Multi-Platform Assessment of DNA Sequencing Performance using Human and Bacterial Reference Materials in the ABRF Next-Generation Sequencing Study

Jonathan Foox[1,2], Scott W. Tighe[3], Charles M. Nicolet[4], Justin M. Zook[5], Marta Byrska-Bishop[6], Wayne E. Clarke[6], Michael M. Khayat[7,8], Medhat Mahmoud[7,8], Phoebe K.Laaguiby[3], Zachary T. Herbert[9], Derek Warner[10], George S. Grills[11], Jin Jen[12], Shawn Levy[13], Jenny Xiang[1], Alicia Alonso[1], Xia Zhao[14,15], Wenwei Zhang[14], Fei Teng[14], Yonggang Zhao[14,16], Haorong Lu[14,17], Gary P. Schroth[18], Giuseppe Narzisi[6], William Farmerie[19], Fritz J. Sedlazeck[7,8 *], Don A. Baldwin[20 *], Christopher E. Mason[1,2,21,22 *]

[1] Department of Physiology and Biophysics, Weill Cornell Medicine, New York, New York, USA
[2] The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medicine, New York, New York, USA
[3] University of Vermont Cancer Center, Vermont Integrative Genomics Resource, University of Vermont, Burlington, Vermont, USA
[4] Keck School of Medicine, University of Southern California, Los Angeles, California, USA
[5] Biosystems and Biomaterials Division, National Institute of Standards and Technology, Gaithersburg, Maryland, USA
[6] New York Genome Center, New York, NY, 10013, USA
[7] Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA;
[8] Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas;
[9] Molecular Biology Core Facilities, Dana-Farber Cancer Institute, Boston, Massachusetts, USA
[10] DNA Sequencing Core, University of Utah, Salt Lake City, Utah, USA
[11] Sylvester Comprehensive Cancer Center, University of Miami, Miami, FL, USA
[12] Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN
[13] HudsonAlpha Institute for Biotechnology, Huntsville, Alabama, USA
[14] BGI-Shenzhen, Shenzhen 518083, China
[15] MGI, BGI-Shenzhen, Shenzhen 518083, China
[16] Department of Biotechnology and Biomedicine, Technical University of Denmark, Denmark
[17] Guangdong Provincial Key Laboratory of Genome Read and Write, Shenzhen 518120, China
[18] Illumina, Inc., San Diego, CA, USA
[19] Interdisciplinary Center for Biotechnology Research, University of Florida, Gainesville, Florida, USA
[20] Department of Pathology, Fox Chase Cancer Center, Philadelphia, Pennsylvania, USA
[21] The Feil Family Brain and Mind Research Institute, New York, New York, USA
[22] The WorldQuant Initiative for Quantitative Prediction, Weill Cornell Medicine, New York, NY, USA
[*] Corresponding authors. Send correspondence to D.A.B (donald.baldwin@fccc.edu), F.J.S (fritz.sedlazeck@bcm.edu), or C.E.M. (chm2042@med.cornell.edu)

Abstract

Massively parallel DNA sequencing is a critical tool for genomics research and clinical diagnostics, but there is a variegated landscape of technologies, platforms, and chemistries. Here, data from the Association of Biomolecular Resource Facilities (ABRF) Next-Generation Sequencing (NGS) Study was used to delineate the reproducibility, accuracy, and utility of both current and emerging NGS platforms. Human and bacterial reference DNA samples were sequenced on Illumina HiSeq/NovaSeq and ThermoFisher Ion Torrent instruments, the Pacific Biosciences and Oxford Nanopore long read sequencers, and the recently released BGI/MGISEQ platform, GenapSys GS110 sequencer, and the Illumina paired-end 2x250bp chemistry. Each platform showed variable reference-based mapping rates, coverage disruptions in high/low GC regions (the lowest in Pacific Biosciences), sequencing mismatch and insertion/deletion rates, and variable variant detection of single-nucleotide variants (SNVs) and structural variants (SVs). The long-read platforms showed the best mapping in repeat-rich areas and across homopolymers, whereas some short-read platforms (e.g. GenapSys) had very distinct base composition profiles, both of which are critical for modeling variant calling. As chemistries, methods, and platforms continue to evolve for NGS, this study serves as a benchmark for current and future genomic technological development, as well as a resource to inform experimental design and NGS variant-calling.

Introduction

High-throughput next-generation DNA sequencing (DNA-seq) is an essential method for clinical and basic biomedical research [1, 2]. DNA-seq has numerous experimental applications, including but not limited to genotyping and variant discovery within individuals [3], population- and species-level characterization of genomes [4], and revealing taxonomic diversity within a metagenomic mixture [5]. Genome sequencing has become ubiquitous, owing to the significant decrease in cost [6], which has led to diversification of sample collection, library preparation, sequencing chemistries, and downstream bioinformatic pipelines. Rapid advancement of DNA-seq has also enabled clinical standards to emerge and proficiency tests to be established that are routinely run by medical organizations [7, 8]. Prior studies have provided valuable reference sets for various modalities of sequencing, including amplicons [9], multilocus/core genome bacterial typing [10], and DNA-seq within then-emerging instruments [11]. The Microarray Quality Control (MAQC) Consortium has led several large-scale studies of RNA-seq reproducibility [12, 13] [14], RNA-seq quality control [15], concordance with microarrays [16], and best practices for data processing [17] and normalization [18] but there is not yet an analogous study for DNA-seq reproducibility. Signicant studies have laid the groundwork for proficiency trials and accreditation of NGS devices for clinical use that have leverged large cohorts [19] across large collections of participating laboratories [20, 21]. As sequencing technologies continue to evolve, a broad collection of DNA-seq data can serve as a robust benchmarking resource to facilitate further standardization of clinical applications, as well as to evaluate new methods, chemistries, and protocols.

The Genome In A Bottle (GIAB) Consortium has enabled genomics benchmarking by developing a series of reference materials (RM) [22], benchmarking tools [23], ultra-deep sequence data [24], and benchmarking variant reference sets [25]. Here, the ABRF NGS Phase II DNA-seq Study leverages reference materials (NIST RM 8392, known as the Ashkenazi Trio; Mother (HG004), Father (HG003), and Son (HG002), a family trio consented through the Personal Genome Project [26]) to provide insight into currently common sequencing instruments. Inter- and intra-lab DNA-seq replicates of the Ashkenazi Trio are analyzed, as well as three individual bacterial strains and a metagenomic mixture of ten bacterial species to study the effects of GC content and library complexity. These replicates were generated across six Illumina and three ThermoFisher Ion Torrent platforms, the BGI-SEQ 500 and MGISEQ-2000 platforms, the GenapSys GS110 platform, and using Oxford Nanopore Flongle, MinION, and PromethION flow cells, as well as publicly available PacBio Circular Consensus Sequence (CCS) data for HG002. These data are tested within the most "difficult" regions of the genome, represented by the UCSC RepeatMasker regions, to highlight the differences between each instrument. All data generated by this consortium were examined for performance and reproducibility over a range of base compositions and GC-content profiles. Collectively, these data provide a robust benchmarking resource for human and bacterial DNA-seq NGS across a wealth of sequencing instruments.

Results

Data Quality

Human and bacterial genomic and targeted exomic libraries were sequenced across an array of platforms, including five Illumina platforms, three Ion Torrent platforms, Oxford Nanopore MinION (R9.4 and Flongle flow cells) and PromethION, BGI-SEQ 500, MGI-SEQ 2000, and GenapSys GS110 (Figure 1A). PacBio datasets generated using Circular Consensus Sequencing (CCS) as well as additional Oxford PromethION flow cells were downloaded from the NCBI Genome database to ensure full representation of commonly used platforms. Multiple inter- and intra-lab replicates per library were prepared for the majority of instruments in this study (see Supplementary Table 1 for an exhaustive list of replicates generated by each sequencing facility).

Depth of sequencing varied across experiment type, ranging from ultra deep genomic coverage of bacterial taxa (nearing 1000x mean coverage) to shallow genomic coverage (<1x mean global coverage). Most WGS libraries were sequenced to between 25-80x mean coverage (Figure 1B), and subsequent analyses were performed on alignments downsampled to 25x mean coverage (see below).

The overall quality of sequence data was consistently high across all libraries, including base quality scores, GC distributions, balanced sequence content, low N content, and low sequence duplication levels (complete FASTQC quality control reports for every replicate are available in Supplementary Data 1). Insert size distributions were highly library-specific (Figure S1). Human data were aligned against GRCh38 with decoy contigs (see methods) which successfully deflected 1% of Illumina and GenapSys reads, 0.5% of BGI- and MGI-SEQ reads, nearly no ThermoFisher reads, and 2-5% of long read data (Figure S2).

Mapping rates were consistent within instruments but highly variable between (Figure 1C). BGI-SEQ 500 and GenapSys had the lowest short read unique mapping efficiency and highest multi-mapping rate, possibly owing to 2x100bp and 150bp single-end chemistries, respectively. ThermoFisher mapping rates were slightly better than Illumina and MGI technologies, reflecting fewer regions in the exome that are difficult to align. PacBio CCS had the most accurate mapping rate compared to Oxford platforms. Not pictured are PromethION replicates, whose mapping rates were around 85%, far lower than other platforms due to the significant fraction of shorter reads within these datasets that do not map. BGI- and MGI-SEQ had lower optical duplicate and unmapped read rates than Illumina platforms, although both data types were very efficient. Total mapping rates are available in Supplementary Table 2. All replcates showed highly consistent capture per GC bin with no platform-specific effect, although whole genome and targeted exome capture revealed differences in GC composition (Figure S3). For AmpliSeq Exome panels used on IonTorrent instruments, the rate of on-target mapping was high, ranging from 84.6–96.6%, with little variation between replicates, showing high consistency for this assay (Supplementary Table 3).

Three individual bacterial species and one metagenomic mixture comprising ten bacterial species were sequenced on Illumina, Ion Torrent, Oxford Nanopore, and GenapSys platforms (Figure 1D). The species chosen for individual and metagenomic sequencing comprised a wide variety of genome sizes, GC content, Gram staining responses, ecological niches, and in some cases would provide physiological challenges for capture, such as high saline affinity (Supplementary Table 4), meant to challenge each platform's ability to overcome these factors. The mappability of reads from each library was found to be directly related to the species sequenced, with high variability between species and high consistency within each instrument.

Normalized Coverage Analysis

Evenness of coverage across the genome was calculated per instrument, using only replicates that had sufficient coverage (mean depth of coverage >=10x with a mapping quality cutoff of MQ20), and with alignments normalized to a global mean of 25x coverage per replicate. Note that replicates from GenapSys and the Flongle and R9.4 MinION flow cells (as two replicates from the HiSeq2500 platform) were excluded here due to inadequate coverage.

Coverage distributions were very consistent among technologies, including short and long reads (Figure 2A). However, within each context, certain platforms out-covered the collective mean of others, based on a one-tailed Wilcoxon base versus mean test. HiSeq2500, BGISEQ500, and MGISEQ2000 consistently under-covered these regions, with HiSeq2500 only out-covering the mean in Low Complexity regions, and BGISEQ500 and MGISEQ2000 only out-covering the mean in Alus (and LTRs for MGISEQ2000). Notably, the HiSeq4000 and HiSeqX10 performed well, with high coverage in L2s, LTRs, and simple repeat regions. NovaSeq replicates performed well in the most regions among short read platforms, particularly using 2x250bp chemistry. Overall, PacBio and PromethION (i.e. long read) technologies outperformed the other platforms in every context. The direct comparison of each platform versus all others summarizes performance across contexts (Figure 2B), and all-vs-all comparisons provide a more detailed profile of any one platform's coverage capture versus any other (Figure S4).

Although the instruments could be stratified by coverage performance, intra-platform varaibility was low, demonstrated by the even distribution of coefficient of variation in all contexts for all platforms (Figure 2C). One notable exception is within satellite regions, where a bimodal distribution of coverage was observed. A sub-set of satellite regions had near-zero coverage across all platforms, primarily on the Y chromosome. The genomic coordinates of each bin of coverage (high and near-zero) for satellite regions is available as Supplementary Data 2.

Sequencing Mismatch Rate

Rates of inconsistency of aligned reads against the reference genome (i.e. mismatch rate) were characterized against UCSC RepeatMasker region to evaluate sequencing performance in difficult regions (Figure 3A). Overall, short read platforms had lower mismatch rates in every context compared to Nanopore. However, PacBio CCS reads had mismatch rates equal to or even lower than short reads in every context except for satellite regions. BGISEQ500 outperformed HiSeq 2500, 4000, and X10, though MGISEQ2000 trailed behind. GenapSys had greater mismatch rates than all other short read platforms except for satellite regions. Notably, NovaSeq 2x250bp had a greater mismatch rate than 2x150bp chemistry.

Mismatches were also stratified by GC% content (Figure 3B) and base position per read (Figure 3C). All platforms showed elevated rates of substitution and insertion/deletion events in low (<25%) and high (>75%) GC contexts in the same manner as above, including PacBio, which otherwise had the lowest rate across GC%. GenapSys showed more INDEL mismatches than point substitutions. All short read platforms and CCS reads had increased error rates toward their 3' end, while Nanopore reads (here the Flongle, MinION R9.4, and PromethION R9.4 flow cells are combined) had flat (though high) rates across their reads.

Reads were then stratified against the UCSC Table Browser Simple Repeat Schema as defined by Tandem Repeat Finder [27]. Repeats were split into true homopolymers (stretches of poly-N in the reference genome) (Figure 3D) and other short tandem repeats (STRs), ordered by their entropy, a measurement of complexity of the STR motif (Figure 3E). Within both homopolymer and STR classes, PacBio CCS showed the lowest mismatch rate. Within short reads, BGISEQ500 and MGISEQ2000 performed better than Illumina instruments in shorter homopolymer stretches, while GenapSys performed worse, although surprisingly returning lower error rates with increasing homopolymer rates. All short reads returned roughly the same performance in homopolymer regions longer than 25bp. GenapSys reads were consistently more erroneous in STR regions. Though all platforms performed worse in longer homopolymer regions or areas of lower entropy, all Nanopore reads had a flat (though high) mismatch rate.

SNV and INDEL Detection

In addition to calculating error rates, mismatches were identified as variants against the human reference genome using benchmarking call sets. Variants, including short nucleotide polymorphism (SNP) and insertion/deletion (INDEL) events, were characterized against the Genome in a Bottle (GIAB) high confidence truth set (v4.1 [28]) for every replicate of the Ashkenazi Son (HG002) genome with adequate depth of coverage (minimum 10x), and each alignment normalized to a mean of 25x coverage. Note that replicates from GenapSys and the Flongle and R9.4 MinION flow cells (as two replicates from the HiSeq2500 platform) were

excluded here due to inadequate coverage.

Several common germline variant callers were compared across instruments, incluidng DeepVariant, GATK HaplotypeCaller, Sentieon Haplotyper, and strelka2 for short reads, as well as Clair2 for long reads (Figure 4A). BGISEQ500, MGISEQ2000, and NovaSeq 2x250bp had the highest precision and recall rates compared to other platforms, with HiSeq2500 and 4000 performing the worst. PacBio CCS reads called with Clair2 performed highly comparably to all short read data for global SNP and INDEL detection. DeepVariant consistently had the highest accuracy rates (except for a few HiSeq2500 replicates). Strelka2 was as precise as DeepVariant, but not as sensitive. Both GATK and Sentieon haplotype callers were less precise, and Sentieon was marginally less sensitive than GATK. Moving forward, all analyses were conducted using the DeepVariant call sets for short read data (normalized to mean 25x coverage for all samples).

Like coverage and mismatches before, the variants were stratified by UCSC RepeatMasker class to look for accuracy and reproducibility in difficult regions (Figure 4B). Sequencing instruments performed similarly against one another as in the global analysis. L1s, L2s, and LTRs were the "easiest" to capture, having the most accurate calls across instruments. Satellites and Alus were the second most "difficult" contexts, followed by low complexity regions and simple repeats as the least accurate across all technologies. Variants within the Askhenazi Son (HG002) genome in particular were harder to capture than the Father (HG003) and Mother (HG004) genome, although within satellites Mother variants were captured with less sensitivity than Father and Son.

Beyond measuring specificity and sensitivity, the total number of variants captured within each context was recorded, as well as the overlap between platforms, for SNPs (Figure 4C) and INDELs (Figure 4D). Within SNP regions, HiSeq2500 and Nanopore captured the fewest true positive variants. MGISEQ2000 and both NovaSeq chemistries captured the greatest number of true positive SNPs. Within INDELs, Nanopore failed to capture the majority of true positives across each context, followed by PacBio CCS, then HiSeq2500, 400, and X10. Again, MGISEQ2000 and both NovaSeq chemistries successfully captured the greatest number of true positive variant calls. Capture of true positive INDELs was also visualized by mutation size (Figure 4E). This showed a similar pattern, with Nanopore capturing the fewest sites. Interestingly, insertions showed a different pattern than deletions. Although NovaSeq and MGISEQ2000 captured the greatest number of large insertions, followed by other Illumina platforms and then BGISEQ500, there was more consistency between platforms to capture deletions, with every platform but Nanopore showing the same capture rate.

SNVs and short INDELs were also captured within genes from the CLINVAR [29] and Online Mendelian Inheritance in Man (OMIM) [30] databases as a measure of confidence in accessing variants in clinically relevant regions, stratified by high confidence regions for each cell line (Figure S5). The NovaSeq chemistries achieved the greatest accuracy in these medically relevant genes, while PacBio CCS achieved the highest

precision, with lowered sensitivity. Sequencing instruments were generally less able to detect variants in OMIM genes than in CLINVAR genes. To incorporate ThermoFisher targeted exome samples, variant call sets in genomic data were filtered to exomic regions and compared (Figure S6). Again here, short read platforms including NovaSeq and BGI/MGISEQ had the greatest sensitivity and precision in these regions, followed by other Illumina platforms, then PacBio. Proton and S5 replicates showed lower ability to accurately detect variants, with some S5 replicates falling below the accuracy provided by PromethION data.

Genomes were further compared to one another by aggregating and merging all calls across the entire trio, revealing strong clustering of replicates by cell line (Figure S7). Relatively little missing data was seen within short read and PacBio replicates, and much more frequent missing data within Nanopore data. Leveraging the trio relationship of these genomes, rates of Mendelian violations were calculated across SNPs, insertions, and deletions of varying sizes. All platforms showed some violations, with BGISEQ500, HiSeqX10, and NovaSeq 2x150 returning the most violations in SNP regions, and BGISEQ500 showing elevated violation rates in INDEL regions (Figure S8). These violations are mostly platform specific, tend to be less than 1% of all variants called, and are likely technical artifacts specific to each platform (Supplementary Table 5).

Structural Variant Detection

Creating a reference set: To enable a detailed analysis of structural variants (SV), a high quality reference SV set was constructed using three ONT and three PacBio CCS data sets (see methods). A high concordance SV set was identified across these long read-based calls (Figure S9) by requiring at least two call sets out of the six to agree on a SV [31]. This high confidence set is hereafter referred to as the HG002 Reference (or HG002 Ref) SV set.

Across all long read data sets, an average of 22,000 SVs were identified per sample, which matches the current expected number of genomic SVs [32]. Interestingly, a slight increase in SVs was observed within Nanopore (22,905) data sets compared to PacBio CCS (22,330) data (see Supp Figure x2), despite the fact that one Nanopore replicate showed a lowered number (21,591 SV).

Insertions and deletions that only overlapped with high confidence regions were investigated (see methods). Note that only replicates from HiSeq2500, HiSeq4000, and HiSeqX10 could be included in all analyses, as multiple replicats were required per instrument. The examined technologies showed a high concordance, with only 3.94% (442) SVs (26.47% deletions and 73.53% insertions) specific to PacBio CCS and 1.69% (190) SVs (63.16% deletions and 36.84% insertions) specific to Nanopore (Figure S10). This was again impacted by the one Nanopore sample that underperformed.

Capturing Structural Variants: An average of 12,435 SVs were detected across 32 short read HG002 sam-

ples. The majority (95.21%) of these SVs overlapped with the lifted over GIAB high confidence regions [25] (see methods). The SV calls followed the expected distribution in size and type, with the majority of events being deletions (7315), followed by translocations (3454), duplications (978), inversions (686) and finally insertions (2). Translocations were ignored as they are often false positives [33]. An average of 6965 SVs were captured that overlap with the filtered data set, 27.59% (1921) of which constitute deletions and insertions that overlap with the established reference set. Figure 5A shows the overall statistics among all data sets, as well as the distribution of SV calls per sample. No significant correlation was found between the total number of SVs and an increase in the average coverage or insert size (see Supplemental results). However, when restricting to true positives, a positive correlation was observed with coverage (mean: 27.06, cor: 0.56, p-value: 0.0008116, standard deviation: 51.60, cor:0.64, p-value: 7.435e-05), insert size (mean: 351.07, cor: 0.59, p-value: 0.0003852, standard deviation: 122.44, cor: 0.64, p-value: 6.996e-05) and read length(mean: 142.97, cor: 0.86, p-value:3.707e-10, standard deviation: 0 cor: NA, p-value: NA).

The different sequencing and analysis steps were analyzed one at a time in order to detect the cause of variability. This included stratifying results by SV callers, sequencing instruments, and by library replicates. Overall, the SV callers contributed the most to individual variability (527 SVs, 41.59%), followed by sequencers (237 SVs, 18.71%), and lastly by replicates (226 SVs, 17.84%). SV call sets overlapped the HG002 reference set for SV callers (82.54%), platforms (40.08%), and replicates (78.32%). Thus, interestingly, false negatives (i.e. calls missed by others) were predominantly observed, rather than the expected false positive. SV call sets did not show any clustering in a particular region of the genome and seemed to be distributed throughout (Figure 5E).

The majority of SV calls that are specific to Delly or Manta are in fact true positives. In parallel to this, it is evident that most false positives from SV caller variability are attributed to SV calls from Lumpy, followed by Delly and Manta (Figure 5B, Figure S12). Supplementary Table 6 summarizes the results for all strategies in terms of false positive, negative and true positive. Within platforms, HiSeqX10 has the largest number of SVs (3751), followed by HiSeq4000 (3714) and HiSeq2500 (3294). We observe that HiSeqX10 produces the largest number of unique false positive SVs (249) followed by HiSeq4000 (223) and HiSeq2500 (208) Interestingly, 14.43% (42) of unique HiSeqX10 SVs are false negatives compared to HiSeq4000 13.90% (36 SVs) and HiSeq2500 8.77% (20 SVs) (Figure 5C, Figure S13). Within replicates, 47.51% of unique replicate SVs are false positives that are not concordant with the HG002 reference SV set. Overall, 73.17% of non-unique SVs overlapped with HG002 reference set, indicating a smaller number of false positives and high concordance between the replicates (Figure 5E, Figure S14).

Bacterial Genome Capture

In addition to the relatively GC-balanced human genome, analysis of sequencer performance at high and low GC content genomes was evaluated. In addition to bacterial isolates, a metagenomic mixture of ten bacterial species was included in order to assess reproducibility of genomic sequencing with variable GC content, Gram stain, ecology, and physiology in a single sample. In particular for the metagenomic pool (ATCC MSA-3001 mix), taxonomic composition was found to be quite variable both within and between platforms (Figure 6A). Replicates within platforms were highly similar to one another, with the exception of the PGM, which had two outlier samples. Still, platform-specific compositions were detected (Figure 6B). Correlation of composition between instruments showed that Flongle and MinION R9.4 flow cells clustered closest to one another, and interestingly most closely to Illumina HiSeq. Also notably the GenapSys and PGM systems had a closer relationship than PGM to its ThermoFisher counterpart in the S5 system, which was most dissimilar to other platforms.

Irrespective of sequencer, taxonomic composition was clearly impacted by GC content of each taxon (Figure 6C). In particular, low-GC (*S. epidermidis* and *E. faecalis*) and high-GC taxa (*H. volcanii* and *M. luteus*) were underrepresented. These taxa were also Gram positive, showing that physiology had a direct impact on genomic sequencing. Taxa with middling GC contents and Gram-negative cell walls were overrepesented, in particular *P. fluorescens*, which averaged nearly double the representation expected from the equamolar mixture.

In addition to the metagenomic mixture, coverage of individual strains was highly consistent among all replicates from all instruments (Figure S11A). Coverage matched the expected GC range per taxon. Calculation of entropy across GC contexts showed the highest in the metagenomic mixture, followed by *E. coli*, then *S. epidermidis*, and finally *P. fluorescens* as the most consistently sequenced isolate (Figure S11B).

Discussion

The ABRF-NGS Phase II study is a comprehensive DNA-seq resource, providing a wealth of whole genome and exome sequencing data across multiple established and emerging instruments. This work adds to the data available for the well-characterized and publicly accessible human cell lines within RM 8392 that have become standard use cases for genomic technology research, as well as bacterial genomes that span a diversity of genome sizes and nucleotide compositions. Analyses of these data provide insight into the relative strengths and weaknesses of each instrument across genomic contexts, offering a valuable resource for benchmarking and experimental design. As expected, long read technologies are better suited to provide coverage in difficult regions of the genome. However, among short read platforms, Illumina HiSeqX10

and Illumina HiSeq4000 excel, and can perform as well as Nanopore and Pacbio Circular Consensus Sequencing (CCS) reads in most regions. Telomeric and centromeric regions are the most highly variable, with a subset of masked satellites poorly covered across all technologies. Oxford Nanopore provides the least variable coverage irrespective of genomic context. Beyond coverage, all platforms demonstrate increased mismatch rates at high and low GC content regions as well as toward the 3' end of reads, though CCS provides the highest nucleotide accuracy against the reference genome in all contexts. This is also true in homopolymer stretches, whereas all short reads show elevated error rates, and an expected increase in error as homopolymers get longer (though worth noting that GenapSys may be more reliable in longer homopolymer stretches). While considerably improved over time, Oxford Nanopore platforms still lag behind others in accuracy across all sequence compositions and genomic contexts, though it is worth noting that enough Nanopore data to achieve 25x mean genomic coverage may be much cheaper than the same cost for equivalent PacBio data. It is also worth noting that, for several instruments, only one replicate was available per cell line or at all (including GenapSys, Flongle and MinION flow cells, and NovaSeq 2x250bp reactions), which made it impossible to estimate intra-platform reproducibility. Additional data will be critical for future assessment of the performance of these platforms.

DeepVariant provided the highest sensitivity and specificity metrics against Genome in a Bottle (GIAB) v4.1 benchmark reference set. This machine learning-based variant caller was highly robust for all genomic contexts from all platforms. It is worth noting that deep learning tools are trained specifically on these single-ethnicity, B-lymphocyte-derived cell line genomes, which may lead to some overfitting to training samples and may perform differently in other use cases [34]. Strelka2 was generaly as precise as DeepVariant, while GATK HaplotypeCaller was generally as sensitive. Sentieon Haplotyper lags slightly behind, but is considerably faster to implement than other callers [35] and has performed comparably in the PrecisionFDA 2016 challenge [22]. It is also worth noting that Sentieon is an implementation of GATK, which makes it applicable to standard GATK variant-calling practices. Although these outcomes portray a current snapshot of variant accuracy, methods for both short and long read variant calling are under continual development and continue to improve beyond the results presented here, particularly in difficult regions, as seen in the recent precisionFDA Truth Challenge V2 [34].

Turning to the cell lines themselves, the Ashkenazi Son (HG002) provided the lowest precision and sensitivity across complex genomic features than the Father (HG003) and Mother (HG004), revealing underlying differences in complexity of each genome irrespective of platform. Sequencing platforms were also not the primary factor influencing structural variant (SV) detection, instead primarily driven by the SV caller which had the largest effect on detection of true positive events. These results highlight the need for continually improved methods to resolve disagreement beyond that of bias introduced by each platform.

The distribution of reads in a DNA-seq reaction was highly reproducible when sequencing an individual genome, including all three members of the Ashkenazi Trio as well as within bacterial strains. Across laboratories and platforms, error rates were consistent, including in repetitive and low complexity regions. In particular, emerging platforms from BGI, GenapSys, and Oxford Nanopore performed comparably to well-established platforms, providing promising results as the genomics landscape continues to grow and diversify. More complex metagenomic samples were less consistent, showing compositional bias and elevated variance of normalized coverage, indicating a challenge for future metagenomic studies. Notably, all platforms were able to identify all strains in each mix, and showed robustness in identifying the presence of each expected taxon within metagenomic samples. Mappability was also highly taxon-specific, with *S. epidermidis* mapping more poorly than all other individual bacterial strains, underlining the importance of high quality reference genomes for any alignment. Overrepresentation of negative Gram staining bacteria also points to DNA extraction as a critical factor for species distribtuion, even within a mock community standard, though it should be noted that a small sample of only ten species may lead to some randomness of representation. At the same time, the degree of variability within metagenomic sequencing remains a clear confounding variable that should be tracked and examined in future work, along with the other components of metagenomics analysis [36].

Building on the resources provided by GIAB, the Global Alliance for Genomic Health (GA4GH), and UCSC, the data made publicly available and results presented within this study provide a resource for benchmarking genomics data as well as an unbiased evaluation of current and emerging sequencing technologies. These findings can inform the evolution of new best practices in sequencing and analysis, serving as highly characterized reference material data designed to support a variety of genomic analyses and methods, which will be essential as new methods emerge.

Summary Box

1. **Mapping efficiency rates** are both platform-specific and species-specific. Illumina instruments are most comparable to one another. BGISEQ500 and GenapSys GS110 instruments return the lowest uniquely mapping rate and highest multi-mapping rate. BGI/MGISEQ libraries have the lowest duplicate read rate. PacBio CCS datasets have the highest rates of unique mapping and lowest non-mapping rate. Short fragments in Nanopore data bring down overall mapping efficiency.

2. **Alignments (BAM files)** can be normalized by calculating mean autosomal coverage using mosdepth and then downsampling using Picard DownsampleSam. However, even within normalized data, coverage

dramatically varies within repetitive and low complexity regions, even among replicates sequenced on the same instrument. Long read technologies provide the highest coverage within these genomic regions. For short read platforms, HiSeq 4000 and X10 provide the most consistent, highest coverage.

3. **Sequencing error** can be calculated with BBMap reformat.sh and comparing mismatch histogram tables. All instruments have some level of sequencing error ranging from 0.1% up to 20% in poorly defined satellite regions. BGI/MGISEQ provide the lowest sequencing error rates among short read technologies. PacBio Circular Consensus Sequencing provides the lowest error rate out of all technologies. Although the error rate is highest of all platforms, Nanopore technologies perform highly consistently from the smallest throughput (Flongle) flow cell to the largest (PromethION R9.4).

4. **Mismatch rates** are elevated in areas of high and low GC content, and by base to a lesser extent. Errors are more frequent in regions with larger repeat sizes of homopolymers and lower entropy of short tandem repeats, except for Nanopore which shows flat (though currently still high) error rates irrespective of sequence content. PacBio CCS has the lowest error rate in these contexts, while GenapSys has elevated STR error rates compared to other short read platforms.

5. **Variant calling** - DeepVariant is the most sensitive and precise software for calling known variants, though this software is trained on immortalized B-lymphocyte cell line data and may be overfitted. Strelka2 is as precise as DeepVariant, while GATK HaplotypeCaller is as sensitive. Sentieon Haplotyper is very nearly as sensitive as GATK HaplotyperCaller, while by far being the most computationally efficient. Default parameters may be used for each caller.

6. **Sensitivity and specificity** of variant detection can be assessed with RTG VcfEval. Among known variants, true positives in L1/L2/STR regions are recalled the most easily, while variants in simple repeats and low complexity are the hardest to capture. Read length makes an impact on the ability to call true positives, since data with shorter read lengths (HiSeq2500 2x125bp and BGISEQ500 2x100bp) capture the lowest proportion of true positives across RepeatMasker regions examined.

7. **The length of insertion/deletions (INDELs)** captured by each platform can be evaluated using RTG vcfstats with the –allele-lengths flag. INDEL detection is highly platform-specific, in particular for insertions (deletions are more comparable between platforms). Nanopore captures the lowest proportion, followed by BGISEQ500, Illumina HiSeq platforms, and then PacBio CCS. The NovaSeq 6000 using 2x250bp read

chemistry is the most robust instrument for capturing known INDELs.

8. **Structural Variant (SV)** calling consistency is most impacted by the variant caller used. This can be evaluated by calling SVs with Delly, Manta, and Lumpy, and then consolidating calls with SURVIVOR. Sequencing instrument is the second highest source of variability, followed by within-instrument replicates. The majority of unique SVs are likely due to sequencing artifacts and can be considered false negatives.

9. **A genome-wide distribution of roughly 20,000 SVs** is common with a given genome, which is slightly higher than previous estimates and benefits from longer reads. Within those, the majority (70%) will be called as deletions, followed by translocations (14%), insertions (6%), duplications (5%), and inversions (4%). No significant clustering of SVs is seen within the genomes examined in this study, indicating that overlapping SVs between replicates or instruments can be considered true positives rather than mapping artifacts.

10. **In mixed metagenomic samples**, the rate of mapping is significantly linked to the GC-content of the reference genome for each taxon. High- and low-GC content taxa tend to be underrepresented in reference-based alignment. This can be determined using mosdepth with the -F 3844 flag to assess the number of reads uniquely mapping to each genome within the mixed reference set.

Methods

Human Genomic DNA

DNA from cell lines derived from a family trio in the Personal Genome Project (PGP) are distributed as National Institutes for Standards and Technology (NIST) reference material RM 8392, which serves as source material for genomic DNA sequencing. These DNA samples were developed for the Genome in a Bottle (GIAB) Consortium to create a set of highly characterized standards for genomic analysis, and are approved for all research uses under the terms of the PGP. Standardized human genomic DNA samples were obtained from NIST, and whole genome sequencing (WGS) libraries were prepared at a single laboratory site (HudsonAlpha Institute for Biotechnology, Huntsville, AL), then distributed to individual laboratories for sequencing on respective instruments.

In a few cases, libraries were prepared at the facility where sequencing was done. All libraries were prepared using the same NIST stock and synthesis kits as at the central site above. This included both sets of NovaSeq 6000 data (one site preparing 2x150bp data and a second site providing both 2x150 and a novel

2x250bp reaction), two laboratories synthesizing and sequencing GenapSys GS110 data, one lab synthesizing and sequencing BGISEQ500 and MGISEQ2000 data, and two labs synthesizing and sequencing Oxford Nanopore data (one using PromethION R9.4 flow cells and the other utilizing Flongle and MinION R.94 flow cells). Oxford Nanopore PromethION R9.4 replicates were prepared using the PCR-free Ligation Sequencing Kit (LSK109). Libraries were prepared for the Flongle using the PCR-free Ligation Sequencing Kit (LSK109) and the native barcoding kit for the MinION R9.4 flow cell. Finally, all replicates of PacBio Circular Consensus Sequencing (CCS), as well as two Oxford Nanopore PromethION replicates, were downloaded from the public repository generated by the Genome in a Bottle (GIAB) Consortium and hosted by the National Center for Biotechnology Information (NCBI).

Bacterial Genomic DNA

Microbial reference gDNA was prepared from bacteria obtained from the American Type Culture Collection (ATCC-Manassas, VA). Pure agar cultures were grown to early log phase and harvested prior to gDNA extraction using the Omega Metagenomics DNA kit (Omega BioTek Norcross GA. M5633-00). Briefly, cell mass was resuspended in dPBS pH 7.5 and digested with Metapolyzyme (MAC4L Millipore Sigma, St. Louis, MO) for 8 hours before dual detergent lysis with CTAB and SDS, and farther lysis and clean up was done using phenol chloroform + isoamyl alcohol and RQ magnetic beads. DNA was evaluated using Qubit spectrofluorometry (Thermo Fisher, Waltham, MA), Agilent Bioanalyzer 2100 (Santa Clara, CA), RTqPCR (Applied Biosystems, Foster City, CA), and Nanodrop spectrophotometry (Thermo Fisher). Sequencing QC was performed using both Sanger sequencing of the entire 16s rDNA (Primer 27f and 1492r) as previously described (Innis et al. 2012). DNA for the 10 species combined mixtures was combined as an equimolar pool at approximately 10% each. This gDNA material is deposited at ATCC as product MSA-3001 and is publicly available.

Library Synthesis

Illumina: For human and bacterial samples, TruSeq PCR-free libraries were prepared according to manufacturer's protocols. The high molecular weight (HMW) genomic DNA from NIST was fragmented using an LE Series Covaris sonicator (Woburn MA) with a targeted average size of 350 bp. Libraries were then synthesized at HudsonAlpha Biotechnology Institute robotically using 1ug of DNA. Library quality was evaluated by Qubit quantification and Agilent Bioanalyzer 2100. After passing QC, libraries were shipped to different sites (core facilities) for sequencing.

ThermoFisher: For non-exomic libraries, each laboratory used the Ion Xpress Fragment Library kit (part 4471269) per the manufacturer's protocol, using 100ng of input DNA. For Ion Ampliseq exome sequencing,

DNA was amplified through a massively multiplexed PCR reaction to create the library following the Ion Ampliseq Exome protocol (kit 4489061).

All libraries were templated onto beads (Ion PI Hi-Q Template OT2 kit A26434 for Proton, Ion PGM Hi-Q OT2 200 Kit A27739 and S5 (part A27751 for bacterial libraries and A27753 for Exome libraries). The Exome libraries were sequenced on either the Ion S5 or Ion Proton instruments used standard 200bp chemistries and protocols (Proton kit A26771, S5 kit A27753). The bacterial libraries were sequenced on the Ion PGM or Ion S5 using 400bp chemistries (Ion PGM Hi-Q Seq Kit A25592 and Ion S5 kit A27751).

GenapSys GS110: Library synthesis was performed using a two step approach by first synthesizing a standard NGS library followed by a GenapSys clonal amplified library. 100 ng of microbial gDNA was fragmented using Covaris S2 instrument to a mean size of 250 bp and used as input to the NEBNext Ultra II kit (E7645 New England Biolabs Ipswich, MA) and checked for quality using the Agilent Bioanalyzer 2100 and Qubit spectrofluotometer. This NEBNext library was used as input to the version 1 chemistry of the fully manual GenapSys clonal amplification kit (1002000) which required 1.0 x 10ȇ molecules (33 pol) before hybridizing to the G3 electronic sequencing chip (1000737 GenapSys Redwood City, CA) and sequencing on the GS111 Genius Sequencing Platform.

Bacterial Nanopore Sequencing: Microbial gDNA was prepared for Nanopore sequencing using two library methods. For the Flongle flow cell runs, the direct ligation sequencing library kit (LSK-109 Oxford Nanopore UK) was used on individual bacteria and sequenced on dedicated flow cells. For the R9.4 flow cells runs, the individual bacteria strains as well as the 10 species mix was prepared using the LSK109 method with the native barcoding expansion kit (EXP-NBD104) and combined into one final library pool and sequenced together on a single flow cell. This ligation sequencing method is a non-PCR based library method that allows for direct sequencing of native DNA. Briefly, gDNA is "repaired" using the NEBNext FFPE DNA Repair reagents (M6630, New England Biolabs Ipswich Ma) followed by dA-tailed using the NEBNext End Repair/dA-tailing module, and ligated to nanopore specific sequencing adapters. Sequencing was performed immediately after library synthesis.

DNA Sequencing

TruSeq PCR-Free libraries were sequenced on the Illumina HiSeq 2500, HiSeq 4000, HiSeq X10, MiSeq, and NovaSeq 6000 with Xp loading. ThermoFisher kibraries were run separately on the PGM and were multiplexed on the Proton PI and S5 540 chips. Standard protocols were used for 400bp read lengths on the PGM and S5 520/530 chips. The bacterial libraries were run using 200bp reads on the Proton and S5 540

chip using standard protocols. The different read lengths were due to the availability of 400bp chemistry on the smaller chips for both PGM and S5 whereas the larger PI and 540 chips run 200bp chemistry. All libraries were run in triplicate. All libraries were synthesized using 1ug of DNA.

The exomes were run only on the Proton PI and S5 540 chips because of the read numbers requirements. Briefly, the exomes were amplified in a massively mulitplexed PCR reaction and the resulting libraries were sequenced per standard sequencing protocols. Samples were run in triplicate with two samples per chip to accommodate read numbers needed for analysis.

For GenapSys sequencing, successful clonal libraries were loaded onto the G3 electronic sequencing chip according to manufacture protocol (GS111 User Guide 1000698, Rev C Oct 2019) following an initial priming step including buffer washes. The electronic flow cell was injected with 35ul of the sequencing bead library followed by 40ul of a DNA polymerase solution. Sequencing was initiated on the GS111 Genius sequencer and run for 48 hours to achieve 15 million reads of single end 150 bp data.

Oxford Nanopore sequencing was performed using the PromethION for the human samples with standard use R9.4 flow cells. For bacterial genomes, the MinION MK1B sequencer was used with both Flongle and R9.4 flow cells. Flongles were injected with 20 fmol of each library on the with the slight modification of a 20% reduction of loading beads to increase Q-score performance. Sequencing was perfomed up to 48hrs. R9.4 flow cells were injected with 50 fmol of the pooled native barcoded library according to the manufactures example protocol (NBE_9065_v109_revJ_23May2018) and allowed to sequence for 72hrs.

Alignment and Variant Processing

Reference Genome: Whole genome human samples were aligned against GCA_000001405.15_GRCh38 retrieved from the NCBI FTP resource. This includes the GRCh38 primary assembly (including canonical chromosomes plus unlocalized and unplaced contigs), the rCRS mitochondrial sequence (AC:NC_012920), Human herpesvirus 4 type 1 (AC:NC_007605), and concatenated decoy sequences to improve variant calling.

Alignment: Short read Illumina datasets were aligned using bwa mem with default scoring parameters. INDEL realignment and base quality score recalibration was performed using the DNASeq workflow within Sentieon build 201808.0329 with default parameters. ThermoFisher datasets were aligned with Torrent Suite v5.10 tmap mapall (tmap mapall -f $reference -r $input -n 20 -v -u -o 1 stage1 map4). Nanopore datasets were aligned using minimap2 (v2.13-r850) [37]) with the –MD, -a, and -x ont flags. Aligned BAMs were sorted with sambamba (https://lomereiter.github.io/sambamba/) and optical duplicates (plus PCR duplicates for non-PCR-free libraries) were marked with Picard v2.10.10-SNAPSHOT. For bacterial data, all reads were aligned to genome builds of respective species derived from the NCBI Genome portal (Supplementary Table 4).

Base quality distributions, insert size distributions, and GC bias metrics were calculated using default values within Picard. Read mapping metrics, on-target mapping rates, species distributions in metagenomic mixtures, conversion of BAMs to FASTQs, BAM indexing, and BAM header alterations were performed using samtools v1.930. Depth of coverage per contig was calculated using mosdepth [38]) with the -n flag. BAMs were downsampled to a normalized 25x coverage using Picard, with the fraction to retain calculated based on mosdepth-inferred depth.

Variant Calling: Genomic germline variants were called using Sentieon Haplotyper [35], GATK HaplotypeCaller [39], Strelka2 [40], and DeepVariant [39], all using default parameters. ThermoFisher alignments had variants called using variant_caller_pipeline.py within tvc, using default parameters. For long reads, SNVs were called with Clair (v2) [41]) while structural variants were called using a multi-algorithmic approach (Delly [42]), Lumpy [43]) and Manta [44]), and validated with SURVIVOR [31]).

Variant Call Set Processing: VCF statistics were summarized using vcftools v0.1.1532, and merging was done with bcftools v1.633. Variant allele frequency marix generation was done with bcftools using the –012 flag. UpSet plots were generated with the UpSetR package [45]. Heatmaps with colored annotation tracks were created using the ComplexHeatmap R library [46]. Mismatch rates across GC content and base number were calculated using mhist tables generated by BBtools (https://sourceforge.net/projects/bbmap/). Mendelian Violations were estimated with VBT [47].

High confidence variants were analyzed using RTG vcfeval (https://github.com/RealTimeGenomics/rtg-tools) against the GIAB truth variant sets for each of the RM 8392 genomes (see Supplementary Methods for RTG vcfeval analysis of SNPs and INDELs). Conversion of VCF data to allele frequency matrices, extraction of mapping/mismatch/variant statistics, generating UpSet matrices, and homopolymer detection and SNP/indel assignment were all performed using Python 3.7.0 scripts, and all visualizations were performed using R 3.6.3.

All custom scripts and R markdown notebooks are available at `https://www.github.com/jfoox/abrfngs2`.

Acknowledgments

We thank Illumina and ThermoFisher for providing reagents allowing the study to take place. We also thank NIST for providing the Genome in a Bottle DNA samples necessary to carry out the study. We acknowledge the HudsonAlpha Institute of Biotechnology for expert assistance in Illumina DNA library prepa-

ration. The Association of Biomolecular Resource Facilities also provided funding, logistical support and project oversight. We are particularly grateful for the assistance provided by multiple core facilities spending their own time and resources in order to participate in this research. We would like to thank the Epigenomics Core Facility and Scientific Computing Unit at Weill Cornell Medicine, as well as the Starr Cancer Consortium (I9-A9-071) and funding from the Irma T. Hirschl and Monique Weill-Caulier Charitable Trusts, Bert L and N Kuggie Vallee Foundation, the WorldQuant Foundation, The Pershing Square Sohn Cancer Research Alliance, NASA (NNX14AH50G, NNX17AB26G), the National Institutes of Health (R25EB020393, R01NS076465, R01AI125416, R01ES021006, 1R21AI129851, 1R01MH117406), the Bill and Melinda Gates Foundation (OPP1151054), TRISH (NNX16AO69A:0107, NNX16AO69A:0061), the Leukemia and Lymphoma Society (LLS) grants (LLS 9238-16, Mak, LLS-MCL-982, Chen-Kiang) the Alfred P. Sloan Foundation (G-2015-13964). Certain commercial equipment, instruments, or materials are identified to adequately specify experimental conditions or reported results. Such identification does not imply recommendation or endorsement by the National Institute of Standards, nor does it imply that the equipment, instruments, or materials identified are necessarily the best available for the purpose. FJS and MM are supported by NIH (UM1 HG008898).

Author Contributions

C.E.M., S.W.T., C.M.N, D.A.B. conceived and designed the study. C.E.M., S.W.T., Z.H., W.F., G.S.G., S.L., P.L., D.W., X.Z, W.Z, F.T, Y.Z, and H.L implemented the protocols. J.M.Z., W.E.C, M.B.B, and G.N assisted with analysis design. J.F. aggregated and processed data, led data analysis and figure generation, and wrote the manuscript. W.E.C, M.B.B., G.N., M.M.K, M.M., S.T performed data analysis, figure generation, and manuscript editing. The ABRF-NGS Study members contributed to the design and execution of the project.

Competing Financial Interests

G.P.S is employed by Illumina Inc. X.Z, W.Z, F.T, Y.Z, and H.L are employees of MGI Inc. All other authors declare no competing financial interests.

Data Availability

The genome sequences in this study are available as EBV-immortalized B-lymphocyte cell lines (from Coriell) as well as from DNA (from Coriell and NIST). All data generated within this study from these genomes are publicly available on NCBI Sequence Read Archive (SRA) under the BioProject PRJNA646948, within

accessions SRR12898279-12898354. All code used within this study is publicly available at https://www.github.com/jfoox/abrfngs2. This repository includes scripts to run heavy lifting such as alignment and variant calling (SLURM), shell scripts to do post-processing calculations (bin), and R scripts used to create figures (Rmds).

References

1. Schuster, S. C. Next-generation sequencing transforms today's biology. *Nature methods* **5,** 16–18 (2008).
2. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nature biotechnology* **26,** 1135 (2008).
3. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43,** 491 (2011).
4. Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends in genetics* **24,** 133–141 (2008).
5. MacLean, D., Jones, J. D. & Studholme, D. J. Application of'next-generation'sequencing technologies to microbial genetics. *Nature Reviews Microbiology* **7,** 96–97 (2009).
6. Glenn, T. C. Field guide to next-generation DNA sequencers. *Molecular ecology resources* **11,** 759–769 (2011).
7. Aziz, N. *et al.* College of American Pathologists' laboratory standards for next-generation sequencing clinical tests. *Archives of Pathology and Laboratory Medicine* **139,** 481–493 (2015).
8. Schlaberg, R. *et al.* Validation of metagenomic next-generation sequencing tests for universal pathogen detection. *Archives of Pathology and Laboratory Medicine* **141,** 776–786 (2017).
9. Zhou, J. *et al.* Reproducibility and quantitation of amplicon sequencing-based detection. *The ISME journal* **5,** 1303–1313 (2011).
10. Mellmann, A. *et al.* High interlaboratory reproducibility and accuracy of next-generation-sequencing-based bacterial genotyping in a ring trial. *Journal of clinical microbiology* **55,** 908–913 (2017).
11. Quail, M. A. *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics* **13,** 341 (2012).
12. Shi, L. *et al.* The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature biotechnology* **24,** 1151 (2006).
13. Shi, L. *et al.* The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature biotechnology* **28,** 827 (2010).
14. Li, S. *et al.* Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nature Biotechnology* **32,** 915 (2014).
15. Su, Z. *et al.* A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature biotechnology* **32,** 903 (2014).
16. Wang, C. *et al.* The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nature biotechnology* **32,** 926 (2014).
17. Li, S. *et al.* Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nature biotechnology* **32,** 888 (2014).
18. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature biotechnology* **32,** 896 (2014).

19. Merker, J. D. *et al.* Proficiency testing of standardized samples shows very high interlaboratory agreement for clinical next-generation sequencing–based oncology assays. *Archives of pathology & laboratory medicine* **143,** 463–471 (2019).
20. Mahamdallie, S. *et al.* The ICR639 CPG NGS validation series: A resource to assess analytical sensitivity of cancer predisposition gene testing. *Wellcome open research* **3** (2018).
21. Zhong, Q. *et al.* Multi-laboratory proficiency testing of clinical cancer genomic profiling by next-generation sequencing. *Pathology-Research and Practice* **214,** 957–963 (2018).
22. Zook, J. M. *et al.* An open resource for accurately benchmarking small variant and reference calls. *Nature biotechnology* **37,** 561–566 (2019).
23. Krusche, P. *et al.* Best practices for benchmarking germline small-variant calls in human genomes. *Nature biotechnology* **37,** 555–560 (2019).
24. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific data* **3,** 1–26 (2016).
25. Zook, J. M. *et al.* A robust benchmark for detection of germline large deletions and insertions. *Nature biotechnology* (2020).
26. Ball, M. P. *et al.* A public resource facilitating clinical use of genomes. *Proceedings of the National Academy of Sciences* **109,** 11920–11927 (2012).
27. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **27,** 573–580 (1999).
28. Wagner, J. *et al.* Benchmarking challenging small variants with linked and long reads. *BioRxiv* (2020).
29. Landrum, M. J. & Kattman, B. L. ClinVar at five years: delivering on the promise. *Human mutation* **39,** 1623–1630 (2018).
30. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM. org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic acids research* **43,** D789–D798 (2015).
31. Jeffares, D. C. *et al.* Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature communications* **8,** 1–11 (2017).
32. Mahmoud, M. *et al.* Structural variant calling: the long and the short of it. *Genome biology* **20,** 246 (2019).
33. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nature methods* **15,** 461–468 (2018).
34. Olson, N. D. *et al.* precisionFDA Truth Challenge V2: Calling variants from short-and long-reads in difficult-to-map regions. *bioRxiv* (2020).
35. Freed, D. N., Aldana, R., Weber, J. A. & Edwards, J. S. The Sentieon Genomics Tools-A fast and accurate solution to variant calling from next-generation sequence data. *BioRxiv,* 115717 (2017).
36. McIntyre, A. B. *et al.* Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome biology* **18,** 182 (2017).
37. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34,** 3094–3100 (2018).
38. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34,** 867–868 (2018).
39. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv,* 201178 (2018).
40. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nature methods* **15,** 591–594 (2018).
41. Luo, R. *et al.* Exploring the limit of using a deep neural network on pileup data for germline variant calling. *Nature Machine Intelligence* **2,** 220–227 (2020).
42. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28,** i333–i339 (2012).

43. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome biology* **15,** R84 (2014).

44. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32,** 1220–1222 (2016).

45. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33,** 2938–2940 (2017).

46. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32,** 2847–2849 (2016).

47. Toptaş, B. Ç., Rakocevic, G., Kómár, P. & Kural, D. Comparing complex variants in family trios. *Bioinformatics* **34,** 4241–4247 (2018).

48. Cleary, J. G. *et al.* Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. *BioRxiv,* 023754 (2015).

49. Chaisson, M. J. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature communications* **10,** 1–16 (2019).

50. Ho, S. S., Urban, A. E. & Mills, R. E. Structural variation in the sequencing era. *Nature Reviews Genetics,* 1–19 (2019).

Figures

Figure 1: Experimental design and mapping results. (a) Three standard human genomic DNA samples from the NIST Reference Material 8392 were used to prepare libraries, including TruSeq PCR-Free whole genome libraries and AmpliSeq exome libraries, for sequencing on an array of platforms. Three bacterial species (*E. coli*, *S. epidermidis*, *P. fluorescens*) and one metagenomic mixture of ten bacterial species (Metagenomic Pool) were also sequenced. (b) Mean depth of coveage of replicate, colored by platform, and stratified by sample type. Depth is calculated by dividing total bases sequenced by size of respective genome. Squares indicate Father replicates, circles indicate Mother replicates, triangles indicat Son replicates. (c) Mapping rate for every replicate for each instrument, including uniquely mapped reads, reads that mapped to multiple places in the genome, reads marked as duplicates, and reads that did not map. (d) The same as (c), but for bacterial species sequenced, colored by sequencing platform.

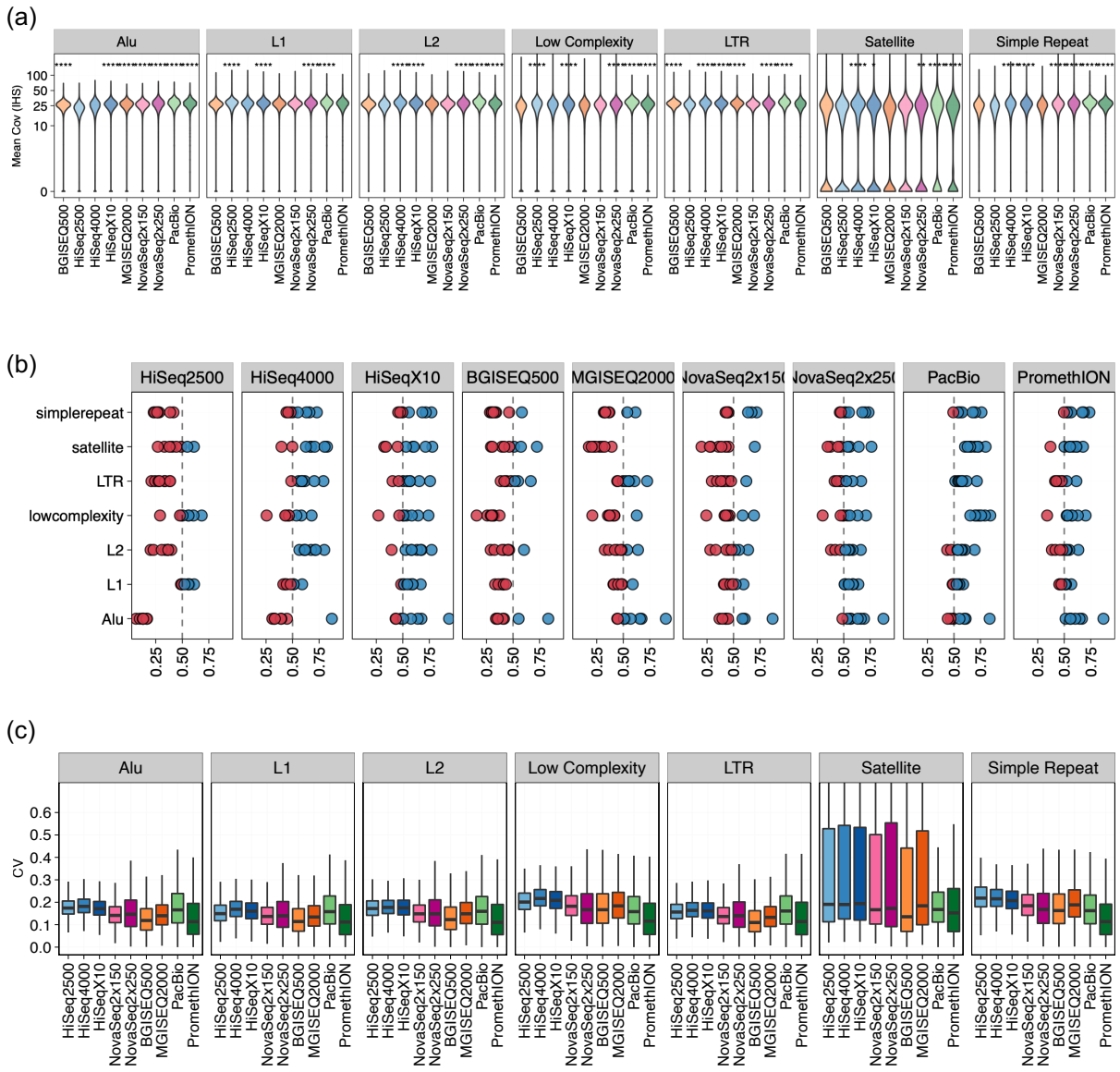

Figure 2: Distribution of genomic coverage across sequencing technologies for all replicates. (a) Aligned BAMs were downsampled to 25x mean read depth, and the distribution of coverage of each locus in the UCSC RepeatMask regions was plotted. Asterisks indicate significantly higher coverage for a given platform compared to the global mean, as measured by a one-tailed Wilcoxon test. (b) Comparison of each platform against all other platforms in each UCSC RepeatMasker context. Blue dots indicate >50% of shared sites are better represented in a given platform versus some other platform. Red dots indicate that the other platform out-covered the given platform. (c) Coefficient of variation of coverage per platform per UCSC RepeatMasker region. Coverage was calculated for all bases within a region and variation was calculated among all replicates per platform.

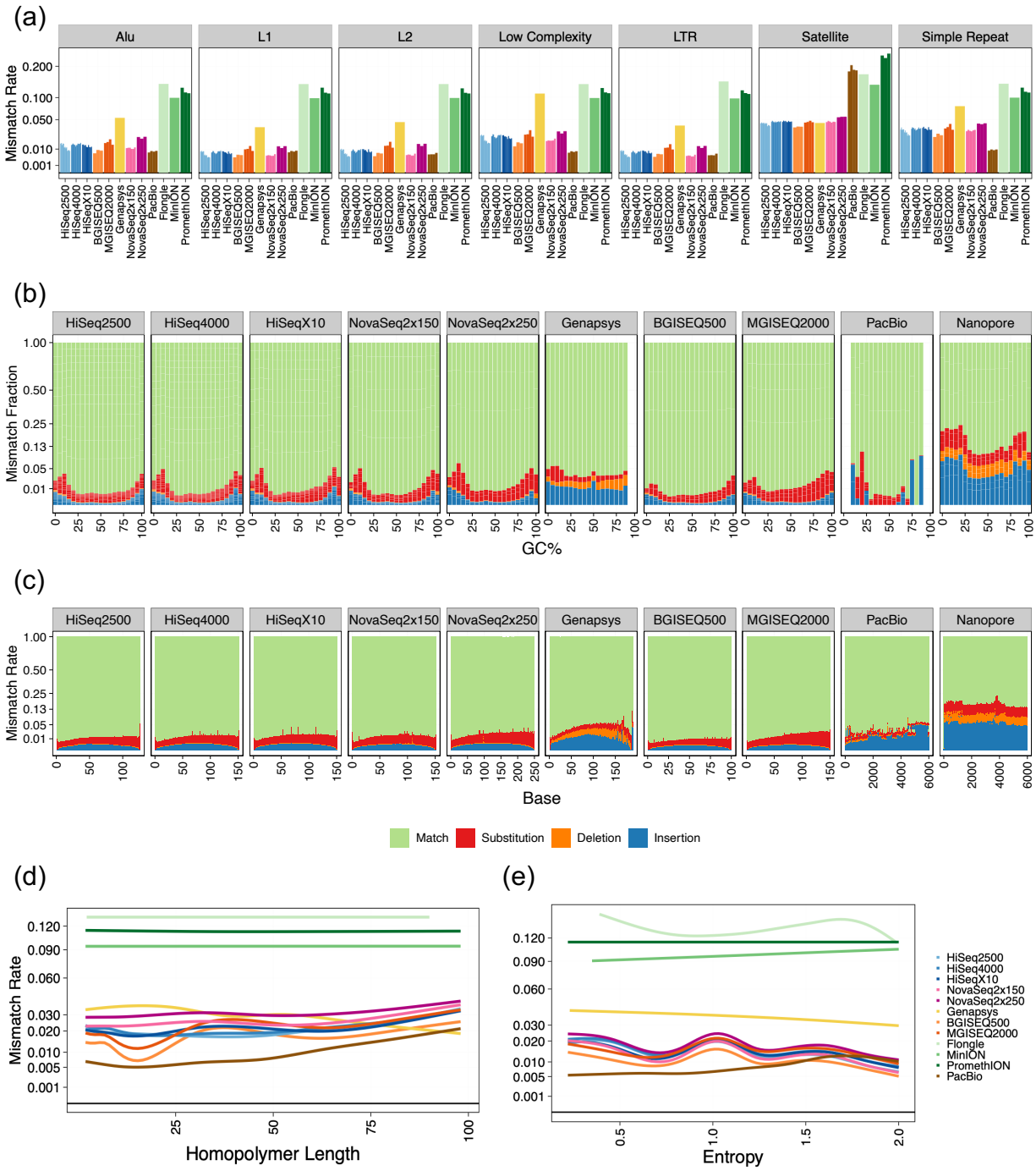

Figure 3: Estimating rates of sequencing error per platform. (a) Bar plot showing total average error rate within each UCSC RepeatMasker context. Individual replicates per platform are shown as separate bars. Values are averaged across all bases covering a given context. Y-axis is plotted as square root transformed. (b,c) Proprortional mismatch rates across GC windows and base number. Values at each window are averaged across all reads from all replicates. For long read platforms, read length is capped at 6kbp. Y-axis is plotted as square root transformed. (d,e) Error rate in homopolymer (n=72,687) and short tandem repeat (n=928,143) (STR) regions, respetively. On the left plot, true homopolymers are shown at increasing copy number. On the right, STRs are plotted by entropy, a measure of complexity of the motif. Y-axis is plotted as square root transformed.

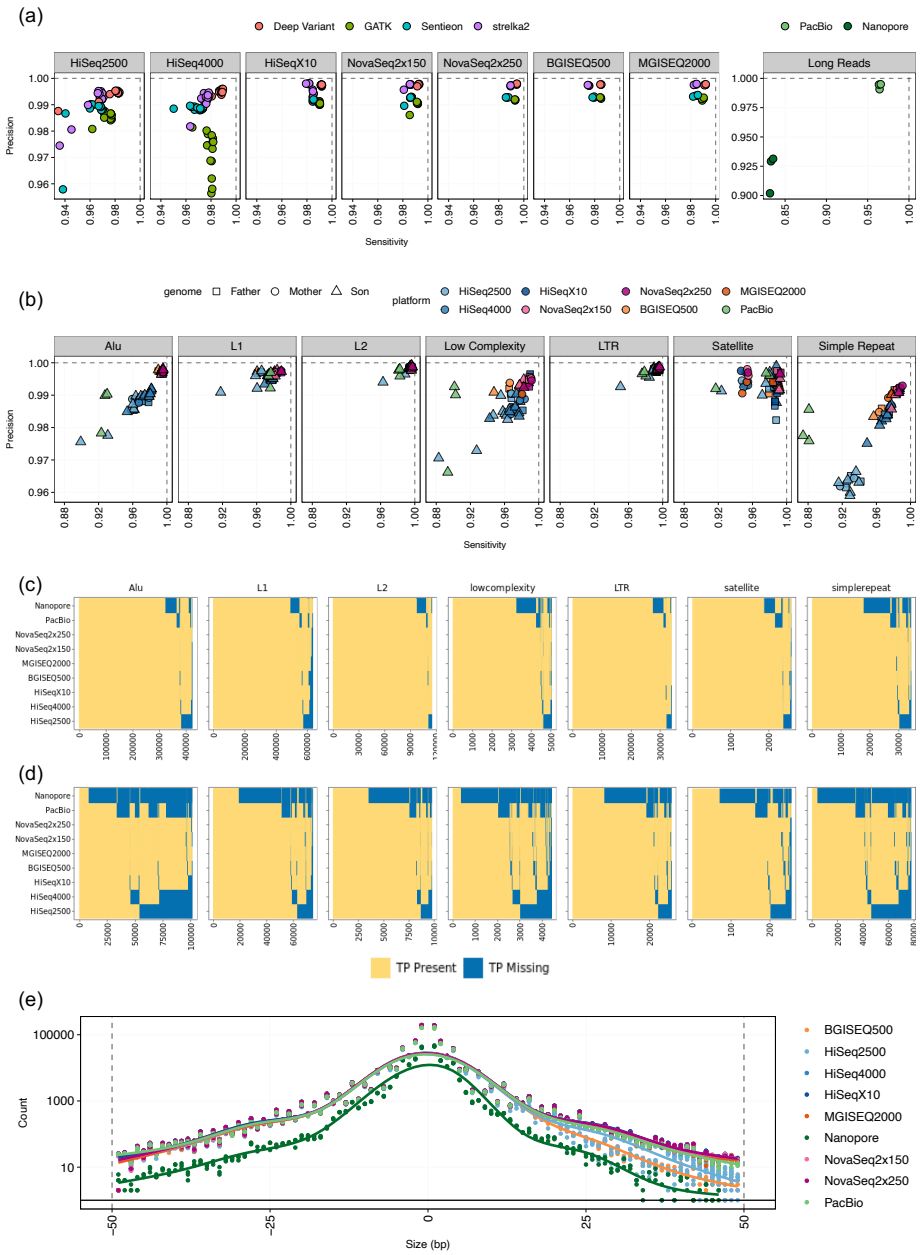

Figure 4: Validating short nucleotide polymorphisms (SNPs) and insertion/deletion (INDEL) events from short read datasets against the Genome in a Bottle (GIAB) high confidence truth set as determined by RTG vcfeval. (a) Common germline haplotype variant callers were compared for each sequencing platform across the entire genome, showing sensitivity and specificty achieved by each, for every replicate. (b) Overall sensitivity and specificity plotted for variants in each UCSC RepeatMasker region, overlapped with high confidence regions for each cell line respectively. (c) Presence matrix of true positive SNP variants within each UCSC RepeatMasker region. Each column is one variant. A yellow value indicates that the majority of replicates for that platform captured that variant, whereas blue indicates that variant was missed. (d) Same as above, but for INDELs. (e) Distribution of sizes of INDELs capture per sequencing platform. Values below zero on the x-axis indicate deletions; values to the right indicate insertions. Number of true positive INDELs is plotted per mutation size and colored by platform.

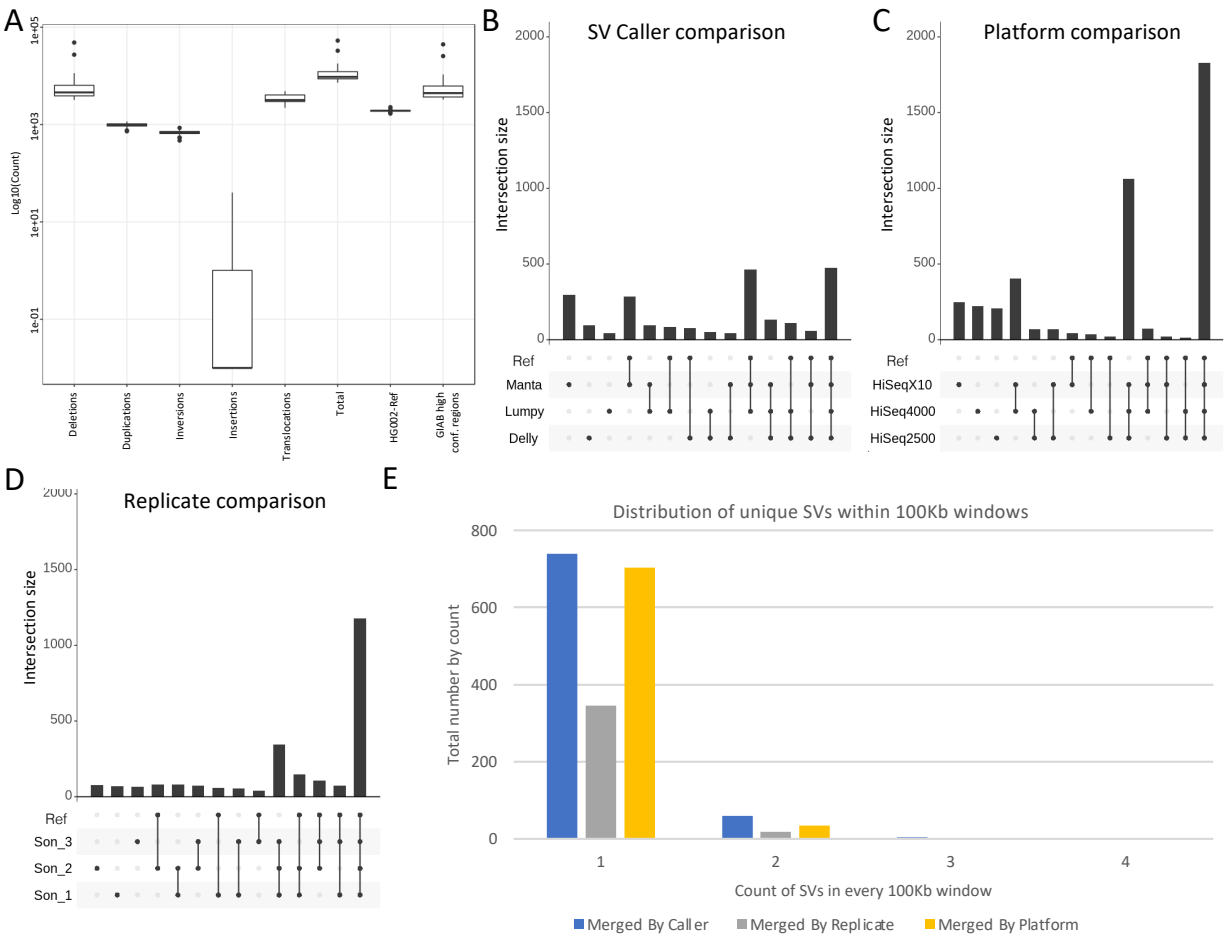

Figure 5: Assessing variability for Son (HG002) across HiSeqX10, HiSeq2000 and HiSeq4000, platforms which had more than one replicate per cell line to enable this analysis. (a) Number of SVs across sequencing reactions for HG002 replicates including deletions, duplications, inversions, insertions, translocations, total, SVs overlapping with the HG002 reference set, and SVs overlapping with GIAB high confidence regions. Variability is shown that can be attributed to callers (b), platforms (c), and replicates (d). (e) The distribution of single support (unique) SVs in 100kb windows across the different stratifications strategies.

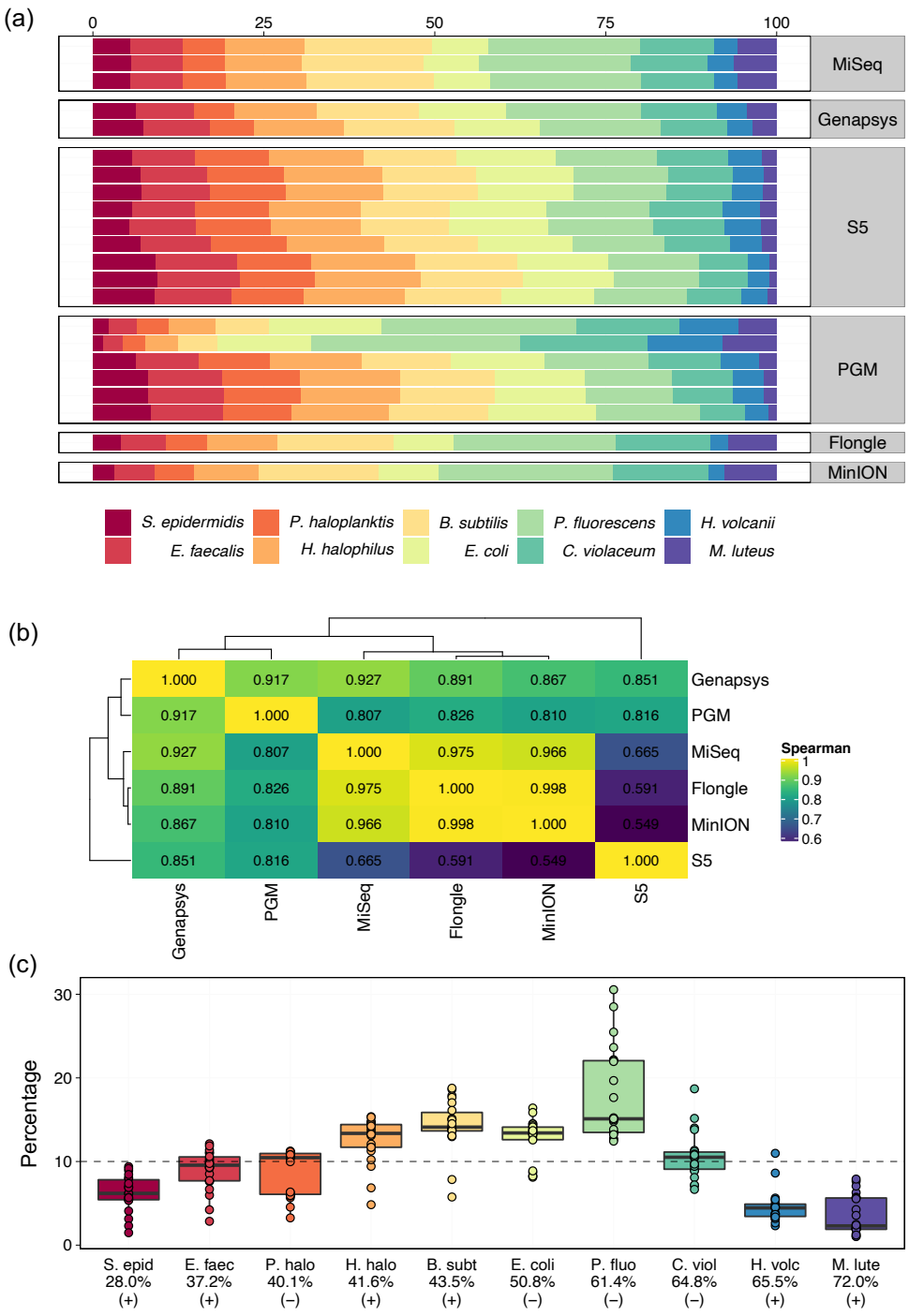

Figure 6: Reproducibility of sequencing of bacterial genomes in a complex metagenomic mixture. (a) Distribution of taxonomic assignment of strains present in the metagenomic mixture, stratified per replicate per sequencing platform. (b) Heatmap showing the spearman correlation of the average coverage within all instruments of each strain in the mixture. (c) Distribution of presence of each taxon across all replicates from each sequencing insrument. The taxa are ordered by GC content and have their Gram-stain status indicated.

Supplementary Methods

SNP and INDEL Analysis

Sensitivity vs. precision analyses were performed using the short nucleotide variant (SNV) and insertion / deletion (INDEL) call sets generated by each of the sequencing platforms (based on downsampled BAMs; see above) for the HG002 sample. Each replicate was compared against the GIAB SNV/INDEL HG002 truth set (v4) [22] using the Real Time Genomics (RTG) vcfeval tool [48]. True positive, false positive, and false negative calls were identified within the high confidence regions of the genome (as defined by GIAB) using Genotype Quality (GQ) as a Receiver-Operator Curve (ROC) score, stratified by variant type. For long read datasets (PacBioCCS and PromethION) processed through the Clair pipeline, we only report sensitivity and precision value based on all SNVs/INDELs (equivalent of GQ >=0) because Clair does not output normalized Phred-scaled GQ scores.

To compare genome-wide INDEL size distribution across sequencing platforms and replicates of sample HG002, we used the RTG vcfstats tool with the option –allele-length, which outputs a histogram of variant length for a given VCF file. Vcfstats increments counts for each called allele, therefore a heterozygous call increases a count of an appropriate size bin by 1, while a homozygous alternate call increases a count by 2. To ensure that differences in distribution of INDEL sizes across platforms are not driven by the differences in mean coverage, we used INDEL call sets generated using alignment files downsampled to mean coverage of 25X.

In addition to comparing the distribution of INDEL sizes genome-wide (i.e. including all INDELs called by the SNV/INDEL Sentieon or Clair pipeline), we also restricted the analysis to high confidence genomic regions, as well as high confidence true positive INDEL calls, as defined by the GIAB the SNV/INDEL truth set (v4.1) for sample HG002 [22, 25]. True positive calls made by each platform were again identified using the RTG vcfeval tool.

To facilitate a more detailed comparison, we also generated genome-wide, as well as high confidence true positive pairwise comparisons of INDEL size distributions for all possible pairs of sequencing platforms and replicates of the HG002 sample, stratified by shared and unique INDEL calls. Shared and unique INDEL calls for each pair of datasets were identified using the RTG vcfeval tool by treating one of the datasets as a truth set and the other as an evaluation set. High confidence true positive subsets were identified using the merged GIAB SNV/INDEL and SV truth set, as described above.

To compare numbers of SNV and INDEL calls that fall within different classes of repetitive and low complexity regions of the genome across all sequencing platforms we first restricted the analysis to the HG002 true positives that match the GIAB high confidence calls. We then annotated the true positive SNV and INDEL

calls from each platform using the UCSC RepeatMask BED files.

Structural Variant Calling

Defining High Confidence Regions: NCBI Genome Remapping Service was used to re-map the GIAB v0.6 high confidence regions (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/NIST_SVs_-Integration_v0.6/HG002_SVs_Tier1_v0.6.bed) from GRCh37 to GRCh38. Subsequently, bedtools intersect was used to filter all SV call sets with the aforementioned regions before they were compared across family members and strategies.

Illumina SV calling: The aligned reads (see above) were analyzed using Delly (v0.8.2) [42], Lumpy (v0.2.13) [43], and Manta (v1.4.0) [44], each with default parameters. The generated SV call sets per sample were merged using SURVIVOR (v1.0.7) [31] using the following parameter: 1000 2 1 0 0 0; requiring a maximum of 1kbp allowance on the start or stop breakpoint; requiring that the SV to be merged have the same SV type, and that at least two out of the 3 caller need to agree on a SV to keep it. Overlaps with GIAB high confidence regions (v.0.6) were captured using bedtools (v2.29.2).

Long read processing: Princess (https://github.com/MeHelmy/princess) was used to analyze samples with default parameters agains the human genome (GRCh38). SVs are identified with a minimum of two reads of support. Using bcftools view (version 1.9), SV were subsequently filtered for a minimum size of 50bp and a minimum read support of 5. Subsequently, SVs were further filtered using bedtools intersect (version 2.29.2) for their overlap with the SV GIAB regions lifted over from GRCh37 to GRCh38. SURVIVOR version 1.0.7 with a maximum of 1kbp between SVs was used to merge and compare these six data sets.

Supplementary Results

Structural Variant Disribution

On average, 12,602 structural variants (SVs) were detected per sequencing run for HG002. The SVs were considered passing if they were concordant between two or more SV call sets produced using Delly, Lumpy and Manta SV callers. The SURVIVOR [31] merge function was used to evaluate such concordance between the three different SV call sets. SVs were merged if their start and end breakpoints were less than 1kbp apart and had the same SV type. Subsequently, the SVs were filtered to be at least 50bp long following the most common definition of SVs [32, 49, 50].

Supplementary Table 6 gives an overview of the SVs identified in HG002. When filtered using the GIAB high confidence regions, 6964.9375 SVs overlapped on average across the samples. Out of these the vast majority (7315.25 SVs) are deletions followed by 3453.75 translocations. The latter were ignored for the subsequent analysis since they are often associated with false positives5. On average we identified 1921.4062 SVs per sequencing reaction that overlapped with the GIAB consortium released HG002 SV call set (v0.6).

Structural Variant Variability Stratification

Sources of Variability: Supplementary Table 7 shows the identifiable SVs across all sequenced samples and the variability of SV call sets between different samples. Not surprisingly, the variability is impacted by multiple factors such as SVs caller, sequencing platforms, biological and replicates (see Supplementary Table 8). Subsequently, the variability observed among the data sets was analyzed in greater detail to highlight the false positive (i.e. artificial SV calls) vs. false negative (i.e. often missed SV calls) rates per variability. For this, the GIAB high confidence regions and SV call sets [25] was used.

Variability Due to SV Callers: For HG002, the variability attributed to SV callers was estimated while stratifying for variability from platforms (HiSeqX10, HiSeq2000 and HiSeq4000), replicates, and centers. To stratify for other variabilities, calls were merged across platforms, replicates, and then centers, requiring that SVs be concordant at each step between the SV call sets respectively. The stratification strategy is seen in Figure S12A. SURVIVOR was used for each step, requiring SVs to be of the same type, have a pairwise overlap smaller than 1kbp and the length of the SV larger than 30bp. In the last step, a union merge was performed using SURVIVOR across the different SV callers (Manta, Lumpy, Delly), requiring SVs to be larger than 50bp. The SV call set was then filtered using the GIAB high confidence regions and compared to the SV calls from GIAB (Figure S12B). After stratification for other sources of variability and filtering, the resulting set had a total of 2303 SVs that showed variability due to the SV callers. The majority of these SV calls were deletions

(72.12%), most of which were between 100-1000bp in size (Figure S12C). Other SVs were spuriously called, including translocations (13.59%), insertions (4.95%), duplications (5.17%), and inversions (4.17%). The SV call set was further filtered to only include deletions and insertions, for a total of 1,775 SVs. This made the SV call set comparable to HG002 reference SV set and eased the interpretation of the majority of SVs observed to be variable due to the SV caller.

Next, the overlap with GIAB SV calls was investigated to identify potential false positives (i.e. identified by a SV caller but absent in HG002 reference SV set) vs. false negatives (i.e. missed by one or more SV callers but present in HG002 reference SV set). A total of 527 (29.69%) insertions and deletions were observed that were unique to a single SV caller. Interestingly, the majority of these unique SVs in the short read SV call set (82.54%) overlapped with the HG002 reference SV set. This indicates that the majority of unique SVs were false negatives that were missed in the SV call sets of different SV callers. Additionally, 17.46% of the unique SV call set did not overlap with HG002 reference SV set and were likely sequencing artifacts (i.e. false positives). Specifically, Manta (582) was observed to have the most unique SV calls, followed by Delly (171) and Lumpy (128). Of the SV call sets, 44.44% (unique by Delly), 66.41% (unique by Lumpy), and 48.97% (unique by Manta) of these overlapped with the HG002 reference SV set. Thus, it is interesting to note that a large fraction of SV calls that are specific to Delly or Manta are in fact true positives. In parallel to this, it is evident that most false positives from SV caller variability are attributed to SV calls from Lumpy, followed by Delly and Manta. Further, most non-unique SVs called were found to be concordant with the GIAB SVs call set. On average, these are 77.55% concordant with HG002 reference SV set compared to unique caller SVs (55.62%).

Variability Due to Sequencing Platforms: For HG002, the variability attributed to different short read sequencing platforms (HiSeqX10, HiSeq2000 and HiSeq4000) was investigated while stratifying for variability from SV callers, replicates, and sequencing centers. SV call sets were merged across SV callers per sequencing run, requiring agreement of two or more SV callers for an SV to pass (Figure S13A). Stratification was then performed by center and replicate variability by merging sequentially, requiring SVs to be concordant in each merge respectively. SURVIVOR was used for each step, requiring SVs to be of the same type, have pairwise overlap smaller than 1kbp and be > 30 bp in size. A final SV call set was then created using a union merge across platforms followed by GIAB high confidence regions filter.

A total of 4319 SVs was observed in the GIAB filtered SV call set, including 2578 deletions, 271 duplications, 314 inversions, 0 insertion and 1156 translocations. For this SV call set, the HiSeqX10 had the highest number of SVs (3751 SVs), followed by HiSeq4000 (3714 SVs), and HiSeq2500 (3294 SVs.) Tthe overlap between the sequencing platforms and GIAB SV call set can be seen in Figure S12B, while the sizes and types

of SVs that were observed can be seen in Figure S12C. Across platforms, 54.07% of HiSeqX10 (52.41%), HiSeq4000 (52.61%), and HiSeq2500 (57.19%) SV call sets overlapped with HG002 reference SV set. Most non-unique SV calls (54.70%) overlapped the HG002 reference SV set. Thus, 45.30% of non-unique SVs were false positives in the SV call set, while the remaining majority are true positives or false negatives missed by a platform's SV call set. HiSeqX10 was seen to produce the largest number of unique false positive SVs (249), followed by HiSeq4000 (223), and HiSeq2500 (208). Interestingly, 14.43% (42) of unique HiSeqX10 SVs were false negatives compared to HiSeq4000 13.90% (36) and HiSeq2500 8.77% (20).

The SV call set was further filtered to include only insertions and deletions overlapping wiht the GIAB SV call set. The majority of SVs in the filtered set were deletions sized between 100-1000bp, which is comparable to other studies including GIAB. A total of 237 insertions and deletions were observed that were supported by a single platform, 95 (40.08%) of which overlap the HG002 reference SV set. This is consistent with the SVs being false negatives missed by the two other platforms respectively. 59.92% of SV calls do not overlap HG002 reference SV set and were thus false positives due to individual platforms.

In the analysis so far, only platforms that included replicates and were used by all centers were considered. The overlap of all the sequencing platforms with each other can be seen in Figure S12D. An increase in translocations and overall variability was observed, likely due to the relaxed filtering strategy (Figure S12E).

Variability Among Sequencing Replicates: The variability among replicates for HG002 was investigated (using HiSeqX10, HiSeq2000 and HiSeq4000 SV call sets) after stratifying for variability from SV callers, platforms and centers (Figure S14A). SV call sets were filtered across SV callers per sequencing run, requiring agreement of greater than two for an SV to pass. Subsequently, these SV sets were merged across the three platforms, requiring the SVs to be concordant to stratify for platform variability. The SV call sets were then merged across centers, requiring an overlap between all three to stratify for center variability. A union merge using SURVIVOR was used across resulting replicates SV call sets with 50bp as the cutoff for SV size while maintaining previously described parameters. The SV call set was then filtered using the GIAB high confidence regions and overlapped with GIAB SV call set (Figure S14B).

A total of 2435 SVs were identified, including 1975 deletions, 107 duplications, 127 inversions, 0 insertions, and 226 translocations (Figure S14C).The SV call sets were filtered to only include insertions and deletions (1975 SVs) for comparisons with HG002 reference SV set. A total of 226 SVs were supported by a single replicate, of which 177 SVs (78.32%) overlap with the HG002 reference SV set. The majority of the variability observed is thus due to false negatives in other replicate SV call sets compared to the individual one that captured a variant. 21.68% of unique replicate SVs were designated false positives that were not concordant with HG002 reference SV set. Overall, 73.17% of non-unique SVs overlapped with HG002 reference SV set,

indicating a smaller number of false positives and high concordance between the replicates.

Structural Variant Clustering across the Genome

A subsequent analysis was performed to look for an enrichment of outlier SVs in certain regions. For each stratification strategy, only the unique SVs that were supported only by one SV caller (blue), one replicate (gray) or one sequencing platform (yellow) were considered (Figure 5E). The number of SVs starting in 100kbp windows was counted across the genome in the GIAB high confidence regions. Overall, no significant clustering was observed, further supporting that the majority of the unique calls were often true positives rather than false positives. Only very few SVs seem to cluster together, likely a result of technical noise. Only one window showed eight SVs unique to a certain SV caller that clustered together within 100kbp.

Factors Impacting Structural Variant Calling

Previous components investigated variability with respect to SV caller, platform, and replicates. However, variant SV calling (or more specifically SV calling) is often discussed to be impacted by coverage, read length (e.g. impacting the mappability) and insert size of the paired end reads. Thus, the contribution of these three factors on SVs calling was investigated specifically across multiple replicates and with respect to true positive calls based on GIAB.

First, the impact of varying coverage on the ability to detect SVs was investigated. For this, the mean coverage along the genome (as well as the standard deviation) was computed and compared both to the number of SVs (Figure S15 top row), and true positives based on GIAB SV calls (Figure S16 top row). Increased coverage has a clear positive effect on the number of SVs detected (Figure S15 top row). This is also reflected in the number of true positives based on the GIAB SV calls, but not as clearly as the overall calls. While it is obvious why the true positive increases with the average coverage, given more evidence can be found to support each SV, it is not clear how and if the standard deviation for coverage has a direct impact on the SV calling. This can be explained by the improvement in calling of SVs from paired-end reads with higher coverage.

Next, the possible impact of insert size (and variability of the insert size) was analyzed for SV detection. It was hypothesized that the variability of insert sizes plays an important role for the detection of SVs, since SV callers leverage the abnormal spacing of paired end reads to detect deletions. However, there was no evidence to support this within the SV data, neither with respect to mean insert size nor for standard deviation of insert sizes (Figure S15 middle row). No trend was observed when focusing only on the true positive

SVs based on the overlap with the GIAB SV call set (Figure S16 middle row). Thus, the insert size variability does not seem to impact the ability of an SV caller to identify SVs, likely due to the ability of the SV callers to leverage split read information.

Finally, the impact of read length was analyzed. It was hypothesized that an increase in read length, better confidence in mapping, and the ability of the mapper to characterize a split in the alignment should improve overall SV calling. The bottom row of Figure S15 shows the trend observed with respect to the average read length and the standard deviation. Since these are all Illumina short read data sets with fixed read lengths, the standard deviation of the read length was ignored. For the average read length, three categories (100bp, 150bp and 250bp) were available. Interestingly, there was no cle pattern for the total number of SVs identified based on the average read length (Figure S15 bottom row). Nevertheless, when filtered for the GIAB SV calls, a clear improvement for true positive SV calls compared to the increase in read length can be seen (Figure S16 bottom row).

Supplementary Figures

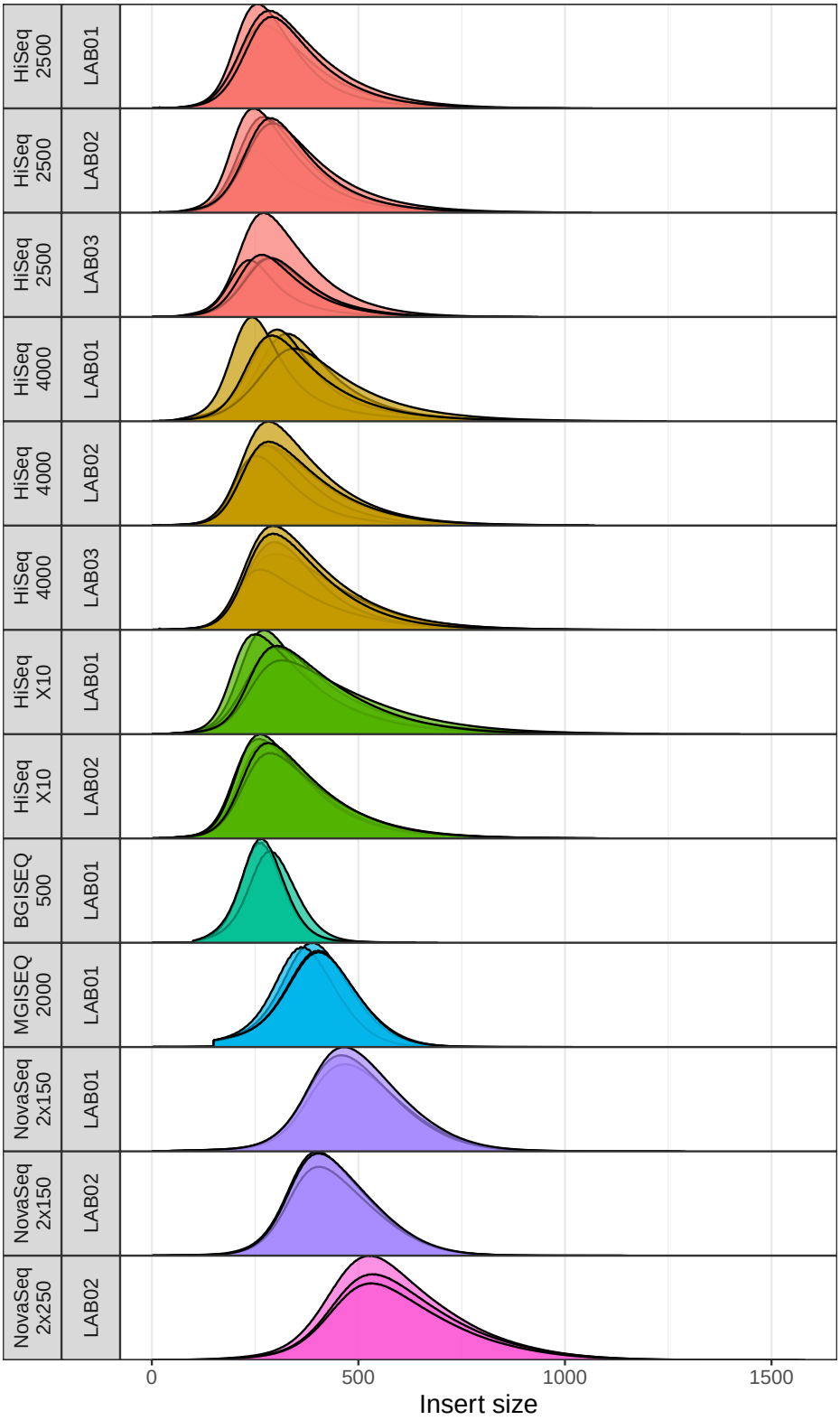

Figure S1: The insert Size distribution of every replicate, stratified by sequencing instrument.

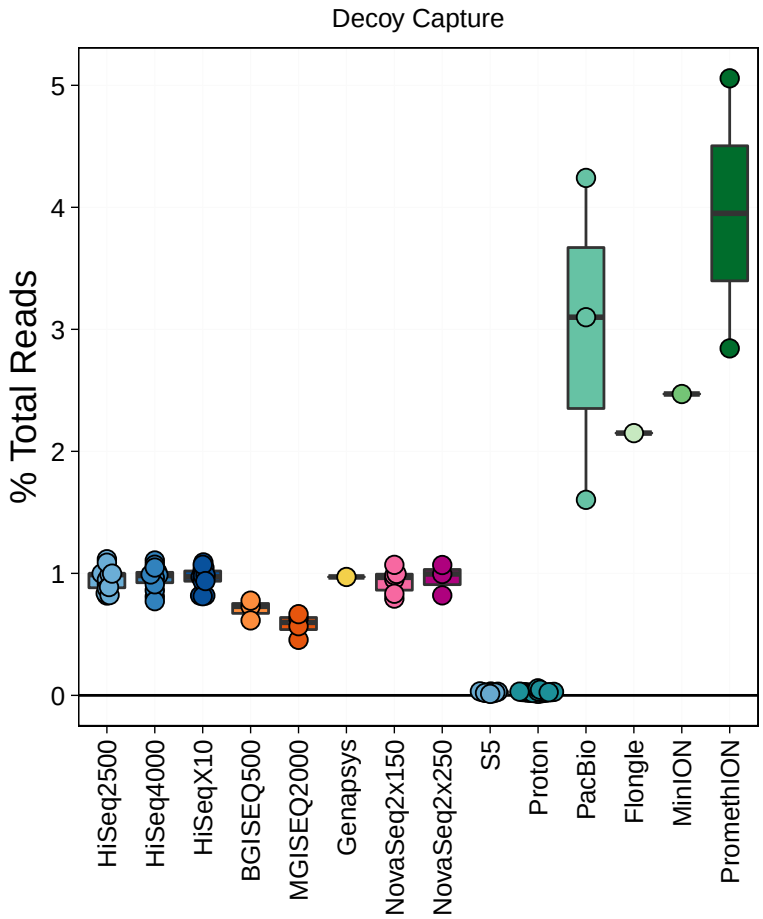

Figure S2: The percentage of total reads that were mapped to decoy contigs within the GRCh38 reference genome.

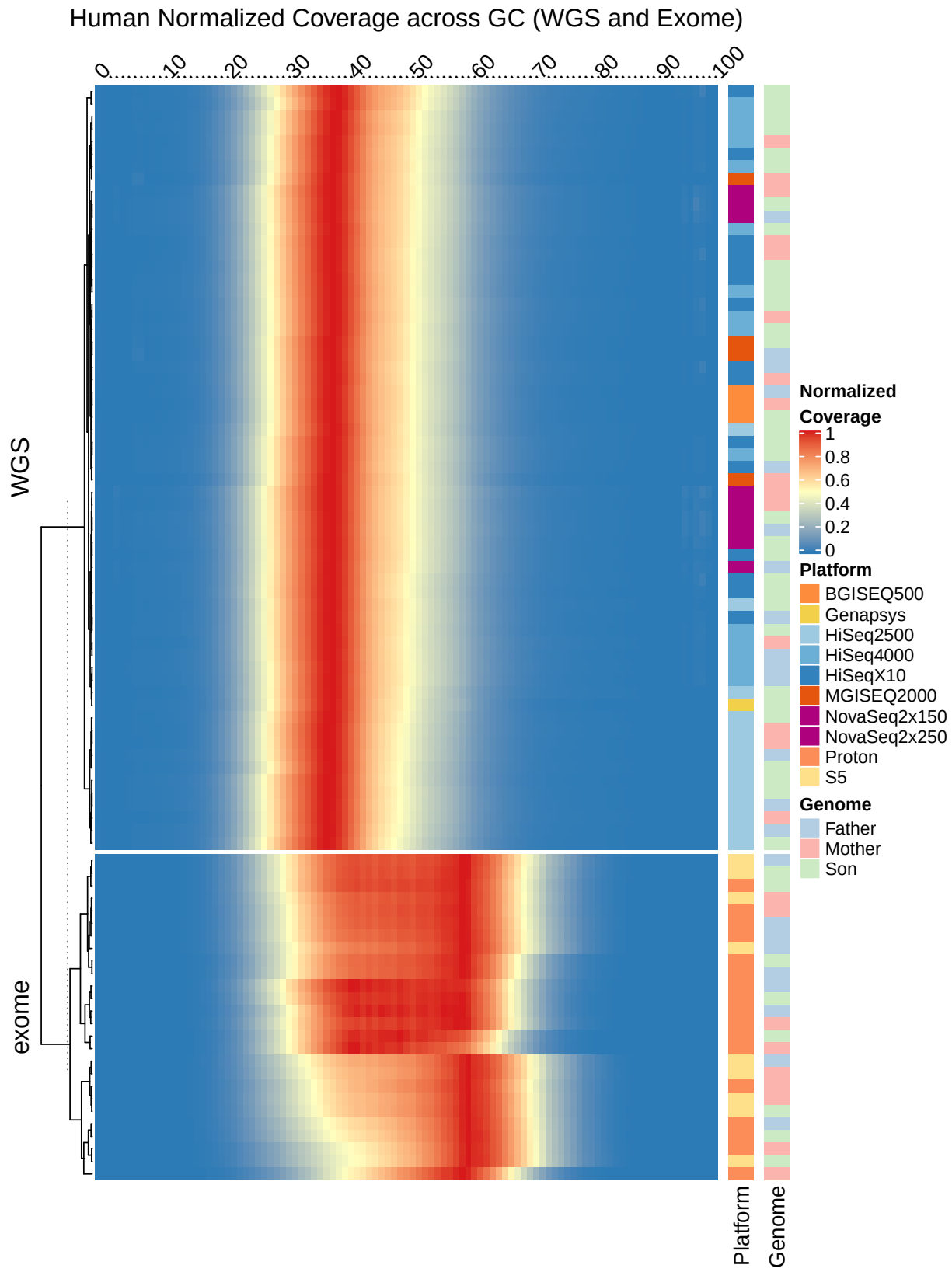

Figure S3: Heatmap showing the distribution of read counts per library (rows) by GC content (columns) across human whole genome and exome samples. Read count values are normalized by total reads per replicate, such that a value of 1 matches maximum value for a given replicate. Annotation tracks on the right indicate the sequencing platform and cell line genome for that replicate.

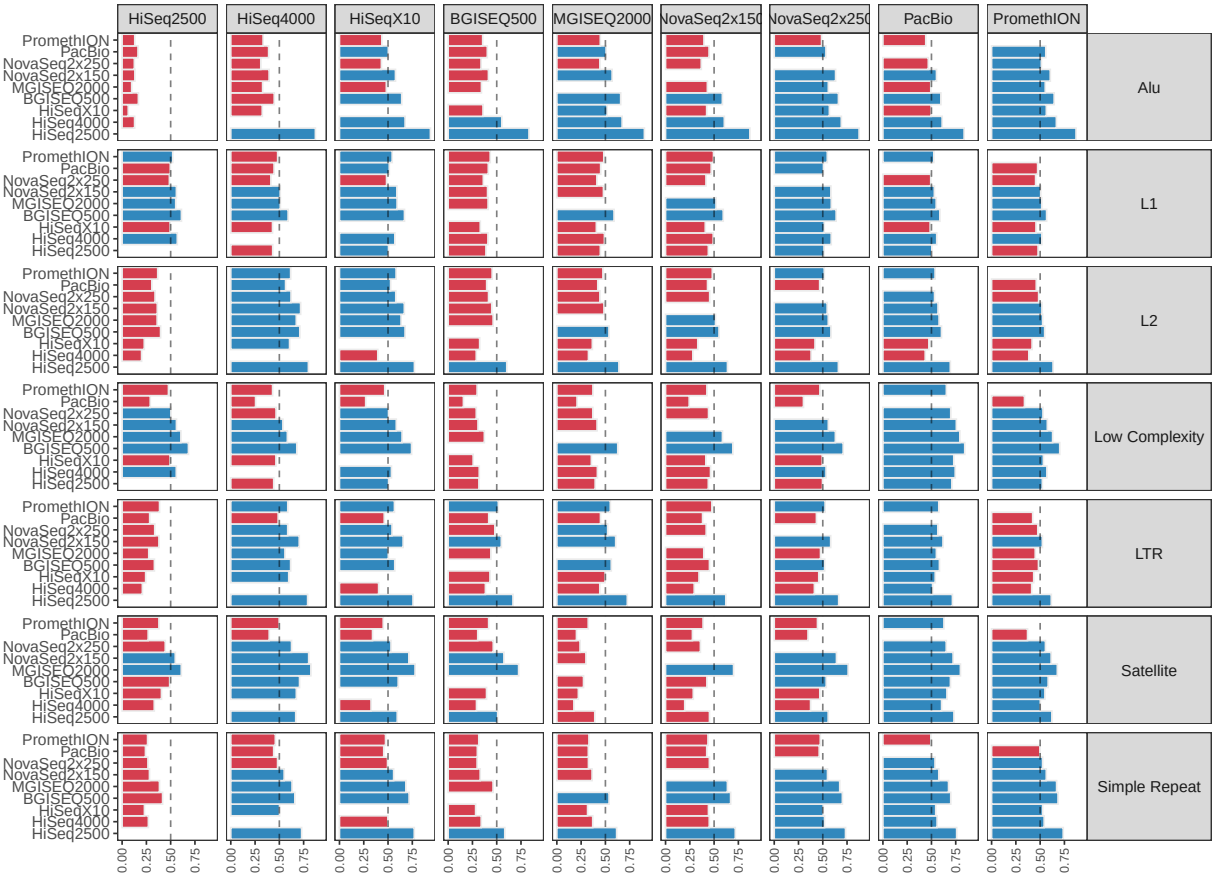

Figure S4: All-versus-all coverage comparisons for every platform within each UCSC RepeatMasker region. Blue bars indicate >50% of shared sites are better represented in the given platform (column) versus all other platforms (rows). Red bars indicate that the other platform out-covered the given platform.

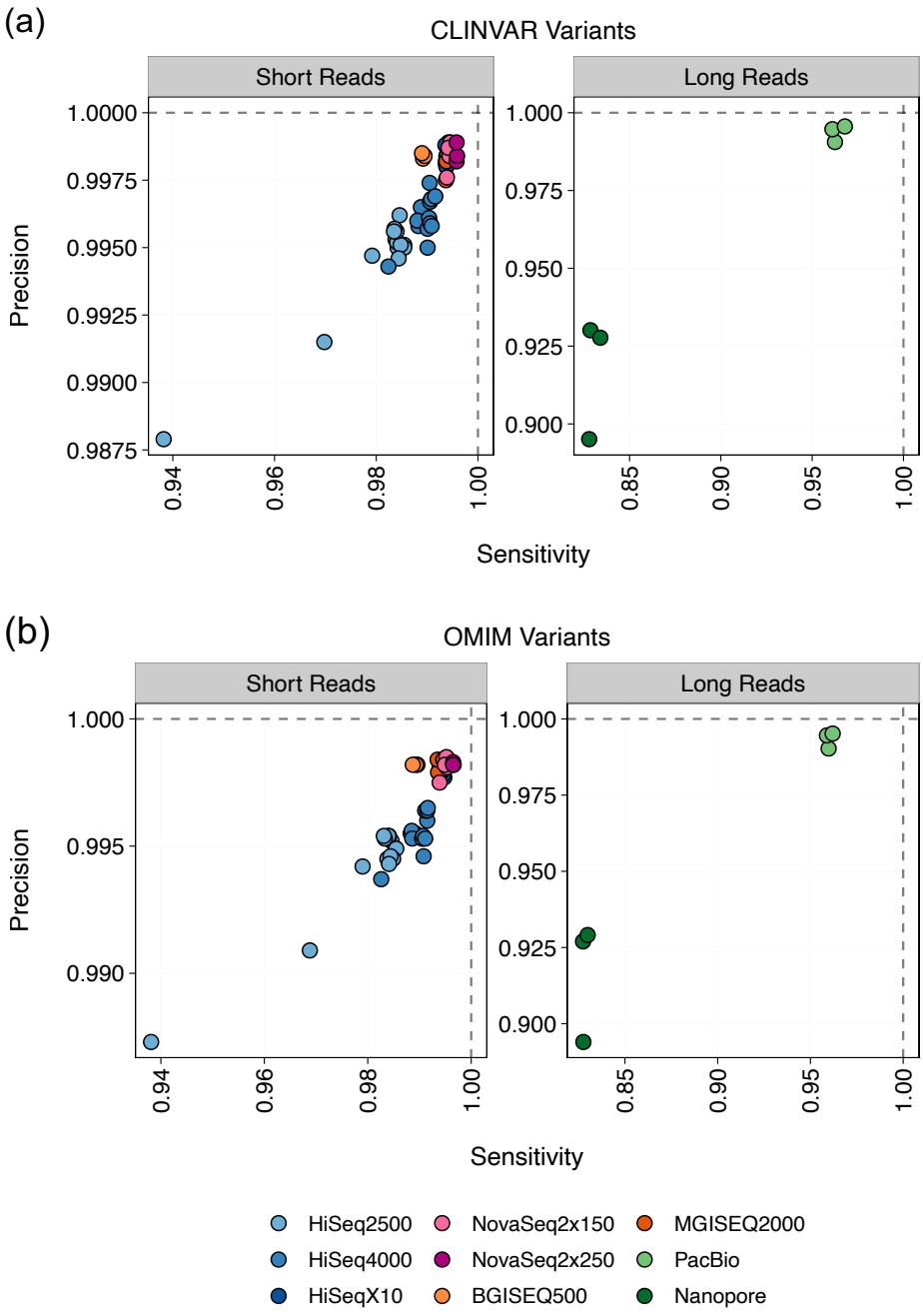

Figure S5: Precision and sensitivity scores as derived from rtg vcfeval analysis, stratified by regions in (a) the CLINVAR database and (b) the OMIM database. For each of the cell lines, genes from each database were overlapped with high confidence regions for variant calling.

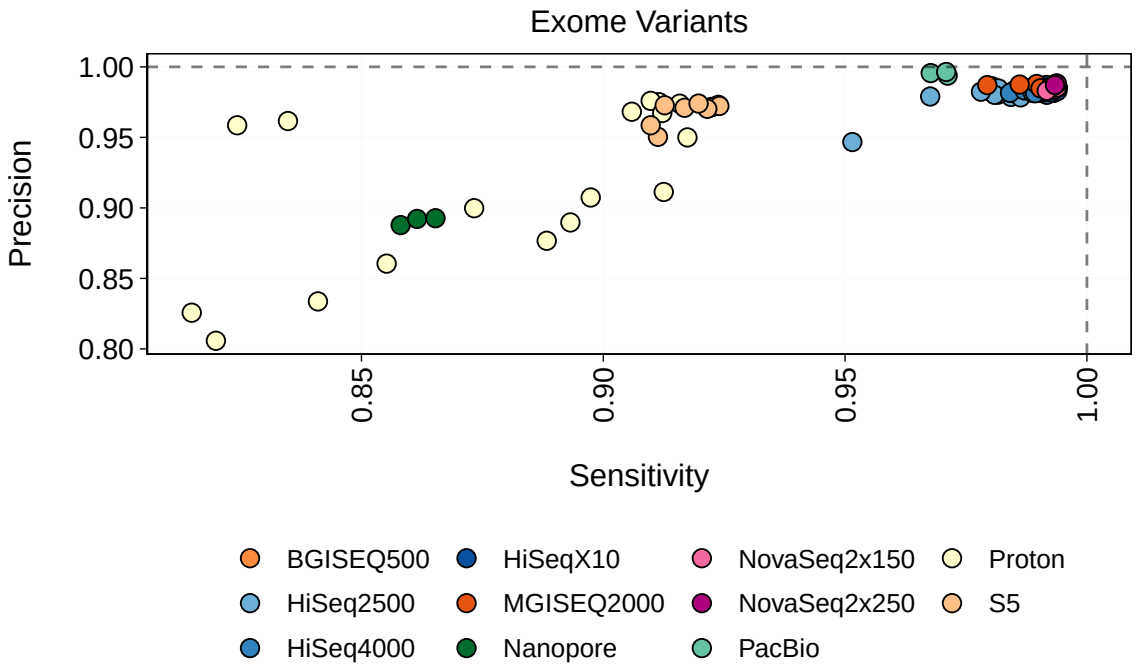

Figure S6: Precision and sensitivity scores as derived from rtg vcfeval analysis, stratified by regions in the exome, as defined by the AmpliSeq target capture regions file. For each of the cell lines, exomic regions were overlapped with high confidence regions for variant calling.

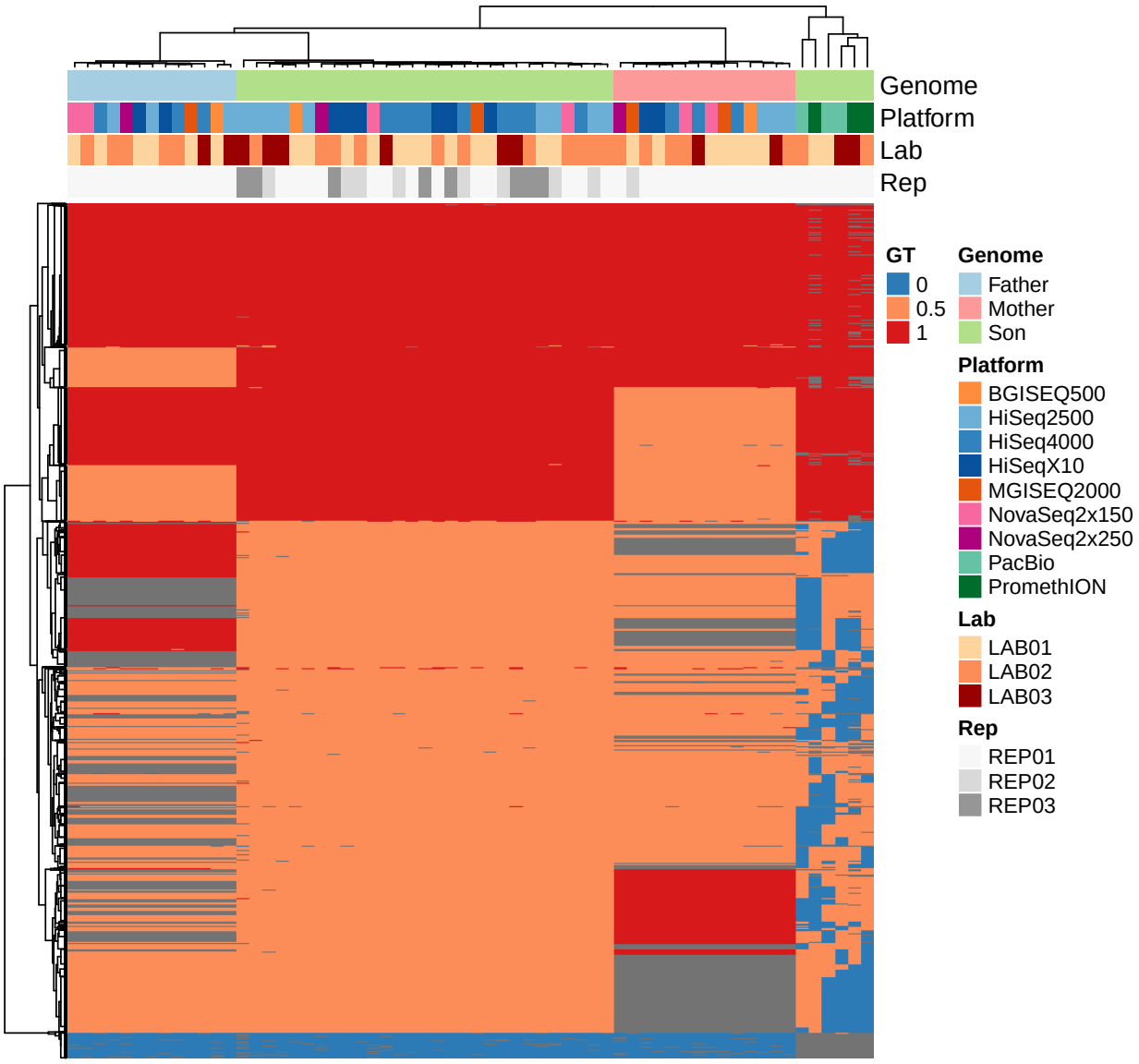

Figure S7: Heatmap of genotype (GT) of variant alleles on chromosome 1, across all human replicates across within sequencing platforms, as measured against the Genome in a Bottle high confidence variant call sets for each genome. Heterozygous variant alleles are shaded in orange (0.5), homozygous variants in red (1), missing data in blue (0), and inapplicable sites (sites outside of the GIAB high confidence region in one cell line but present in another) in gray. Hierarchical clustering reveals strong grouping by cell line, followed by less clear grouping within platforms and inter- and intra-lab replicates.

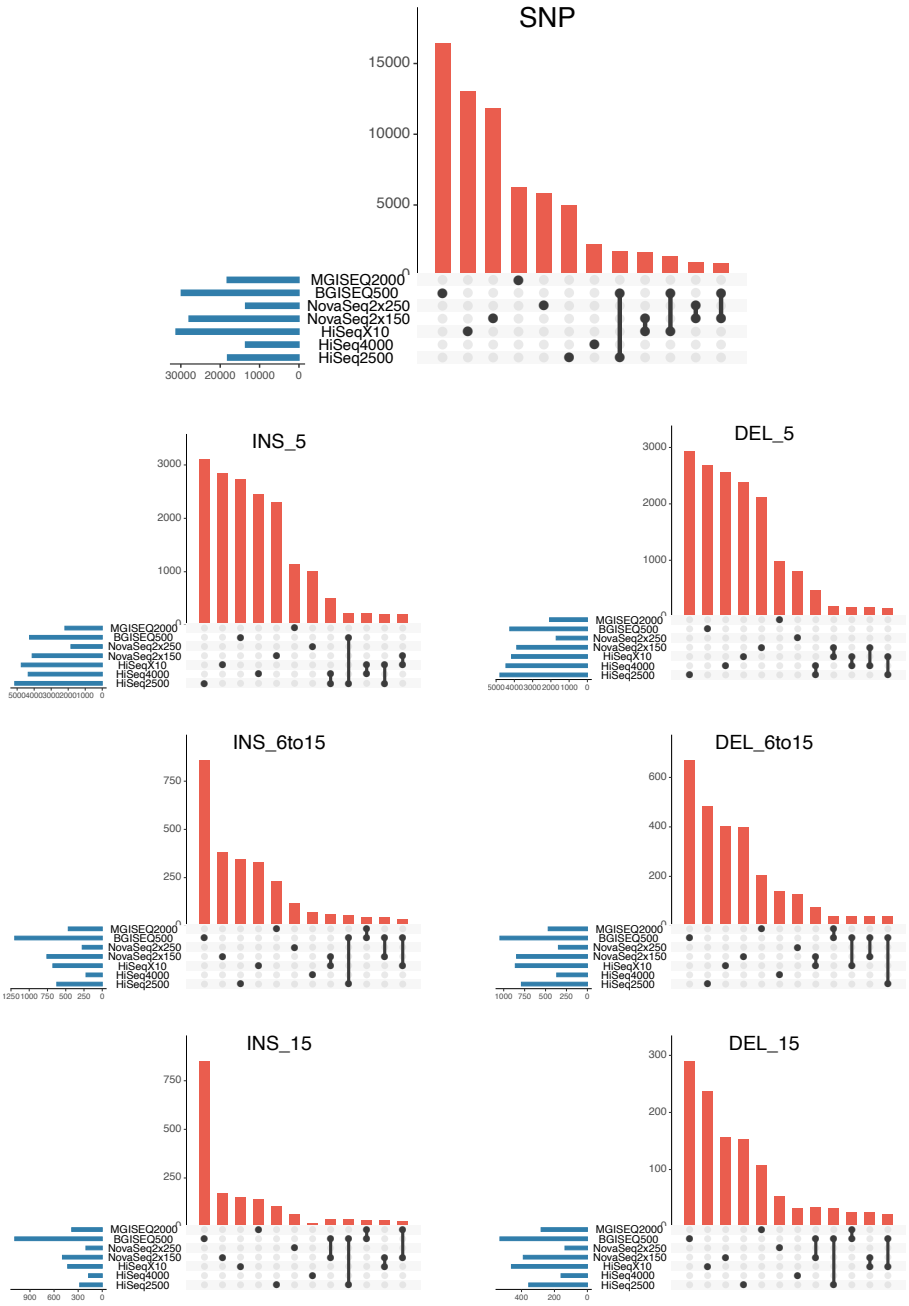

Figure S8: UpSet intersections of Mendelian violations. Each plot is stratified by variant type (SNPs on top, followed by INDELs; INS_5 = insertions 0-5bp in size, INS_6to15 = insertions 6 to 15bp in size, INS_15 = insertions >15bp in size; same for deletions, "DEL"). Events were recorded within high confidence regions for the Ashkenazi Son (HG002).

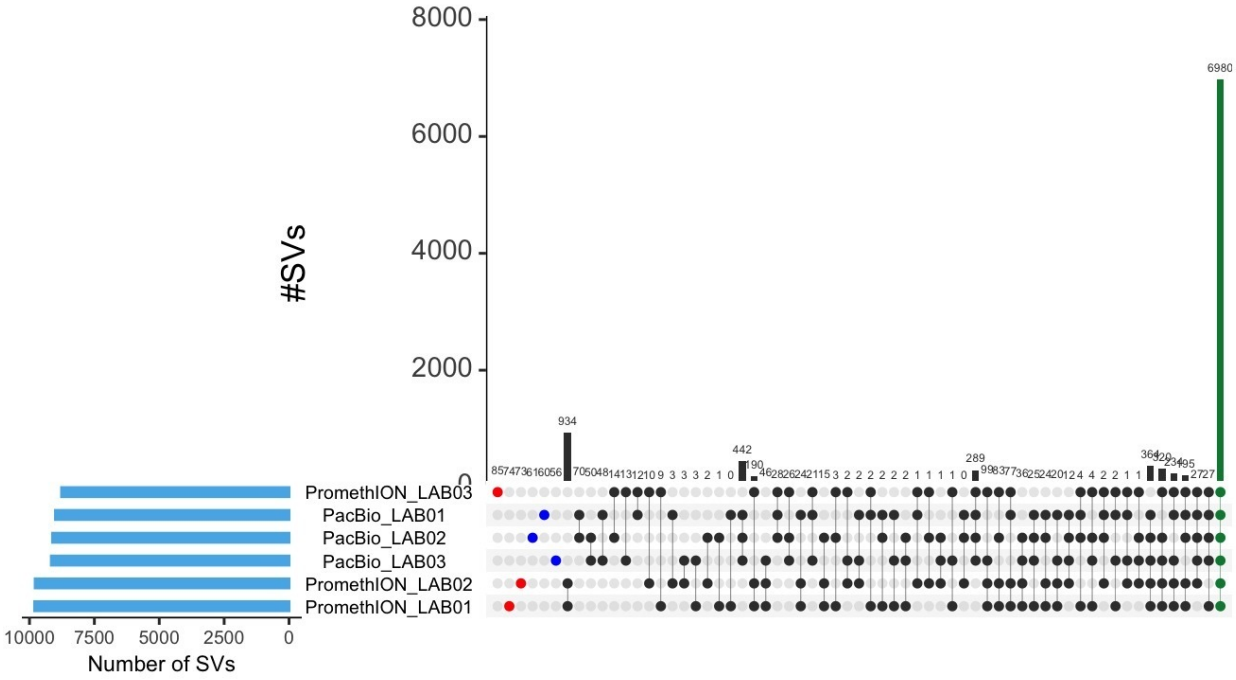

Figure S9: Comparison between the identified SVs in the six samples showing agreement of 6,980 SVs between samples (green column).

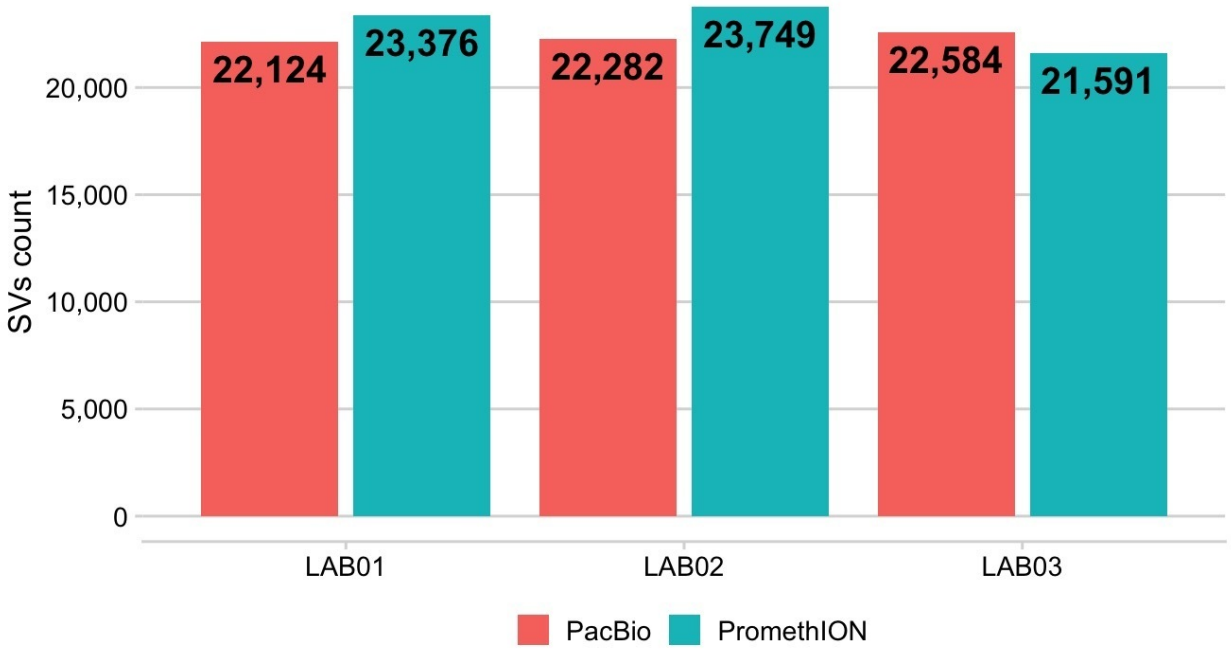

Figure S10: Identified SVs between samples.

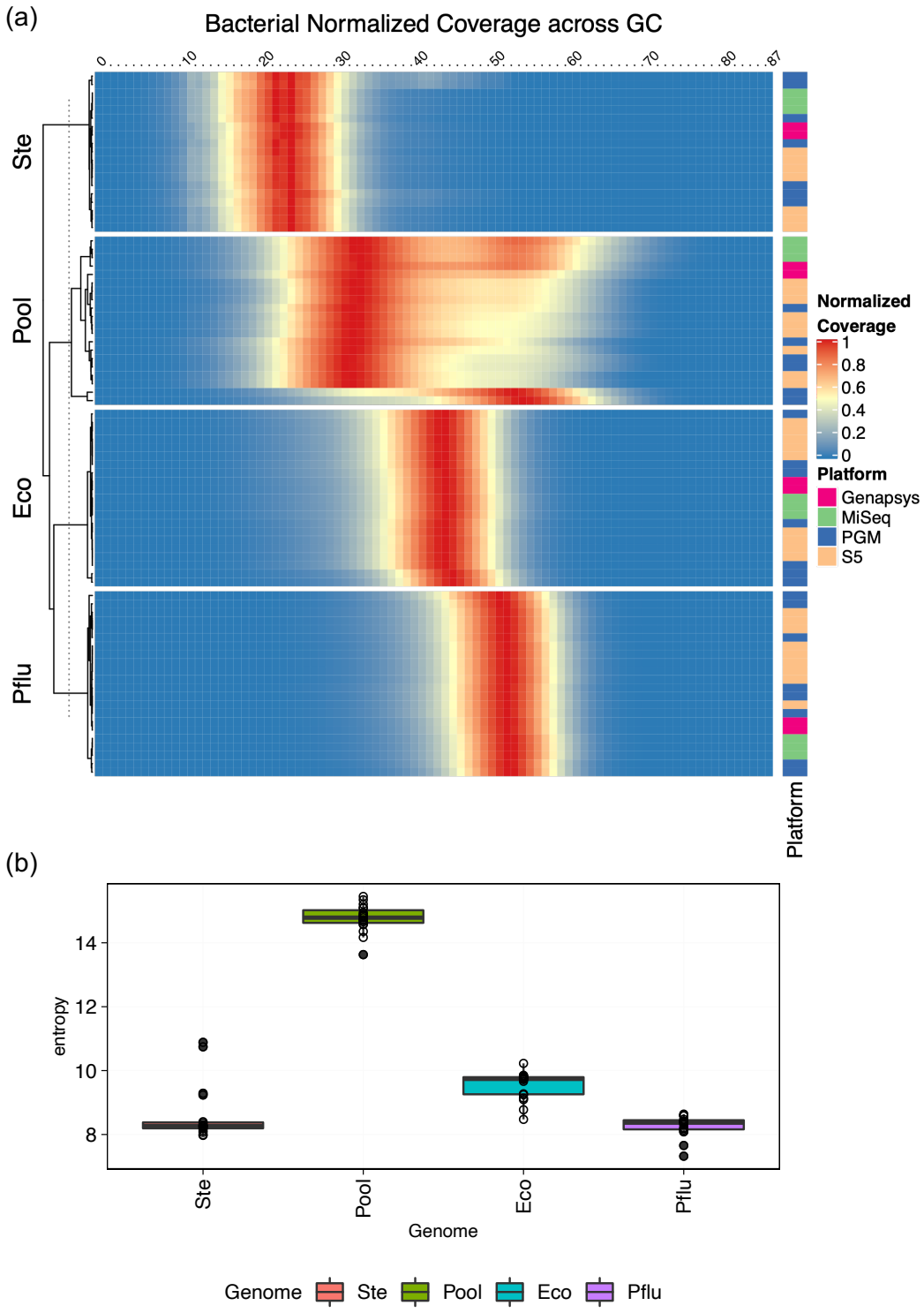

Figure S11: (a) Heatmap showing the distribution of read counts per library (rows) by GC content (columns) across bacterial genomes and the metagenomic mixtrue. Read count values are normalized by total reads per replicate, such that a value of 1 matches maximum value for a given replicate. Annotation tracks on the right indicate the sequencing platform and cell line genome for that replicate. (b) Calculations of entropy per genome/metagenomic mixture. Entropy was measured across all GC windows for all replicates for a given sample, $rowSums(-(p * log(p))$.

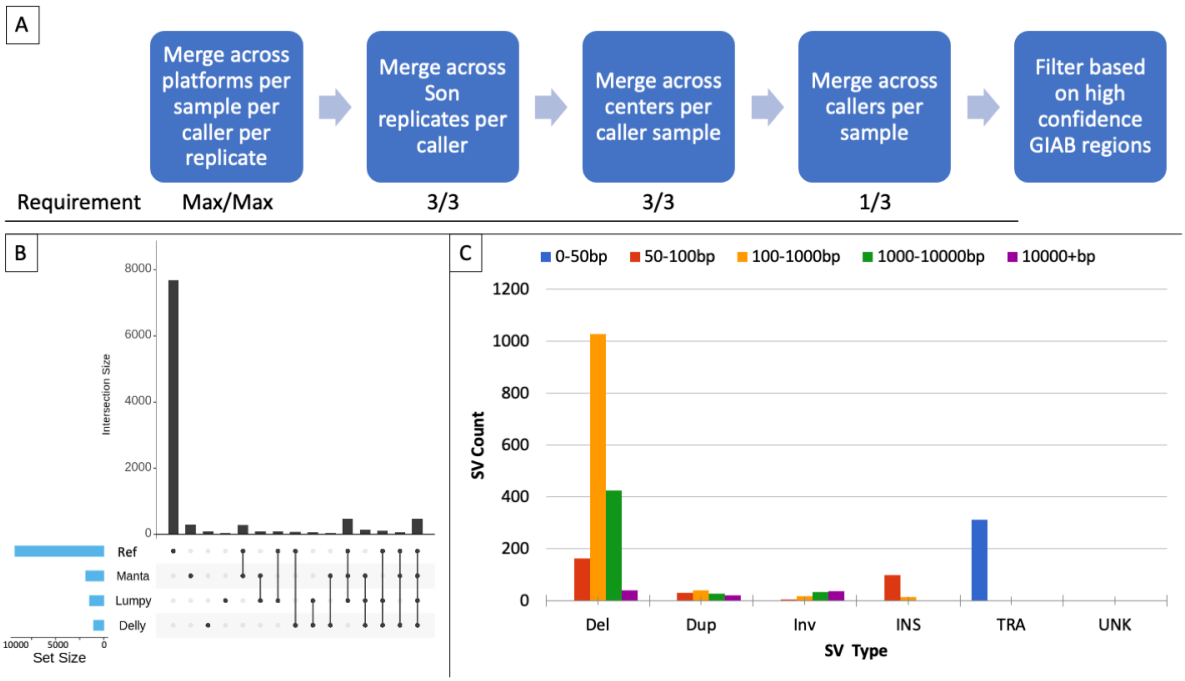

Figure S12: Insights into SV variability by caller. (a) The strategy used to examine SV caller variability after stratifying for platforms, replicates and centers variability. (b) Shows SV call set sizes and overlap with the GIAB SV call set for the SV caller variability set of HG002. (c) Types and sizes of SVs in the SV caller variability set of HG002 (translocations are set to size 50 by default in the SURVIVOR parameters for visualization purposes)

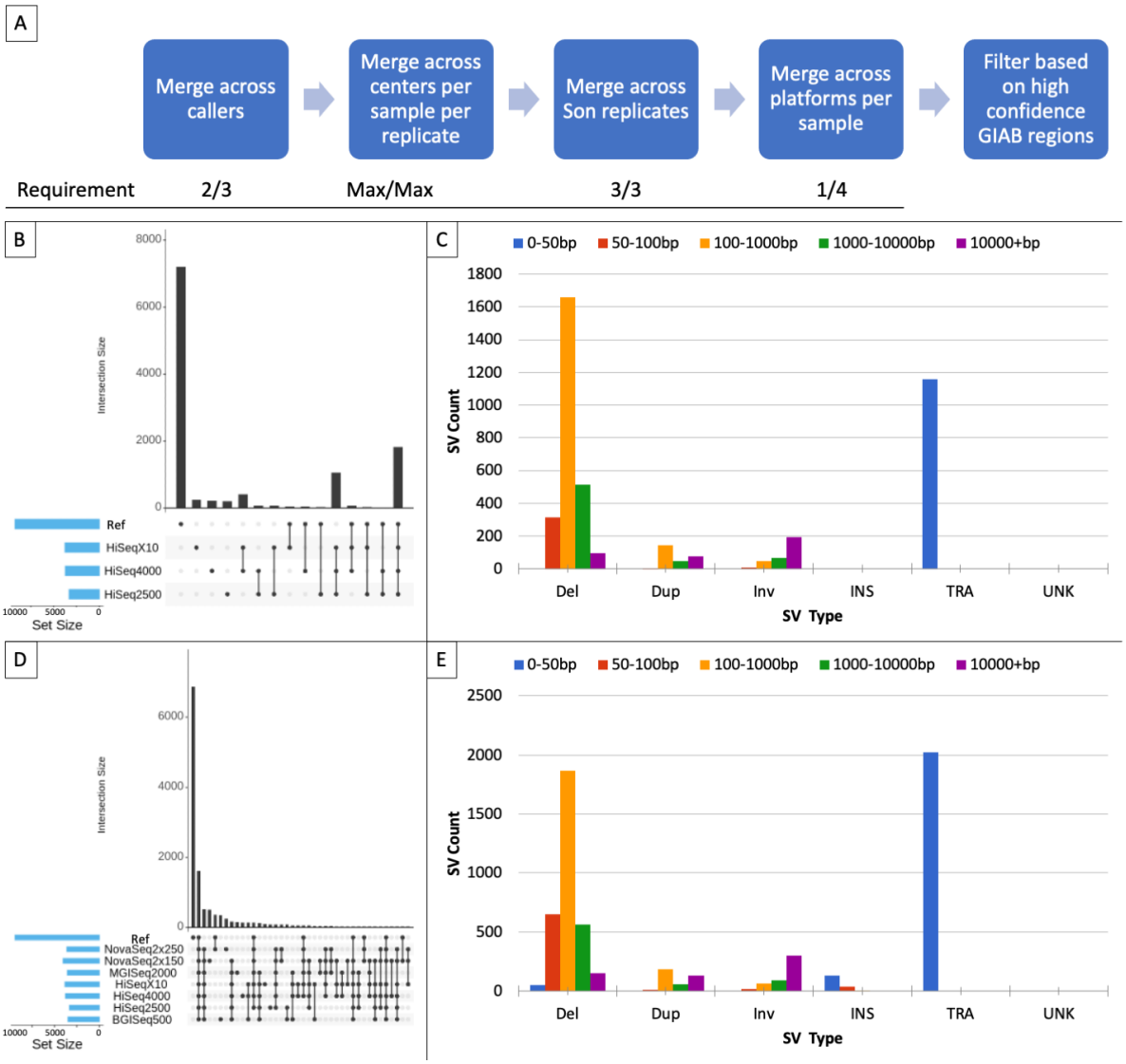

Figure S13: Insights into SV variability by platform. Diagrams (a,b,c) utilize sequencing runs from HiSeqX10, HiSeq2000 and HiSeq4000 whereas (d,e) characterize all platforms available. (a) The strategy used to examine platform variability after stratifying for SV callers, centers and replicates variability. (b) SV call set sizes and overlaps with the GIAB SV call set for the platform variability SV call set of HG002. (c) Types and sizes of SVs in the platform variability SV call set of HG002. Panels (d) and (e) include HiSeqX10, HiSeq2000, HiSeq4000, NovaSeq, BGIS, and MGI for visualization purposes. The NovaSeq, BGI, and MGI SV call sets were not integrated into the analyses strategy because sequencing runs with replicates for each sample at different centers on different platforms were not available. (d) SV call set sizes and overlap with the GIAB SV call set for the platform variability SV call set of HG002. (e) Types and sizes of SVs in the platform variability SV call set of HG002. (Translocations are set to size 50 by default in the SURVIVOR parameters for visualization purposes)

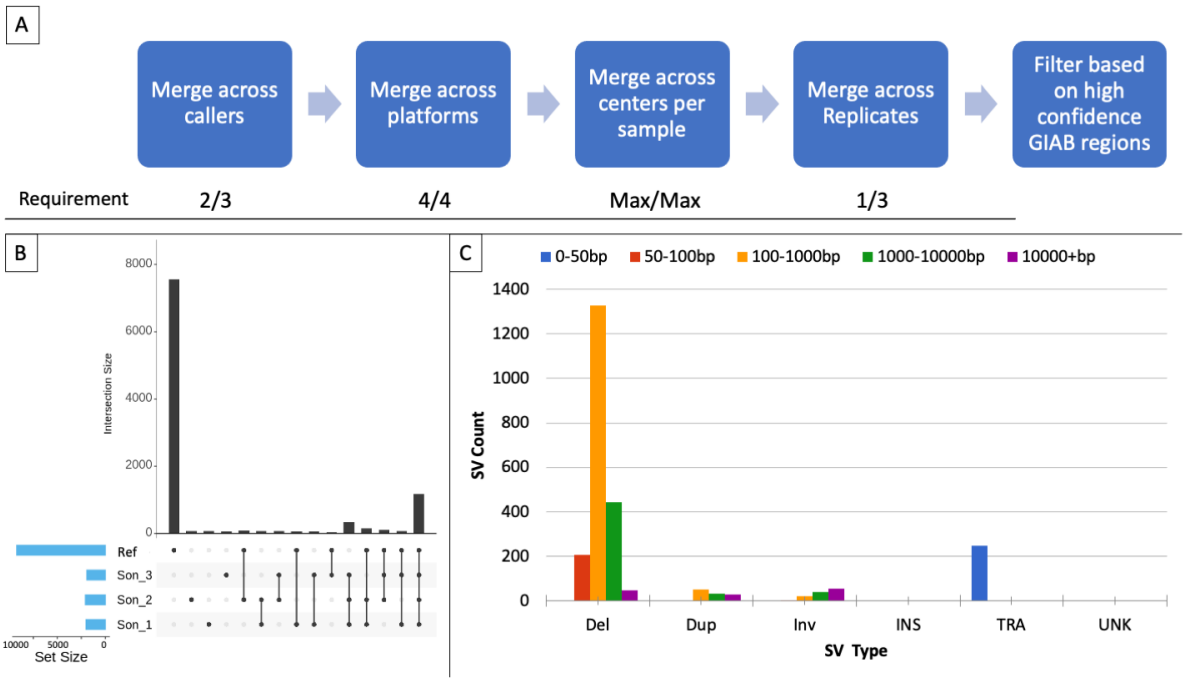

Figure S14: Insights into SV variability by replciate. Diagrams (a,b,c) utilize sequencing runs from HiSeqX10, HiSeq2000 and HiSeq4000. (A) The strategy used to examine replicate variability after stratifying for SV callers, platforms and centers variability. (b) SV call set sizes and overlap with the GIAB SV call set for the replicate variability SV call set of HG002. (c) Shows the types and sizes of SVs in the replicate variability SV call set of HG002 (translocations are set to size 50 by default in the SURVIVOR parameters for visualization purposes).

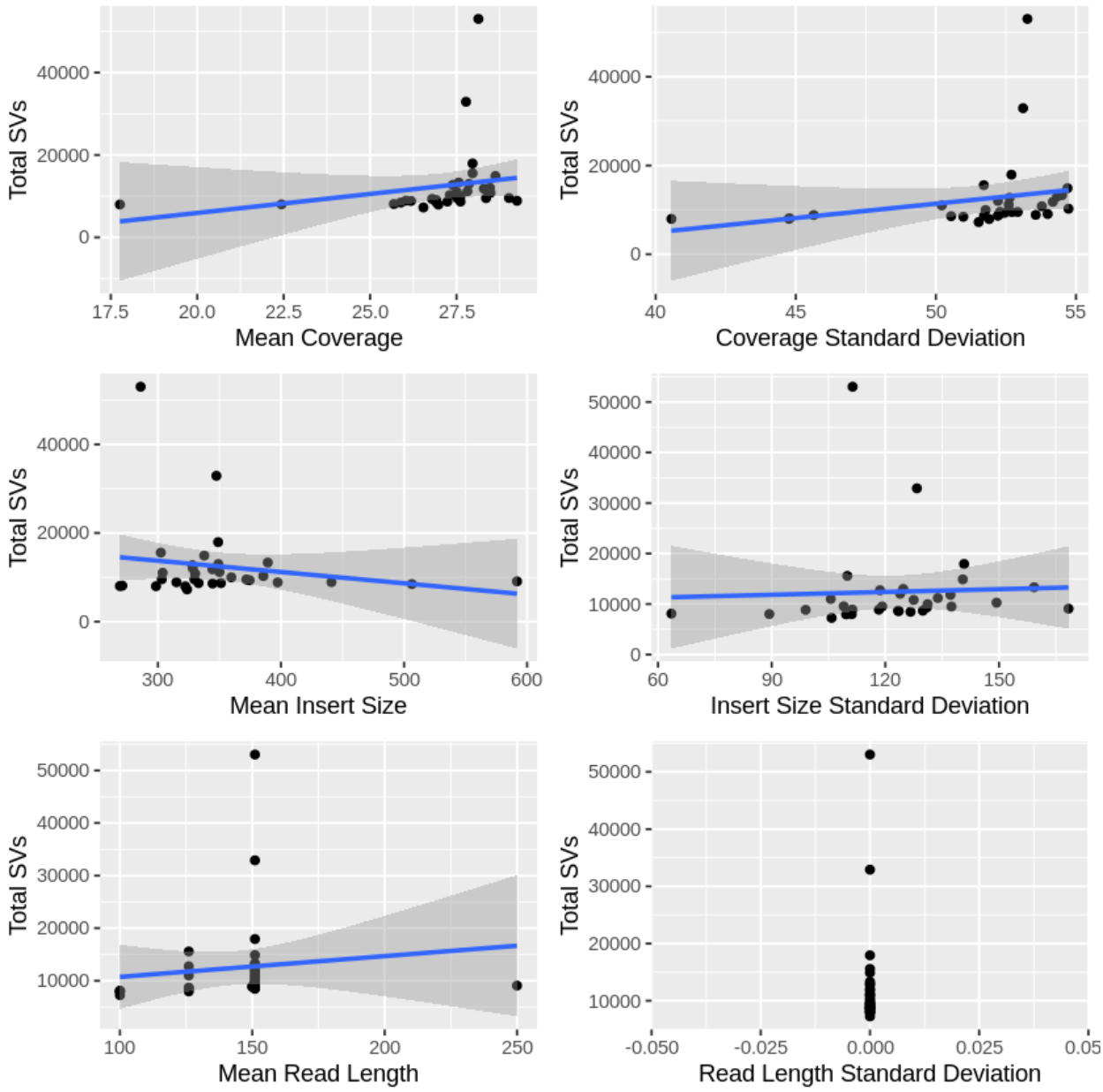

Figure S15: Coverage, insert size, and read length mean and standard deviation across total SVs in sequencing runs.

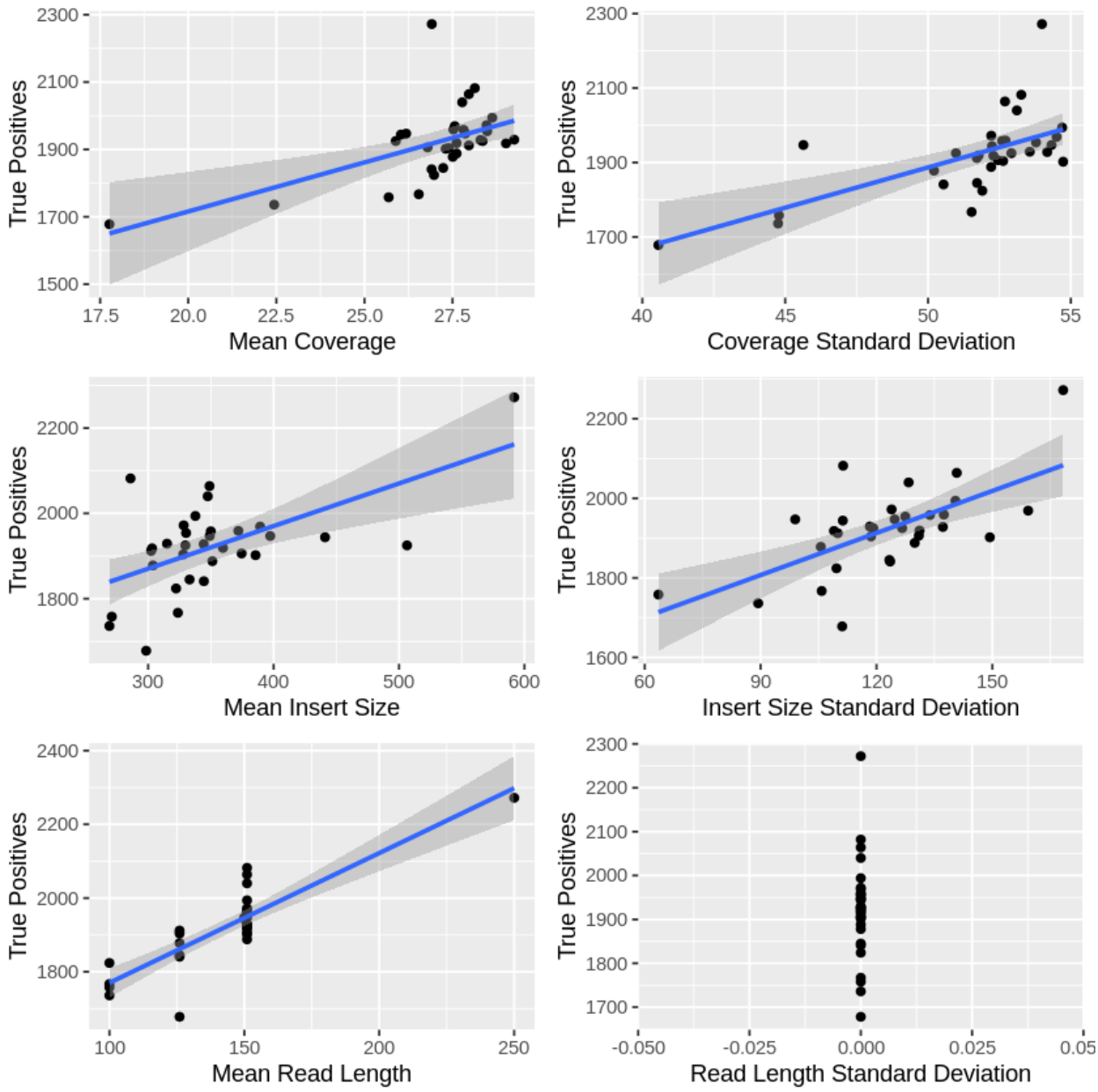

Figure S16: Coverage, insert size, and read length mean and standard deviation across true positives (overlapping with the HG002 reference SV set) in sequencing runs.