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[2,4,5], Hyeon Park[5], Carlos Pabon[7], 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 Teneyck[7], Douglas Roberts[7], Jorge Monge[8], Francesca Demichelis[9], Wayne Tam[10], Madhu M Ouseph[5], Alexandros Sigaras[1,2,3], Himisha Beltran[11], Hannah Rennert[5], Neal Lindeman[5], Wei Song5, James Solomon[5], Juan Miguel Mosquera[2,5], Rob Kim[2], Jeffrey Catalano[2,5], Duane C Hassane[2,3,6], Michael Sigouros[1,2], Olivier Elemento[1,2,3,12], Alicia Alonso[2,6,12,*], Andrea Sboner[2,3,4,5,12,*]

Affiliations

1. Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY
2. Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY
3. Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY
4. Field Applications Department, Illumina, Ann Arbor, MI
5. Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY
6. Department of Medicine, Weill Cornell Medicine, New York, NY
7. Diagnostics and Genomics Group, Agilent Technologies, Santa Clara, CA
8. WCM Myeloma Center, Weill Cornell Medicine, New York, NY
9. Department of Cellular, Computational, and Integrative Biology, Interdepartmental Center of Medical Sciences (CISMed), University of Trento, Italy
10. Department of Anatomic Pathology, Hematopathology, Molecular Genetic Pathology, Northwell Health, New York, NY
11. Department of Medicine, Dana-Farber Cancer Institute, Boston, MA,
12. Meyer Cancer Center, Weill Cornell Medicine, New York, NY

* Alicia Alonso and Andrea Sboner jointly supervised this work

Acknowledgements

We acknowledge help from the following institutions and individuals. Giuseppe Giaccone, MD/PhD, for the GTF2I and MTAP customization discussion. The Genomics and Epigenomics Core Research Facility at Weill Cornell Medicine for the library preparation and sequencing. The Englander Institute of Precision Medicine for funding this project.

EXaCT-2: Combining depth and breadth in a single whole-exome sequencing test

Corresponding Author

Andrea Sboner, PhD
Department of Pathology and Laboratory Medicine,
Englander Institute of Precision Medicine,
Institute for Computational Biomedicine,
Meyer Cancer Center,
Weill Cornell Medicine, New York, NY
email: ans2077@med.cornell.edu
Mastodon handle: @asboner@genomic.social

Abstract

With the rapid advances in cancer research, the list of variants and genes that drive human diseases is constantly expanding. Moreover, the FDA has approved more cancer therapies that incorporate a broader set of genomic features than simple gene variants such as Tumor Mutation Burden (TMB), microsatellite instability status (MSI), and fusion events in gene families such as the NTRK receptors. These features currently require multiple testing methods (IHC/FISH/etc.). With the cost of NGS testing dropping, it is now possible to envision an NGS assay capable of reliably detecting these features without the need for additional testing. The EIPM multidisciplinary team has developed EXaCT-2: a whole exome sequencing (WES) assay that gives the coverage of a targeted assay on cancer genes and the breadth to detect copy number events, cancer-related fusions, and viruses, which can facilitate diagnostic and therapeutic decisions for cancer patients.

We evaluated EXaCT-2 on 250 matched tumor/normal pairs and compared its performance with orthogonally validated results. We show the assay achieves the expected coverage of critical cancer genes, provides a better characterization of somatic copy number alterations, detects common cancer rearrangements and viruses, and enables the accurate estimation of global molecular metrics, such as tumor mutational burden and microsatellite instability.

Introduction

Advances in our understanding of the molecular basis of tumor development and progression have led to the identification of many alterations in the cancer genome, resulting in the development of targeted therapeutics[1–4]. The list of genes that drive the disease continues to expand, as does the number of potential treatment options available to patients, bringing us closer to fulfilling the promise of precision medicine: "administering the right drug to the right patient at the right time". Indeed, response in phase I clinical trials is higher when treatment is based on biomarker eligibility[5–10]. Next-generation sequencing (NGS) provides a comprehensive evaluation of genomic alterations and inform diagnosis, prognosis, and therapeutic avenues and it is routinely available at many healthcare facilities and molecular laboratories. However, most oncology tests are based on a limited set of genes and variants (targeted panels), which leads to the following problems: 1) they need to be updated when new cancer variants are discovered. This a non-trivial issue as much of our knowledge of the disease is biased by using datasets enriched with people of European descent [11–13]. As large-scale projects such as ICGC-ARGO, H3Africa, AllofUs[14–17] work to address this bias, the field will need to quickly consider new additional diagnostically or therapeutically relevant cancer genes; 2) they tend to overestimate important features such as tumor mutational burden (TMB) due to the limited genomic territory they interrogate[18]. As new immunotherapy treatments become more available, accurate TMB estimation is important as an indicator of responsiveness to immunotherapy[19–22]; 3) they are limited in evaluating mutational signatures, a genomic feature increasingly relevant to the understanding of the underlying mechanisms leading to cancer development as well as the results of therapeutic intervention[23]; 4) impaired evaluation of somatic copy number alterations (SCNAs) due to limited information from

broader genomic regions. Determining a single gene copy number status is possible, but characterizing it as 'focal' (localized) or 'broad' (large areas or entire chromosomal arm) is difficult, with different disease implications[24–27]; 5) identification of gene rearrangements is limited, unless the associated breakpoints is targeted by the assay. As gene rearrangements are increasingly important biomarkers[28–30], their omission may prove deleterious for many patients. Currently, they are identified as gene fusions via RNA profiling, hence requiring a different workflow in the lab, with potential delays in returning results[31, 32].

Using broader NGS assays, such as WES and Whole Genome Sequencing (WGS) can overcome some of these limitations[33]. However, these assays are unable of achieving the depth of coverage that targeted panels can provide[33,34]. Lower coverage can impair detection of important events in tumors with high normal cell contamination or subclonal events.

We here describe EXaCT-2, a new sequencing assay that combines the breadth of a WES assay with features of WGS and the depth of targeted panels (see Figure 1) that has already demonstrated its utility by one recent study[36]. Designed with cancer genomics in mind, EXaCT-2 aims for a coverage of ~400x for nearly 1,400 cancer genes, similar to that of targeted panels, while still providing an average 100x coverage for the remainder of the exome (Table S1). Additionally, 32,700 probes on common SNPs, spread evenly across the genome, enable accurate estimation of both tumor purity and the boundaries of SCNAs. Furthermore, selected intronic areas are included to capture common rearrangement breakpoints, such as BRAF, and ALK. Similarly, immunoglobulin regions are almost entirely covered. Finally, EXaCT-2 can interrogate the presence of four common viruses involved in hematological cancers. In summary, we here present an assay that can interrogate multiple oncogenic events in one single assay, thus reducing the need for serial testing.

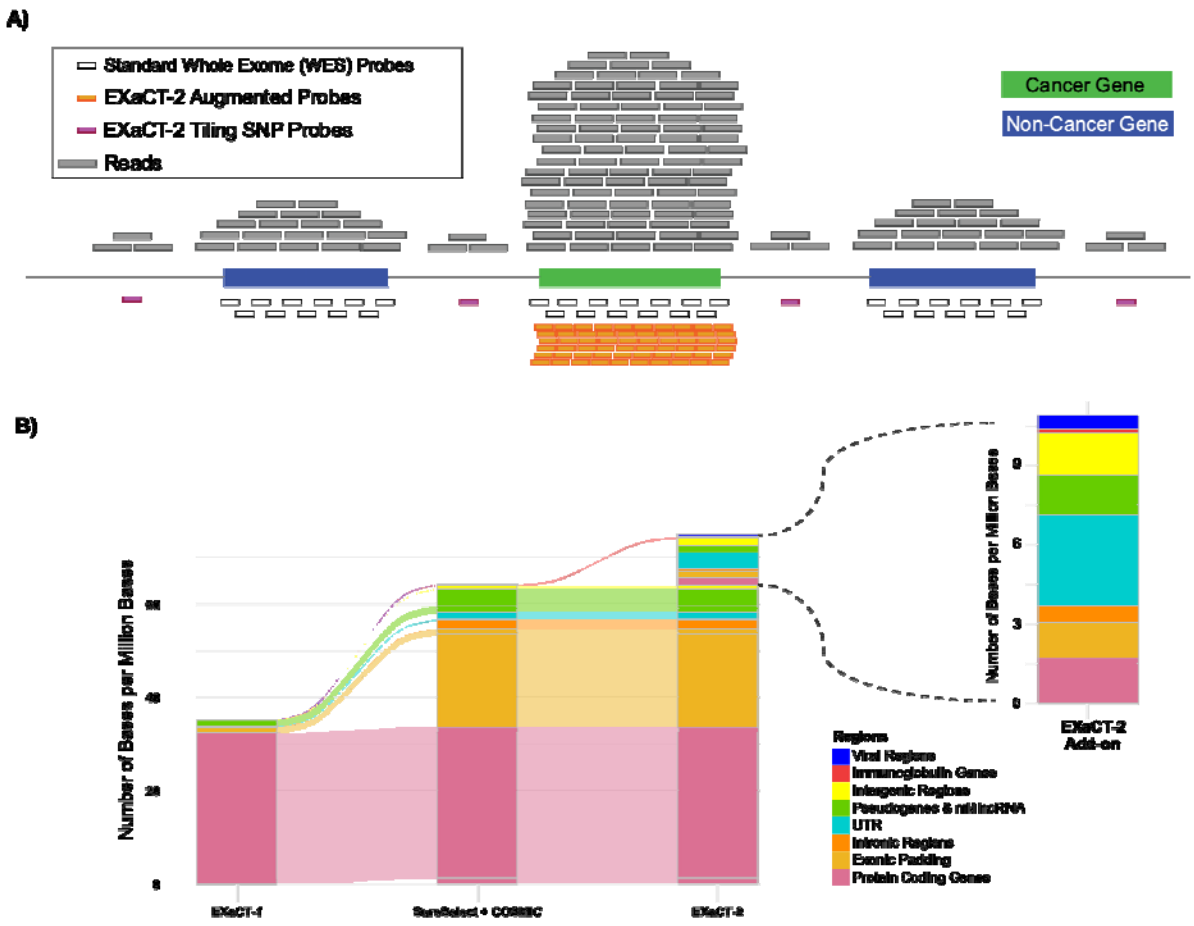


Figure 1: EXaCT-2 Assay Design. A) *Schematic of the probe design strategy.* ***B)*** *Representation of the bases covered by EXaCT-2, compared with the standard whole exome assays EXaCT-1 and SureSelect + Cosmic. EXaCT-2 includes additional bases in specific regions such as intergenic and intronic areas, aiming to improve molecular profiling of specimens with enhanced resolution.*

Results

Evaluation of novel features incorporated into EXaCT-2

The features relevant to cancer characterization as per FDA-approved drugs reflect a wide range of event types, from single nucleotide mutations to microsatellite instability, to complex rearrangements. They can be broadly categorized into the enhancement of both the *Breadth of Coverage* as well as the *Depth of Coverage* of the EXaCT-2 assay. Breadth of coverage improvements are achieved by the inclusion of probes for intronic regions (to allow for rearrangement detection), for intergenic regions (to improve on somatic copy number alteration detection), for comprehensive characterization of BCR rearrangements, and the inclusion of probes for known blood cancer-associated viruses. Depth of coverage improvements include modifications to boost the coverage of clinically relevant genes, such as EGFR and IDH2, as well as cancer genes difficult to assay due to GC-rich regions, such as

CEBPA, KRAS, and PTEN. In the following sections, we use these categorizations to organize the discussion of our evaluation of these improvements.

Breadth of Coverage

Improved genomic territory assayed by EXaCT-2

The EXaCT-2 assay we introduce here is an hybridization assay that builds upon the Agilent SureSelect + Cosmic assay (Santa Clara, CA) [37]. To illustrate the novelty of EXaCT-2 we also compared it with EXaCT-1, a WES amplicon-based assay. Figure 1 illustrates the difference of the assays with respect to the coverage of the human genome (Table S2). The three assays remain relatively unchanged in terms of protein-coding regions (~21K genes – 35MB). Only a 3.6% increase in protein coding bases is targeted by EXaCT-2 and the SureSelect assays, compared to EXaCT-1. Due to the hybridization capture method, exonic padding regions, defined as the 50 non-coding bases flanking an exon, account for about 22MB of the assay compared to EXaCT-1.

As shown in Figure 1 - inset, the EXaCT-2 specific 10MB enhanced design includes: several intergenic targets for well-established non-coding SNPs to allow for accurate estimation of copy number boundaries resulting in an expansion of overlapping non-coding RNAs; probes for immunoglobulin regions that were previously missing – the inclusion of which allows for accurate estimation of clonal BCR rearrangements; probes for intronic hotspots of DNA breakpoints, such as the KIAA1549-BRAF; probes for 4 viral agents commonly involved in hematological or soft-tissue tumors, and probes for expanded coverage of the non-coding UTR regions, to improve the detection of mutations that impact mechanisms of post-transcriptional control.

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

As described above, we illustrate in Figure 2 the effect of the EXaCT-2 extented genomic territory on the assessment of the boundaries of somatic copy number alterations (SCNAs) by CNVKit (see Methods). In this example, EXaCT-1 and EXaCT-2 were applied to the same sample, with their respective analytical pipelines, identifying a somatic deletion involving the Desmocollin genes (DSC1, DSC2, DSC3). However, the deletion boundaries identified by the two pipelines were markedly different. The EXaCT-1 pipeline inferred a deletion boundary within the second exon of Cadherin 2 (CDH2; 18:25727685), while the EXaCT-2's inclusion of intergenic SNPs allowed its pipeline to locate the boundary within the intergenic region (18:27646011) (Figure 2B).

EXaCT-2: Combining depth and breadth in a single whole-exome sequencing test

Beyond this one example, a systematic comparison of 4 samples that were analyzed with both EXaCT-2 and EXaCT-1 was performed. This comparison demonstrated that while the segmentations that were generated from the two assays were highly concordant in terms of gene level alterations (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 is possible with an exome-based assay like EXaCT-1 (Figures 2C, Figure S1-S2, Tables S4-S5).

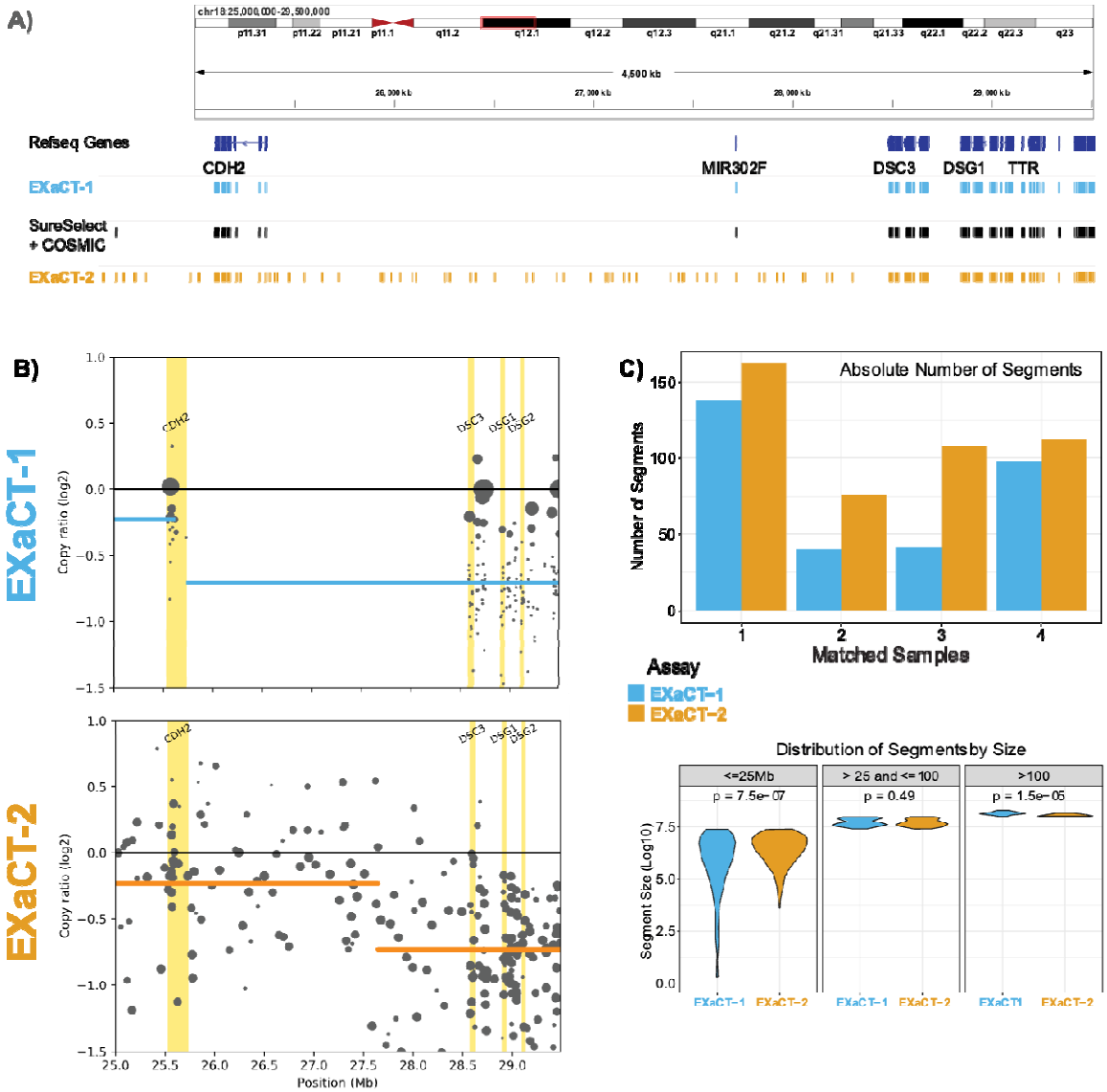


Figure 2: Somatic Copy Number Alteration (SCNAs) comparison between matching samples. A) *Illustration of an example of the same region analyzed with EXaCT-1 and EXaCT-2 on the same sample: the region on chromosome 18 as well as the genes and the probes of 3 different WES assays, EXaCT-1, SureSelect + COSMIC, and EXaCT-2.* ***B)*** *The plots show the segmented, bin-level coverage profiles for EXaCT-1 (top) and EXaCT-2 (bottom). Note, the size of the data point is proportional to the weight of that point in the segmentation performed by CNVkit (see methods).* ***C – top)*** *Absolute Number*

of segments in EXaCT-1 and EXaCT-2 for the 4 matching samples. ***C – bottom)*** *Comparison of segment sizes within the 3 categories of small, medium, and large, with statistical difference assessed by t-test.*

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 (Figure S3).[38] As this is a DNA-based assay, the assumptions are that: 1. genomic rearrangements are the basis for the generation of gene fusions, and 2. rearrangement breakpoints frequently occur within specific introns. With these two assumptions, we designed probes to cover introns that harbor some of the most common breakpoints (Table S7). Despite the limited scope of these non-coding regions, our evaluation indicated that their inclusion was remarkably effective (see Figure 3A for an example: KIAA1549-BRAF, a clinical marker in Pediatric Low-Grade Astrocytoma[28]). Using a set of ten positive control samples identified using traditional RNA-seq-based methods, the EXaCT-2 assay identified all 10 fusions (Tables S8 and S9).

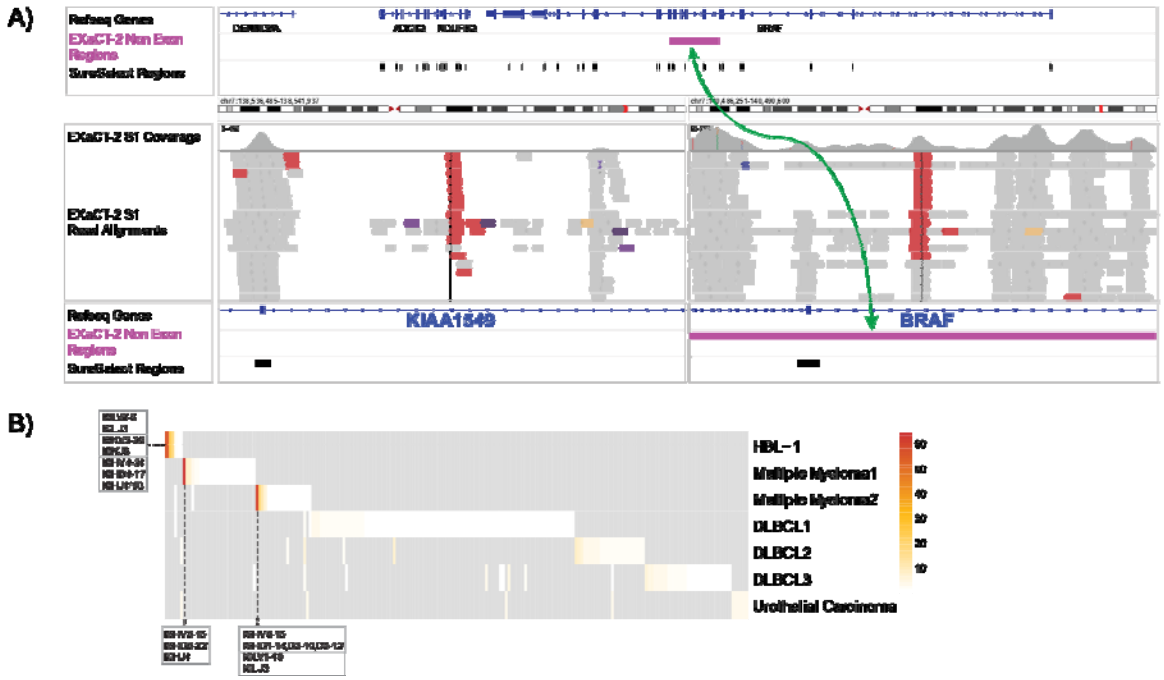


Figure 3: Types of rearrangements captured by EXaCT-2. **A)** *In pink, a region known to be fusion hotspot for BRAF is captured by EXaCT-2. Below that is an IGV representation of BRAF-KIAA1549 rearrangement: Reads involved in the rearrangements are in red. Note, that the reads in the KIAA1549 locus align to an untargeted intronic region and are only recovered due the BRAF intronic region targeted by EXaCT-2.* **B)** *The colors red, orange, yellow, and white indicate the relative read frequency of each BCR clonal rearrangement within the corresponding sample. Red signifies the highest frequency, followed by orange and then yellow, while white denotes the lowest frequency. 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%.*

EXaCT-2: Combining depth and breadth in a single whole-exome sequencing test

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[37]. In Monoclonal Gammopathies[38], 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 (Figure 1). Utilizing the bioinformatics tool MiXCR[39], EXaCT-2 demonstrated effective clonal BCR detection (Figure 3B) on a set of positive control samples assessed with Lymphotrack (Invivoscribe, Inc., USA). For the lymphoma cell line HBL-1[40,41] MiXCR correctly identified the heavy chain rearrangement at a frequency of 17.6% and various light chain rearrangements (top row). Healthy controls (DLBCL-1, -2, -3 and Urothelial Carcinoma) showed no predominant clonotype and displayed a dispersion of BCR clonotypes at a very low frequency of 2.8% (bottom rows). In two Multiple Myeloma patients, two clonal heavy chain rearrangements at 64.4% (second row), and 60.1% (third row), respectively were detected. The EXaCT-2 assay demonstrated sensitivity with just 1ng of input (Figure S5).

Identification of viruses common in hematological and soft-tissue malignancies

Probes for several viruses that are involved in liquid tumors were also included in the assay design, including Epstein-Barr Virus (types 1 & 2 – EBV1/2), Kaposi's sarcoma-associated herpesvirus (KHSV), Human T-cell lymphotropic virus (HTLV), and Merkel cell polyomavirus (MCPyV). For validation, we collected 14 clinical samples, including one negative control sample and 13 viral positive that included 4 samples with KSHV, and 3 each for EBV, HTLV and MCPyV (Table S10). Apart from the MCPyV specimens, all other samples were initially identified using qualitative PCR. In all viral positive cases, the EXaCT-2 assay was able to detect the presence of each virus (Table S10, see Supplementary Materials for further discussion).

Depth of Coverage in Cancer Relevant Genes

The EXaCT-2 assay is designed to achieve higher coverage in specific genomic regions highly relevant to cancer profiling. A typical somatic WES assay has an average coverage ranging from 80-120x. However, regions with high-GC content, repetitive sequences, typically may be overly under-represented. Similarly, contamination of normal cells in a tumor specimen reduces the evidence of somatic mutations in the resulting sequencing reads; analogous to sub-clonal events. Thus, we augmented the depth of coverage by selectively adding probes to nearly 1400 cancer genes identified via expert curation (Table S1).

We target to achieve a coverage of about 380-400x for those genes to allow EXaCT-2 to achieve a coverage depth that enables the detection of variants at 5% allele frequency with

a probability of 90%[39]. Moreover, an effective 400x coverage is comparable to those provided by targeted panels such as MSK-IMPACT[34].

Figure 4A illustrates the average coverage in the augmented regions compared to the rest of the assays for 484 samples, including cases and controls. We observed that the expected 400x average coverage in the augmented regions was exceeded (mean = 420, SD=100) and determined that DNA quality has no significant impact on the average coverage (Figure S6). In addition to the augmented regions, we demonstrate the customizability of the assay to allow users to enhance the coverage of notoriously difficult to assay genes, such as CEBPA, KRAS, or commonly altered genes such as IDH1 and EGFR (Figures 4B, S7). On average, the EXaCT-2 assay provides 15.8x greater coverage for these genes when compared to EXaCT-1.

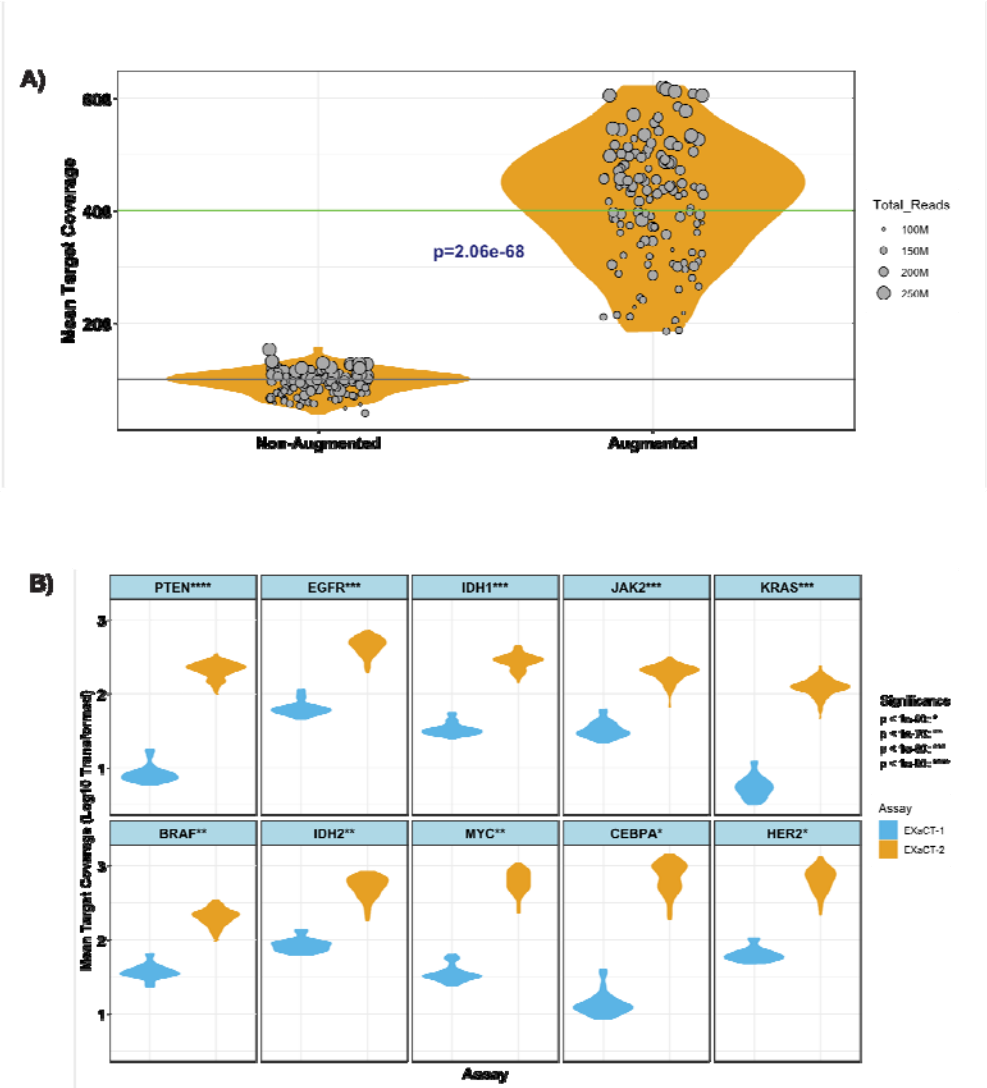


Figure 4: Coverage Analyses. A) Mean target coverage between augmented and non-augmented regions. *Each dot represents a sample where size is proportional to the Total no. of reads per million.* ***B) Comparison of mean target coverage on selected cancer genes.*** *Each plot reports the distribution of log10-transformed mean target coverage on the indicated genes for EXaCT-2 and EXaCT-2.*

As an example of this improved coverage, we observed significantly higher coverage for exons 2-5 of the KRAS protein (Figures 5, S8). With 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 carcinoma and endometrial carcinoma [40–43] (Figures 5A, S8A). KRAS is believed to be one of the most common oncogenic drivers [44–46], heavily researched over the past decade[40]. 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, except when using targeted panels such as the MSK-IMPACT [34] assay (Figure 5B). By way of comparison, the augmentation of the EXaCT-2 assay can achieve a coverage nearly equivalent to that of a targeted panel (Figure 5B).

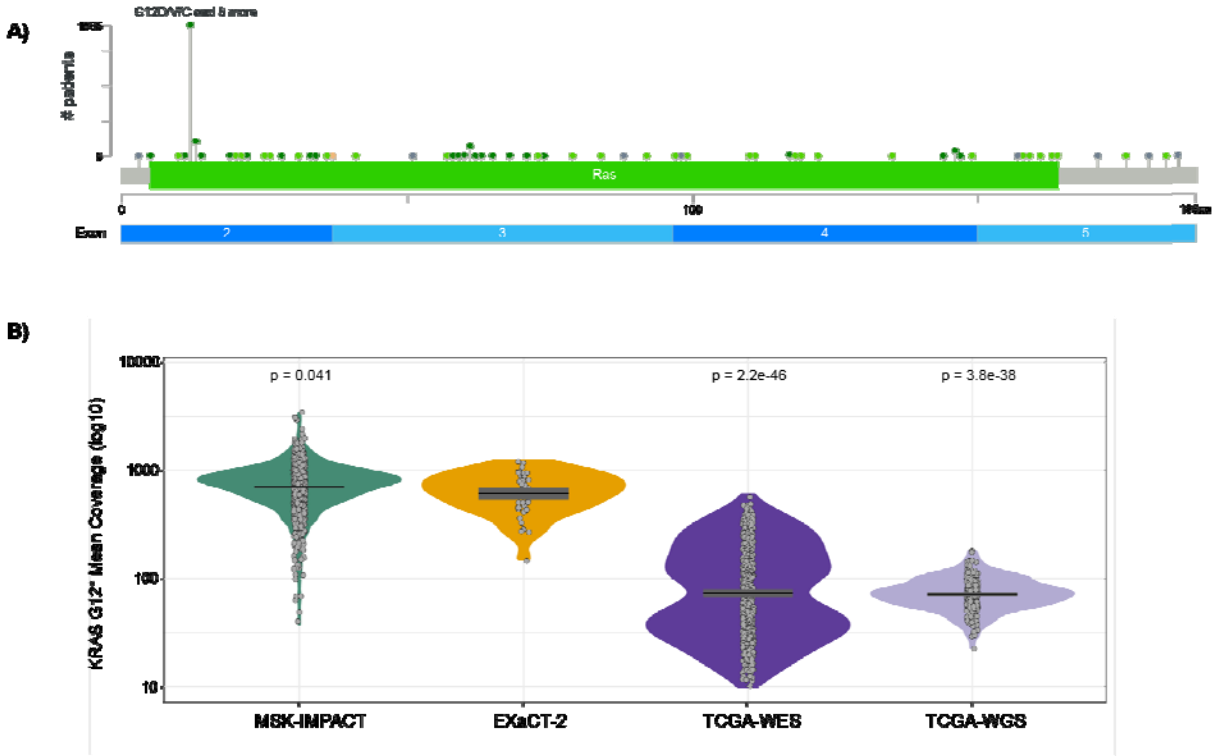


Figure 5: KRAS coverage across protein coding exons. A) *Schematic of the locations of common KRAS mutations from cBioPortal, showing the frequency of the KRAS^{G12*} mutations. Image generated using a combined cohort of patients with colorectal, pancreatic and non-small-cell lung cancers (see Supplemental Methods).* ***B)*** *Comparison of the read coverages for the KRAS^{G12D} mutations in KRAS positive samples assayed by EXaCT-2, TCGA Pan-cancer (WES [47] and WGS [48]) , and MSK-IMPACT [34] (non-EXaCT-2 samples were retrieved from cBioPortal).*

Additionally, the EXaCT-2 evaluation of 27 validation samples harboring 9 clinically relevant cancer variants achieved perfect recovery of these mutations with a high degree of variant allele frequencies (VAF) concordance (Figure S9; Pearson correlation: 0.69; p: 5.6e-5).

Discussion

Next-generation sequencing (NGS) has revolutionized cancer research and treatment by identifying a broad spectrum of genomic alterations that drive the disease. This has facilitated the development of targeted treatments, thus improving survival and reducing toxicity for various tumor types. Targeted assays excel in precision, detecting sub-clonal mutations in a limited number of known cancer driver genes. However, broader assays have also revealed a long-tail of cancer mutations, with only a few genes exhibiting common alterations [49]. Given this complexity of the cancer landscape, an accurate and comprehensive characterization of clinical samples is required. WES assays detect coding mutations on all genes but cannot accurately identify intergenic events. WGS assays cover the broadest genomic territory but have limited coverage especially in GC-rich regions and remain expensive, especially considering the hidden costs of processing and long-term data storage.

EXaCT-2 enhances *depth* of coverage based on known cancer genes and expand the *breadth* by interrogating intronic and intergenic regions to detect common rearrangements and improve copy number boundary detection. Additionally, it examines the presence of viruses, expands the evaluation of the BCR repertoire, and is customizable, allowing rapid adaptation to new discoveries, as demonstrated by the increasing coverage of GTF2I in thymoma or CEBPA (Figures S5, S10)[50].

Overall, EXaCT-2 provides a more comprehensive picture of tumor alterations, particularly important as non-European populations exhibit broader genomic diversity, which is not yet fully represented in current targeted panels.[13, 51–53]

Global metrics such as tumor mutation burden (TMB) and microsatellite instability (MSI) have emerged as crucial prognostic biomarkers for patient outcomes and responses to immunotherapy[19–22, 54–57]. Whole-exome sequencing (WES) is considered the gold standard for accurately estimating TMB.[58, 59, 18, 60,61] EXaCT-2, which evaluates both coding and non-coding regions, can further explores correlations between TMB and MSI (Figures S11, S12), confirming that higher TMB does not necessarily correlate with higher MSI.[62, 63]

The utility of a new exome panel is assessed by comparing the biological information it yields against the effort and costs involved. EXaCT-2 requires 3 days for library preparation and a sequencing depth of 125M reads per sample, comparable with targeted panels and only ~15% more expensive than TSO500 and ~24% cheaper than EXaCT-1 (see Table S14 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. Limitations are: the need for matched tumor/normal specimens, increasing costs and logistical complexity, and the ability to detect only selected rearrangements (n=66 genes). However, the former is common to other targeted assays and the latter is comparable to the RNA component of the TSO500 assay (n=55 genes) [34, 64]. As a DNA-based assay, it cannot determine if rearrangements are expressed, so complementing it with a transcriptome-based assay provides a comprehensive molecular picture.[65]

In summary, EXaCT-2 is a novel and customizable WES assay that balances the strengths of targeted panels and WGS assays, providing comprehensive coverage and detecting clinically relevant rearrangements and viruses.

Methods

Sample Collection, DNA Preparation, Library Construction and Sequencing

Samples from 120 cancer patients were collected (IRB1305013903, IRB0107004999, IRB1008011221, IRB1011011386). H&E-stained slides were annotated by board-certified pathologists and tumor DNA was excised from fresh frozen tissues or FFPE. Germline DNA was obtained from blood, buccal swabs, or benign tissue. DNA was isolated using Promega Maxwell 16 (Promega) and evaluated using Qubit dsDNA HS Assay (ThermoFisher Scientific) and TapeStation System (Agilent). DNA from 2 Multiple Myeloma patients was derived from CD8+ bone marrow biopsies (IRB10004608). FFPE from 28 patients for fusions and viral content, and DNA from 27 patients for clinical mutations studies, were provided by the WCM Clinical Lab; HBL-1 DNA by Dr. Tam, WCM HemOnc. DNA with Integrity Number (DIN) <4 was repaired using NEBNext FFPE DNA Repair MIX M6630 (NewEngland BioLabs). Libraries were prepared with 50ng (DIN>-4) or 200ng (DIN<4) of DNA using SureSelect Enzymatic Fragmentation and XT Low Input kits following manufacturer instructions (Agilent, Cat 5191 and G9703A). Indexed libraries (~1500ng) were hybridized with 5µl of the EXaCT-2 Target Enrichment probes (Agilent, Cat G9496D). EXaCT-2 probes were designed using high-density tiling (5X), with a booster probe set for cancer regions with <50X coverage and an augmented probe set for a subset of these gene. Probes are approximately 120bp size, which also is about the median exon size of human genes.[66] EXaCT-2 composition: SureSelect V6 + COSMIC (1X), Non-Augmented Regions (1X), Augmented Targets (5X), Booster Probe Set (5X), CEBPA (300X), and Viral Genomes (1/30X). Libraries obtained were sequenced for 150 cycles on a pair-end flowcell (Illumina NovaSeq6000) to obtain ~125M read per sample.

Bioinformatics processing

Data were processed using Illumina's RTA and bcl2fastq software to generate and demultiplex FASTQ files. Quality Control analysis of the sequencing run was performed with fastp[67]. Reads passing Illumina's purity filter are adapter trimmed and aligned to a modified version of the GRCh37 human genome that includes the viruses included in the assay. Quality control analysis is performed upon the result BAM files, and the matched tumor-normal samples are analyzed with SPIA[68] to ensure that the samples are from the same individual.

The tools used for the different analyses for each matched tumor-normal sample pair are listed hereafter (details provided in Supplementary Methods): An updated version of the MC3 algorithm (SNV & Indel Identification)[69], and Nirvana (annotation)[70]; CNVkit (SCNA Identification)[71]; MSISensor-Pro (Microsatellite instability)[72, 73]; Delly (Rearrangement Detection)[74]; in-house calculation for Tumor Mutation Burden (TMB)[61, 75]; MiXCR (Immuno-globulin (Ig) rearrangements)[76]; in-house tool for Viral Detection.

References

1. Hanahan D, Weinberg RA: Hallmarks of Cancer: The Next Generation. Cell 144:646–674, 2011

2. Sondka Z, Bamford S, Cole CG, et al: The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer 18:696–705, 2018

3. Tate JG, Bamford S, Jubb HC, et al: COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res 47:D941–D947, 2019

4. Nocini R, Tondulli L, D'Ambrosio C, 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

5. Planchard D: Identification of driver mutations in lung cancer: first step in personalized cancer. Target Oncol 8:3–14, 2013

6. Aguirre AJ, Nowak JA, Camarda ND, et al: Real-time Genomic Characterization of Advanced Pancreatic Cancer to Enable Precision Medicine. Cancer Discov 8:1096–1111, 2018

7. Ecker BL, Tao AJ, Janssen QP, et al: Genomic Biomarkers Associated with Response to Induction Chemotherapy in Patients with Localized Pancreatic Ductal Adenocarcinoma. Clin Cancer Res 29:1368–1374, 2023

8. Aung KL, Fischer SE, Denroche RE, et al: Genomics-Driven Precision Medicine for Advanced Pancreatic Cancer: Early Results from the COMPASS Trial. Clin Cancer Res Off J Am Assoc Cancer Res 24:1344–1354, 2018

9. Beg A, Parveen R, Fouad H, et al: Unravelling driver genes as potential therapeutic targets in ovarian cancer via integrated bioinformatics approach. J Ovarian Res 17:86, 2024

10. Zhang SX, 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

11. Hindorff LA, Bonham VL, Brody LC, et al: Prioritizing diversity in human genomics research. Nat Rev Genet 19:175–185, 2018

12. Fatumo S, Chikowore T, Choudhury A, et al: A roadmap to increase diversity in genomic studies. Nat Med 28:243–250, 2022

13. Wang T, Antonacci-Fulton L, Howe K, et al: The Human Pangenome Project: a global resource to map genomic diversity. Nature 604:437–446, 2022

14. ICGC ARGO [Internet][cited 2024 Jul 9] Available from: https://platform.icgc-argo.org/

15. H3Africa – Human Heredity & Health in Africa [Internet][cited 2024 Jul 9] Available from: https://h3africa.org/

16. The "All of Us" Research Program. N Engl J Med 381:668–676, 2019

17. All of Us Research Hub [Internet][cited 2024 Jul 9] Available from: https://www.researchallofus.org/

18. Fang H, Bertl J, Zhu X, et al: Tumour mutational burden is overestimated by target cancer gene panels. J Natl Cancer Cent 3:56–64, 2022

19. Wu H-X, Wang Z-X, Zhao Q, et al: Tumor mutational and indel burden: a systematic pan-cancer evaluation as prognostic biomarkers. Ann Transl Med 7:640, 2019

20. Sha D, Jin Z, Budzcies J, et al: Tumor Mutational Burden (TMB) as a Predictive Biomarker in Solid Tumors. Cancer Discov 10:1808–1825, 2020
21. Huang T, Chen X, Zhang H, 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
22. Griffin R, Wenzl K, Sarangi V, et al: Tumor Mutational Burden As a Prognostic Factor in Diffuse Large B-Cell Lymphoma. Blood 142:1633, 2023
23. Alexandrov LB, Nik-Zainal S, Wedge DC, et al: Signatures of mutational processes in human cancer. Nature 500:415–421, 2013
24. Harbers L, Agostini F, Nicos M, 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
25. Buccitelli C, Salgueiro L, Rowald K, et al: Pan-cancer analysis distinguishes transcriptional changes of aneuploidy from proliferation. Genome Res 27:501–511, 2017
26. Davoli T, Uno H, Wooten EC, et al: Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. Science 355:eaaf8399, 2017
27. Taylor AM, Shih J, Ha G, et al: Genomic and Functional Approaches to Understanding Cancer Aneuploidy. Cancer Cell 33:676-689.e3, 2018
28. Hawkins C, Walker E, Mohamed N, et al: BRAF-KIAA1549 Fusion Predicts Better Clinical Outcome in Pediatric Low-Grade Astrocytoma. Clin Cancer Res 17:4790–4798, 2011
29. Vellichirammal NN, Chaturvedi NK, Joshi SS, et al: Fusion genes as biomarkers in pediatric cancers: A review of the current state and applicability in diagnostics and personalized therapy. Cancer Lett 499:24, 2021
30. Bao Z, Chai R, Liu X, et al: Fusion genes as diagnostic and predictive biomarkers for tumor. Glob Transl Med 1:54, 2022
31. TruSight Oncology 500 | Enable CGP utilizing DNA and RNA from FFPE [Internet][cited 2024 Jul 9] Available from: https://www.illumina.com/products/by-type/clinical-research-products/trusight-oncology-500.html
32. Scientific TF: Oncomine Next-Generation Sequencing Solutions for Precision Oncology Research [Internet][cited 2024 Jul 9] Available from: https://www.oncomine.com
33. GENEWIZ from Azenta Life Sciences [Internet][cited 2024 Aug 20] Available from: https://www.genewiz.com/
34. Cheng DT, Mitchell TN, Zehir A, et al: Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT). J Mol Diagn JMD 17:251–264, 2015
35. Rose Brannon A, Jayakumaran G, Diosdado M, et al: Enhanced specificity of clinical high-sensitivity tumor mutation profiling in cell-free DNA via paired normal sequencing using MSK-ACCESS. Nat Commun 12:3770, 2021
36. Altorki NK, Borczuk AC, Harrison S, et al: Global evolution of the tumor microenvironment associated with progression from preinvasive invasive to invasive human lung adenocarcinoma. Cell Rep 39:110639, 2022

37. Agilent Technologies: SureSelect Human All Exon V6 [Internet], 2015[cited 2024 Dec 2] Available from: https://www.agilent.com/cs/library/datasheets/public/SureSelect%20V6%20DataSheet%205991-5572EN.pdf
38. FusionGDB 2.0: fusion gene annotation updates aided by deep learning | Nucleic Acids Research | Oxford Academic [Internet][cited 2024 Sep 3] Available from: https://academic.oup.com/nar/article/50/D1/D1221/6424754
39. Shin H-T, Choi Y-L, Yun JW, et al: Prevalence and detection of low-allele-fraction variants in clinical cancer samples. Nat Commun 8:1377, 2017
40. Hofmann MH, Gerlach D, Misale S, et al: Expanding the Reach of Precision Oncology by Drugging All KRAS Mutants. Cancer Discov 12:924–937, 2022
41. Lu D, Chen Y, Jiang M, et al: KRAS G12V neoantigen specific T cell receptor for adoptive T cell therapy against tumors. Nat Commun 14:6389, 2023
42. Zeissig MN, Ashwood LM, Kondrashova O, et al: Next batter up! Targeting cancers with KRAS-G12D mutations. Trends Cancer 9:955–967, 2023
43. Qunaj L, May MS, Neugut AI, et al: Prognostic and therapeutic impact of the KRAS G12C mutation in colorectal cancer. Front Oncol 13:1252516, 2023
44. Prior IA, Hood FE, Hartley JL: The Frequency of Ras Mutations in Cancer. Cancer Res 80:2969–2974, 2020
45. Cerami E, Gao J, Dogrusoz U, et al: The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discov 2:401–404, 2012
46. The AACR Project GENIE Consortium, The AACR Project GENIE Consortium, André F, et al: AACR Project GENIE: Powering Precision Medicine through an International Consortium. Cancer Discov 7:818–831, 2017
47. Hoadley KA, Yau C, Wolf DM, et al: Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell 158:929–944, 2014
48. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium: Pan-cancer analysis of whole genomes. Nature 578:82–93, 2020
49. Armenia J, Wankowicz SAM, Liu D, et al: The long tail of oncogenic drivers in prostate cancer. Nat Genet 50:645–651, 2018
50. Petrini I, Meltzer PS, Kim I-K, et al: A specific missense mutation in GTF2I occurs at high frequency in thymic epithelial tumors. Nat Genet 46:844–849, 2014
51. Sirugo G, Williams SM, Tishkoff SA: The Missing Diversity in Human Genetic Studies. Cell 177:26–31, 2019
52. Venner E, Patterson K, Kalra D, et al: The frequency of pathogenic variation in the All of Us cohort reveals ancestry-driven disparities. Commun Biol 7:174, 2024
53. Valle LF, Li J, Desai H, 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
54. Li K, Luo H, Huang L, et al: Microsatellite instability: a review of what the oncologist should know. Cancer Cell Int 20:16, 2020
55. Chang L, Chang M, Chang HM, et al: Microsatellite Instability: A Predictive Biomarker for Cancer Immunotherapy. Appl Immunohistochem Mol Morphol AIMM 26:e15–e21, 2018

56. Morihiro T, Kuroda S, Kanaya N, 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
57. Rantanen P, Keränen A, Barot S, et al: The prognostic significance of microsatellite instability in colorectal cancer: a Swedish multi-center study. Int J Colorectal Dis 38:197, 2023
58. Rizvi NA, Hellmann MD, Snyder A, et al: Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science 348:124–128, 2015
59. Fancello L, Gandini S, Pelicci PG, et al: Tumor mutational burden quantification from targeted gene panels: major advancements and challenges. J Immunother Cancer 7:183, 2019
60. Furtado LV, Bifulco C, Dolderer D, 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
61. Fernandez EM, Eng K, Beg S, et al: Cancer-Specific Thresholds Adjust for Whole Exome Sequencing–Based Tumor Mutational Burden Distribution. JCO Precis Oncol 1–12, 2019
62. Salem ME, Puccini A, Grothey A, 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
63. Goodman AM, Sokol ES, Frampton GM, et al: Microsatellite-Stable Tumors with High Mutational Burden Benefit from Immunotherapy. Cancer Immunol Res 7:1570–1573, 2019
64. Wei B, Kang J, Kibukawa M, 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
65. Beg S, Bareja R, Ohara K, 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
66. Mokry M, Feitsma H, Nijman IJ, 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
67. Chen S, Zhou Y, Chen Y, et al: fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884–i890, 2018
68. Demichelis F, Greulich H, Macoska JA, et al: SNP panel identification assay (SPIA): a genetic-based assay for the identification of cell lines. Nucleic Acids Res 36:2446–2456, 2008
69. Ellrott K, Bailey MH, Saksena G, et al: Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. Cell Syst 6:271-281.e7, 2018
70. Stromberg M, Roy R, Lajugie J, et al: Nirvana: Clinical Grade Variant Annotator [Internet], in Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics. New York, NY, USA, Association for

Computing Machinery, 2017, p 596[cited 2024 Nov 6] Available from: https://doi.org/10.1145/3107411.3108204

71. Talevich E, Shain AH, Botton T, et al: CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLOS Comput Biol 12:e1004873, 2016

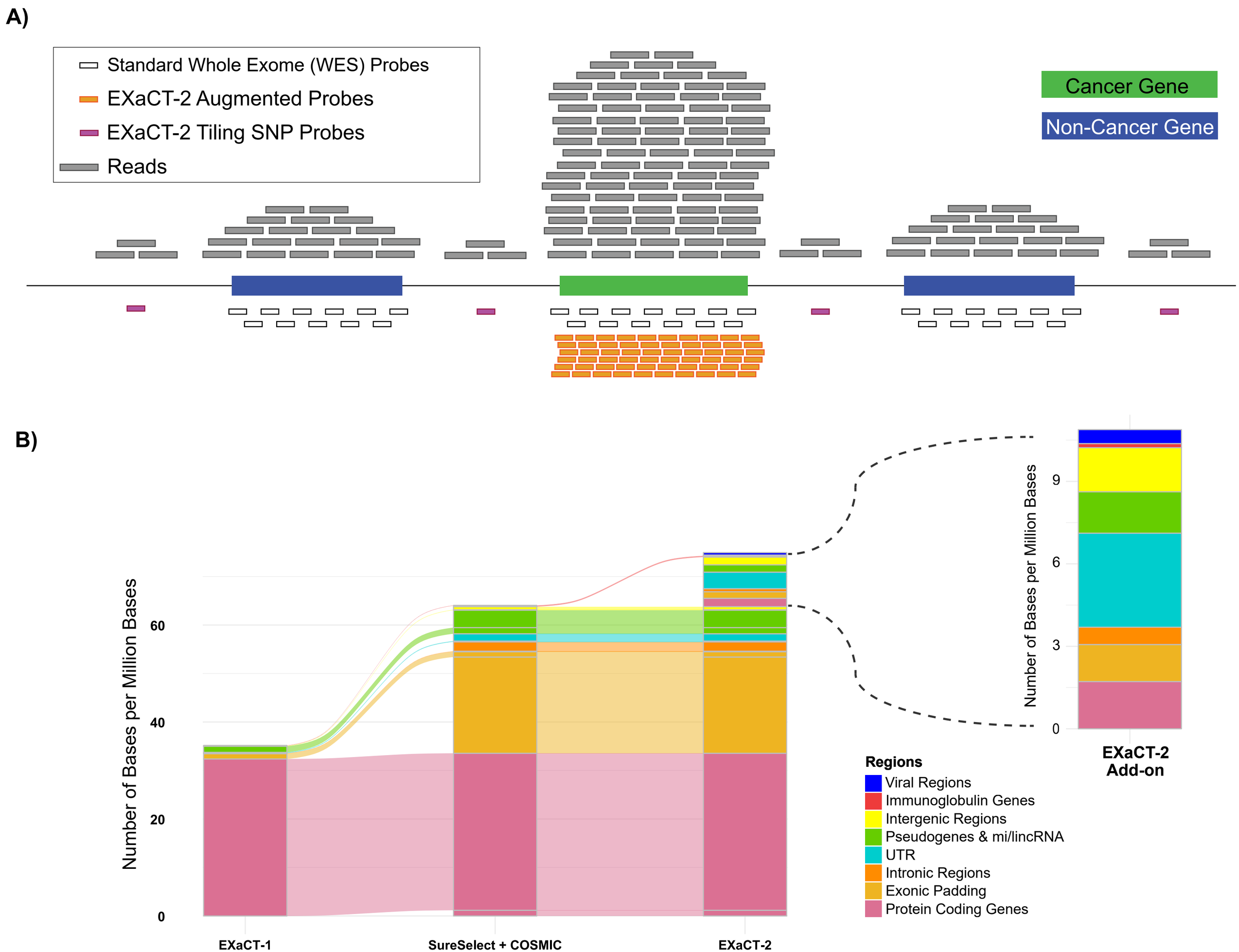
72. Niu B, Ye K, Zhang Q, et al: MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. Bioinformatics 30:1015–1016, 2014

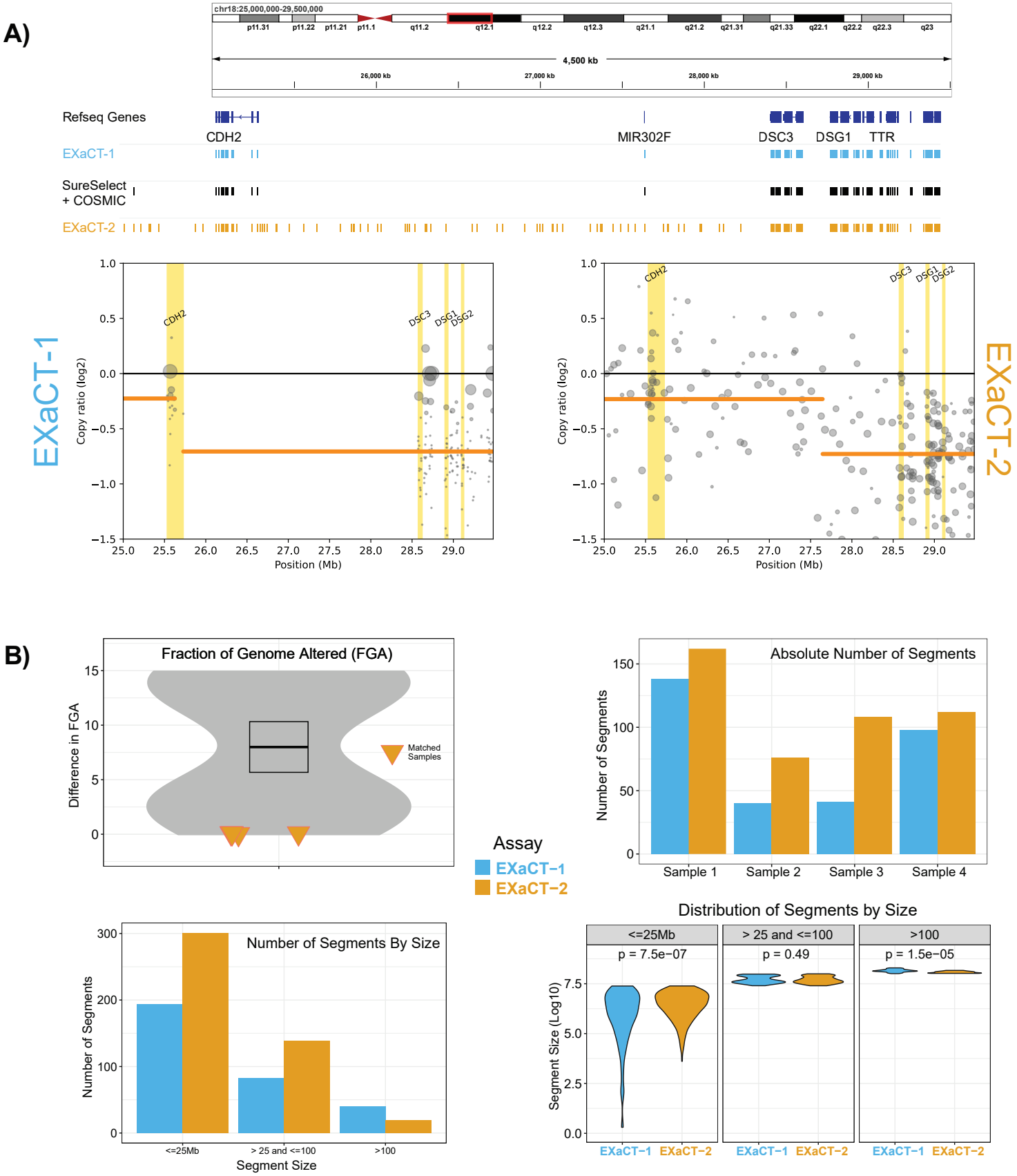
73. Jia P, Yang X, Guo L, et al: MSIsensor-Pro: Fast, Accurate, and Matched-Normal-Sample-Free Detection of Microsatellite Instability. Genomics Proteomics Bioinformatics 18:65–71, 2020

74. Rausch T, Zichner T, Schlattl A, et al: DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28:i333–i339, 2012

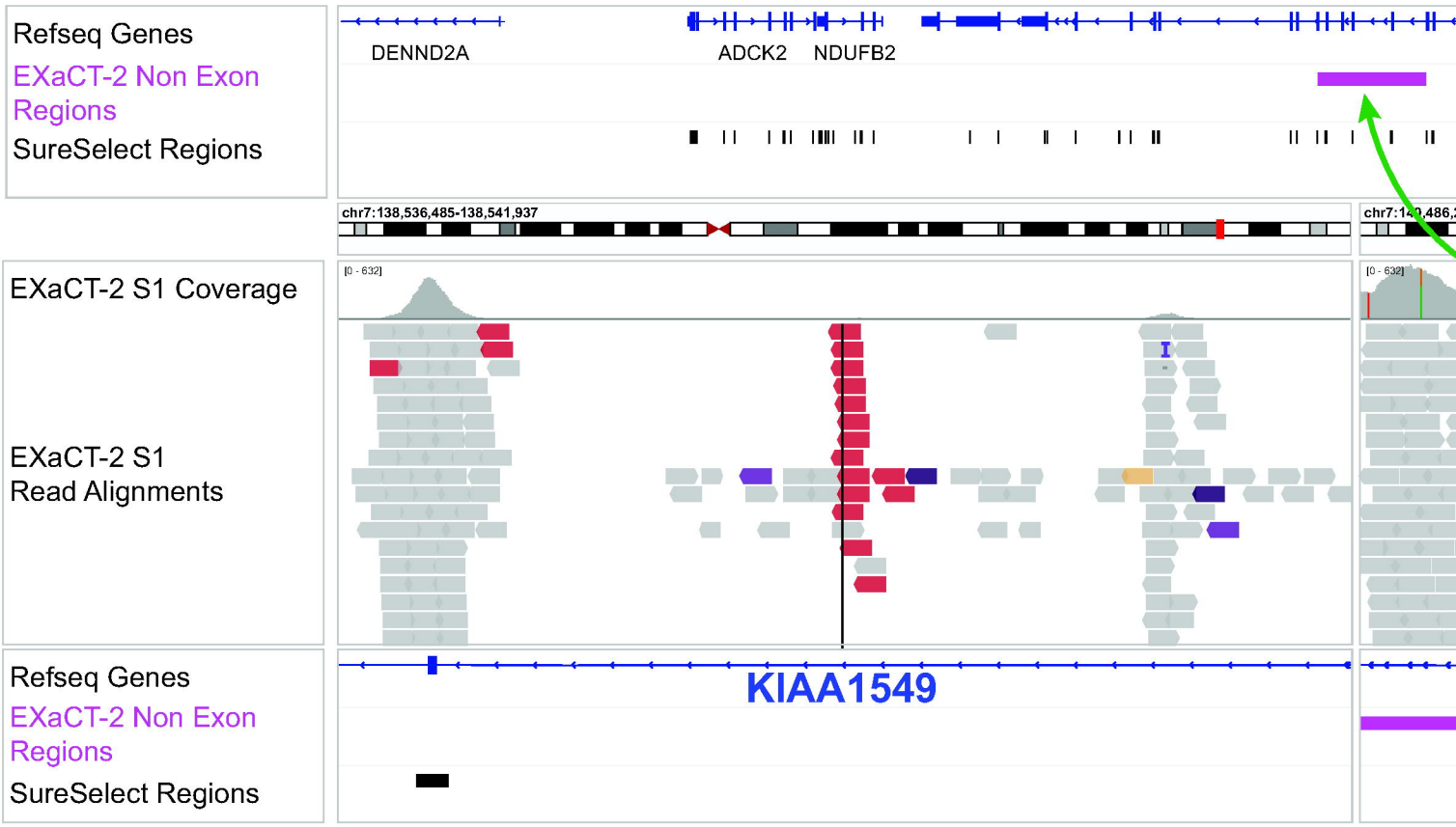
75. Chalmers ZR, Connelly CF, Fabrizio D, et al: Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. Genome Med 9:34, 2017

76. Bolotin DA, Poslavsky S, Mitrophanov I, et al: MiXCR: software for comprehensive adaptive immunity profiling. Nat Methods 12:380–381, 2015

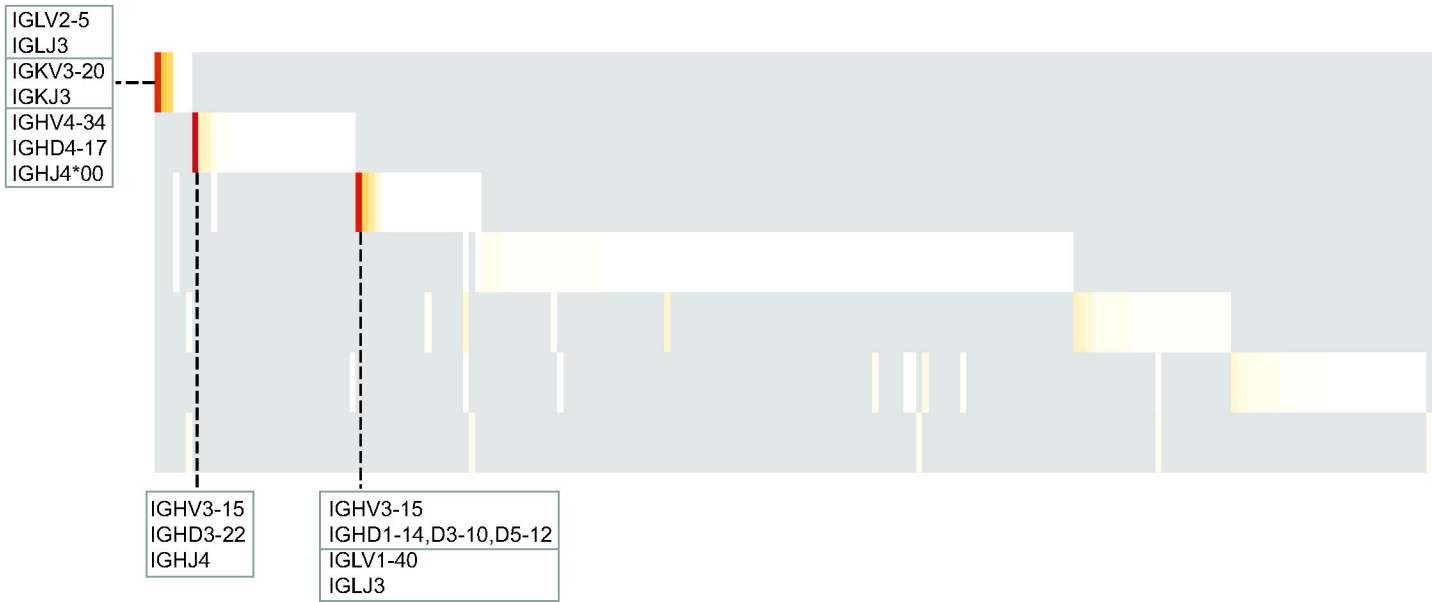

A)
Standard Whole Exome (WES) Probes
EXaCT-2 Augmented Probes
EXaCT-2 Tiling SNP Probes
Reads
Cancer Gene
Non-Cancer Gene
B)
Number of Bases per Million Bases
0
20
40
60
EXaCT-1
SureSelect + COSMIC
EXaCT-2
Number of Bases per Million Bases
0
3
6
9
EXaCT-2
Add-on
Regions
Viral Regions
Immunoglobulin Genes
Intergenic Regions
Pseudogenes & mi/lincRNA
UTR
Intronic Regions
Exonic Padding
Protein Coding Genes

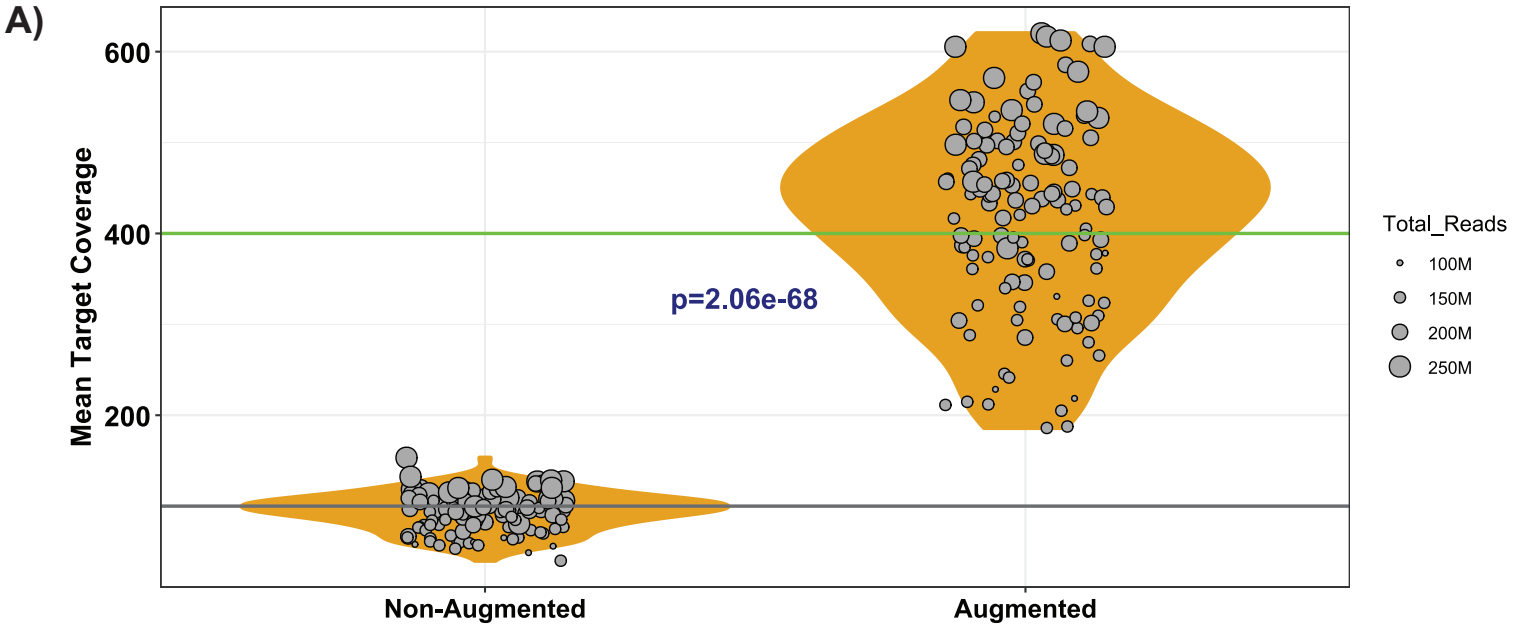

A)
chr18:25,000,000-29,500,000
4,500 kb
Refseq Genes
CDH2
MIR302F
DSC3
DSG1
TTR
EXaCT-1
SureSelect + COSMIC
EXaCT-2
Copy ratio (log2)
Position (Mb)
B)
Fraction of Genome Altered (FGA)
Difference in FGA
Matched Samples
Absolute Number of Segments
Number of Segments
Sample 1
Sample 2
Sample 3
Sample 4
Assay
EXaCT−1
EXaCT−2
Number of Segments By Size
<=25Mb
> 25 and <=100
>100
Segment Size
Distribution of Segments by Size
p = 7.5e−07
p = 0.49
p = 1.5e−05
Segment Size (Log10)

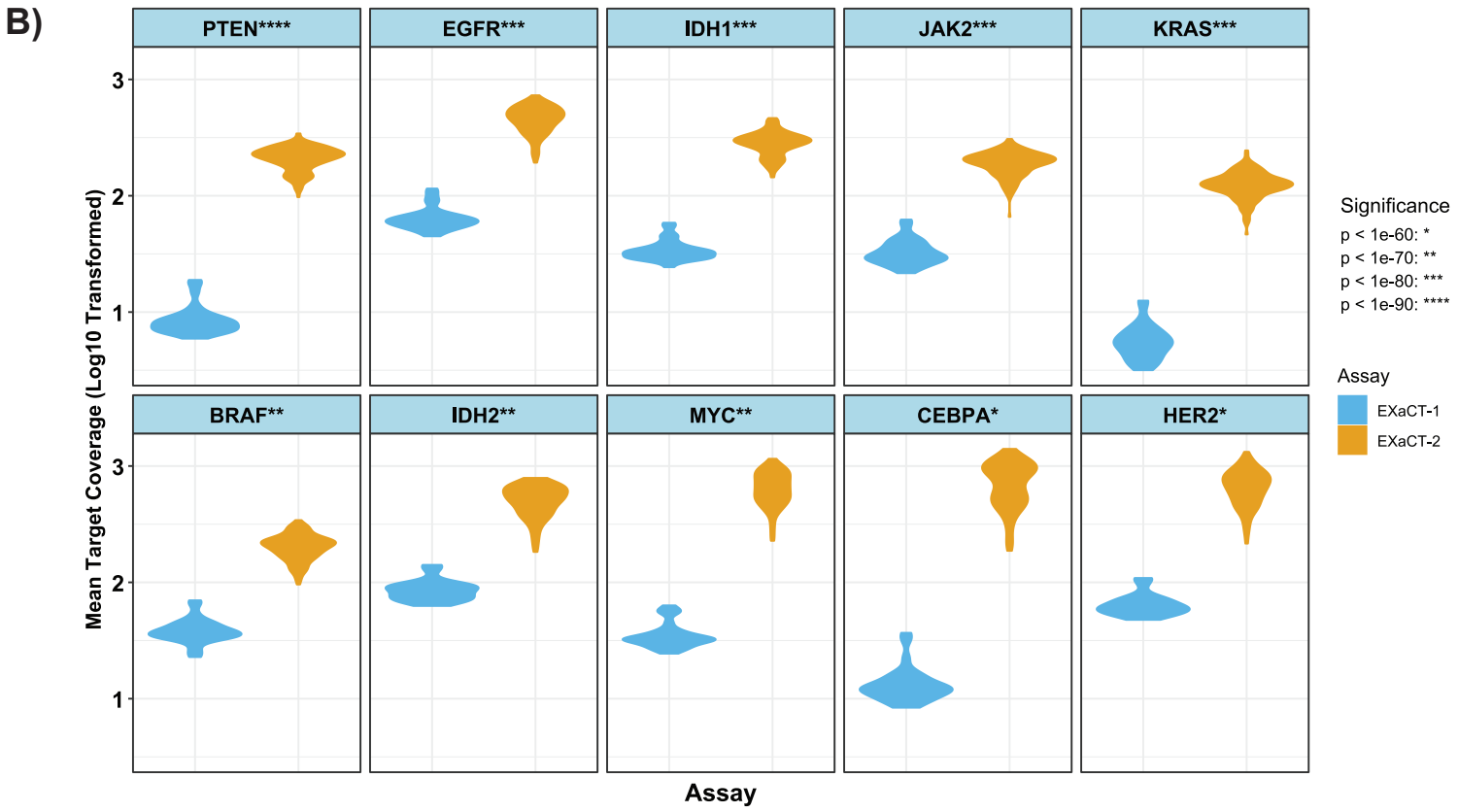
A)

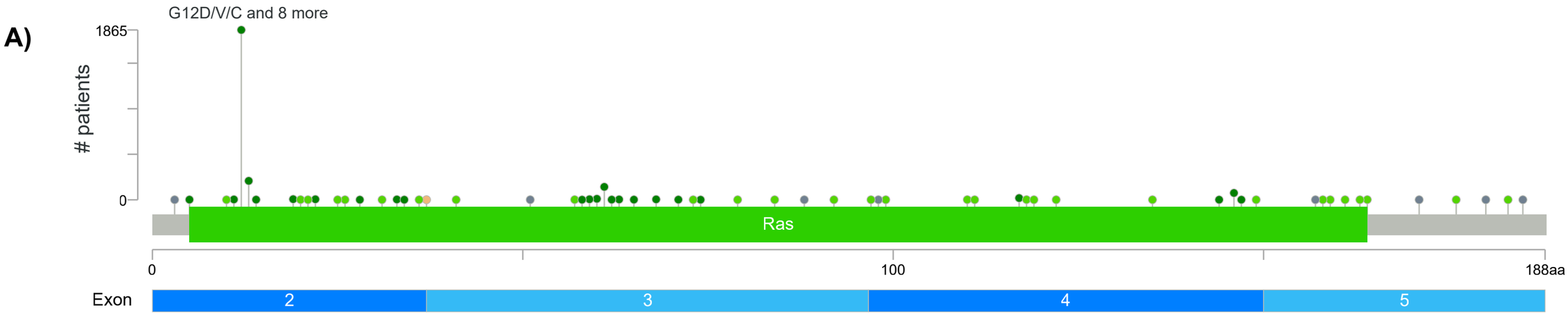

Refseq Genes
EXaCT-2 Non Exon Regions
SureSelect Regions
DENND2A
ADCK2
NDUFB2
chr7:138,536,485-138,541,937
EXaCT-2 S1 Coverage
EXaCT-2 S1 Read Alignments
KIAA1549
Refseq Genes
EXaCT-2 Non Exon Regions
SureSelect Regions

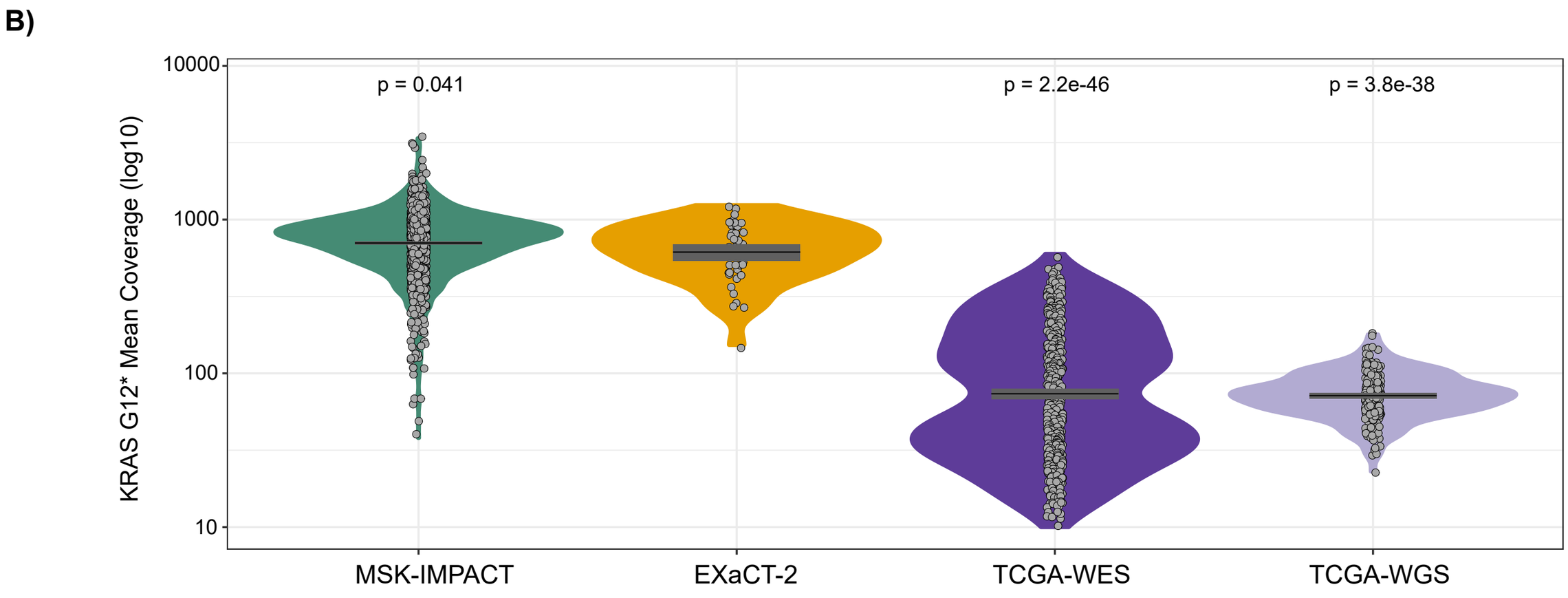
B)


IGLV2-5
IGLJ3
IGKV3-20
IGKJ3
IGHV4-34
IGHD4-17
IGHJ4*00
IGHV3-15
IGHD3-22
IGHJ4
IGHV3-15
IGHD1-14,D3-10,D5-12
IGLV1-40
IGLJ3


A)
Mean Target Coverage
600
400
200
p=2.06e-68
Non-Augmented
Augmented
Total_Reads
100M
150M
200M
250M


B)
PTEN****
EGFR***
IDH1***
JAK2***
KRAS***
BRAF**
IDH2**
MYC**
CEBPA*
HER2*
Mean Target Coverage (Log10 Transformed)
3
2
1
Assay
Significance
p < 1e-60: *
p < 1e-70: **
p < 1e-80: ***
p < 1e-90: ****
Assay
EXaCT-1
EXaCT-2


A)
G12D/V/C and 8 more
1865
patients
0
Ras
0
100
188aa
Exon
2
3
4
5


B)
10000
1000
100
10
KRAS G12* Mean Coverage (log10)
p = 0.041
p = 2.2e-46
p = 3.8e-38
MSK-IMPACT
EXaCT-2
TCGA-WES
TCGA-WGS