNVs and Indels, following predefined filters. Only mutations falling into variant classes, including Frameshift Insertions/Deletions, Inframe Insertions/Deletions, Start and Stop codon Mutations, Splice-Site Mutations, Missense mutations, Splice mutations, Non-Coding, and Synonymous Mutations, are considered.

Moreover, these mutations are subjected to additional filters, requiring Tumor Total Depth > = 30, Control Total Depth > = 30 and Tumor VAF > = 0.1. Furthermore, Genome Aggregation Database (gnomAD) Allele Frequency < = 0.01, and Trans-Omics for Precision Medicine (topmed) Allele Frequency < = 0.01, ensuring the exclusive inclusion of rare variants. Additionally, variants must have proper HGVSP and HGVSC notations. Genes and variants identified as artifacts, based on a predefined list, are also excluded from the analysis.

Microsatellite Instability (MSI) Detection. MSI serves as a predictive and prognostic biomarker indicating DNA mismatch repair deficiency, a condition frequently observed in cancer. Here, MSISensor-Pro[90,91] is used which utilizes the length distributions of microsatellites at specific loci in both the tumor and the matched-normal from the BAM files followed by statistical comparisons on the observed distributions in both the samples.

Using the tool, it identifies the total number of sites with sufficient data (a minimum of 20 reads spanning both tumor and normal samples) and then determines the number of somatic sites. The MSI score is then computed, representing the percentage of somatic sites. According to prior literature, samples with an MSI score of > = 0.4 are classified as MSI-high.

Data availability

Mutational data is or will be made available on a private cBioPortal server. Credentials for access can and will be made available upon request of the editors. Raw, read data is in the process of being deposited to the database of Genotype and Phenotype (dbGaP).

Code availability

The source code for the pipeline used for the analysis will be made available as a private Github repository. Eventually, this will become a publicly available release. Credentials for access can and will be made available on the request of the editors. Tools and Versioning information are provided in Table S19.

Received: 7 May 2025; Accepted: 16 March 2026;
Published online: 21 April 2026

References

1. Hindorff, L. A. et al. Prioritizing diversity in human genomics research. *Nat. Rev. Genet.* **19**, 175–185 (2018).
2. Fatumo, S. et al. A roadmap to increase diversity in genomic studies. *Nat. Med.* **28**, 243–250 (2022).
3. Wang, T. et al. The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437–446 (2022).
4. Fang, H. et al. Tumour mutational burden is overestimated by target cancer gene panels. *J. Natl. Cancer Cent.* **3**, 56–64 (2022).
5. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
6. Harbers, L. et al. Somatic Copy Number Alterations in Human Cancers: An Analysis of Publicly Available Data From The Cancer Genome Atlas. *Front. Oncol.* **11**, 700568 (2021).
7. Buccitelli, C. et al. Pan-cancer analysis distinguishes transcriptional changes of aneuploidy from proliferation. *Genome Res.* **27**, 501–511 (2017).
8. Davoli, T., Uno, H., Wooten, E. C. & Elledge, S. J. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science* **355**, eaaf8399 (2017).
9. Taylor, A. M. et al. Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell* **33**, 676–689.e3 (2018).
10. Kim, R. et al. Clinical application of whole-genome sequencing of solid tumors for precision oncology. *Exp. Mol. Med.* **56**, 1856–68 (2024).
11. Cheng, D. T. et al. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT). *J. Mol. Diagn. JMD* **17**, 251–264 (2015).
12. Beltran, H. et al. Whole-Exome Sequencing of Metastatic Cancer and Biomarkers of Treatment Response. *JAMA Oncol.* **1**, 466–474 (2015).
13. Rennert, H. et al. Development and validation of a whole-exome sequencing test for simultaneous detection of point mutations,

indels and copy-number alterations for precision cancer care. *NPJ Genom. Med*. **1**, 16019 (2016).
14. Agilent Technologies. SureSelect Human All Exon V6. https://www.agilent.com/cs/library/datasheets/public/SureSelect%20V6%20DataSheet%205991-5572EN.pdf (2015).
15. Shin, H.-T. et al. Prevalence and detection of low-allele-fraction variants in clinical cancer samples. *Nat. Commun.* **8**, 1377 (2017).
16. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLOS Comput. Biol.* **12**, e1004873 (2016).
17. FusionGDB 2.0: fusion gene annotation updates aided by deep learning | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/50/D1/D1221/6424754.
18. Hawkins, C. et al. BRAF-KIAA1549 Fusion Predicts Better Clinical Outcome in Pediatric Low-Grade Astrocytoma. *Clin. Cancer Res.* **17**, 4790–4798 (2011).
19. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
20. Briney, B., Inderbitzin, A., Joyce, C. & Burton, D. R. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature* **566**, 393–397 (2019).
21. Dogan, A. et al. B-cell maturation antigen expression across hematologic cancers: a systematic literature review. *Blood Cancer J* **10**, 1–13 (2020).
22. Bolotin, D. A. et al. MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods* **12**, 380–381 (2015).
23. Nozawa, Y. et al. Establishment and characterization of an Epstein-Barr virus negative B-cell lymphoma cell line and successful heterotransplantation. *Tohoku J. Exp. Med.* **156**, 319–330 (1988).
24. Abe, M., Nozawa, Y., Wakasa, H., Ohno, H. & Fukuhara, S. Characterization and comparison of two newly established Epstein-Barr virus-negative lymphoma B-cell lines. Surface markers, growth characteristics, cytogenetics, and transplantability. *Cancer* **61**, 483–490 (1988).
25. Prior, I. A., Hood, F. E. & Hartley, J. L. The Frequency of Ras Mutations in Cancer. *Cancer Res*. **80**, 2969–2974 (2020).
26. Cerami, E. et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discov.* **2**, 401–404 (2012).
27. The, A. A. C. R. et al. AACR Project GENIE: Powering Precision Medicine through an International Consortium. *Cancer Discov.* **7**, 818–831 (2017).
28. Hofmann, M. H., Gerlach, D., Misale, S., Petronczki, M. & Kraut, N. Expanding the Reach of Precision Oncology by Drugging All KRAS Mutants. *Cancer Discov.* **12**, 924–937 (2022).
29. Lu, D. et al. KRAS G12V neoantigen specific T cell receptor for adoptive T cell therapy against tumors. *Nat. Commun.* **14**, 6389 (2023).
30. Zeissig, M. N., Ashwood, L. M., Kondrashova, O. & Sutherland, K. D. Next batter up! Targeting cancers with KRAS-G12D mutations. *Trends Cancer* **9**, 955–967 (2023).
31. Qunaj, L., May, M. S., Neugut, A. I. & Herzberg, B. O. Prognostic and therapeutic impact of the KRAS G12C mutation in colorectal cancer. *Front. Oncol.* **13**, 1252516 (2023).
32. Zook, J. M. et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* **3**, 160025 (2016).
33. Perez, G. et al. The UCSC Genome Browser database: 2025 update. *Nucleic Acids Res*. **53**, D1243–D1249 (2025).
34. Hoadley, K. A. et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929–944 (2014).
35. Baechmann, S. et al. Switch in KRAS mutational status during an unusual course of disease in a patient with advanced pancreatic adenocarcinoma: implications for translational research. *BMC Cancer* **17**, 374 (2017).
36. Luo, J. KRAS mutation in Pancreatic Cancer. *Semin. Oncol.* **48**, 10–18 (2021).
37. Stickler, S., Rath, B. & Hamilton, G. Targeting KRAS in pancreatic cancer. *Oncol. Res.* **32**, 799–805 (2024).
38. Nusrat, F. et al. The Clinical Implications of KRAS Mutations and Variant Allele Frequencies in Pancreatic Ductal Adenocarcinoma. *J. Clin. Med.* **13**, 2103 (2024).
39. Ogino, S. et al. Sensitive Sequencing Method for KRAS Mutation Detection by Pyrosequencing. *J. Mol. Diagn. JMD* **7**, 413–421 (2005).
40. Risques, R. A. & Kennedy, S. R. Aging and the rise of somatic cancer-associated mutations in normal tissues. *PLoS Genet*. **14**, e1007108 (2018).
41. Martincorena, I. Somatic mutation and clonal expansions in human tissues. *Genome Med*. **11**, 35 (2019).
42. De, S. Somatic mosaicism in healthy human tissues. *Trends Genet* **27**, 217–223 (2011).
43. Krimmel, J. D., Salk, J. J. & Risques, R. A. Cancer-like mutations in non-cancer tissue: towards a better understanding of multistep carcinogenesis. *Transl. Cancer Res*. **5**, https://doi.org/10.21037/tcr.2016.11.67 (2016).
44. Salem, M. E. et al. Comparative molecular analysis between microsatellite instability-high (MSI-H) tumors with high tumor mutational burden (TMB-H) versus MSI-H tumors with TMB-intermediate/low. *Ann. Oncol.* **29**, viii650–viii651 (2018).
45. Goodman, A. M., Sokol, E. S., Frampton, G. M., Lippman, S. M. & Kurzrock, R. Microsatellite-Stable Tumors with High Mutational Burden Benefit from Immunotherapy. *Cancer Immunol. Res.* **7**, 1570–1573 (2019).
46. Kadri, S., Sboner, A., Sigaras, A. & Roy, S. Containers in Bioinformatics: Applications, Practical Considerations, and Best Practices in Molecular Pathology. *J. Mol. Diagn.* **24**, 442–454 (2022).
47. Di Tommaso, P. et al. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
48. Roy, S. et al. Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. *J. Mol. Diagn.* **20**, 4–27 (2018).
49. Kadri, S. et al. Clinical Bioinformatician Body of Knowledge—Bioinformatics and Software Core: A Report of the Association for Molecular Pathology. *J. Mol. Diagn.* **27**, 566–582 (2025).
50. Lavrichenko, K. et al. Recommendations for bioinformatics in clinical practice. *Genome Med*. **17**, 124 (2025).
51. Belova, V. et al. Comparative evaluation of four exome enrichment solutions in 2024: Agilent, Roche, Vazyme and Nanodigmbio. *BMC Genomics* **26**, 76 (2025).
52. Hosea, R., Hillary, S., Naqvi, S., Wu, S. & Kasim, V. The two sides of chromosomal instability: drivers and brakes in cancer. *Signal Transduct. Target. Ther.* **9**, 75 (2024).
53. Thompson, J. S. et al. Predicting resistance to chemotherapy using chromosomal instability signatures. *Nat. Genet.* **57**, 1708–1717 (2025).
54. Nocini, R. et al. ICGC ARGO precision medicine: genomic profiling-informed prediction of immunotherapy response in two patients with metastatic head and neck squamous cell carcinoma. *Lancet Oncol.* **24**, 851–853 (2023).
55. The All of Us Research Program Investigators The "All of Us" Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).
56. H3Africa Consortium et al. Research capacity. Enabling the genomic revolution in Africa. *Science* **344**, 1346–1348 (2014).

57. He, Y. et al. Longitudinal genomic analysis of five cases of recurrent thymomas. *J. Thorac. Oncol.* **0**, 103521 (2025).
58. Altorki, N. K. et al. Global evolution of the tumor microenvironment associated with progression from preinvasive invasive to invasive human lung adenocarcinoma. *Cell Rep.* **39**, 110639 (2022).
59. Kuo, H.-H. et al. Pan-Cancer PDOs Preserve Tumor Heterogeneity and Uncover Therapeutic Vulnerabilities. 2025.04.10.647635 Preprint at https://doi.org/10.1101/2025.04.10.647635 (2025).
60. Wu, H.-X. et al. Tumor mutational and indel burden: a systematic pan-cancer evaluation as prognostic biomarkers. *Ann. Transl. Med.* **7**, 640 (2019).
61. Sha, D. et al. Tumor Mutational Burden (TMB) as a Predictive Biomarker in Solid Tumors. *Cancer Discov.* **10**, 1808–1825 (2020).
62. Huang, T. et al. Prognostic Role of Tumor Mutational Burden in Cancer Patients Treated With Immune Checkpoint Inhibitors: A Systematic Review and Meta-Analysis. *Front. Oncol.* **11**, 706652 (2021).
63. Griffin, R. et al. Tumor Mutational Burden As a Prognostic Factor in Diffuse Large B-Cell Lymphoma. *Blood* **142**, 1633 (2023).
64. Li, K., Luo, H., Huang, L., Luo, H. & Zhu, X. Microsatellite instability: a review of what the oncologist should know. *Cancer Cell Int*. **20**, 16 (2020).
65. Chang, L., Chang, M., Chang, H. M. & Chang, F. Microsatellite Instability: A Predictive Biomarker for Cancer Immunotherapy. *Appl. Immunohistochem. Mol. Morphol. AIMM* **26**, e15–e21 (2018).
66. Morihiro, T. et al. PD-L1 expression combined with microsatellite instability/CD8+ tumor infiltrating lymphocytes as a useful prognostic biomarker in gastric cancer. *Sci. Rep.* **9**, 4633 (2019).
67. Rantanen, P. et al. The prognostic significance of microsatellite instability in colorectal cancer: a Swedish multi-center study. *Int. J. Colorectal Dis.* **38**, 197 (2023).
68. Rizvi, N. A. et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348**, 124–128 (2015).
69. Armenia, J. et al. The long tail of oncogenic drivers in prostate cancer. *Nat. Genet.* **50**, 645–651 (2018).
70. Fancello, L., Gandini, S., Pelicci, P. G. & Mazzarella, L. Tumor mutational burden quantification from targeted gene panels: major advancements and challenges. *J. Immunother. Cancer* **7**, 183 (2019).
71. Furtado, L. V. et al. Recommendations for Tumor Mutational Burden Assay Validation and Reporting: A Joint Consensus Recommendation of the Association for Molecular Pathology, College of American Pathologists, and Society for Immunotherapy of Cancer. *J. Mol. Diagn.* **26**, 653–668 (2024).
72. Fernandez, E. M. et al. Cancer-Specific Thresholds Adjust for Whole Exome Sequencing-based Tumor Mutational Burden Distribution. *JCO Precis. Oncol.* **3**, PO.18.00400 (2019).
73. Wei, B. et al. Evaluation of the TruSight Oncology 500 Assay for Routine Clinical Testing of Tumor Mutational Burden and Clinical Utility for Predicting Response to Pembrolizumab. *J. Mol. Diagn. JMD* **24**, 600–608 (2022).
74. Beg, S. et al. Integration of whole-exome and anchored PCR-based next generation sequencing significantly increases detection of actionable alterations in precision oncology. *Transl. Oncol.* **14**, 100944 (2021).
75. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).
76. Sondka, Z. et al. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).
77. Tate, J. G. et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res*. **47**, D941–D947 (2019).
78. Planchard, D. Identification of driver mutations in lung cancer: first step in personalized cancer. *Target. Oncol.* **8**, 3–14 (2013).
79. Aung, K. L. et al. Genomics-Driven Precision Medicine for Advanced Pancreatic Cancer: Early Results from the COMPASS Trial. *Clin. Cancer Res: J. Am. Assoc. Cancer Res.* **24**, 1344–1354 (2018).
80. Ecker, B. L. et al. Genomic Biomarkers Associated with Response to Induction Chemotherapy in Patients with Localized Pancreatic Ductal Adenocarcinoma. *Clin. Cancer Res: J. Am. Assoc. Cancer Res.* **29**, 1368–1374 (2023).
81. Zhang, S. X., Fergusson, D. & Kimmelman, J. Proportion of Patients in Phase I Oncology Trials Receiving Treatments That Are Ultimately Approved. *JNCI J. Natl. Cancer Inst.* **112**, 886–892 (2020).
82. Beg, A., Parveen, R., Fouad, H., Yahia, M. E. & Hassanein, A. S. Unravelling driver genes as potential therapeutic targets in ovarian cancer via integrated bioinformatics approach. *J. Ovarian Res.* **17**, 86 (2024).
83. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).
84. Venner, E. et al. The frequency of pathogenic variation in the All of Us cohort reveals ancestry-driven disparities. *Commun. Biol.* **7**, 174 (2024).
85. Valle, L. F. et al. Oncogenic alteration rates, race, and prostate cancer specific mortality in Veterans with metastatic prostate cancer undergoing somatic tumor next generation sequencing. *J. Clin. Oncol.* **42**, 5017–5017 (2024).
86. Mokry, M. et al. Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries. *Nucleic Acids Res*. **38**, e116 (2010).
87. Demichelis, F. et al. SNP panel identification assay (SPIA): a genetic-based assay for the identification of cell lines. *Nucleic Acids Res* **36**, 2446–2456 (2008).
88. Ellrott, K. et al. Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst.* **6**, 271–281.e7 (2018).
89. Prandi, D. et al. Unraveling the clonal hierarchy of somatic genomic aberrations. *Genome Biol.* **15**, 439 (2014).
90. Niu, B. et al. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics* **30**, 1015–1016 (2014).
91. Jia, P. et al. MSIsensor-Pro: Fast, Accurate, and Matched-Normal-Sample-Free Detection of Microsatellite Instability. *Genomics Proteom. Bioinforma.* **18**, 65–71 (2020).
92. Stromberg, M. et al. Nirvana: Clinical Grade Variant Annotator. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* 596 (Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3107411.3108204 (2017).
93. Chalmers, Z. R. et al. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med*. **9**, 34 (2017).
94. Fernandez, E. M. et al. Cancer-Specific Thresholds Adjust for Whole Exome Sequencing–Based Tumor Mutational Burden Distribution. *JCO Precis. Oncol*. 1–12 https://doi.org/10.1200/PO.18.00400 (2019).
95. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
96. Koboldt, D. C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. **22**, 568–576 (2012).
97. Larson, D. E. et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**, 311–317 (2012).
98. Benjamin, D. et al. Calling Somatic SNVs and Indels with Mutect2. 861054 Preprint at https://doi.org/10.1101/861054 (2019).
99. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
100. Radenbaugh, A. J. et al. RADIA: RNA and DNA integrated analysis for somatic mutation detection. *PloS One* **9**, e111516 (2014).

101. Ji, S., Montierth, M. D. & Wang, W. MuSE: A Novel Approach to Mutation Calling with Sample-Specific Error Modeling. *Methods Mol. Biol. Clifton NJ* **2493**, 21–27 (2022).
102. 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).
103. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
104. Chen, S. et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).
105. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).

Acknowledgements

We acknowledge help from the following institutions and individuals. Giuseppe Giaccone, MD/PhD, for the GTF2I and MTAP customization discussion. Alessandro Romanel, PhD, University of Trento - Italy, for technical input and discussions. The Genomics and Epigenomics Core Research Facility at Weill Cornell Medicine for the library preparation and sequencing. The Englander Institute of Precision Medicine (EIPM) and the Joint Clinical Genomics Initiative (JCGI) of Weill Cornell Medicine and New York Presbyterian Hospital for funding this project.

Author contributions

Analysis of the data, A.A.; A.S.; B.B.; J.M.M.; J.P.S.; K.E.; P.C.; P.W.; W.S.; Code review, K.G.; Data/Sample collection/provision, A.A.; H.B.; H.R.; J.M.; J.M.; J.M.M.; J.P.S.; K.E.; M.A.A.; M.M.O.; M.S.; N.L.; O.E.; W.S.; W.T.; Data/Sample processing, A.K.; D.C.W.; E.M.F.; H.J.P.; J.M.M.; K.E.; M.S.; P.C.; P.D.; P.W.; T.K.; Design of the study, A.S.; D.C.H.; D.C.W.; HR.;. M.S.; O.E.; R.K.; W.S.; Editing/reviewing the manuscript, A.A.; A.S.; C.P.; D.C.H.; D.C.W.; D.R.; F.D.; H.B.; H.J.P.; J.C.; J.M.; J.M.M.; J.P.S.; K.G.; M.A.A.; M.M.O.; M.S.; N.L.; P.W.; T.T.E.; Funding, J.C.; N.L.; O.E.; R.K.; Infrastructure, A.S.; J.T.; P.Z.; Probe and Assay Design, C.P.; D.C.H.; D.R.; F.D.; T.TE.; Testing and Verification, C.P.; J.P.S.; W.S.; Writing of the manuscript, A.A.; A.S.; D.C.W.; P.C.; P.W.

Competing interests

Theresa Ten Eyck, Douglas Roberts, and Carlos Pabon are or were employees of Agilent Technologies, the maker/producer of the EXaCT-2 assay. The other authors do not have a competing interest.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41698-026-01390-5.

Correspondence and requests for materials should be addressed to Andrea Sboner.