



## OPEN ACCESS

## EDITED BY

Honey V. Reddi,  
Medical College of Wisconsin,  
United States

## REVIEWED BY

Shrey Sukhadia,  
Dartmouth Hitchcock Medical Center,  
United States  
Qian Nie,  
Medical College of Wisconsin,  
United States

## \*CORRESPONDENCE

Lei Sun,  
✉ sunlei@genemind.com

<sup>†</sup>These authors have contributed equally to this work

RECEIVED 14 September 2023

ACCEPTED 14 December 2023

PUBLISHED 04 January 2024

## CITATION

Feng B, Lai J, Fan X, Liu Y, Wang M, Wu P, Zhou Z, Yan Q and Sun L (2024), Systematic comparison of variant calling pipelines of target genome sequencing cross multiple next-generation sequencers. *Front. Genet.* 14:1293974. doi: 10.3389/fgene.2023.1293974

## COPYRIGHT

© 2024 Feng, Lai, Fan, Liu, Wang, Wu, Zhou, Yan and Sun. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Systematic comparison of variant calling pipelines of target genome sequencing cross multiple next-generation sequencers

Baosheng Feng<sup>1†</sup>, Juan Lai<sup>1†</sup>, Xue Fan<sup>2</sup>, Yongfeng Liu<sup>1</sup>, Miao Wang<sup>1</sup>, Ping Wu<sup>1</sup>, Zhiliang Zhou<sup>1</sup>, Qin Yan<sup>1</sup> and Lei Sun<sup>1\*</sup>

<sup>1</sup>GeneMind Biosciences Company Limited, Shenzhen, China, <sup>2</sup>Clinical Research Institute, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

Targeted genomic sequencing (TS) greatly benefits precision oncology by rapidly detecting genetic variations with better accuracy and sensitivity owing to its high sequencing depth. Multiple sequencing platforms and variant calling tools are available for TS, making it exruciating for researchers to choose. Therefore, benchmarking study across different platforms and pipelines available for TS is imperative. In this study, we performed a TS of Reference OncoSpan FFPE (HD832) sample enriched by TSO500 panel using four commercially available sequencers, and analyzed the output 50 datasets using five commonly-used bioinformatics pipelines. We systematically investigated the sequencing quality and variant detection sensitivity, expecting to provide optimal recommendations for future research. Four sequencing platforms returned highly concordant results in terms of base quality (Q20 > 94%), sequencing coverage (>97%) and depth (>2000x). Benchmarking revealed good concordance of variant calling across different platforms and pipelines, among which, FASTASeq 300 platform showed the highest sensitivity (100%) and precision (100%) in high-confidence variants calling when analyzed by SNVer and VarScan 2 algorithms. Furthermore, this sequencer demonstrated the shortest sequencing time (~21 h) at the sequencing mode PE150. Through the intersection of 50 datasets generated in this study, we recommended a novel set of variant genes outside the truth set published by HD832, expecting to replenish HD832 for future research on tumor variant diagnosis. Besides, we applied these five tools to another panel (TargetSeq One) for Twist cfDNA Pan-cancer Reference Standard, comprehensive consideration of SNP and InDel sensitivity, SNVer and VarScan 2 performed best among them. Furthermore, SNVer and VarScan 2 also performed best for six cancer cell lines samples regarding SNP and InDel sensitivity. Considering the dissimilarity of variant calls across different pipelines for datasets from the same platform, we recommended an integration of multiple tools to improve variant calling sensitivity and accuracy for the cancer genome. Illumina and GeneMind technologies can be used independently or together by public health laboratories performing tumor TS. SNVer and VarScan 2 perform better regarding variant detection sensitivity for three typical tumor samples. Our study provides a standardized target sequencing resource to benchmark new bioinformatics protocols and sequencing platforms.

## KEYWORDS

HD832, TSO500, NGS, target genome sequencing, SNVer, Varscan 2

## 1 Introduction

Targeted genomic sequencing (TS) is an effective next-generation sequencing (NGS) method that focuses on a panel of genes or targets implicated in the pathogenesis or clinical relevance. It greatly benefits precision oncology by rapidly detecting genetic variations, providing a resource for clinicians to help them interpret genetic profiles, and implement personalized anticancer recommendations. TS shows better accuracy and sensitivity in identifying targeted variations owing to greater sequencing depth at the same sequencing cost and data burden when compared with whole-genome sequencing (WGS) or whole-exome sequencing (WES) (Bewicke-Copley et al., 2019). Thus, it allows for identifying mutations presenting low variant allele frequencies (VAFs) with high confidence, especially for low-quality and fragmented clinical DNA samples. Clinical targeted sequencing has revolutionized tumor surveillance and diagnosis and facilitated the development of precision oncology (Nagahashi et al., 2019; Tan et al., 2020a).

TruSight™ Oncology 500 (TSO500, Illumina) is a comprehensive target-sequencing panel covering more than eight cancer types and 523 cancer-related genes (1.94 Mb) to identify relevant genomic variants and signatures in a single assay (Zhao et al., 2020). It uses a hybridization capture-based target-enrichment strategy to detect diverse variants with high specificity and sensitivity, including single nucleotide polymorphisms (SNPs), InDels, copy number variations, splice variants and gene fusions, especially for mutants with low VAFs via tactfully suppressing technical noise and excluding germline variants (Zhao et al., 2020; Conroy et al., 2021). Notably, TSO500 is also robust for tumor mutational burden (TMB) detection with a relatively high concordance compared with WES in different solid tumor samples (Wei et al., 2022).

Currently, TSO500 panel has been benchmarked and verified against orthogonal validated NGS assays (such as WGS, WES) using diverse clinical tissue specimens across different tumor types (Pestinger et al., 2020; Sahajpal et al., 2020; Conroy et al., 2021; Ramos-Paradas et al., 2021; Wei et al., 2022). The high concordance with the reference methods demonstrates that the TSO500 assay is reliable for accurately detecting gene alterations to support the precision therapeutics in oncology. While NGS has revolutionized the understanding of disease diagnosis and prediction, the magnitude of sequencing data points toward the potential challenges for advancing large-scale executions—the appropriate platform and bioinformatics pipeline capable of handling these data efficiently in a timely and accurate manner. In this regard, the need for benchmarking across different sequencing platforms and bioinformatics analysis tools for TS supporting precision diagnostics is imperative. Regrettably, the systematical benchmarking studies of the TSO500 panel on different sequencing platforms and bioinformatics pipelines are still blank. In addition, the Twist Pan-cancer Reference Standard is a high-quality, standardized control for researching and developing NGS-based liquid biopsy assays, primarily used to track the quality of an NGS assay workflow and assess the fidelity of the assay process (Cherry, 2022). Besides our study has also included the published panel sequencing data of six ovarian and breast cancer cell lines (Shi et al., 2022).

In this study, we comprehensively compared six commercially available sequencers (NA: NovaSeq 6000, NS: NextSeq 550, MGI: MGISEQ-2000, GL: GenoLab M, SF: SURFSeq 5000, and FS: FASTASeq 300) and five commonly-used bioinformatics pipelines for

tumor variants detection. These pipelines were highly cited or could detect low-frequency variation by adjusting the appropriate parameters. HaplotypeCaller (HC) (McKenna et al., 2010) and Mutect2 (Benjamin et al., 2019) were GATK-related tools. SiNVICT (Kockan et al., 2017) and SNVer (Wei et al., 2011) with short release time and long release time, respectively, and VarScan 2 (Koboldt et al., 2012) had been well recognized with more than 4,500 citations (Supplementary Table S1). The consistency and dissimilarity of different platforms and pipelines were evaluated in terms of SNP, InDel and TMB calling. By comparing with the truth set, we expected to provide objective insights into platforms and variant callers to achieve higher sensitivity, which can be crucial in cancer research and personalized medicine.

## 2 Materials and methods

### 2.1 Sample

The DNA sample used in this study is OncoSpan FFPE (Catalog ID: HD832, Horizon Discovery, United States), a well-characterized, cell line-derived Reference Standard containing 386 variants in 152 cancer genes (<https://horizondiscovery.com/en/reference-standards/products/oncospan-gdna>). In this study, the mutation cohort captured by the TSO500 panel theoretically includes 212 variants (194 SNPs and 18 InDels), with 24 confirmed by droplet digital PCR (ddPCR). This dataset was designed as a truth set for following benchmarking analyses. The VAF varies between 1% and 100%. The cfDNA sample used is Twist Pan-cancer Reference Standard (MineBio Life Sciences Ltd., Beijing, China), covering 458 individual mutations with 132 clinically actionable variants across 84 genes associated with cancer.

### 2.2 DNA extraction and quality assessment

WuXi Nextcode LTD. performed DNA isolation from OncoSpan scroll using the SEQPLUS FFPE DNA Isolation Kit following the manufacturer's protocol. The DNA quality with OD 260/280 value between 1.7 and 2.2 was qualified by a Nanodrop spectrophotometer. The integrity and concentration of the extracted DNA were confirmed by agarose electrophoresis and Qubit dsDNA HS Assay (Thermo Fisher Scientific). About 40 ng DNA sample was used for library preparation.

### 2.3 Library preparation and target enrichment

The extracted DNA passing quality control (QC) was sheared on the Covaris E220 evolution (Covaris Ltd., United States) to form 90–250 bp dsDNA fragments. The size of the fragments was chosen *via* TapeStation 2200 (Agilent, Cheshire, UK) after shearing. Then, the sequencing libraries were prepared and enriched using the hybrid capture-based TSO500 library preparation kit (#20028213, TruSight Oncology 500 DNA Kit, Illumina, San Diego, CA, United States) following the manufacturer's instruction.

In brief, the sheared DNA was treated with end repair and A-tailing reagents to convert the 3' and 5' overhangs into blunt

ends. UMI adapters containing unique molecular indexes were ligated to identify the unique sequence. After cleaning up the excessive ligation reagents and unligated adapters, the library fragments were amplified using primers that added index sequences for sample multiplexing. Next, the libraries were enriched through two rounds of hybridization capture. A pool of oligos specific to 523 genes (TSO500 panel) were used to hybridize to the DNA libraries, which were later captured with SMB (Streptavidin Magnetic Beads)-conjugated biotin probes. Subsequently, the enriched libraries were amplified, quantified with the Qubit dsDNA HS Assay Kit (#Q32854, Invitrogen, United States), and bead-based normalized for sequencing. The Twist cfDNA Pan-cancer Reference Standard was enriched by TargetSeq One kit (iGeneTech, Inc, Beijing, China.) referred to a previous study (Nie et al., 2021).

## 2.4 Sequencing, data preparation and quality control

Before sequencing, the size distribution of the sequencing library HD832 was characterized using Agilent TapeStation 4,200. WuXi Nextcode LTD. carried out HD832 library preparation and sequencing on Illumina NS and NA platforms in Shanghai. Meanwhile, sequencing on the FS and GL platforms was performed at the lab of GeneMind LTD. in Shenzhen. Besides, Twist cfDNA was sequenced by SF, NA, and MGI at the lab of GeneMind LTD. The sequencing mode is PE150 bp.

Raw data from FFPE, cfDNA, and cancer cell samples was filtered by fastp software to trim sequencing adapters, and low-quality reads with default parameters. The reads were filtered out if the proportion of low-quality bases (<Q20) was higher than 15%. The sequencing reads were mapped using to the human reference genome (hg19) with default parameters via Burrows-Wheeler Aligner (BWA v0.7.17-r1188)-Maximal Exact Match (MEM) algorithm. Hereafter, the samtools and GATK “Mark Duplicates (Picard)” module were used to implement the following processes, including indexing, sorting, and duplicates removal in BAM files. The deduplicated BAM files were subjected to different pipelines for variant calling. Quality metrics were generated from these BAM files by fastQC and bamdst tools.

## 2.5 Benchmarking analyses of variant calling among five variant callers

The tumor mutations were called with five popular pipelines: HC, Mutect2, SiNVICT, SNVer, and VarScan 2 with default parameters. To reduce false-positive calls, eligible mutations include only coding variants with VAF  $\geq 5\%$ , coverage  $\geq 8$  reads, and allele read depth  $\geq 3\times$ . Next, the variants were annotated for impact prediction with snpEFF software. Broadening to the whole panel range, we calculated the somatic mutations through aligning against four common population databases: Exome Aggregation Consortium, 1000 Genomes Project, Database of Short Genetic Variations, and Exome Sequencing Project v. 6500 to remove the common germline variants (VAF  $> 0.1\%$ ). Then, variants were labeled germline if they were regarded as a benign or likely

benign variant in either the Human Gene Mutation Database or ClinVar. In addition to these common frequency databases, the depth of somatic variants required  $DP \geq 100$ ,  $AD \geq 8$  for SNVs, and  $AD \geq 5$  for InDels (Tang et al., 2020b). TMB was computed as a ratio between the numbers of somatic mutations with the target region size of the panel (Merino et al., 2020). We further implemented the correlation analysis by comparing relative variant allele frequency (r-VAF), defined as the ratio of detected value to the theoretical value of the truth set, to excavate the concordance of 50 datasets.

## 2.6 Statistical analysis

All datasets were analyzed using the R statistics package (v4.1.2; R: The R-project for statistical computing). The similarity of datasets was calculated using Pearson's correlation coefficient. Precision, recall (sensitivity), and F-score were calculated based on the true sets of 212 mutations in the TSO500 panel, 442 mutations in TargetSeq One, and true mutations in ovarian and breast cancer cell lines referred to Pan et al., 2022. The following formulas were used.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}).$$

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN}).$$

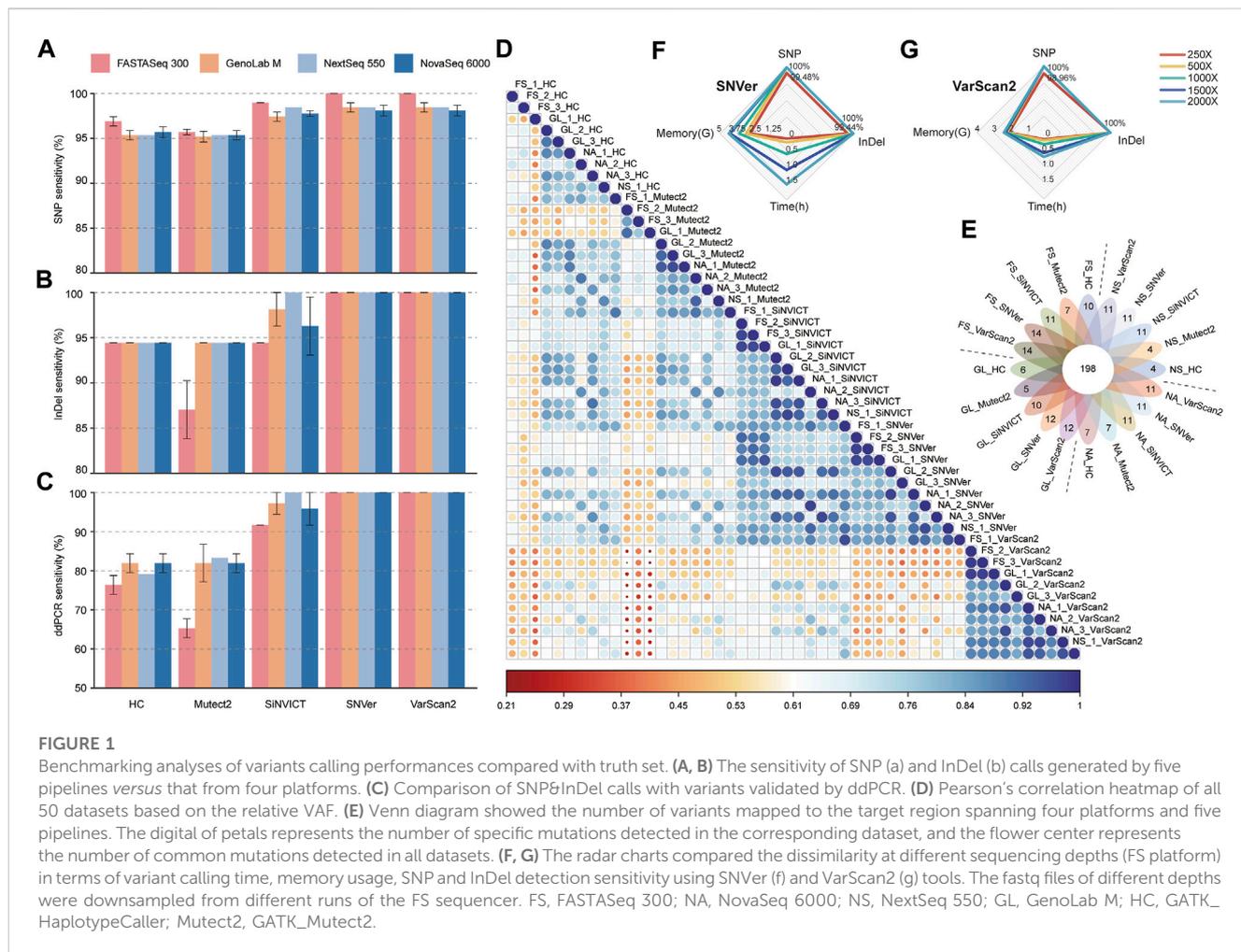
$$\text{F-score} = 2 * \text{Sensitivity} * \text{Precision} / (\text{Sensitivity} + \text{Precision}).$$

TP: true positive, FN: false negative, FP: false positive.

## 3 Result

### 3.1 Sequencing data summary

We aliquot the same DNA sequencing library (HD832 captured by TSO500 panel) to each platform to avoid the possibility of inconsistency caused by library construction differences. After filtering, aligning and deduplicating, 10 datasets (FS 3, GL 3, NA 3, and NS 1) were subjected to five analysis pipelines for variant calling: HC, Mutect2, SiNVICT, SNVer and VarScan 2 (Supplementary Figure S1). FASTQ and BAM quality statistics were calculated for the final 50 datasets, as shown in Supplementary Table S2. Benchmarking showed that results from GeneMind platforms (FS and GL) are compatible with those using Illumina platforms (NA and NS) in terms of the data yield and data quality. All sequencers presented comparable high base quality (over Q20) base percentages, with an average of 96.49% (FS), 97.40% (NA), 97.01% (GL) and 94.05% (NS). Each dataset from four sequencing platforms has a trustworthy average depth, higher than 2000 $\times$ . The short target sequence length and great sequencing depth contributed to the high duplication. To be compatible with relatively low coverage regions, we downsampled data with a depth  $\geq 4\times$ ,  $\geq 10\times$ ,  $\geq 30\times$ , and  $\geq 100\times$ , and the coverage at different levels for all datasets was over 97%. The average depth for each gene across multiple sequencers ranged from 31.56 to 4,823.42  $\times$  (Supplementary Table S3). Overall, the datasets from four sequencing platforms returned highly concordant quality, achieving the requirement of the high depth of panel sequencing with approving uniformity.



### 3.2 High-confidence variants calling spanning four platforms and five pipelines

Subsequently, we investigated the SNP&InDel calling rate of four platforms with five popular analysis algorithms at 2000× depth by comparing with the truth set of the HD832 sample captured by the TSO500 panel. As shown in Figures 1A, B, four platforms presented nearly concordant results in terms of SNP&InDel sensitivity under the same pipeline, albeit with some minor differences in the InDel calling using SiNVICT and Mutect2 callers. Encouragingly, the FS platform showed perfect SNP&InDel calling (100%) using SNVer and VarScan2 pipelines. For F-score and precision, SNVer and VarScan2 pipelines also performed best on SNP&InDel calling (Supplementary Table S4). Similarly, high agreement of variant calling was observed when compared with the truth set validated by ddPCR, especially under the SNVer and VarScan2 tools (Figure 1C). For TargetSeq One panel, SNVer and VarScan2 exhibited values close to 100% in the F-score, recall, and precision of SNP&InDel (Supplementary Figure S2) under the almost equal sequencing depth (Supplementary Table S5). For panel sequencing of six breast cancer and ovarian cell lines samples, all five tools performed flawlessly in three samples, while, in the remaining three samples, the detection results of

SNVer and VarScan2 were closest to the true set (Supplementary Table S6).

Pearson correlation coefficient (*r*) heatmap (Figure 1D) revealed that under the same software, similar overall performance was presented across-platforms, with slightly higher consistency from SiNVICT with average *r*: 0.86 (0.68–0.97), SNVer with 0.86 (0.71–0.97), and VarScan2 with 0.90 (0.75–0.98) approaches. Intriguingly, data from SiNVICT and SNVer pipelines showed high correlation across all 50 datasets (average *r*: 0.84). Less accordance was observed in data from five pipelines under the same platform, especially for HC (average *r*: 0.71) and Mutect2 (average *r*: 0.78), consistent with the variant calling results (Figures 1A–C). In addition, the data of three technical replicates on the same platform indicated high correlation except for a few outliers, such as FS\_HC samples. Hereafter, we combined three replicates to investigate the detection number of high-confidence variants among tools and platforms (Figure 1E). Up to 93.4% (198/212) concordance was obtained from all datasets, of which, the FS platform performed better with more unique mutations, particularly under the SNVer and VarScan2 callers, reproducing the above-described results.

Overall, the results of SNP&InDel calling suggested that software may have a greater impact on variant calling than platform for the target sequencing datasets, and FS performed better under SNVer and VarScan2 approaches.

### 3.3 Assessment of variant calling at different sequencing depths

Sequencing to greater depths increases the reproducibility of variant detection but at the expense of longer turnaround times, increased computational complexity, and greater cost. Focusing on the best-performed FS platform and the SNVer and VarScan2 pipelines, we compared the differences in variant calling time, memory usage, SNP\_recall and InDel\_recall with radar chart at different sequencing depths, expecting to recommend a balance between sequencing depth and variant detection sensitivity (Figures 1F,G). As expected, the calling time and memory usage progressively increased with increasing depth of coverage, from 250× to 2000×. Additionally, we saw no apparent improvement in SNP&InDel recall percentage with increasing depth from 500×. At the given depth of 250×, the SNVer pipeline identified a median recall of 99.48% for SNPs and 99.44% for InDels compared with a recall of 98.96% for SNPs and 100% for InDels using the VarScan2 pipeline. It suggests that the lowest depth to discover all high-confidence mutations within the truth set is 500× for the FS datasets analyzed by the SNVer or VarScan2 pipeline. Concerning turnaround time and memory usage, the VarScan2 performed better than SNVer tool.

### 3.4 Concordance of annotated genes from different datasets

To further explore the effects of gene annotation on different platforms and software tools, we compared the concordance of the annotated genes based on their rVAF. Supplementary Figure S3 showed the visual difference between the VAF of all datasets and truth set, and white represents no difference. The heatmap returned highly concordant results for datasets analyzed by different pipelines, while some perceptible differences can be noticed when using SNVer and VarScan2 callers. Specifically, noticeable increases in rVAF were observed by VarScan2, conversely, obvious decreases for the same genes were detected by SNVer. Focusing on variants with low mutation frequency (*MET*, *KIT* and *NRAS* genes), we were delighted to find that SiNVICT, SNVer and VarScan2 approaches were friendly to all platforms with high consistency, including the novel FS platform. HC caller performed worst across different platforms with more notable dissimilarity. Notably, for *EGFR* mutation with very low VAF (4.5%), five platforms failed to discern it under HC and Mutect2 pipelines. The absolute VAF (Supplementary Tables S7, S8) also indicated high concordance for datasets compared to the truth set.

### 3.5 Benchmarking somatic mutations and TMB distributing in the complete panel

Furthermore, we broadened to the global patterns of somatic variations observed in the panel range to obtain comprehensive insight. The common germline variants and benign variants were filtered out when VAF >0.1%. Detailed information on the somatic mutation spectrum in each dataset is represented as a waterfall plot (Supplementary Figure S4). Overall, the driver somatic variants, including missense, stop-gained, frame-shift, and splice mutations, showed very high agreement among 50 datasets from different

platforms and pipelines. Nevertheless, HC pipeline exhibited more discordance, such as genes *NSD1*, *BRAF*, *STAT3*, and *GRM3*. With respect to TMB, all sequencers and pipelines returned comparable concordance, with higher TMB number when using the SiNVICT and VarScan2 tools, which may be explained by the higher number of the somatic mutations (Supplementary Figure S5). The number of somatic mutations from three technical replicates indicated that the highest shared-variants for FS platform were obtained from VarScan2 (93.67%), for GL platform from SiNVICT (92.33%), for NA platform from SiNVICT (82.30%). In contrast, HC shared inferior proportion of common variants for the FS (42.50%), consistent with the results from waterfall plot. While Mutect2 performed worst for GL (55.53%) and NA (28.13%) platforms. Need to note, we detected a novel set of somatic variant sites and genes outside the truth set published by HD832 through the intersection of our 50 datasets (Table 1). These high-confidence variants were expected to be a supplement to the published truth set of HD832, offering new insights into the tumor variant research.

## 4 Discussion

NGS-based targeted sequencing has gained prominence in assessing genetic alterations in cancer, influencing the ongoing development of personalized medicine. Compared to WGS and WES, TS could minimise manpower, running time, sequencing data, storage space and computational demand by merely probing into the interesting targets, thus being economical and cost-effective (Bewicke-Copley et al., 2019). The excellent coverage depth is the icing on the cake for its sensitivity and specificity. Whereas, an insufficient clinical consensus on sequencing platforms and variant calling tools confuses researchers and clinicians. We evaluated five variant calling tools on four sequencing platforms to identify the possibilities and limits of variant screening.

Turn-around-time and sensitivity are foundational to good clinical diagnosis and essential to achieving optimal patient outcomes. Specifically, early and reliable tests of genetic biomarkers available in patients may decrease the use of more invasive testing, under- and overtreatment, and reduces the adverse events associated with inappropriately targeted therapies (Singh et al., 2019; Wurcel et al., 2019). In this study, we observed high concordance of mutation calling across different sequencers and software tools generally, albeit with some outlier values and minor differences (especially for HC and Mutect2 pipelines). Of note, the FS platform showed higher sensitivity (100%) in high-confidence variants calling than the other three platforms, when analyzed by SNVer and VarScan 2 software algorithms (Figures 1A–C). Considering the dissimilarity of variant calls across different bioinformatics tools for datasets from the same sequencer (Figures 1D,E), we appraised that software appears to contribute greater to the variant detection. In addition, comparing the sequencing time of four platforms in this study, the FS platform shows dominance with the shortest sequencing time (FS, ~21 h; NA, ~25 h; NS, ~26 h; GM, ~43 h) at the sequencing mode PE150.

Due to the high-depth nature of TS, special attention needs to be taken to the relationship between the sequencing depth and precision/sensitivity of mutation detection. However, there is no consensus on the minimum required depth in a clinical setting

**TABLE 1 Novel somatic variant genes identified in our study that are not included in the truth set published by HD832.**

GRCh37 (hg19)		Ref	Alt	Gene	Variant type	Cosmic ID (v97)	Amino acid	Sequence ontology term	Average allele frequency (%)	Depth (x)
Chr	Location									
chr1	36,933,199	T	C	<i>CSF3R</i>	SNP	COSV58970806	p.Thr640Ala	missense_variant	29.05	809.66
chr2	99,154,426	G	A	<i>INPP4A</i>	SNP	COSV50811933	p.Val190Ile	missense_variant	30.35	730.08
chr2	202,137,387	G	A	<i>CASP8</i>	SNP	COSV51857659	p.Met205Ile	missense_variant	30.56	532.12
chr3	49,936,300	C	A	<i>MST1R</i>	SNP	COSV56572804	p.Gln516His	missense_variantandsplice_region_variant	31.82	538.80
chr3	187,446,153	C	T	<i>BCL6</i>	SNP	COSV51656257	p.Ser512Asn	missense_variant	30.31	658.08
chr5	39,002,766	A	G	<i>RICTOR</i>	SNP		p.Leu88Ser	missense_variantandsplice_region_variant	13.74	346.00
chr6	111,983,012	T	C	<i>FYN</i>	SNP	COSV57622174	p.Tyr515Cys	missense_variant	28.68	723.56
chr7	50,444,272	G	A	<i>IKZF1</i>	SNP	COSV58792485	p.Gly68Arg	missense_variant	21.94	708.74
chr8	93,017,471	G	GT	<i>RUNX1T1</i>	InDel		p.Gln264fs	frameshift_variant	13.71	780.74
chr10	90,767,575	A	T	<i>FAS</i>	SNP	COSV58239879	p.Arg105Ser	missense_variant	23.66	575.32
chr12	6,709,179	C	A	<i>CHD4</i>	SNP			splice_acceptor_variantandintron_variant	11.67	511.06
chr13	102,375,201	T	C	<i>FGF14</i>	SNP	COSV65936716	p.Asn247Asp	missense_variant	8.62	609.44
chr15	45,003,808	C	T	<i>B2M</i>	SNP	COSV62564195	p.Gln22*	stop_gained	30.25	813.56
chr15	45,008,529	C	T	<i>MIR10393</i>	SNP	COSV62562883	p.Arg117*	stop_gainedandsplice_region_variant	10.32	360.76
chr15	66,782,085	A	G	<i>MAP2K1</i>	SNP	COSV57235242, COSV57235143	p.Asp351Gly	missense_variant	31.95	528.86
chr15	73,994,743	T	C	<i>CD276</i>	SNP	COSV59205301	p.Val76Ala	missense_variant	31.09	767.76
chr15	93,480,818	C	CA	<i>CHD2</i>	InDel		p.Val175fs	frameshift_variant	20.62	562.08
chr17	30,310,119	G	T	<i>SUZ12</i>	SNP		p.Gly340Val	missense_variant	31.03	449.58
chr19	36,214,360	G	T	<i>KMT2B</i>	SNP	COSV55867628	p.Cys1005Phe	missense_variant	29.18	552.34
chr19	50,906,813	T	C	<i>POLD1</i>	SNP		p.Phe401Leu	missense_variant	11.40	747.88
chr20	31,379,499	G	A	<i>DNMT3B</i>	SNP	COSV52422156	p.Met302Ile	missense_variant	23.90	593.92
chr20	3,975,901	G	T	<i>TOP1</i>	SNP		p.Trp754Cys	missense_variant	28.17	668.46

Note: These variants were not recorded in both ClinVar and GnomAD., Chr: Chromosome, Ref: Reference, Alt: Alteration. Gene names are listed in italics.

for TS. Previous studies attempted to leverage the power of Illumina technologies to recommend an appropriate depth for TS. Nevertheless, there is still no consistent standard presently, ranging from 100× to 1,000× (Cortini et al., 2017; Bewicke-Copley et al., 2019; Koboldt, 2020). Further studies are thus needed to define the proper sequencing depth, especially for novel platforms. For datasets sequenced on the FS platform and called by SNVer and VarScan 2 algorithms, the sufficient sequencing depth for the TSO500 panel was 500×, which enables 100% detection of high-confidence variants (Figures 1F,G).

Accurate variant calling in NGS data is well known to have a critical impact on the downstream analysis and interpretation processes, for which, a proper variant detection algorithm is important for high sensitivity and specificity (Sandmann et al., 2017). Through the benchmarking analysis of five popular variant calling tools, we noted that the choice of software algorithm represented an important factor for variant detection. For example, SiNVICT, SNVer and VarScan 2 had high tool-specific “unique” calls when compared with the truth set (Figure 1E), while HC seemed to have the fewest concordant calls (Figures 1D,E). However, focusing on the VAF of the annotated genes, we noticed some perceptible differences between the truth set and the values analyzed by SNVer and VarScan2 callers (Supplementary Figure S3). These findings indicated that multiple analysis algorithms should be used to obtain a consensus to reduce false-discovery when diagnosing cancer-associated mutations.

While we provide a useful benchmark for the researcher to choose sequencing platforms and analysis pipelines, there are some limitations. Because of time and budget constraints, we only analyzed two standard cancer samples (OncoSpan FFPE (HD832) and Twist cfDNA Pan-cancer Reference Standard) with five pipelines. Although these samples provide reliable variant information, FFPE-based positive and negative patient samples would generate more complete evaluations. To acquire a comprehensive profile of sequencing platforms and pipelines, multiple samples from the HD832 cell line and clinical cancer FFPE samples should be enrolled for future research to enable ultimate clinical applications of the sequencers and variant callers.

## 5 Conclusion

There was good concordance of variant calling across different platforms and pipelines. FS platform showed highest sensitivity (100%) in high-confidence variants calling when analyzed by SNVer and VarScan 2 algorithms. Our study provides a standardized target sequencing resource to benchmark new bioinformatics protocols and sequencing platforms. Moreover, the variant calling performances of five software on three typical tumor samples provide a valuable reference for future tumor variant research.

## Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession

number(s) can be found below: <https://db.cngb.org/search/project/CNP0004275/>, CNP0004275.

## Ethics statement

Ethical approval was not required for the studies on humans in accordance with the local legislation and institutional requirements because only commercially available established cell lines were used.

## Author contributions

BF: Methodology, Writing–original draft. JL: Writing–original draft. XF: Methodology, Writing–original draft. YL: Writing–original draft, Resources. MW: Methodology, Software, Visualization, Writing–original draft. PW: Writing–review and editing, Resources. ZZ: Writing–review and editing, Resources. QY: Writing–review and editing, Resources. LS: Project administration, Supervision, Writing–review and editing.

## Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

## Acknowledgments

We would like to thank all current and past members of the GeneMind team who contributed to the development of the sequencing technology.

## Conflict of interest

Authors LS, BF, JL, YL, MW, PW, ZZ, and QY were employed by GeneMind Biosciences Company Limited.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2023.1293974/full#supplementary-material>

## References

- Benjamin, D., Sato, T., Cibulskis, K., Getz, G., Stewart, C., and Lichtenstein, L. (2019). Calling somatic SNVs and indels with Mutect2. *bioRxiv*, 861054. doi:10.1101/861054
- Bewicke-Copley, F., Kumar, E. A., Palladino, G., Korfi, K., and Wang, J. (2019). Applications and analysis of targeted genomic sequencing in cancer studies. *Comput. Struct. Biotechnol. J.* 17, 1348–1359. doi:10.1016/j.csbj.2019.10.004
- Cherry, P. D. (2022). Abstract LB110: twist pan-cancer synthetic reference materials for cell-free DNA (CfDNA) assay development. *Cancer Res.* 82, LB110. doi:10.1158/1538-7445.am2022-lb110
- Conroy, J. M., Pabla, S., Glenn, S. T., Seager, R., Van Roey, E., Gao, S., et al. (2021). A scalable high-throughput targeted next-generation sequencing assay for comprehensive genomic profiling of solid tumors. *PLoS One* 16, e0260089. doi:10.1371/journal.pone.0260089
- Cortini, F., Marinelli, B., Pesatori, A. C., Seia, M., Seresini, A., Giannone, V., et al. (2017). Clinical application of NGS tools in the diagnosis of collagenopathies. *Explor. Res. Hypothesis Med.* 2, 57–62. doi:10.14218/erhm.2017.00010
- Koboldt, D. C. (2020). Best practices for variant calling in clinical sequencing. *Genome Med.* 12, 91–13. doi:10.1186/s13073-020-00791-w
- Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McEellan, M. D., Lin, L., et al. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 22, 568–576. doi:10.1101/gr.129684.111
- Kockan, C., Hach, F., Sarrafi, I., Bell, R. H., McConnelly, B., Beja, K., et al. (2017). SiNVICT: ultra-sensitive detection of single nucleotide variants and indels in circulating tumour DNA. *Bioinformatics* 33, 26–34. doi:10.1093/bioinformatics/btw536
- Mckenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi:10.1101/gr.107524.110
- Merino, D. M., McShane, L. M., Fabrizio, D., Funari, V., Chen, S. J., White, J. R., et al. (2020). Establishing guidelines to harmonize tumor mutational burden (TMB): *in silico* assessment of variation in TMB quantification across diagnostic platforms: phase I of the Friends of Cancer Research TMB Harmonization Project. *J. Immunother. cancer* 8 (1), e000147. doi:10.1136/jitc-2019-000147
- Nagahashi, M., Shimada, Y., Ichikawa, H., Kameyama, H., Takabe, K., Okuda, S., et al. (2019). Next generation sequencing-based gene panel tests for the management of solid tumors. *Cancer Sci.* 110, 6–15. doi:10.1111/cas.13837
- Nie, D. A., Xia, C. R., Huang, K. C., Liu, J., Gan, T., Wen, C., et al. (2021). Identification of a novel pathogenic COL4A3 gene mutation in a Chinese family with autosomal dominant Alport syndrome: a case report. *Biomed. Rep.* 15 (5), 90–96. doi:10.3892/br.2021.1466
- Pan, B., Ren, L., Onuchic, V., Guan, M., Kusko, R., Bruinsma, S., et al. (2022). Assessing reproducibility of inherited variants detected with short-read whole genome sequencing. *Genome Biol.* 23, 2–26. doi:10.1186/s13059-021-02569-8
- Pestinger, V., Smith, M., Sillo, T., Findlay, J. M., Laes, J.-F., Martin, G., et al. (2020). Use of an integrated pan-cancer oncology enrichment next-generation sequencing assay to measure tumour mutational burden and detect clinically actionable variants. *Mol. Diagnosis Ther.* 24, 339–349. doi:10.1007/s40291-020-00462-x
- Ramos-Paradas, J., Hernández-Prieto, S., Lora, D., Sanchez, E., Rosado, A., Caniego-Casas, T., et al. (2021). Tumor mutational burden assessment in non-small-cell lung cancer samples: results from the TMB2 harmonization project comparing three NGS panels. *J. Immunother. Cancer* 9, e001904. doi:10.1136/jitc-2020-001904
- Sahajpal, N. S., Mondal, A. K., Ananth, S., Njau, A., Ahluwalia, P., Jones, K., et al. (2020). Clinical performance and utility of a comprehensive next-generation sequencing DNA panel for the simultaneous analysis of variants, TMB and MSI for myeloid neoplasms. *PLoS One* 15, e0240976. doi:10.1371/journal.pone.0240976
- Sandmann, S., De Graaf, A. O., Karimi, M., Van Der Reijden, B. A., Hellström-Lindberg, E., Jansen, J. H., et al. (2017). Evaluating variant calling tools for non-matched next-generation sequencing data. *Sci. Rep.* 7, 43169. doi:10.1038/srep43169
- Shi, Z., Lopez, J., Kalliney, W., Sutton, B., Simpson, J., Maggert, K., et al. (2022). Development and evaluation of ActSeq: a targeted next-generation sequencing panel for clinical oncology use. *PLoS One* 17 (4), e0266914. doi:10.1371/journal.pone.0266914
- Singh, H., Graber, M. L., and Hofer, T. P. (2019). Measures to improve diagnostic safety in clinical practice. *J. Patient Saf.* 15, 311–316. doi:10.1097/PTS.0000000000000338
- Tan, A. C., Lai, G. G., San Tan, G., Poon, S. Y., Doble, B., Lim, T. H., et al. (2020a). Utility of incorporating next-generation sequencing (NGS) in an Asian non-small cell lung cancer (NSCLC) population: incremental yield of actionable alterations and cost-effectiveness analysis. *Lung Cancer* 139, 207–215. doi:10.1016/j.lungcan.2019.11.022
- Tang, Y., Li, Y., Wang, W., Lizaso, A., Hou, T., Jiang, L., et al. (2020b). Tumor mutation burden derived from small next generation sequencing targeted gene panel as an initial screening method. *Transl. Lung Cancer Res.* 9 (1), 71–81. doi:10.21037/tlcr.2019.12.27
- Wei, B., Kang, J., Kibukawa, M., Arreaza, G., Maguire, M., Chen, L., et al. (2022). Evaluation of the trusight oncology 500 assay for routine clinical testing of tumor mutational burden and clinical utility for predicting response to pembrolizumab. *J. Mol. Diagnostics* 24, 600–608. doi:10.1016/j.jmoldx.2022.01.008
- Wei, Z., Wang, W., Hu, P., Lyon, G. J., and Hakonarson, H. (2011). SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res.* 39, e132. doi:10.1093/nar/gkr599
- Wurcel, V., Cicchetti, A., Garrison, L., Kip, M. M., Koffijberg, H., Kolbe, A., et al. (2019). The value of diagnostic information in personalised healthcare: a comprehensive concept to facilitate bringing this technology into healthcare systems. *Public Health Genomics* 22, 8–15. doi:10.1159/000501832
- Zhao, C., Jiang, T., Hyun Ju, J., Zhang, S., Tao, J., Fu, Y., et al. (2020). TruSight oncology 500: enabling comprehensive genomic profiling and biomarker reporting with targeted sequencing. *bioRxiv*, 349100. doi:10.1101/2020.10.21.349100