Advanced Whole Genome Sequencing Using a Complete PCR-free Massively Parallel Sequencing (MPS) Workflow

Hanjie Shen[1,2,3#] , Pengjuan Liu[1,2,3#], Zhanqing Li[1,2,3#], Fang Chen[1,2,3],Hui Jiang[1,2,3], Shiming Shi[1], Yang Xi[1,2,3], Qiaoling Li[1,2,3], Xiaojue Wang[2,3], Jing Zhao[1,2,3], Xinming Liang[1], Yinlong Xie[1], LinWang[4],Wenlan Tian[4], Tam Berntsen[4], Yinling Luo[1,2,3], Meihua Gong[1,2,3], Jiguang Li[1,2,3], Chongjun Xu[1,2,3,4],Sijie Dai[1],Zilan Mi[2,3], Han Ren[2,3], Zhe Lin[2], Ao Chen[2,3], Wenwei Zhang[2,3], Feng Mu[1],Xun Xu[2,3], Xia Zhao[1,2,3]* Yuan Jiang[2,3,4] * Radoje Drmanac[1,2,3,4]*

1. MGI, BGI-Shenzhen, Shenzhen 518083, China
2. BGI-Shenzhen, Shenzhen 518083, China
3. China National Genebank, BGI-Shenzhen, Shenzhen 518120, China
4. Complete Genomics Inc., 2904 Orchard Pkwy, San Jose, CA 95134 USA

# These authors contributed equally to this work.

*Correspondence:

Xia Zhao: zhaoxia@genomics.cn, Yuan Jiang: yjiang@completegenomics.com, and Radoje Drmanac: rade@completegenomics.com

Abstract

Systematic errors could be introduced by amplification during MPS library preparation and cluster/array formation. Polymerase Chain Reaction (PCR)-free library preparation methods have previously demonstrated improved sequencing quality with PCR-amplified read-clusters,however we hypothesized that some some InDel errors are still introduced by the remaining PCR step. Here we sequenced PCR-free libraries on MGI's PCR-free DNBSEQ™ arrays to obtain for the first time a true PCR-free WGS (Whole Genome Sequencing). We used MGI's PCR-free WGS library preparation kits as recommended or with some modifications to make several NA12878 libraries. Reproducibly high quality libraries where obtained with low bias and less than 1% read duplication for both ultrasonic and enzymatic DNA fragmenting.In a triplicate analysis, over 99% SNPs and about 98% indels in each library were found in at least one of the other two libraries. Using machine learning (ML) methods (DeepVariant or DNAscope), variant calling performance (SNPs F-measure>99.94%, InDels F-measure>99.6%) exceeded the widely accepted standards. The F-measure of 15X PCR-free ML-WGS was comparable to or even better than 30X PCR WGS analyzed with GATK. Furthermore, PCR-free WGS libraries sequenced on PCR-free DNBSEQ™ platform have up to 55% less InDel errors compared to NovaSeq platform confirming that DNA clusters have PCR-generated errors.Enabled by the new PCR-free library kits , super high-thoughput sequencer and ML-based variant calling, DNBSEQ ™ true PCR-free WGS provides a powerful solution to improve accuracy while reducing cost and analysis time to facilitate future precision medicine,cohort studies and large population genome project.

1 **Keywords**

2 WGS, PCR-free, DNBSEQ™, Machine learning based variant calling

Introduction

MPS (also known as next-generation sequencing, NGS) technology has revolutionized basic biology and precision medicine during the past decade. Comparing to partially sequencing the genome via targeted panels such as whole-exome sequencing (WES), an entire genome sequencing, also known as whole genome sequencing (WGS), achieves better coverage uniformity and more reliable detection in single-nucleotide polymorphisms (SNPs), insertions and deletions (InDels), structural variants (SVs) and copy number variants (CNVs) at both coding and non-coding regions. There is an increasing clinical demand to apply WGS as one single test, especially in conditions when WES could potentially fail to detect all pathogenic variants in a large fraction of mendelian disorder cases [1,2,3] A variety of studies have already demonstrated the feasibility of WGS to investigate rare diseases such as inherited retinal disease [4], cancers such as liver cancer [5] and infectious diseases such as Mycobacterium tuberculosis (MTB) [6]. More importantly, according to the data from the NHGRI Genome Sequencing Program,sequencing cost has rapidly dropped to $0.012 per Mb in July 2017 [7]. WGS costs no more than thousand dollars nowadays, with faster turnaround

time and at greater depth, making it more economic to conduct large-scale projects such as the Genomics England 100,000 Genome Project.

Standard WGS workflow includes MPS library preparation, array/cluster generation, on chip sequencing, reads filtering, mapping, and variant calling (aka secondary analysis). Many efforts have been undertaken to further reduce the cost and turnaround time while improving the WGS data performance. For example, it is well known that PCR amplification in regular WGS library preparation causes uneven amplification and copy errors, resulting in coverage bias, GC bias and nucleotide misincorporation [8,9,10,11]. To address this issue, an optimized WGS library protocol that eliminates the need for PCR called PCR-free WGS is developed. The PCR-free WGS improves read mapping quality and aids in *de novo* assembly of genomic regions containing extreme GC content (the percentage of G and C bases) [12]. Moreover, PCR-free protocol removes duplicated reads from library prep, which increases the read utility and variant calling confidence [13]. The last benefit of excluding PCR step in the WGS library prep is obviously shorter turnaround time and lower cost.

In addition to the new library construction chemistry, different innovative bioinformatics tools are applied to expedite data analysis without sacrificing accuracy. The steps of a standard analysis workflow are typically as following: 1) to trim and filter read data; 2) to align raw data to a reference genome, 3) to call germline or somatic variants, 4) to conduct tertiary analysis and generate report. The best Practices pipeline for germline short variant discovery developed at the Broad Institute with the Genome Analysis Toolkit (GATK) [14] is currently industry standard pipeline for variant calling on WGS. However, traditional GATK pipeline takes around two days for whole genome data processing on a standard 24 thread machine [15]. To achieve better detection accuracy and account for systematic errors in the WGS workflow, researchers have explored machine learning or deep learning based algorithms and developed several new analysis pipelines such as GATK CNNScoreVariants from the Broad Institute [16], Deepvariant from the Google Brain team [17],DNAscope from Sentieon [18] and Clairvoyante from R. Luo et al [19].

In this study, we presented a completely PCR-free MPS workflow by constructing PCR-free WGS libraries and sequencing on PCR-free on-chip arrays. Both PCR-free WGS sets (mechanic fragmentation with MGIEasy PCR-Free DNA Library Prep Set and enzymatic fragmentation with MGIEasy FS PCR-Free DNA Library Prep Set) demonstrated highly reproducible data quality even with 200ng or 50ng genomic DNA as input. More importantly, a significant improvement was achieved in InDels calling with GATK from an average F-score of 95.43% in three PCR libraries to that of 99.05% in three PCR-free WGS libraries. By incorporating machine learning based algorithms, the F-score of 15x PCR-free analyzed with DNAscope or DeepVariant can outperform that of 30x PCR analyzed with GATK in some scenarios. The complete PCR-free WGS significantly decreased FP and FN InDel reads, leading to a better InDel calling accuracy (99.3-99.6% F-score) than that of the Illumina's PCR-free libraries (98.0-99.3%) with lower genomic DNA as input. Additionally, an incredibly low duplication rate, less than 1%, was achieved in the DNBSEQ™ PCR-free WGS, resulting in better read utility and further reducing the sequencing cost for a standard 30x coverage. In summary, the advanced PCR-free WGS reported herein could lead to wider adoption of WGS in the genomic research and gradually in the clinical practice for rare disease urgent diagnosis.

Results

A complete PCR-free MPS workflow

PCR is frequently used to increase template amount during MPS library construction. It's also an essential step of 'bridge amplification' to generate identical copies on a flow cell surface [20]. Here, we describe an MPS workflow called DNBSEQ™ PCR-free which completely eliminates PCR amplification in both library and array preparations (Figure 1a). DNBSEQ™ PCR-free starts with DNA fragmentation, followed by size selection using solid-phase reversible immobilization (SPRI) beads. A single-tube protocol is applied to conduct sequential multiple enzymatic reactions to attach a barcoded adapter to the DNA of interests. After removing excess adapters, single-stranded circle (ssCir) is formed and served as the template in rolling circle amplification (RCA) for DNB preparation. DNBs are then loaded onto the patterned flow slides and sequenced. In contrast to bridge amplification, RCA is a linear amplification from the original ssCir template and therefore errors would not accumulate[21,22]. Different from DNBSEQ™ PCR with additional amplification after adapter ligation and bridge PCR or emulsion PCR

based dsDNA library sequencing, DNBSEQ™ PCR-free can completely avoid the enrichment of PCR errors in template amplification and library cloning process, and faithfully restores the original landscape of the genome.

Two sets from MGI (MGIEasy PCR-Free DNA Library Pre Set and MGIEasy FS PCR-Free DNA Library Pre Set) are used to prepare the DNBSEQ™ PCR-free libraries. The PCR-Free set is used with acoustic fragmented samples, while the FS PCR-Free one includes sequential reactions of enzymatic fragmentation, end repair/dA-tailing and adapter addition in a single tube. We compared the performance of both sets with nine libraries constructed from 1 μg NA12878 reference genomic DNA and sequenced with pair-end 150 bp read length. Summarized in figure 1b were the QC statistics including GC Content, Duplication Rate, Median Insert Size, and regions with > 10X coverage. Both sets showed highly reproducible performance with PCR-Free set having slightly smaller variation than FS PCR-Free set. We also tested different input amounts (1 μg, 500 ng,200 ng , and 50ng only for FS PCR-Free Set) with both sets and observed comparable performance (Figure 1S).

Minimal GC bias for genomes with different GC contents

The relationship between GC content and read coverage across a genome, namely GC bias, can be greatly affected by MPS library prep, cluster/array amplification, and sequencing. To evaluate the performance of the DNBSEQ™ PCR-free MPS workflow on GC bias, DNA samples from bacteria strains with GC contents of 38% and 62% were processed with both sets mentioned above. Libraries were made following by kit instruction and sequenced on MGISEQ-2000 with paired-end sequencing (2x150 bp). From Figure 2, the number of reads covering different regions was normalized by the mean read coverage and was then plotted against genomic GC contents of corresponding regions. Normalized coverage lower or higher than 1 indicates certain GC bias. The high GC and low GC genomes demonstrated fairly even coverage of reads across the genome with either acoustic shearing or enzymatic shearing. Overall, the DNBSEQ™ PCR-free MPS workflow demonstrated minimal GC bias in genomes with a variety of GC contents.

Low duplicate rate and high variant calling F-score for GIAB with DNBSEQ™ PCR-free

We compared the sequencing accuracy of MPS libraries with or without PCR using DNBSEQ™ sequencing technology [23]. After acoustic fragmentation, NA12878 DNA was processed following instructions of MGIEasy PCR-Free DNA Library Prep Set to construct three PCR-free WGS libraries. Meanwhile three PCR WGS libraries were prepared from the same DNA sample. Each library was sequenced individually on one lane of MGISEQ-2000 with paired-end sequencing (2x150 bp). (Methods and Materials) The total raw data of each lane was greater than 120G with GC content ranging from 41.46% to 41.78% (Table 1) .

The raw reads were down sampled from original full lane (approximately 46x) to create additional 30x and 15x depth dataset. After raw reads filtering, clean reads were aligned to the human reference genome with decoy sequence, hs37d5. The mapping quality matrix was summarized in Table 1. We compared all 3 depth datasets from the 6 libraries (18 dataset in total). In 30x depth data, WGS PCR and PCR-free libraries had similarly high mapping rate at 99.8% and overall coverage higher than 99.1%. PCR-free

libraries showed slightly lower duplication rate and mismatch rate, around 1% and 0.7% respectively on average, while they were 1.5% and 0.9% in PCR libraries. Other depth dataset showed similar patterns. Theoretically, because of the "true PCR-free" workflow duplicate rate should be zero, but there are chances that the same DNB being read twice or multiple times caused by optical overflow (aka "optical duplicates") or the same regions from different genome copy produced in DNA fragmentation step (aka "natural duplicates") being sequenced and being marked as duplicate by QC tool incorrectly [24,25].

We next used three variant callers: GATK, DeepVariant, and DNAscope (Methods and Materials), to assess the accuracy of PCR or PCR-free methods (Table2, Table1S, and Figure 3). The GATK HaplotypeCaller has become the industry standard small variant caller due to its high accuracy and has achieved top performance in a variety of public and third-party benchmarks [15,26,27,28], while DeepVariant and DNAscope are two newly developed variant callers based on machine learning method [17,18,29,30]. It

should be noted that both machine learning variant callers used in this study were optimized for the DNBSEQ™ platform through the use of utilizing in-house training data.

Variant calling matrix is highly reproducible in all three replicates for both PCR and PCR-free libraries (Table1S). Table2 summarized the average number of three replicates for different depths. At 30x depth, GATK called and marked Passed Filter (named true positive, or TP for short) for an average of a total of 3,651,696 TP variants for PCR libraries, and 3,674,252 TP variants for PCR-free libraries, while DeepVariant and DNAscope showed higher sensitivity as they both detected additional TP variants for both PCR and PCR-free libraries. From all three callers, PCR-free libraries demonstrated slightly reduction in False Positive (FP) numbers of SNPs (from 2806 to 2586 in GATK, from 2913 to 2111 in DeepVariant, from 1565 to 1254 in DNAscope) as well as reduction in False Negative (FN) numbers of SNPs (from 16671 to 10926 in GATK, from 4266 to 3109 in DeepVariant, from 4355 to 3202 in DNAscope), and dramatic reduction in FP InDels (from 20699 to 2766 in GATK, from 8008 to 2124 in DeepVariant, from 8632 to 1620 in DNAscope) as well as in FN InDels (from 23154 to

6345 in GATK, from 13082 to 3690 in DeepVariant, from 14301 to 3299 in DNAscope). This leads to slight increase of SNP F-score (harmonic mean of recall and precision) and significant increase of InDels F-score for PCR-free libraries, suggesting more precise variant calling for all depths (Figure 3a). As the highest accuracy combination, PCR-free data with DNAscope had the lowest FP SNPs, FP InDels, and FN InDels. As an additional evaluation, in selected "difficult" genome regions such as repeat regions, extreme GC regions, PCR-free generally showed better InDels F-score than PCR libraries, that output more faithful genome sequences for applications (Figure 2S).

To increase the confidence in the performance of PCR-free WGS workflow across a variety of sequencing depth, we generated additional 10x and 20x data point to conduct a 10x-46x low depth test. As being demonstrated in Figure 3B, the reduction in coverage (i.e. 10x, 15x and 20x) clearly affected the quality of variant calling from all methods. Nevertheless, PCR-free method coupled with machine learning-based callers produced more accurate calling than pipelines involving PCR library construction or GATK caller. Importantly, the SNP F-scores of DNAscope and DeepVarient for 15x PCR-free libraries were comparable to GATK for 30x PCR libraries (99.45%, 99.48% vs 99.70%), while InDel F-scores of DNAscope and DeepVarient for 15x PCR-free libraries were

significantly higher than GATK for 30x PCR libraries (97.90%, 98.23% vs 95.43%), indicating the potential application to lower sequencing cost while maintain the variant detection accuracy level. Noticeably, here DeepVariant for 15X PCR-free libraries showed the highest accuracy among all callers.

Reproducibility of PCR-free libraries

We also conducted consistency analysis, in order to determine the level of randomness introduced in library construction and sequencing step, and whether variant callers can help correct them. As a result, variant consistency of three replicates in PCR-free libraries was observed better than PCR libraries, especially for InDel consistency and the trend is similar across the three different variant callers (Figure 4). This is expected since PCR step inevitably introduces random errors during amplification. InDel consistency (represented by portion of variants shared by all replicates) of PCR-free libraries were found at GATK 84.2%, DeepVariant 86.5%, DNAscope 89.1% (three callers in average 86.6%), which were almost 20% higher than that of PCR libraries with GATK 63.2%, DeepVariant 66.9%, DNAscope 68.9% (three callers in average 66.3%).

We also observed more than 99% SNPs and approximately 98% InDels in each library overlapping with at least one of the other two libraries. Consistency of SNPs among all three replicates were similarly high from both PCR and PCR-free libraries (three callers in average 94.6% vs 95.2%).

On the other hand, we investigated the consensus calling on the same library among three callers. In average, the three pipelines for SNPs and InDels are 92.3% and 67.0% for PCR libraries as well as 92.6% and 82.7% for PCR-free libraries, indicating that PCR-free libraries generated more “clear” variants candidates that are less challenging for variant callers (Figure 3S).

Evaluation of PCR-free WGS performance on different sequencing platforms

In addition to library construction kit itself, sequencing platform also brings in additional bias or systematic variation that causes different performance. Here we compared two PCR-free libraries made by both MGIEasy PCR-Free DNA Library Pre Set and MGIEasy FS PCR-Free DNA Library Pre Set and sequenced on MGISEQ-2000, with three datasets downloaded from Illumina Basespace website to represent performance of

TruSeq PCR-free libraries sequenced on HiSeq4000, HiSeqXTen, or Novaseq platforms. Included in the comparison were three additional datasets in order to provide further information: 1) library prepared with MGIEasy PCR-Free DNA Library Pre Set and sequenced on DNBSEQ-T7 [38]; 2) library prepared with MGIEasy PCR-Free DNA Library Pre Set with illumina's adapter and sequenced on Novaseq by a third party sequencing service provider; 3) library prepared with a research protocol that has yet been incorporated into the MGIEasy kit and sequenced on MGISEQ-2000. It should be noted that MGIEasy kit prepared libraries used 1 μg or 250ng DNA input, far less than the input amount of datasets downloaded from Illumina Basespace website. All FASTQ files were processed in a same pipeline for reads trimming/filtering and mapping, as well as variant calling using DNAscope pipeline (Methods and Materials), to minimize bias or variation introduced at secondary analysis stage.

From mapping QC matrix, all three datasets generaged from MGISEQ-2000 showed significant lower duplicate rate at around or less than 1%, comparing to that from all Illumina's platforms with at least 10% duplicate rate, including the hybrid sample with

MGI library prep (Table 3). The "T7"dataset had 3.60% duplication rate with 250ng genomic DNA as input, still lower than ILMN's duplicate rate. This resulted in a more cost-effective read utility from DNBSEQ™. Meanwhile, with DNAscope's DNBSEQ™ and Illumina models applied separately, SNP calling accuracy (represented by F-score) of all samples reached a similarly high level, except for Hiseq4000 (Table 3). For InDels, the two pure MGI pipeline (library prep + sequencing) generated datasets and the "research library dataset" all outperformed the three ILMN datasets, indicating less errors (FP and FN) generated due to the completely elimination of PCR during both library construction and sequencing procedures. Comparing the top performer from both sides: the "research library dataset" and the "TruSeq and Novaseq dataset", the "research library dataset" had similar SNP calling but significant lower false InDels calling - 1304 in FP, making a ~55% reduction from 2879 on "TruSeq and Novaseq dataset", and 2201 FN as ~48% reduction as well. The inclusion of data from new sequencer DNBSEQ-T7 provided us with an opportunity to place its accuracy performance among the sequencer matrix here. Although prioritizing scale and cost in designing this ultra-high throughput sequencer, its sequencing accuracy was not comprimized, still reached the level of "TruSeq and Novaseq dataset".

These data collectively highlighted that the cost as well as data quality such as duplication rate InDel calling accuracy could be tremendously benefited from the complete DNBSEQ™ PCR-free WGS workflow.

Discussion

The widely adoption of WGS gives the credit to dramatic decreasing of the cost of sequencing, as well as the potential downside of WES, such as missing variants in the non-exome regions and incapable of detecting copy number variation. The development of machine-learning based secondary analysis tools also promote the usage of WGS, by shortening the analysis time from days to hours while reaching a higher variant calling accuracy.

Advantages of DNBSEQ™ PCR-free

In this study, we performed the PCR-free library prepare combined with DNBSEQ™ excluding PCR from the entire sequencing process. This leads to a good performance of GC bias and coverage. We showed the PCR-free libraries had more accurate small

variant especially InDel calls at 30x and other sequencing depth dataset. This result confirmed a previous report that the PCR-free library had high quality of InDel calls [31]. And this advantage expanded when the machine learning variant callers, such as DeepVariant and DNAscope, is applied to replace GATK. The fact that InDel F-score of PCR-free library called by both machine learning callers at 15x depth is significantly better than that of PCR library called by GATK at 30x implies a potential to use low depth sequencing data to conduct variant calling while not compromising any accuracy. Furthermore, true PCR-free WGS provided by good quality PCR-free libraries and DNBSEQ™ array is expected to improve detection of low frequency somatic mutations.

It is known that duplicate rate represents the proportion of the duplicate reads from all the sequenced data. In order to ensure the accuracy, the duplicated reads need to be removed for the following bioinformatics analysis. Therefore, for the same amount of raw data, lower duplicate rate means more usable clean data. The average duplicate rate of our PCR-free data on MGISEQ-2000 is about 1% , which is much lower than the PCR-free libraries (10.43%-13.62%) from the Illumina platforms. So for 30x WGS , 90-100G

raw base is normally enough on MGISEQ-2000 , however, about 20% more raw base (110-120G) is requested on Illumina platforms. The superiority of MGISEQ-2000 duplicate rate is due to the “true PCR-free” process, which strictly has no PCR step both in the library construction and sequencing workflow.

Rapid WGS for infant genetic disorders diagnostic

It was reported that genetic disorders are among the top cause of morbidity and mortality in infants, and the newly developed Rapid whole-genome sequencing (rWGS) shows the power to diagnose genetic disorders in a timely manner enabling healthcare providers to generate or change management plan accordingly and thus improves outcomes in acutely ill infants [32]. To facilitate this clinical utility, it is essential for the whole process from sample collection to diagnostic report review and signing can be complete within 2 days. As a result, there is urgent need for pushing turn-around time of library construction, sequencing, secondary analysis down to 24 hours, to open up some space for candidate variant clinical annotation and board review.

Compared to traditional PCR involved library construction method, the PCR-free method skipped amplification and therefore significantly reduced the turn-around-time. Moreover, the enzymatic shearing method makes it more feasible for automation and further save the library construction hands-on time.The new sequencer DNBSEQ-T7 with super high-thouput up to 6T per run can finish within 24h,which greatly shorten the sequencing time compared with MGISEQ-2000 and provides the possibility to realize rWGS. In the secondary analysis part, the most time-consuming steps are reads alignment and variant calling. While traditional BWA GATK pipeline takes more than desired computational time, some alternative pipelines have been proposed and developed to meet the accelerated speed requirement, including CPU based tools such as Strelka2 [33], DNAseq [34], and SpeedSeq [35], and FGPA&CPU based instrument such as DRAGEN system [36] and MegaBOLT system [37], as well as the GPU/TPU implemented Deepvariant [17]. All these advanced bioinformatics tools are compatible with data generated from MGI's library construction kit and MGI sequencing platforms.

Structural Variation (SV) and Copy Number Variation (CNV)

SV and CNV are two critical clinical utilities and for which researchers choose WGS instead of WES. Although not being designed as primary goal for this benchmark study, the data generated also allow us to investigate if PCR-free method improves SV and CNV detection accuracy. SV was called by DNAscope in default parameters and CNV was called by GATK 4.1.2 pipeline. From the result presented at Figure 4S it can be seen that PCR-free method likely helps improve both sensitivity and specificity of SV calling, comparing to PCR involved library construction method, as more SV events were called in PCR-free group while they reached higher consistency among three replicate samples. Key point in detecting SV events is correctly detecting breakpoints, which relies on sufficient coverage across targets and less error to generate false positive. Obviously, PCR-free libraries will benefit this process.

Germline CNV for all these 6 testing samples were called by GATK 4.1.2 and approximately 2500 CNV events (mainly deletions) were identified from each sample. When conducting 3-way comparison to analyze the repeatability among replicates, it can be seen PCR-free group showed slightly higher common called CNV events, but overall not much difference comparing to PCR group (Fig 5S).

Clinical Utility brought by higher accuracy

Clinical WGS starts to show its potential in rare disease diagnostic utility, as this application can quickly cover whole genome area, and finds clinical meaningful variants especially at UTR or promoter regions when panel or WES fail to cover. The increased variant detection accuracy from the utility of PCR-free library construction and machine learning based variant calling pipelines clearly increased WGS variant calling accuracy and therefore will definitely add value to the diagnostic application. On the other words, for the region PCR WGS fail to generate enough reads coverage or consistently generate wrong variant calling the PCR-free WGS will be able to provide with the correct SNP/InDel information. From the six DNBSEQ™ dataset and three ILMN dataset evaluated in this study, we found two example genes where all three PCR-free libraries showed accurate variant callings on the gene coding or UTR regions but the all three PCR libraries generated FN calling. These two genes were ATK1 and GNAS (Figure 6S, 7S). Similarly, we also showed one example gene “MAF”(Figure 8S), where DNBSEQ™ PCR-free method really stood out, as all 3 ILMN dataset and 2 MGI PCR dataset failed

to detect a C to CT insertion. These three genes code clinical meaningful proteins where a failed variant detection could lead to mis-diagnostic result. For example, gene ATK1 (ATK serine / threonine kinase 1) is associated with multiple clinical phenotypes including Breast cancer (MIM #114480), Colorectal cancer (MIM #114500), Cowden syndrome 6 (MIM #615109), and Ovarian cancer (MIM #167000). The T to TC insertion at locus Chr14:105262025, and CG to GC SNP at Chr14:105262041 may cause malfunction and introduce the disease phenotype, and only PCR-free libraries are capable to catch the critical causing variant.

Future improvement

As a benchmark project, this study managed to show the current performance of tools and pipelines used in library prep, sequencing, and data analysis. As expected, emerging techs will continuely push forward the upper limit of sequencing accuracy. For example, performance of new sequencer DNBSEQ-T7 [38] was briefly displayed in this study, which among other in developing sequencers, can bring down cost per genome or provide deeper coverage and further shorten the sequencing time when needed.

Weakness of the Structural Variation detection brought by the nature of short read sequencing can be greatly compensated by applying long fragment read (LFR) barcoding technologies [39,40,41]. For some clinical samples like cfDNA, it's impractical to obtain 200ng for library preparation. With developed method based on the MGI PCR-Free sets with pooling sequencing strategy, we successfully generated good data from 10ng cfDNA in some unpublished studies (Data not shown).

On the analysis side, current GATK best practice pipeline was developed and tuned based on Illumina data and did not officially support DNBSEQ™ generated data at least by the time of December 2018 according to GATK team's response in forum [42]. It is known that certain error correction steps in the pipeline such as BQSR and VQST was not developed for DNBSEQ™ data and thus may generate less than optimized result comparing to applying on Illumina sequencer generated data. Our suspect also comes from two recent published benchmark studies [43,44], where DNBSEQ™ data analyzed by Strelka2 and DeepVariant showed comparable accuracy with Illumina data, but DNBSEQ™ data analyzed by GATK returned a worse accuracy.

Both DeepVariant and DNAscope rely on the proper model training, which requires sufficient sample numbers from the library construction method and sequencing platform same as the testing samples would be generated through. In this study we do not believe the requirement was fully met, especially for DNAscope. For example, only 30x and higher depth dataset was included into training set but testing on 15x depth data was conducted. The negative effect of DNAscope lacking proper 15x depth training dataset can be obvious when comparing its 10x and 15x accuracy result with DeepVariant, whose training dataset included 15x depth data. Another point worth noticing is that using a single model for all library kit/assay and sequencers could sacrifice accuracy for individual cases. It is desirable for users to train individual model for each individual case (i.e., combination of sample prep kit, library construction kit, and sequencing platform), to achieve the optimal variant calling accuracy. With all the above-mentioned improvements for future WGS cohort studies, the unprecedented data generation speed and quality would help to answer difficult genetic questions and move the genomics field into a new era of broad clinical use.

Methods/Experimental

DNA preparation

Genomic DNA of NA12878 cell line (RRID:CVCL_7526) was purchased from the Coriell Institute. Genomic DNA was quantified by using Qubit™ 3 Fluorometer (Life Technologies, Paisley, UK). Size and quality of genomic DNA was confirmed by running 0.8% of agarose gel.

PCR-free library preparation

Acoustic fragmentation PCR-free libraries and enzymatic fragmentation PCR-free libraries were respectively constructed using MGIEasy PCR-Free DNA Library Prep Set (MGI, cat. no. 1000013453) and MGIEasy FS PCR-Free DNA Library Prep Set (MGI, cat. No. 1000013455) respectively.

For acoustic PCR-free libraries, genomic DNA was fragmented to 100-600 bp with peak size at 350-400 bp using LE220 Ultrasonicator (Covaris, Woburn, MA, USA). For FS PCR-free libraries, genomic DNA was fragmented to 100-1000 bp with peak size at 350-

475 bp using enzymatic shearing method. Subsequently, fragmented DNA with a size range of 200-450bp was selected using MGIEasy DNA Clean Beads (MGI, cat. no. 1000005279) and attached with DNBSEQ™ adapters following the set instruction. We also followed the protocols from the kit to make single-stranded DNA (ssDNA) circles and quantified them on Qubit™ 3 Fluorometer (Life Technologies, Carlsbad, CA, USA).

The library preparation procedure for research libraries was similar to that in MGIEasy PCR-Free DNA Library Prep Set, except for size selection and single strand degeneration method.

PCR library preparation

PCR libraries were prepared with the same procedure of DNA fragmentation, end repair and adapter ligation using MGIEasy PCR-Free DNA Library Prep Set (Cat. no. 1000013453), which was as descripted above. After adapter ligation, the reaction product was purified with MGIEasy DNA Clean Beads (MGI, cat. no. 1000005279) and the ligation products were subject for PCR amplification following instruction from KAPA HiFi HotStart ReadyMix (KAPA BIOSYSTEMS , KK2602) . A total of 6 cycles (95°C 3 min; 6 cycles, 98°C 20 s, 60°C 15 s, 72°C 60 s; 72°C 10 min; 4°C Forever) in a volume

of 100 ul was carried out. After beads purification using MGIEasy DNA Clean Beads and quantification using Qubit™ 3 Fluorometer, 1 pmol PCR product was taken for the single strand molecule circularization following the ssCir formation protocol from the MGIEasy kit.

Sequencing

Whole genome sequencing was performed on a DNBSEQ™ platforms MGISEQ-2000 for PE150, and DNBSEQ-T7 for PE150. Before sequencing, 75 fmol ssDNA of PCR-free library or 40fmol single strand circle DNA of PCR library was made into DNB (DNA Nanoball) following kit instruction from MGISEQ-2000RS High-throughput Sequencing Set (FCL PE150) (MGI, cat. no. 1000012555) . 75 fmol ssDNA of PCR-free library was made into DNB following kit instruction from DNBSEQ-T7RS High-throughput Sequencing Set (PE150) (MGI, cat. no. 1000016106). Subsequently, loading DNBs onto the sequencing chip was performed, and PE150 sequencing was conducted on DNBSEQ™ platforms using MGISEQ-2000RS or DNBSEQ-T7RS High-throughput Sequencing Set.

GC bias analysis

To explore GC bias performance, we sequenced two bacteria samples (Table 2S) on MGISEQ-2000 for PE150. Filtered reads were aligned to reference genomes by Burrows-Wheeler aligner (BWA) [45].To investigate the relationship between GC bias and read coverage, we scanned genome with a sliding window of the default size (100 bases). GC content and average read coverage were calculated within each window. Read coverage was normalized to the mean value so that the results would not scale with the total amount of data. In addition, we eliminated the data points whose coverage was more than twice the mean coverage because they likely represented repeats. Finally, we fit the remaining data points by a straight line and defined the slope as the degree of GC bias in the real data.

Reads filtering and mapping

As the first step, raw reads sequenced from PCR or PCR-free libraries were debarcoded by Seqtk [46] with default parameters. Then split reads were pre-processed by

SOAPnuke to generate filtered reads [47]. During this filtering process reads containing more than 10% of 'N' or 50% of the base quality score lower than 12 were removed. Meanwhile, adapters were trimmed off, which was followed by alignment of all reads against human reference genome with decoy sequencing hs37d5 (or reference genome sequences of two bacterias for GC bias analysis) using Burrows-Wheeler aligner (BWA) with default parameters. The output SAM file was converted to BAM file and sorted by Samtools [48]. Lastly, duplicates were marked by Picard [49] to make both BAM files variant calling ready as input for GATK, DeepVariant and DNAscope variant calling pipeline.

Running GATK

SNP and InDel calling were performed according to GATK (version 3.3) Best Practice [50]. Reads around InDel were re-aligned and base quality scores were recalibrated. HaplotypeCaller was used to call variants in gVCF mode on each chromosome. Genotyping on the gVCF files was performed by using GenotypeGVCFs with parameters as following: -stand_call_conf 30 and -stand_emit_conf 10. SNPs and InDels were

separated using SelectVariants tool. Variants quality score recalibrates (VQSR) was carried out to filter low quality variants. SNP annotation --ts_filter_level was used for calculation and filtered at a 99.9% level, while for InDel, --ts_filter_level was used for calculation and filtered at a 99.9% level of the true sensitivity.

Running DeepVariant

Taking advantages of the state-of-the-art deep-learning technique for image classification, DeepVariant (V0.7.0 in this study) can achieve a higher accuracy for bioinformatics analysis. The GIAB (Genome In A Bottle) truth set and corresponding fastq reads were utilized as training dataset to train a Convolutional Neural Network (CNN) model. As an alternative for GATK HaplotypeCaller, DeepVariant accepts aligned reads (e.g. BAM file) as input. In DeepVariant, candidate variants are carefully filtered along the genome and classified into three genotype states, homozygous reference (hom-ref), heterozygous (het) or homozygous alternate (hom-alt), with the previously trained CNN model.

To achieve best calling performance, we fine-tuned the CNN model in DeepVariant using a set of PCR-free data, including 30x and 15x DNBSEQ™ PCR-free sequencing data of HG001 and HG005 samples. The fine-tuned model is accessible at [51].

Running DNAscope

Sentieon DNAscope (version 201808.01 and 201808.05 were used in this study) uniquely combines the well-established methods from haplotype-based variant callers with machine learning to achieve improved accuracy. As a successor to GATK HaplotypeCaller, DNAscope uses an identical logical architecture of active region detection, local haplotype assembly, and read-likelihood calculation (Pair-HMM) to produce variant candidates. DNAscope outputs additional informative variant annotations used by the machine learning model. Annotated variant candidates are then passed to a machine learning model for variant genotyping resulting in improvements in both variant calling and genotyping accuracy.

For this study, DNBSEQ™ model for DNAscope was constructed using publicly available data from the HG001 and HG005 Genome in a Bottle samples downloaded from the NIST GiaB FTP site along with proprietary 30x HG001 samples. Illumina model for DNAscope was trained using a subset of the GiAB HG001 and HG005 data as well. None of the tested samples were used during model training. Training was performed across all chromosomes with the exception of chromosome 20.

Variant accuracy evaluation

All VCF files generated from this benchmark study were collected for accuracy evaluation. Firstly, they were separated into SNP and InDel subgroups and each subgroup was then compared against NIST truth set using Vcfeval from RTGtools [52] to calculate a F-score as representation of accuracy.

Acknowledgements

We would like to acknowledge the ongoing contributions and support of all MGI-Shenzhen and BGI research employees, and also acknowledge Frank Hu from Sentieon for commenting on the manuscript.

Funding

This work was supported by Shenzhen Peacock Plan.No.KQTD20150330171505310, the National Key R&D Program of China (2018YFC0910201), and the Key R&D Program of Guangdong Province (2019B020226001).

Availability of data and materials

The data reported in this study are available in the CNGB Nucleotide Sequence Archive.(https://db.cngb.org/cnsa; accession number CNP0000602,CNP0000466.) All the other data used here are included within the published article and its Additional files.

References

1. Meienberg J, Zerjavic K , Keller I, et al. New insights into the performance of human whole-exome capture platforms. *Nucleic Acids Res*, 2015, 43(11): e76.

2. Belkadi A, Bolze A, Itan Y, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci U S A*, 2015,112(17): 5473-5478.

3. Meienberg J, Bruggman R, Oexle K, et al.Clinical sequencing: is WGS the better WES? *Hum Genet*, 2016, 135: 359–362.

4. Carss KJ, Arno G, Erwood M,Stephens J, Sanchis-Juan A, Hull S, Megy K, Grozeva D, Dewhurst E, Malka S, et al. Comprehensive Rare Variant Analysis via Whole-Genome Sequencing to Determine the Molecular Pathology of Inherited Retinal DiseaseAn integrated map of structural variation. *Am J Hum Genet*, 2017, 100, 75-90.

5. Fujimoto A, Furuta M, Totoki Y, et al. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat Genet*, 2016, 48(5): 500-9.

6. Satta G, Lipman M, Smith GP, et al. Mycobacterium tuberculosis and whole-genome sequencing: how close are we to unleashing its full potential? *Clin Microbiol Infect*,2018, 24(6): 604-609.

7. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

8. Kebschull JM, Zador AM. Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res*, 2015, 43(21): e143.

9. Aird D, Ross MG, Chen WS, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol*, 2011, 12(2): R18.

10. Chen YC, Liu T, Yu CH, et al. Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly. *PLoS One*,2013, 8(4): e62856.

11. Ross MG, Russ C, Costello M, et al. Characterizing and measuring bias in sequence data. *Genome Biology*, 2013, 14(5): R51.

12. Kozarewa I, Ning Z, Quail MA, et al. Amplification-free Illumina sequencing preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods*, 2009, 6(4):291-5.

13. Jones MB,Highlander SK,Anderson EL,et al.Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proc Natl Acad Sci U S A*,2015,112 (45) 14024-14029.

14. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 2010, 20(9): 1297-303.

15. Heldenbrand JR, Baheti S, Bockol MA, et al.Performance benchmarking of GATK3.8 and GATK4. *bioRxiv*, 348565v1; doi: https://doi.org/10.1101/348565.

16. Friedman S. Deep learning in GATK4 [Internet]. GATK | Blog. 2017. Available from: https://software.broadinstitute.org/gatk/blog?id=10996.

17. Poplin R, Chang PC, Alexander D, et al. A universal SNP and small-InDel variant caller using deep neural networks. *Nat Biotechnol*, 2018, 36(10):983-987.

18. https://github.com/Sentieon/sentieon-dnascope-ml

19. Luo R, Sedlazeck FJ, Lam TW, et al. A multi-task convolutional deep neural network for variant calling in single molecule sequencing. *Nat Commun*, 2019, 10(1):998.

20. Adessi C1, Matton G, Ayala G, et al.Solid phase DNA amplification characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res*, 2000, 28(20): E87.

21. Blanco L, Bernad A, Lázaro JM,et al.Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication.*J Biol Chem*, 1989, 264(15): 8935-40.

22. Ali MM, Li F, Zhang Z, Zhang K, et al. Rolling circle amplification: a versatile tool for chemical biology, materials science and medicine. *Chem Soc Rev,* 2014, 43(10): 3324-41.

23. https://en.mgitech.cn/article/detail/285.html

24. Zhou X, Rokas A. Prevention, diagnosis and treatment of high-throughput sequencing data pathologies. *Mol Ecol*, 2014, 23(7): 1679-700.

25. Bansal V. A computational method for estimating the PCR duplication rate in DNA and RNA-seq experiments. *BMC Bioinformatics*, 2017, 18(Suppl 3): 43.

26. Cornish A, Guda C. A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. *Biomed Res Int*, 2015, 2015: 456479.

27. Hwang S, Kim E, Lee I, et al.Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci Rep,* 2015, 5: 17875.

28. Zook JM, Catoe D, McDaniel J, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data*, 2016, 3:160025.

29. Supernat A, Vidarsson OV, Steen VM, et al. Comparison of three variant callers for human whole genome sequencing. *Sci Rep*, 2018, 8(1): 17851.

30. https://precision.fda.gov/challenges/truth/results

31. Fang H, Wu Y, Narzisi G, et al. Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Med*, 2014, 6(10): 89.

32. Willig LK, Petrikin JE, Smith LD, et al. Whole-genome sequencing for identification of Mendelian disorders in critically ill infants: a retrospective analysis of diagnostic and clinical findings. *Lancet Respir Med* , 2015, 5: 377–87.

33. Kim S, Scheffler K, Halpern AL, et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods*, 2018, 15(8): 591-594.

34. Kendig KI, Baheti S, Bockol MA, et al. Sentieon DNASeq Variant Calling Workflow Demonstrates Strong Computational Performance and Accuracy. *Front Genet*, 2019, 10: 736.

35. Chiang C, Layer RM, Faust GG, et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat Methods,* 2015,12(10): 966-8.

36. Goyal A, Kwon HJ, Lee K, et al. Ultra-Fast Next Generation Human Genome Sequencing Data Processing Using DRAGENTM Bio-IT Processor for Precision Medicine. *Open Journal of Genetics,* 2017, 7, 9-19.

37. https://en.mgitech.cn/article/detail/germline.html

38. https://en.mgitech.cn/product/detail/DNBSEQ-T7.html

39. Peters BA, Kermani BG, Sparks AB, et al. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature*, 2012, 487(7406): 190-5.

40. Zheng GXY, Lau BT, Schnall-Levin M, et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotech*, 2016, 34: 303–311.

41. Wang O, Chin R, Cheng XF, et al. Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. *Genome Res*, 2019, 29: 798-808.

42. https://gatkforums.broadinstitute.org/gatk/discussion/23202/does-sequencing-platform-have-effect-on-the-variant-detection

43. https://blog.dnanexus.com/2018-05-31-training-and-applying-genomic-deep-learning-models/

44. Chen J, Li X, Zhong H, et al. Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers. *Sci Rep*, 2019, 9(1):9345.

45. Li H. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv*: 1303. 3997.

46. https://github.com/lh3/seqtk

47. Chen Y, Chen Y, Shi C, et al. SOAPnuke: A MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *GigaScience,* 2018, 7: 1–6.

48. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and samtools. *Bioinformatics,* 2009, 25(16): 2078-9.

49. https://github.com/broadinstitute/picard

50. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 2010, 20: 1297–303.

51. https://github.com/MGI-tech-bioinformatics/MGI_DeepVariant_model

52. https://github.com/RealTimeGenomics/rtg-tools

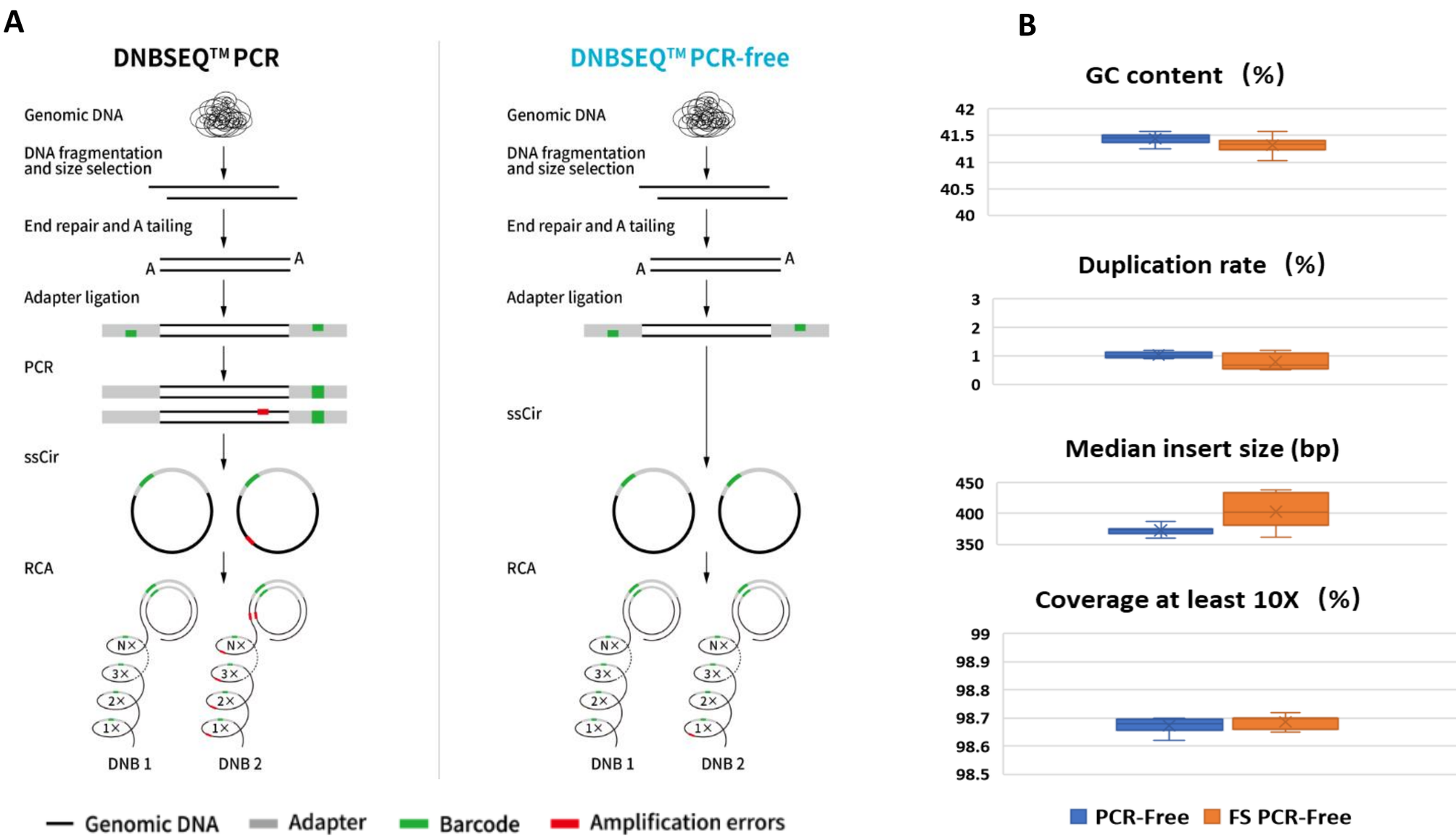


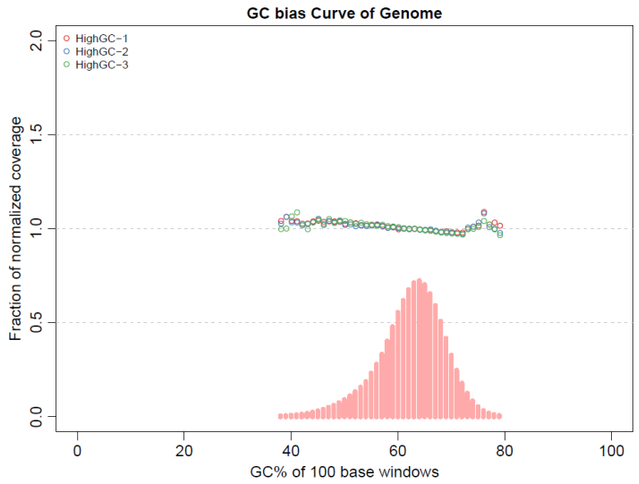
Fig1. DNBSEQ™ WGS PCR vs PCR-free workflows and general performance of PCR-free libraries. (a) MPS library construction workflows of WGS PCR and PCR-free libraries. RCA (rolling circle amplification) is used to increase signal intensity during array formation, which is followed by sequencing DNBs with DNBSEQ™ technology. Individual copies from the same DNB are replicated independently using the same ssCir template. Therefore, amplification errors cannot accumulate. Black, genomic DNA; Gray rectangle, adapter; green, barcode; red, amplification errors. (b) Two sets of 9 replicates from 1 μg NA12878 DNA were processed with MGIEasy PCR-Free DNA Library Prep Set (blue) or MGIEasy FS PCR-Free DNA Library Prep Set (orange). The GC content, Duplication rate, Median Insert size, and regions with >10x Coverage were calculated and plotted. The error bars represent the standard deviations across the replicates.

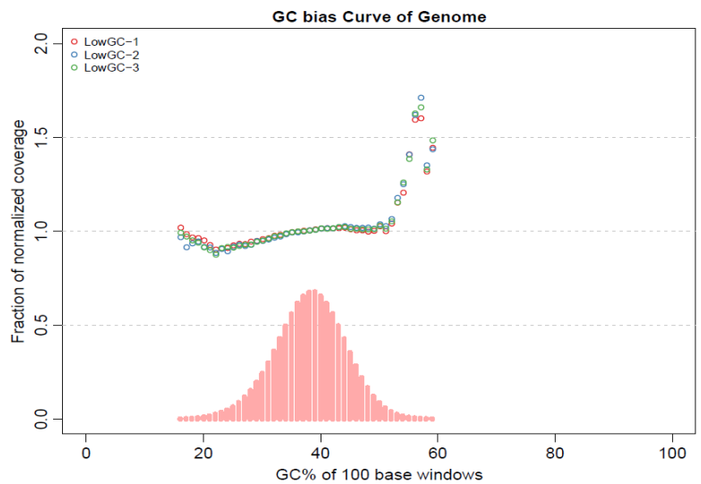
A

High GC (62%)

Low GC (38%)

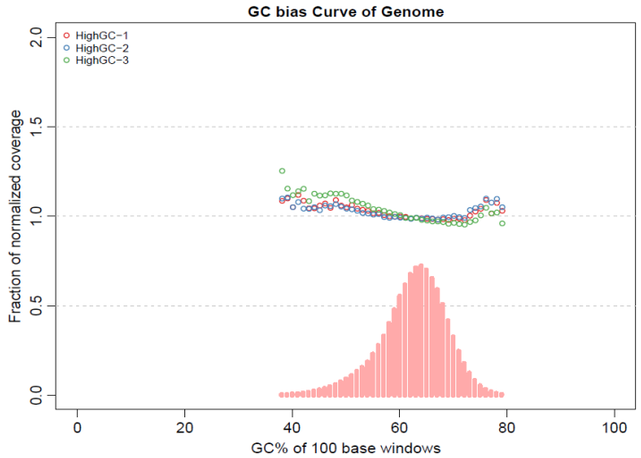
PCR-Free





B

FS PCR-Free



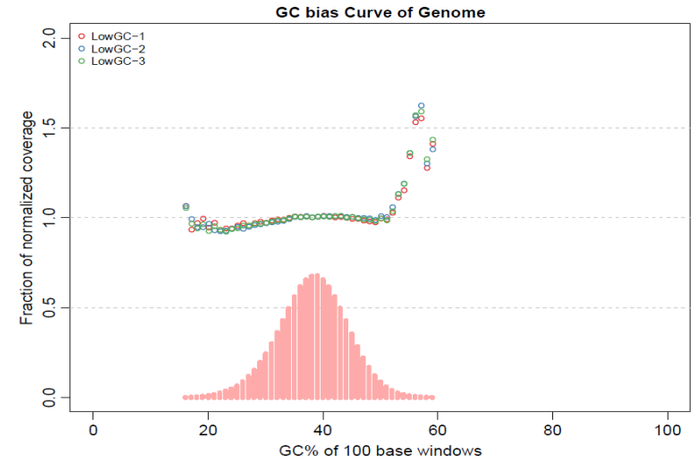
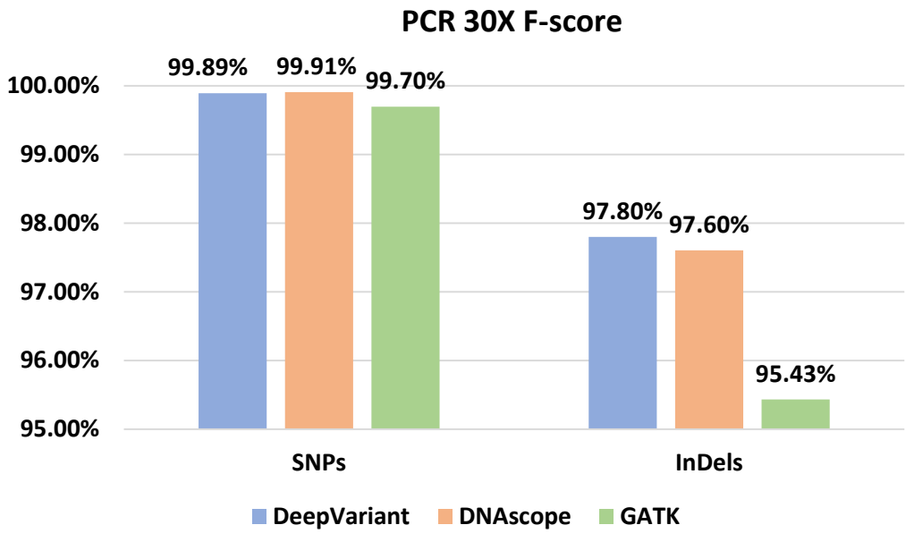
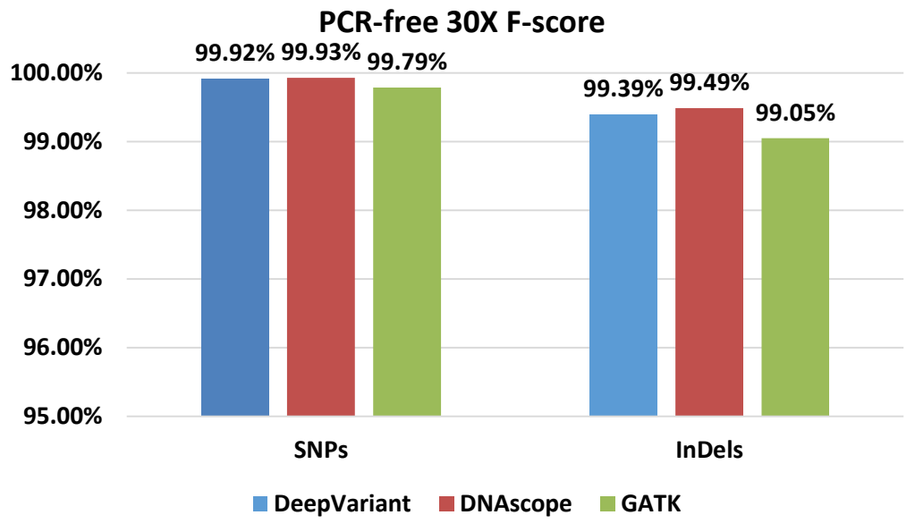


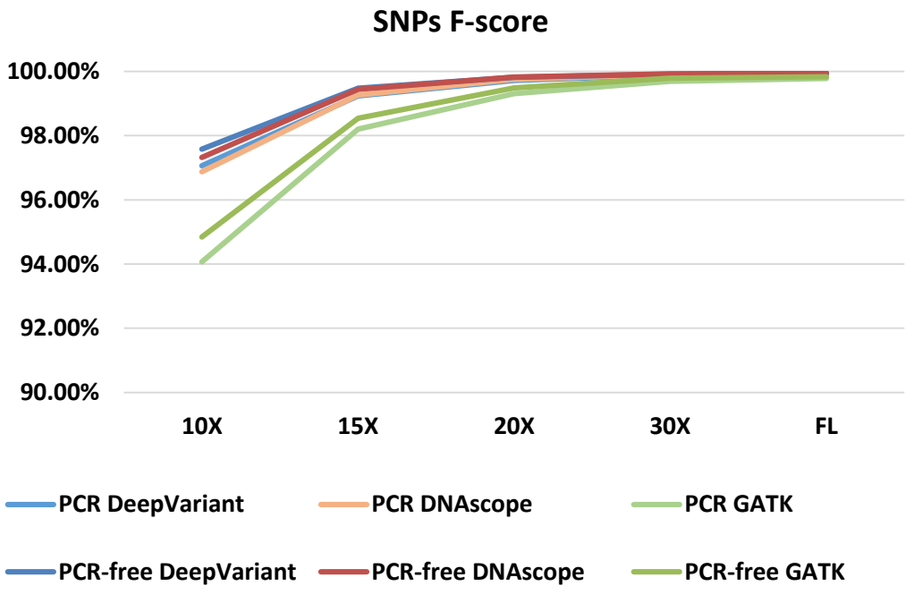
Figure 2. Coverage of the microbial genomes, *Olsenella profusa* (62% GC) (left) and *Bacillus megaterium* (38% GC) (right) with (A) MGIEasy PCR-Free DNA Library Pre Set and (B) MGIEasy FS PCR-Free DNA Library Pre Set. Read coverage across the range of the GC content, calculated in 100 bp windows (pink bars) and normalized coverage (colored dots). 3 replicates (red, blue and green dots) were included in the normalized coverage VS GC content analysis.

A





B



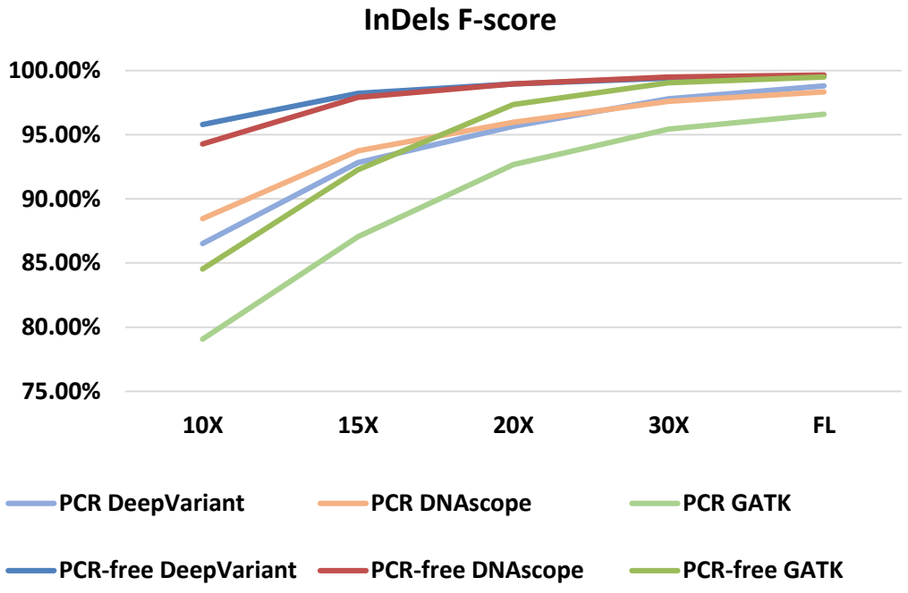


Figure 3. variant calling performance in PCR vs PCR-free libraries (A) F-Scores of 30x sequencing PCR and PCR-free libraries were compared for the accuracy performance of different variant callers. (B) Down-sampled dataset representing different sequencing depth were processed to exam the tolerance to shallow data from different variant callers. PCR-free libraries (Dark color) and machine learning variant callers (blue and red) showed good accuracy at >15X, and even better than PCR+GATK combination at 30X depth.

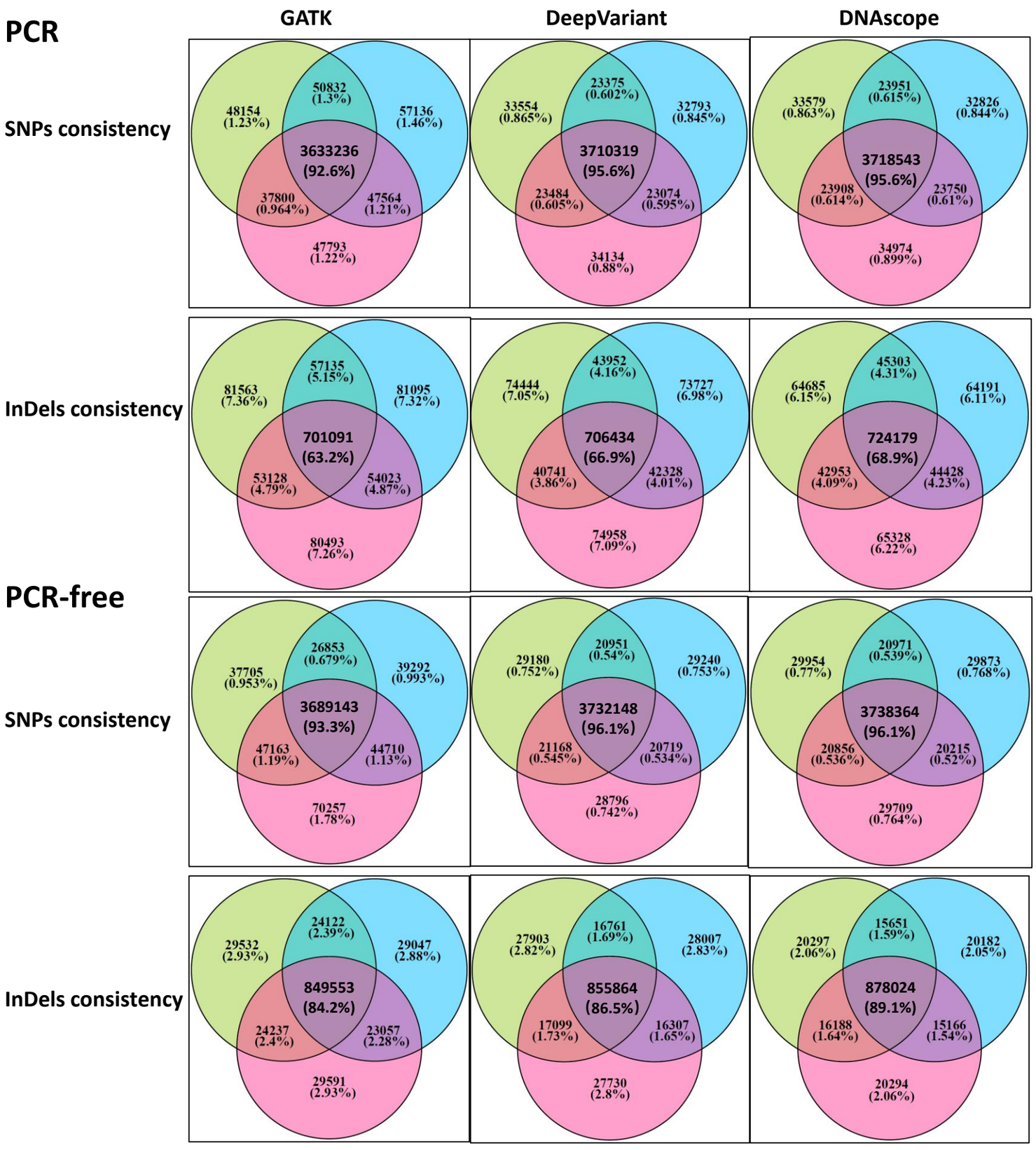


Figure4. Variant consistency of 3 replicates of PCR and PCR-free libraries Consistency analysis were conducted on the 3 libraries generated by the same library kit and variant calling pipelines. Venn-Diagram were generated to show the common shared variants and the unique variants.

Method	PCR-1			PCR-2			PCR-3			PCR-free-1			PCR-free-2			PCR-free-3		
Depth	**15x**	**30x**	**FL**	**15x**	**30x**	**FL**	**15x**	**30x**	**FL**	**15x**	**30x**	**FL**	**15x**	**30x**	**FL**	**15x**	**30x**	**FL**
Clean reads (x10^6)	296.64	633.73	940.37	296.80	633.76	954.92	296.45	632.78	949.34	295.93	631.36	926.38	296.03	632.98	944.92	296.13	631.49	964.32
Clean bases (Gb)	44.49	95.05	141.05	44.52	95.06	143.23	44.46	94.91	142.40	44.39	94.70	138.95	44.40	94.94	141.73	44.42	94.72	144.64
Insert size peak (bp)	383	383	383	374	374	374	375	375	375	374	374	375	375	375	375	375	375	375
GC content (%)	41.64	41.64	41.64	41.78	41.78	41.78	41.49	41.49	41.49	41.53	41.53	41.53	41.48	41.48	41.48	41.46	41.47	41.46
Mapping rate (%)	99.91	99.82	99.82	99.91	99.84	99.84	99.89	99.82	99.82	99.89	99.83	99.83	99.72	99.65	99.65	99.93	99.86	99.86
Unique rate (%)	99.42	95.41	94.78	99.48	95.53	94.92	99.35	95.23	94.51	99.69	95.76	95.38	99.71	95.82	95.44	99.77	95.95	95.61
Duplicate rate (%)	0.58	1.56	2.24	0.52	1.44	2.10	0.65	1.72	2.49	0.31	1.07	1.50	0.29	1.03	1.45	0.23	0.91	1.31
Mismatch rate (%)	0.89	0.87	0.86	0.91	0.89	0.88	0.95	0.93	0.92	0.74	0.74	0.73	0.74	0.74	0.73	0.71	0.70	0.70
Average seq depth (X)	15.37	30.83	45.46	15.39	30.89	46.27	15.34	30.71	45.74	15.37	30.81	45.04	15.35	30.86	45.91	15.4	30.91	47.04
Coverage (%)	99.07	99.13	99.16	99.07	99.13	99.16	99.08	99.14	99.17	99.1	99.16	99.18	99.1	99.16	99.18	99.1	99.16	99.18
Coverage at least 4X (%)	98.65	98.96	99.03	98.64	98.97	99.03	98.65	98.97	99.04	98.76	99.01	99.06	98.75	99.01	99.06	98.74	99.00	99.06
Coverage at least 10X (%)	89.78	98.58	98.83	89.61	98.58	98.84	89.55	98.58	98.84	90.5	98.70	98.89	90.44	98.69	98.89	90.49	98.69	98.90

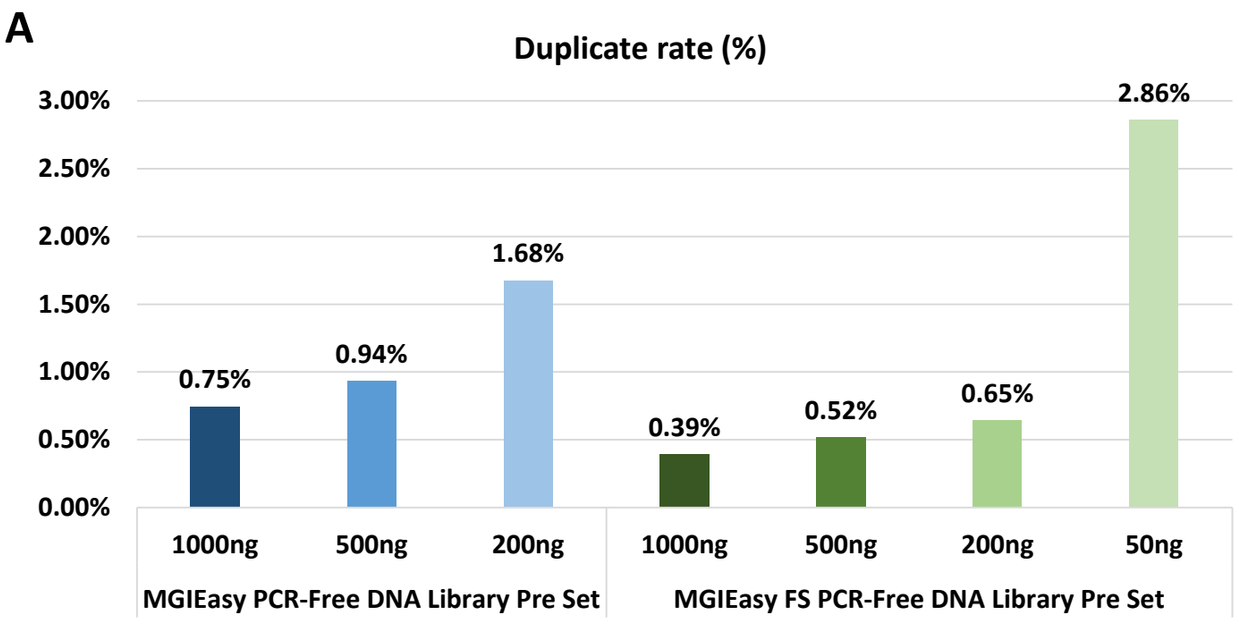
Table1. Mapping performance of three replicates for PCR and PCR-free libraries QC matrix from 3 PCR and 3 PCR-free libraries was collected and showed. Each sample was down sampled to 15x and 30x as well as FL (full lane). "FL" represents Full Lane sequencing data amount approximately equals 46x coverage.

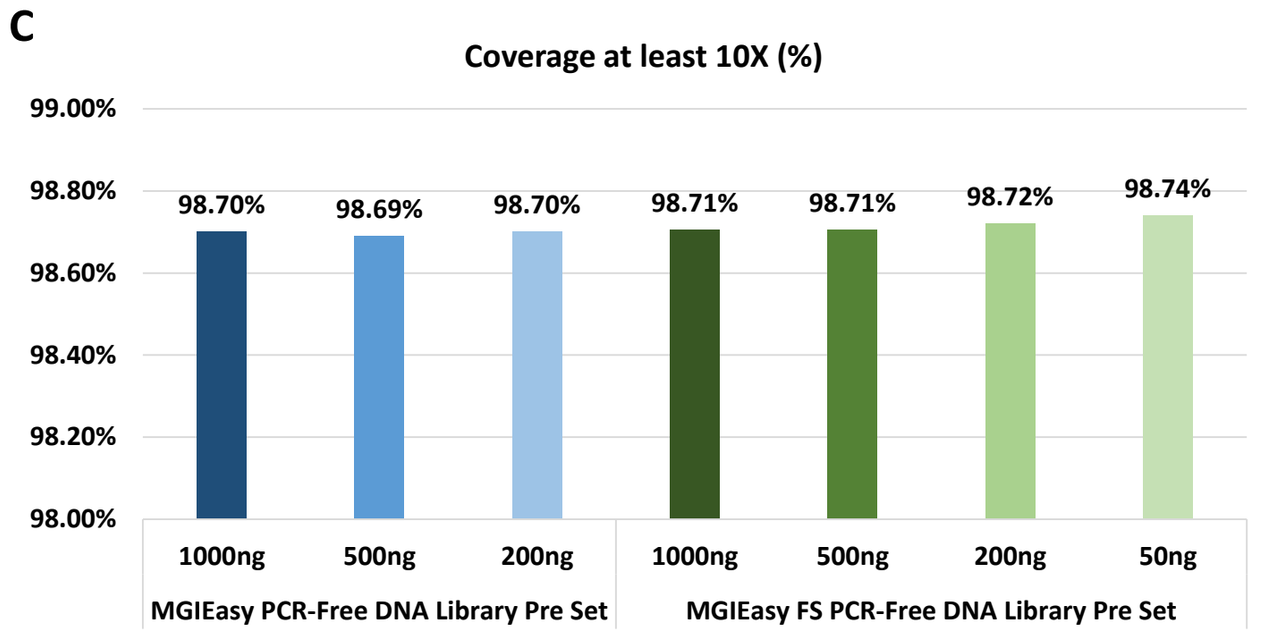
Pipeline	Variant type	Method	PCR			PCR-free		
		Depth	15x	30x	FL	15x	30x	FL
GATK	SNPs	True positive	3106478	3193586	3198933	3126614	3199331	3202461
		False positive	9615	2806	2479	8544	2586	2347
		False negative	103778	16671	11324	83643	10926	7796
		Precision	99.69%	99.91%	99.92%	99.73%	99.92%	99.93%
		Sensitivity	96.77%	99.48%	99.65%	97.39%	99.66%	99.76%
		F-score	98.21%	99.70%	99.79%	98.55%	99.79%	99.84%
	InDels	True positive	397779	458110	466343	429105	474921	477918
		False positive	34787	20699	17891	19730	2766	1686
		False negative	83485	23154	14922	52160	6345	3349
		Precision	91.96%	95.67%	96.30%	95.60%	99.42%	99.65%
		Sensitivity	82.65%	95.19%	96.90%	89.16%	98.68%	99.31%
		F-score	87.05%	95.43%	96.60%	92.27%	99.05%	99.48%
DeepVariant	SNPs	True positive	3174776	3205991	3207021	3184757	3207148	3207769
		False positive	12952	2913	1871	7837	2111	1691
		False negative	35481	4266	3236	25499	3109	2488
		Precision	99.59%	99.91%	99.94%	99.75%	99.93%	99.95%
		Sensitivity	98.90%	99.87%	99.90%	99.21%	99.90%	99.92%
		F-score	99.24%	99.89%	99.92%	99.48%	99.92%	99.93%
	InDels	True positive	436804	468237	474129	470682	477616	478556
		False positive	22906	8008	4389	6310	2124	1615
		False negative	44546	13082	7185	10695	3690	2745
		Precision	95.02%	98.32%	99.08%	98.67%	99.56%	99.66%
		Sensitivity	90.74%	97.28%	98.51%	97.78%	99.23%	99.43%
		F-score	92.83%	97.80%	98.80%	98.23%	99.39%	99.55%
DNAscope	SNPs	True positive	3171728	3205902	3207318	3180864	3207055	3207911
		False positive	8351	1565	1021	6034	1254	901
		False negative	38529	4355	2939	29393	3202	2346
		Precision	99.74%	99.95%	99.97%	99.81%	99.96%	99.97%
		Sensitivity	98.80%	99.86%	99.91%	99.08%	99.90%	99.93%
		F-score	99.27%	99.91%	99.94%	99.45%	99.93%	99.95%
	InDels	True positive	443656	466964	471037	466852	477969	479103
		False positive	21653	8632	5773	5617	1620	1202
		False negative	37609	14301	10229	14413	3299	2164
		Precision	95.35%	98.19%	98.79%	98.81%	99.66%	99.75%
		Sensitivity	92.18%	97.03%	97.88%	97.00%	99.31%	99.55%
		F-score	93.74%	97.60%	98.33%	97.90%	99.49%	99.65%

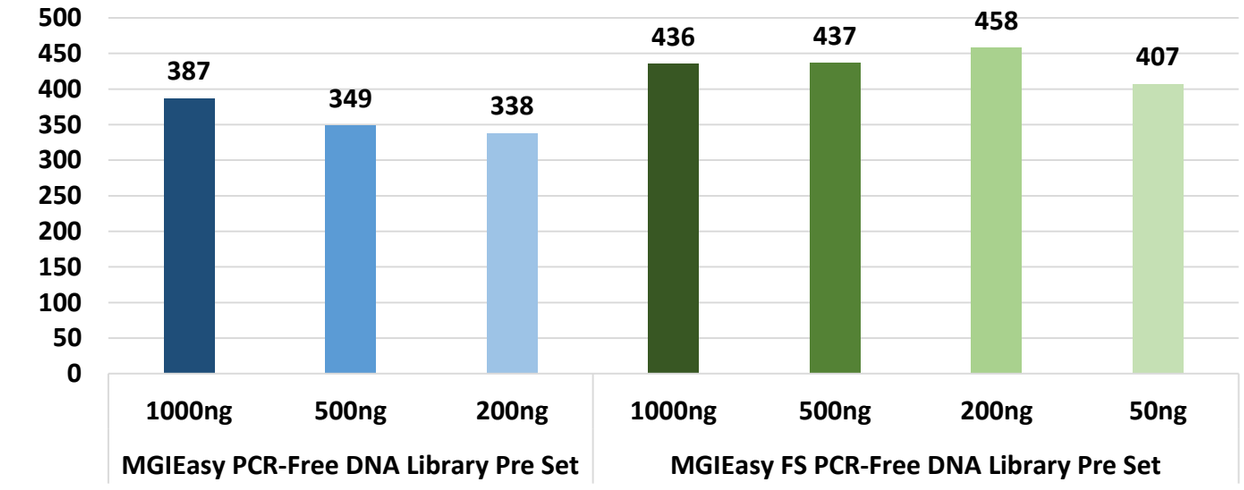
Table2. Average variant calling performance of three replicates for PCR and PCR-free libraries using three variant callers Variant called from each library and variant caller was evaluated by Vcfeval tool in RTGtools against NIST truth set at high confident regions. Then average values from the same library construction method were generated and showed in table. "FL" represents Full Lane sequencing data amount approximately equals 46x coverage.

Library Kit		MGIEasy PCR-Free DNA Library Prep Set	MGIEasy FS PCR-Free DNA Library Prep Set	MGIEasy PCR-Free DNA Library Prep Set	MGIEasy PCR-Free DNA Library Prep Set	PCR-Free research lib	TruSeq DNA PCR-Free Library Prep Kits		
DNA input		250ng	1μg	1μg	1μg	1μg	2μg	2μg	2μg
Sequence platform		DNBSEQ-T7	MGISEQ-2000	MGISEQ-2000	Novaseq	MGISEQ-2000	Hiseq4000	xTen	Novaseq
Average seq depth (X)		30.46	30.68	30.91	30.75	30.6	30.02	31.5	30.43
Mapping rate (%)		99.99	99.89	99.86	99.82	99.83	99.03	99.85	99.94
Duplicate rate (%)		3.6	0.6	0.91	18.19	0.76	13.62	12.31	10.43
Coverage at least 10X (%)		98.64	98.73	98.69	98.83	98.82	98.8	98.91	98.87
SNPs	True positive	3206204	3207301	3207056	3207540	3207839	3206782	3207834	3208037
	False positive	1545	1209	1255	1331	1048	2704	1280	1262
	False negative	4053	2956	3201	2717	2418	3475	2423	2220
	Precision	99.95%	99.96%	99.96%	99.96%	99.97%	99.92%	99.96%	99.96%
	Sensitivity	99.87%	99.91%	99.90%	99.92%	99.92%	99.89%	99.92%	99.93%
	F-score	99.91%	99.94%	99.93%	99.94%	99.95%	99.90%	99.94%	99.95%
Indels	True positive	476673	478047	478004	477283	479066	469353	477420	477017
	False positive	2266	1882	1595	2479	1304	7642	2598	2879
	False negative	4592	3218	3265	3984	2201	11913	3846	4250
	Precision	99.53%	99.61%	99.67%	99.48%	99.73%	98.40%	99.46%	99.40%
	Sensitivity	99.05%	99.33%	99.32%	99.17%	99.54%	97.52%	99.20%	99.12%
	F-score	99.29%	99.47%	99.49%	99.33%	99.64%	97.96%	99.33%	99.26%

Table3. Mapping and Variant calling performance of PCR-free libraries sequenced on different sequencing platforms DNBSEQ-T7,MGISEQ-2000 data was generated following kit instructions. Illumina Hiseq4000, HiSeq xTen, Novaseq data was downloaded from Illumina Basespace website. To compare the sequence platform only，one PCR-free library was generated by MGIEasy FS PCR-Free DNA Library Prep Set with illumina’ s adapter and sequenced on Novaseq by a third party sequencing service provider. To further improve the library preparation method, one additional PCR-free library was prepared in house using a research protocol for the library construction and sequenced on MGISEQ-2000. Variants from each library was called using DNAscope pipeline and accuracy was evaluated by vcfeval tool in RTGtools against NIST truth set at high confident regions.







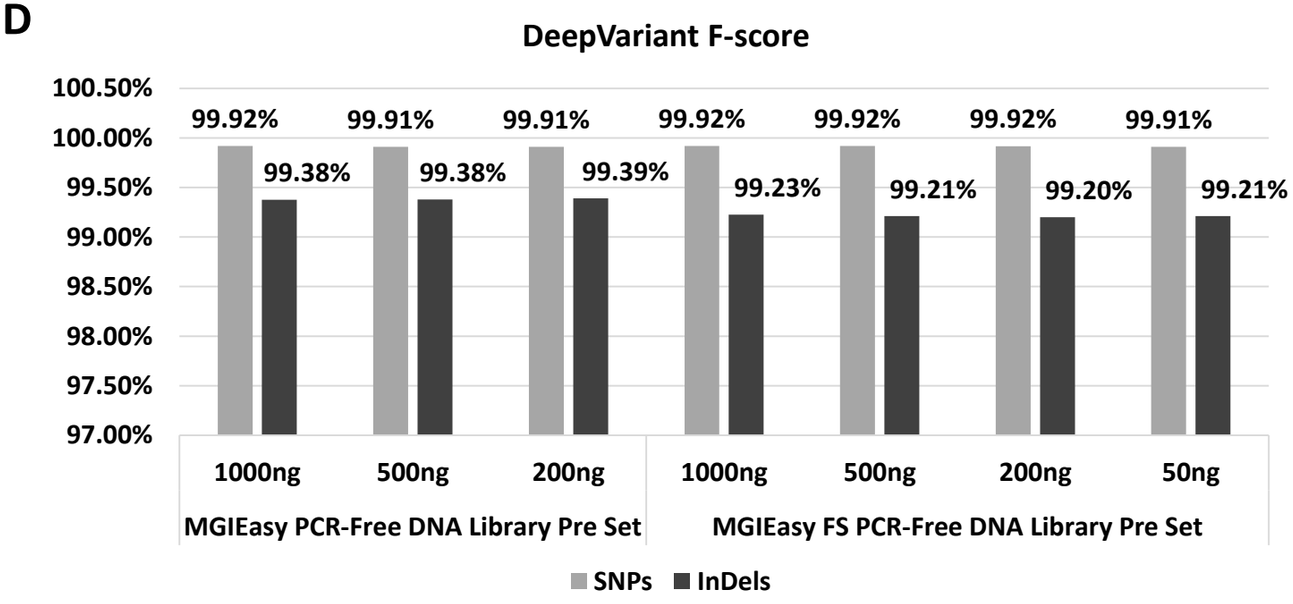


Fig1S. The data performance with different library inputs using either ultrasonic shearing (MGIEasy PCR-Free DNA Library Prep Set) or enzymatic shearing method (MGIEasy FS PCR-Free DNA Library Prep Set) Each condition was repeated three times, and the average value was calculated for display.

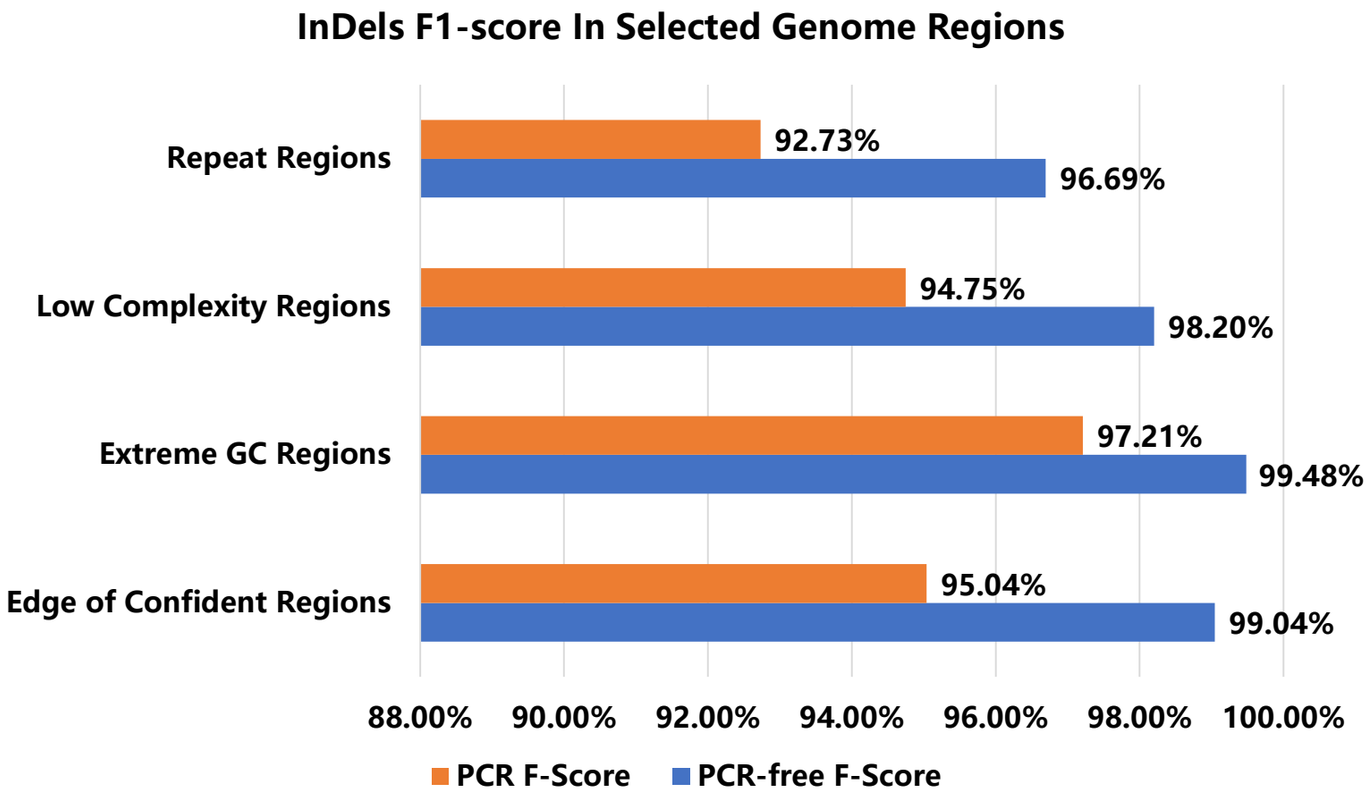


Figure 2S. InDels F-score in Selected Genome Regions In comparison between one PCR sample and PCR-free sample, a number of genome subsets regions were selected and F1-scores generated from Sentieon DNAscope were compared within each region. InDels showed more significant differences between two samples, and especially in 4 categories above. These categories are selected from GA4GH stratification regions with category name changed for easier understand. Repeat Regions matches category “lowcmp_AllRepeats_51to200bp_gt95identity_merged”; Low Complexity Regions matches category “lowcmp_Human_Full_Genome_TRDB_hg19_150331_all_merged”; Extreme GC Regions matches category “gclt30orgt55”; Edge of Confident Regions matches category "TS_boundary”.

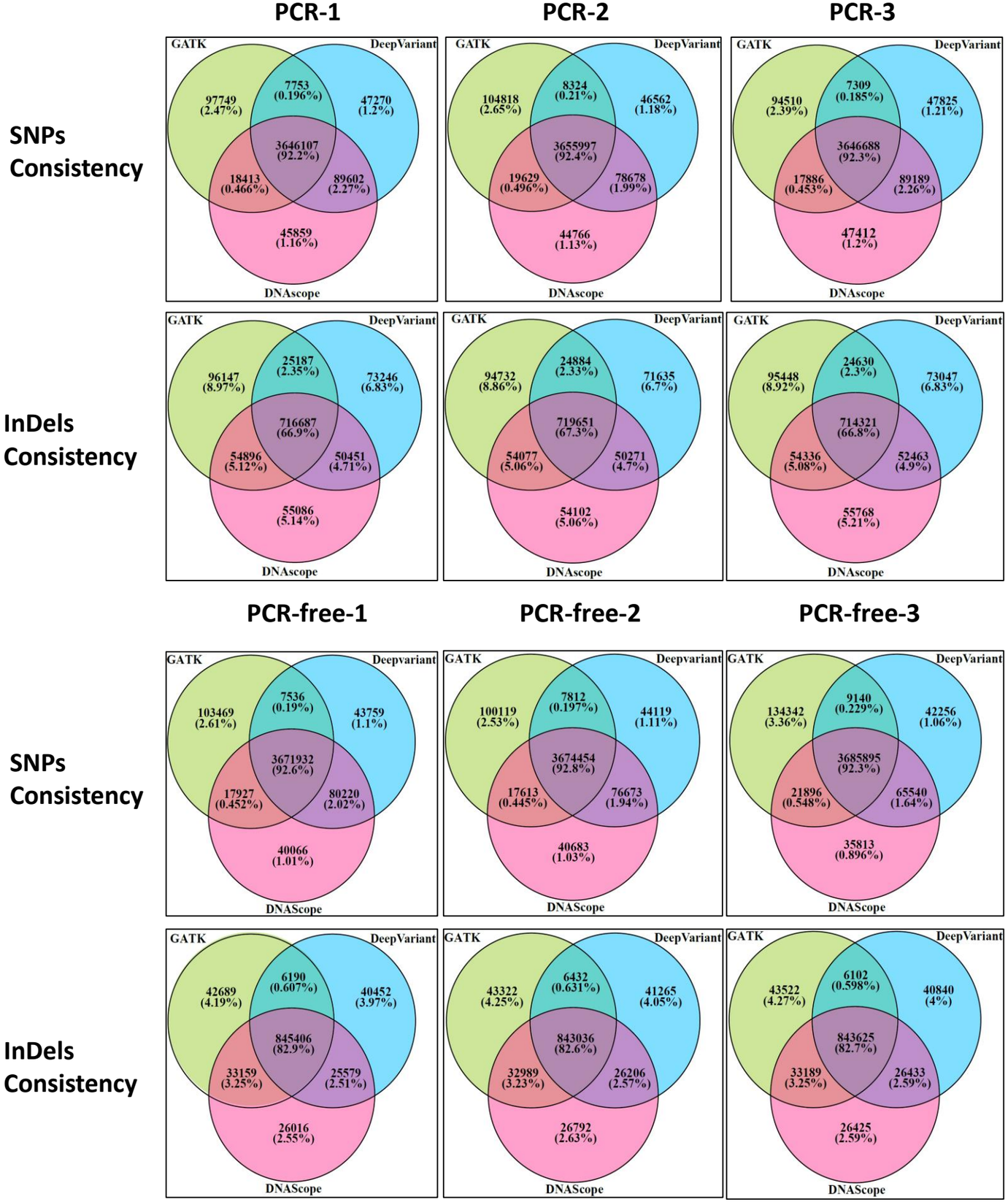


Fig 3S. Consistency of three pipelines for SNPs and InDels. Consistency analysis were conducted on the 3 variant calling pipelines on the same library. Venn-Diagram were generated to show the common shared variants and the unique variants.

A

Samples	TRA	DEL	DUP	INV	INS
PCR-1_30X	203	3528	175	227	99
PCR-2_30X	205	3488	138	224	99
PCR-3_30X	204	3477	172	225	95
PCR-free-1_30X	293	3684	189	231	104
PCR-free-2_30X	288	3625	182	254	106
PCR-free-3_30X	251	3689	168	256	123

B

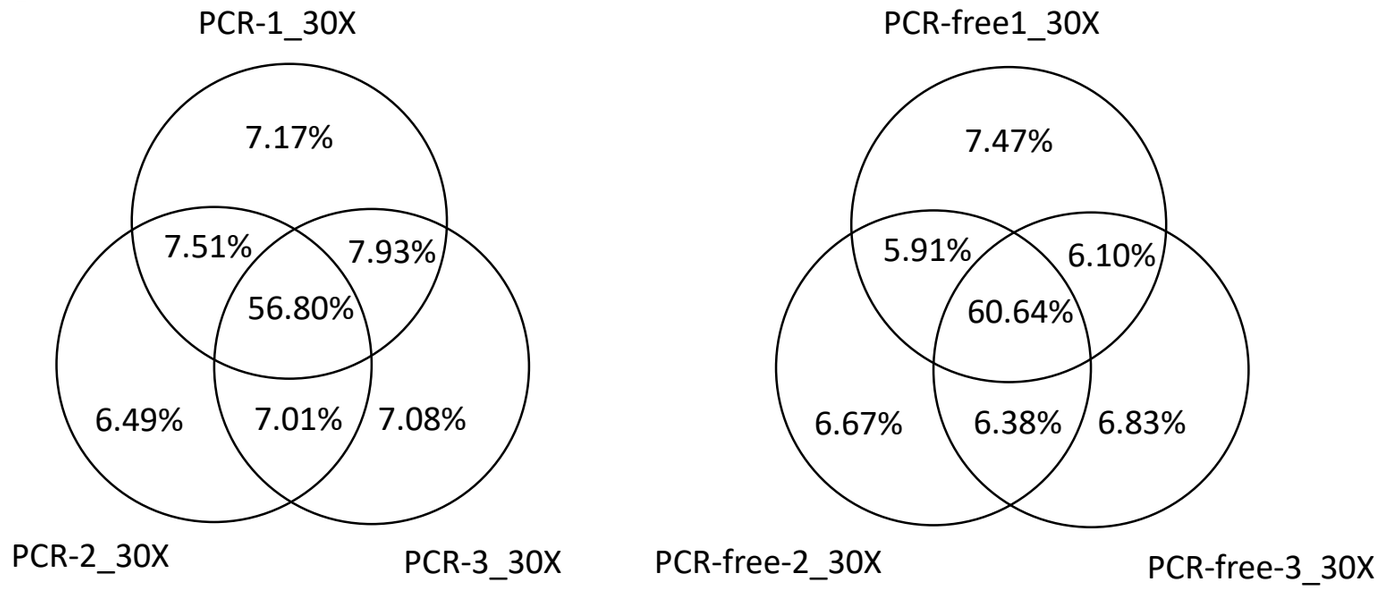


Figure 4S. (A) SV events detected by DNAscope in 6 testing samples. SV events were reported into one of the five categories: Translocation, Deletion, Duplication, Inversion, and Insertion. PCR-free libraries showed higher number of reported SV events, and (B) higher consistency among all 3 samples as well.

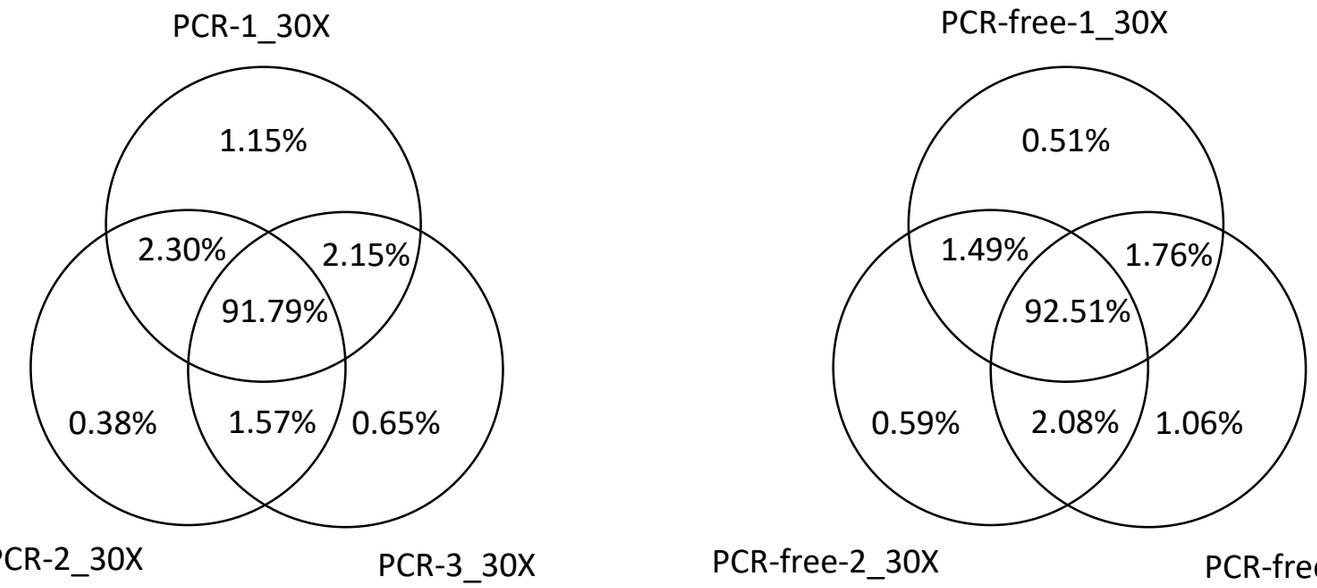


Figure 5S. CNV events detected by DNAscope in 6 testing samples. 3-way comparison was conducted to analysis consistency among replicates.

Pipline	Variant type	Method	PCR-1			PCR-2			PCR-3			PCR-free-1			PCR-free-2			PCR-free-3		
		Depth	15x	30x	FL	15x	30x	FL	15x	30x	FL	15x	30x	FL	15x	30x	FL	15x	30x	FL
GATK	SNPs	True positive	3110618	3193126	3200655	3108081	3195584	3197789	3100734	3192049	3198355	3128229	3198582	3201565	3123297	3198921	3202891	3128317	3200489	3202928
		False positive	9634	2770	2750	9352	2939	2205	9858	2708	2483	8708	2403	2570	8559	2268	2385	8364	3086	2087
		False negative	99638	17131	9602	102175	14673	12468	109522	18208	11902	82028	11675	8692	86960	11336	7366	81940	9768	7329
		Precision	99.69%	99.91%	99.91%	99.70%	99.91%	99.93%	99.68%	99.92%	99.92%	99.72%	99.92%	99.92%	99.73%	99.93%	99.93%	99.73%	99.90%	99.93%
		Sensitivity	96.90%	99.47%	99.70%	96.82%	99.54%	99.61%	96.59%	99.43%	99.63%	97.44%	99.64%	99.73%	97.29%	99.65%	99.77%	97.45%	99.70%	99.77%
		F-score	98.27%	99.69%	99.81%	98.24%	99.73%	99.77%	98.11%	99.67%	99.78%	98.57%	99.78%	99.82%	98.49%	99.79%	99.85%	98.58%	99.80%	99.85%
	InDels	True positive	398324	458096	466055	398912	458728	466804	396100	457507	466171	429199	475007	477924	428632	474850	477933	429485	474906	477897
		False positive	35000	21111	18550	34291	20094	17163	35069	20892	17960	19737	2741	1688	19793	2775	1719	19661	2782	1650
		False negative	82941	23168	15210	82352	22536	14462	85162	23757	15094	52066	6258	3343	52632	6417	3334	51783	6361	3370
		Precision	91.92%	95.59%	96.17%	92.08%	95.80%	96.45%	91.87%	95.63%	96.29%	95.60%	99.43%	99.65%	95.59%	99.42%	99.64%	95.62%	99.42%	99.66%
		Sensitivity	82.77%	95.19%	96.84%	82.89%	95.32%	96.99%	82.30%	95.06%	96.86%	89.18%	98.70%	99.31%	89.06%	98.67%	99.31%	89.24%	98.68%	99.30%
		F-score	87.10%	95.39%	96.50%	87.24%	95.56%	96.72%	86.82%	95.35%	96.58%	92.28%	99.06%	99.48%	92.21%	99.04%	99.47%	92.32%	99.05%	99.48%
DeepVariant	SNPs	True positive	3175227	3206052	3207074	3174871	3205898	3206874	3174231	3206023	3207115	3185071	3207185	3207765	3184356	3207120	3207784	3184844	3207140	3207757
		False positive	12381	2879	1875	13175	2917	1886	13301	2942	1851	7973	2119	1700	7830	2138	1698	7709	2077	1675
		False negative	35030	4205	3182	35386	4359	3383	36026	4234	3142	25185	3072	2492	25900	3137	2473	25413	3117	2500
		Precision	99.61%	99.91%	99.94%	99.59%	99.91%	99.94%	99.58%	99.91%	99.94%	99.75%	99.93%	99.95%	99.75%	99.93%	99.95%	99.76%	99.94%	99.95%
		Sensitivity	98.91%	99.87%	99.90%	98.90%	99.86%	99.89%	98.88%	99.87%	99.90%	99.22%	99.90%	99.92%	99.19%	99.90%	99.92%	99.21%	99.90%	99.92%
		F-score	99.26%	99.89%	99.92%	99.24%	99.89%	99.92%	99.23%	99.89%	99.92%	99.48%	99.92%	99.93%	99.47%	99.92%	99.94%	99.48%	99.92%	99.93%
	InDels	True positive	436833	468201	474042	437513	468409	474219	436066	468101	474127	470823	477670	478584	470466	477597	478546	470756	477582	478538
		False positive	23254	8085	4458	22359	7786	4257	23105	8152	4453	6233	2111	1614	6514	2120	1614	6182	2141	1616
		False negative	44515	13120	7283	43834	12915	7093	45290	13212	7179	10553	3641	2719	10915	3711	2756	10617	3718	2760
		Precision	94.95%	98.30%	99.07%	95.14%	98.36%	99.11%	94.97%	98.29%	99.07%	98.69%	99.56%	99.66%	98.63%	99.56%	99.66%	98.70%	99.55%	99.66%
		Sensitivity	90.75%	97.27%	98.49%	90.89%	97.32%	98.53%	90.59%	97.25%	98.51%	97.81%	99.24%	99.44%	97.73%	99.23%	99.43%	97.79%	99.23%	99.43%
		F-score	92.80%	97.79%	98.78%	92.97%	97.84%	98.82%	92.73%	97.77%	98.79%	98.25%	99.40%	99.55%	98.18%	99.39%	99.55%	98.25%	99.39%	99.54%
DNAscope	SNPs	True positive	3172347	3205898	3207358	3172082	3205807	3207242	3170755	3206001	3207353	3181144	3207111	3207908	3180305	3206997	3207905	3181142	3207056	3207920
		False positive	8291	1522	996	8325	1577	1046	8438	1595	1021	6050	1269	927	5969	1237	888	6083	1255	889
		False negative	37909	4359	2899	38175	4450	3015	39502	4256	2904	29113	3146	2349	29951	3260	2352	29115	3201	2337
		Precision	99.74%	99.95%	99.97%	99.74%	99.95%	99.97%	99.73%	99.95%	99.97%	99.81%	99.96%	99.97%	99.81%	99.96%	99.97%	99.81%	99.96%	99.97%
		Sensitivity	98.82%	99.86%	99.91%	98.81%	99.86%	99.91%	98.77%	99.87%	99.91%	99.09%	99.90%	99.93%	99.07%	99.90%	99.93%	99.09%	99.90%	99.93%
		F-score	99.28%	99.91%	99.94%	99.27%	99.91%	99.94%	99.25%	99.91%	99.94%	99.45%	99.93%	99.95%	99.44%	99.93%	99.95%	99.45%	99.93%	99.95%
	InDels	True positive	443499	466856	470799	444329	467200	471196	443141	466836	471115	467078	477999	479118	466450	477903	479072	467027	478004	479118
		False positive	21902	8738	6002	21250	8438	5540	21807	8721	5777	5615	1617	1194	5775	1648	1217	5460	1595	1195
		False negative	37767	14409	10466	36938	14066	10070	38121	14429	10150	14187	3268	2149	14815	3365	2194	14238	3265	2149
		Precision	95.29%	98.16%	98.74%	95.44%	98.23%	98.84%	95.31%	98.17%	98.79%	98.81%	99.66%	99.75%	98.78%	99.66%	99.75%	98.84%	99.67%	99.75%
		Sensitivity	92.15%	97.01%	97.83%	92.32%	97.08%	97.91%	92.08%	97.00%	97.89%	97.05%	99.32%	99.55%	96.92%	99.30%	99.54%	97.04%	99.32%	99.55%
		F-score	93.70%	97.58%	98.28%	93.85%	97.65%	98.37%	93.67%	97.58%	98.34%	97.92%	99.49%	99.65%	97.84%	99.48%	99.65%	97.93%	99.49%	99.65%

Table1S. Variant calling performance of PCR and PCR-free with three variant callers Variant called from each library and variant caller was evaluated by Vcfeval tool in RTGtools against NIST truth set at high confident regions.

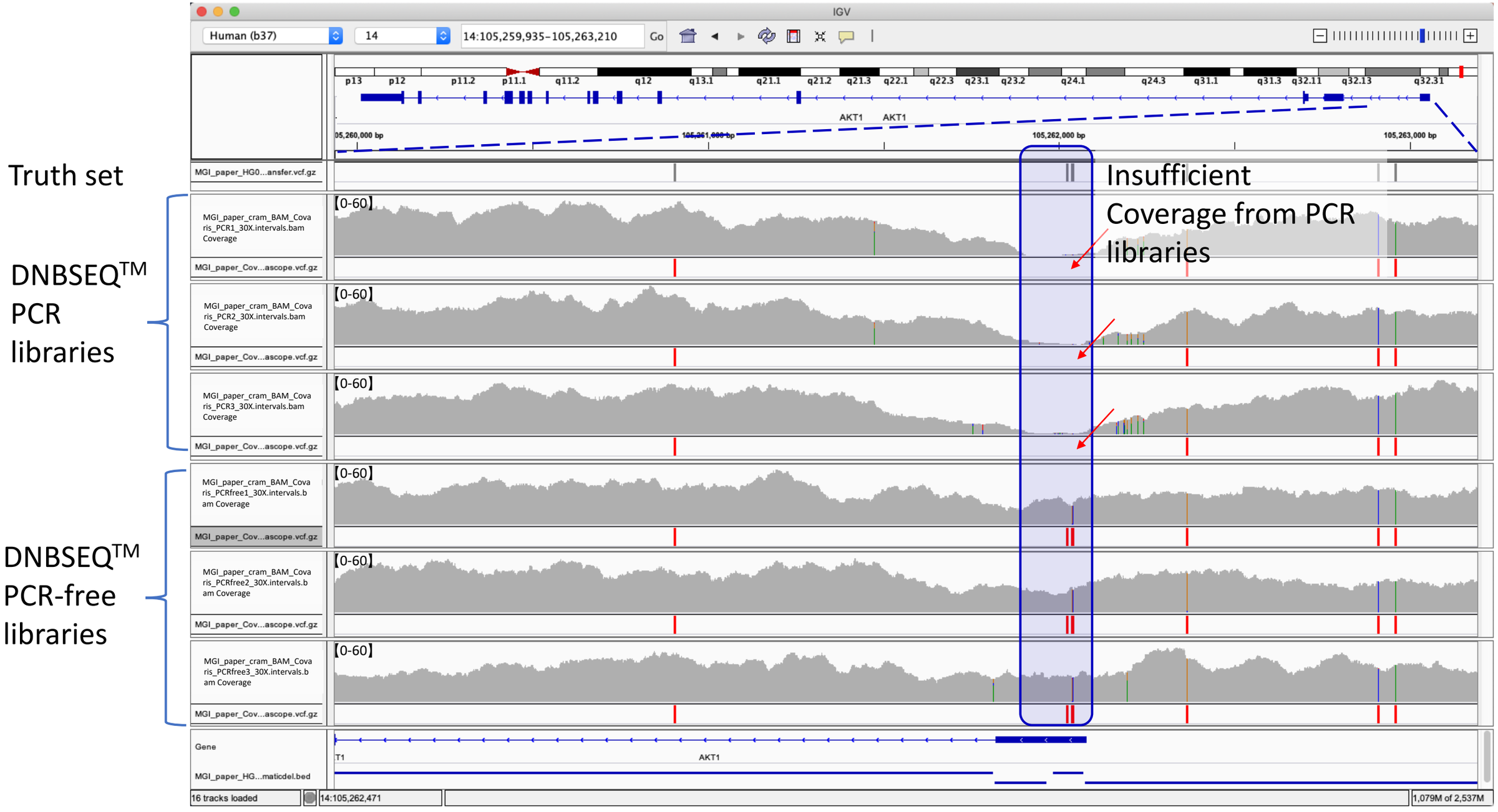


Figure 6S. False negative detection in DNBSEQ™ PCR libraries that may cause mis-diagnostic in clinical utilities related to ATK1 gene. Clinical Phenotype: Cancer, Cowden syndrome 6; All PCR libraries failed to detect: Chr14: 105262025 T to TC insertion at Exon1 UTR, and Chr14: 105262041 CG to GC SNP at Exon1 UTR.

Truth set

DNBSEQ™ PCR libraries

DNBSEQ™ PCR-free libraries

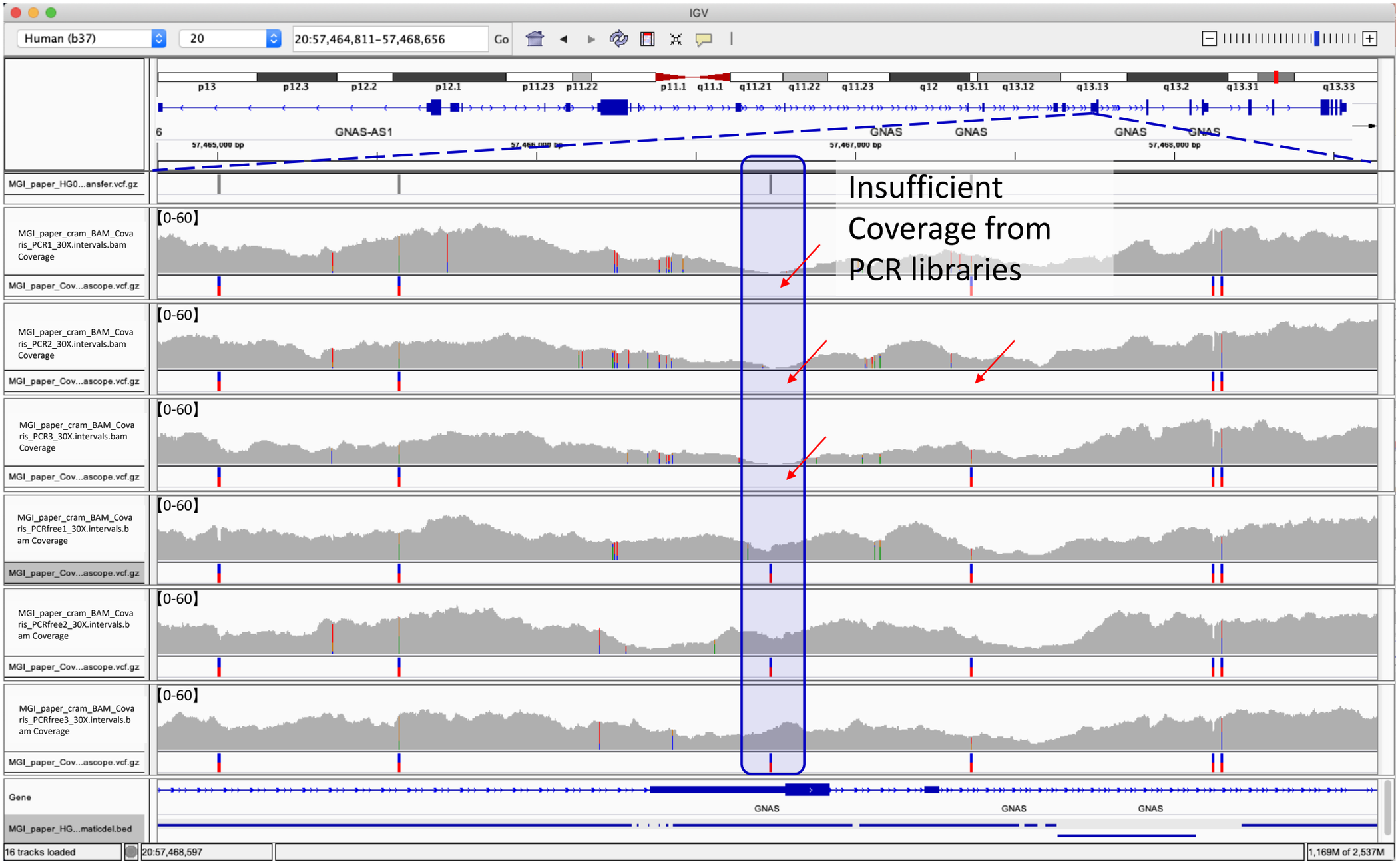


Figure 7S. False negative detection in DNBSEQ™ PCR libraries that may cause mis-diagnostic in clinical utilities related to GNAS gene. Clinical Phenotype: Pseudohypoparathyroidism; All PCR libraries failed to detect: Chr20: 57466734 C to CCCG insertion at Exon1 UTR.

Figure 8S. False negative detection in 3 ILMN libraries and 2 DNBSEQ™ PCR libraries that may cause mis-diagnostic in clinical utilities related to MAF gene. Clinical Phenotypes: Ayme-Gripp syndrome; Cataract 21, multiple types; All ILMN libraries failed to detect: Chr16: 79632062 C to CT insertion at Exon1, which was picked up by all DNBSEQ™ PCR free libraries at even 15x depth.