Tumor Diagnosis of FFPE Sample by Target Genome Sequencing From Multiple Sequencing Platforms

Baosheng Feng
GeneMind Biosciences Company Limited
Juan Lai
GeneMind Biosciences Company Limited
Xue Fan
Longgang District Maternity&Child Healthcare Hospital of Shenzhen City
Xin Guo
Longgang District Maternity&Child Healthcare Hospital of Shenzhen City
Yongfeng Liu
GeneMind Biosciences Company Limited
Miao Wang
GeneMind Biosciences Company Limited
Ping Wu
GeneMind Biosciences Company Limited
Zhiliang Zhou
GeneMind Biosciences Company Limited
Qin Yan
GeneMind Biosciences Company Limited
Lei Sun (sunlei@genemind.com)
GeneMind Biosciences Company Limited

Article

Keywords: HD832, TSO500, GenoLab M, NovaSeq 6000, NextSeq 550

Posted Date: March 11th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1429288/v1

Abstract

Technological innovation and increased affordability have contributed to the widespread adoption of target genome sequencing (TS) technologies in scientific research and clinical diagnostic applications. Cancer research consortia showed increasing acceptance of TS techniques to define the mutation landscape of multiple cancer types. These studies have primarily utilized Illumina's sequencing instruments. In this study, we performed target genome sequencing of Reference OncoSpan FFPE (HD832) sample enriched by TSO500 panel using a new platform, GenoLab M, and compared it with NovaSeq 6000 and NextSeq 550 platforms. The data demonstrated that all three platforms showed high-quality scores, deep coverage, and high concordance for variant detection (94.8%). Besides, the GenoLab M platform displays high accuracy in tumor diagnosis of FFPE sample. Briefly, the GenoLab M platform shows comparable performances with NovaSeq 6000 and NextSeq 550, indicating its potential applicability in cancer TS at a lower cost.

Introduction

Targeted genome sequencing (TS) is an effective NGS method that focuses on the interesting regions of the genome and enables the rapid detection of genetic variations implicated in diseases. Through selective enrichment of the targeted genomic regions, TS can profile a panel of genes or biomarkers with known clinical relevance to discover powerful mutations, rather than screening the whole genome via large-scale DNA sequencing (whole genome sequencing, WGS). 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 WGS or whole exome-sequencing (WES)[1]. Thus, it allows high confidence to identify mutations with low variant allele frequencies (VAFs), 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[2,3]. TruSight™ Oncology 500 (TSO500, Illumina) is a comprehensive target-sequencing panel that covers more than eight cancer types, 523 cancer-related genes (1.94 Mb) to identify relevant genomic variants and signatures in a single assay (https://www.illumina.com/products/by-brand/trusight-oncology-500.html). We can obtain multiple variant information from formalin-fixed paraffin-embedded (FFPE) tissue specimens through the TSO500 assay[4]. These mutations include small nucleotide variants (SNVs), InDels, copy number variations (CNVs), splice variants and gene fusions. In addition, it accurately measures key current immuno-oncology biomarkers: microsatellite instability (MSI) and tumor mutational burden (TMB). TSO500 workflow allows variants detection with high specificity and sensitivity in FFPE samples, especially for mutants with low VAFs via tactfully suppressing technical noise and excluding germline variants. TSO500 assay has displayed exemplary performance in profiling TMB, MSI, and various gene variants in different tumor types[5–7].

Many high-throughput sequencers are compatible with targeted genome sequencing, and have been widely applied in scientific studies and clinical diagnosis. NextSeq550 system (Illumina), a flexible

sequencing platform with multiple read length and output configurations, has been commonly used in WES, transcriptome, as well as targeted sequencing[8,9]. It was reported that the TSO500 panel performed well for variant screening of myeloid neoplasm coupled with the NextSeq550 platform[10]. The high sensitivity, specificity, precision, and accuracy demonstrated the clinical applicability of this assay for SNVs and InDels/duplications detection. Similarly, on the NovaSeq 6000 platform (Illumina), the TSO500 panel also showed excellent performances and good reproducibility in identifying multiple DNA/RNA variant types and molecular signatures from FFPE tissue specimens[11–13]. GenoLab M (GeneMind LTD., China), a novel high-throughput sequencing platform, utilizes the well-recognized sequencing-by-synthesis (SBS) techniques and reversible termination approaches to deliver confident base-calling accuracy and high yield of error-free reads. Moreover, GenoLab M is highly compatible with Illumina Library Preparation Kits, and 2X150 bp paired-end sequencing mode is available. Previously, the performance metric of GenoLab M on transcriptome and LncRNA sequencing for three model species (mouse, bean, and human) has been compared with NovaSeq 6000 platform. This budding platform shows comparable data quality in detecting gene expression, alternatively spliced (AS) events, single nucleotide polymorphism (SNP), and InDel[14].

The study aimed to evaluate whether the GenoLab M platform can accurately detect mutations in tumor standard sample. The OncoSpan FFPE (HD832) sample was captured by the TSO500 panel through hybrid-capture-based target enrichment, and sequenced by NovaSeq 6000, NextSeq 550, and GenoLab M platforms. We compared the performances of three sequencing platforms.

Results

Sequencing data quality validation

To assess the quality of sequencing data generated by three platforms, we used the FastQC software for quality checking. It demonstrated that the data yield and data quality were acceptable and comparable for GenoLab M and NovaSeq 6000 platforms (Table 1; Fig. 1). For datasets from these two platforms, the average percentages of bases quality for Q20 and Q30 were higher than 95% and 88%, respectively. However, the data quality of the start and end positions in the fastq files of GenoLab M was distinctly lower than NovaSeq 6000 and tended to gradually deteriorate towards the end of the read (although it was not lower than Q20) (Figs. 1A and 1B; Figs. 1D and 1E). The curves representing the median base score of each nucleotide position from NovaSeq 6000 datasets remained stable without significant decline. For NextSeq 550, the data yield and data quality were slightly inferior to the other two platforms (Table 1; Figs. 1G and 1H), whose average percentages of bases quality for Q20 and Q30 were 88% and 80%. The GC contents were approximately similar for all three platforms, and the distribution maps were practically identical (Figs. 1C, 1F and 1I).

Table 1
Mapping statistics of the datasets from different platforms. GL, GenoLab M; NA, NovaSeq 6000; NS, NextSeq 550.

Metrics	GenoLab M				NovaSeq 6000			NextSeq 550
	GL_1	GL_2	GL_3	GL_4	NA_1	NA_2	NA_3	NS_1
Clean reads (M)	161.28	167.13	162.37	167.13	165.44	165.56	166.34	156.24
Clean bases (Gb)	20.46	21.23	20.68	25.02	21.58	24.65	24.77	19.69
Mapping rate (%)	98.11	97.46	98.56	98.20	98.15	98.49	98.90	98.22
Read and mate paired rate (%)	96.69	95.14	96.98	98.02	96.43	97.91	98.33	96.31
Duplicate rate (%)	61.99	70.64	70.86	86.03	71.25	85.73	85.9	69.13
Target region reads (M)	133.07	137.08	134.82	138.66	135.38	136.26	137.31	128.6
Target region mapping rate (%)	84.32	84.49	84.56	84.64	83.86	84.06	83.95	84.14
Target mean depth	6117.53	6244.49	6169.68	6315.54	6400.1	6361.59	6455.21	5856.94
Coverage (> 0X, %)	99.72	99.69	99.69	99.68	99.72	99.71	99.71	99.71
>=4X (%)	99.67	99.64	99.63	99.64	99.69	99.69	99.70	99.68
>=10X (%)	99.63	99.57	99.56	99.58	99.66	99.66	99.66	99.63
>=30X (%)	99.54	99.41	99.39	99.42	99.58	99.60	99.59	99.53
>=100X (%)	99.25	98.97	98.92	98.96	99.37	99.38	99.39	99.23

Reads mapping and coverage distribution

First, the raw fastq files were filtered and trimmed to ensure accurate variants calling. Then, the trimmed fastq files were aligned to the reference genome and sorted to remove duplicates. The generated bam files were further analyzed.

Comparing the datasets from GenoLab M, NovaSeq 6000, and NextSeq 550 platforms, the average mapping rates were 98.08%, 98.51%, and 98.22%, respectively; while the mean duplicate rates were 72.38%, 80.96%, and 69.13%, respectively (Table 1). Notably, short target sequence length and great sequencing depth in all datasets might explain the high duplicate rates. To be compatible with relatively low coverage regions, we downsampled data with a depth ≥ 4X, ≥ 10X, ≥ 30X, and ≥ 100X. In this study, the coverage at different levels for all datasets was over 99%, and the mean coverage depth of three platforms was more than 5800X per nucleotide position within the target region (Table 1). Similar distribution graphs of the sequencing coverage indicated that three platforms achieved the requirement of the high depth of panel sequencing with approving uniformity (Fig. 2).

The results highlighted that we can use GenoLab M platform to obtain comparable mapping rate and sequencing depth to NovaSeq 6000 and NextSeq 550 platforms. Besides, all the datasets were available for the following variant calling analysis.

Comparison of variants inter platforms

We adopted the popular pipeline GATK and SAMtools to investigate the performance of variant calling for the three platforms. The number of theoretical mutations in the HD832 sample captured by the TSO500 panel was 212, including 194 SNPs and 18 InDels. Twenty-five variants of them have been confirmed by the ddPCR (Table S1).

The variants detected among three sequencers were grouped and compared to investigate the mutation detection ability at different VAFs. Overall, three platforms showed similar mutation detection rates when the VAF was greater than 10%. For variants with low VAFs (<10%), GenoLab M was comparable in mutation detection compared to NovaSeq 6000, but superior to NextSeq 550 (Fig. 3). The variant overlap between four datasets from GenoLab M, three datasets from NovaSeq 6000, and one dataset from NextSeq 550 was 94.8% (201 in total 212 variants), reflecting relatively high consistency of mutation detection between different platforms (Fig. 4A). Comparing the detected VAFs (actual) with mean VAFs (expected) on three platforms using linear models, we found a high correlation as indicated by the mean R^2 value of 0.9518 (GenoLab M), 0.9647 (NovaSeq 6000), and 0.949 (NextSeq 550) (Fig. 4B). In addition, we observed a solid linear correlation between the mean actual VAF of datasets detected on the GenoLab M platform and on Illumina's platforms (R^2 = 0.9696; Fig. 4C).

In order to further determine the detection accuracy, we compared 25 variants identified by three platforms with expected results confirmed by ddPCR to identify the shared mutants. Overall, these variants were distributed across a range of genes with wide variant allele frequencies (Table S2; Fig. 5). Three platforms demonstrated substantially consistent calling results compared to the validated mutations. And the mutation type and detection numbers were identical among the three platforms

except for the *EGFR* variants (Fig. 5A and 5B). The inconsistent *EGFR* variants were distributed in the very low allele frequency area, which was better identified by the GenoLab M platform. It revealed that the GenoLab M platform displayed a comparable detection rate in most SNPs and InDels calling with Illumina's platforms, with a slight advantage in variant detection with VAFs lower than 10% (Fig. 5C and 5D; Table S2). All three platforms could detect other ddPCR-validated mutations with comparable VAFs compared to expected VAFs (Table S2).

Discussion

We conducted target sequencing of the OncoSpan FFPE sample captured by the TSO500 panel using the GenoLab M, NovaSeq 6000 and NextSeq 550 platforms. The sequencing data from just one sample can function as an intuitive index for researchers to study the performance differences across platforms. Therefore, we aliquot the same DNA sequencing library (HD832) to each platform to avoid the possibility of inconsistency caused by library construction differences. The FastQC results indicated that all three platforms provide comparable levels of quality, sequencing coverage, and GC content. Meanwhile, all datasets generated on these sequencers showed high unique mapping rates and target mean depth, but with a bit of high duplicate rate. Though a high duplicate rate makes it complicated to obtain accurate allele frequency estimates, the ultra-deep targeted sequencing guarantees precise results because the unique reads are enough to provide independent and reliable information[15]. Elevated sequencing depth means an increase number of aligned sequencing reads and a high confidence in the base call at a particular position, indicating high reproducibility and accuracy of variant detection[16].

In comparing SNVs and InDels detected by GenoLab M, NovaSeq 6000, and NextSeq 550 platform, we found that all three sequencing platforms identified the majority of variants, and the actual VAF is comparable to the expected VAF confirmed by ddPCR. However, we observed that one deletion mutation of *EGFR* (expected VAF<2%) was not detected on all three platforms. Intrinsic sequencing error rate and insufficient sequencing depth are bottlenecks for detecting ultra-low-frequency variants. For low VAF variants, it is challenging to distinguish them from artifacts and most variant callers have difficulty in detecting them[17]. In addition, the sensitivity of the bioinformatics pipeline varies based on sequencing depth, while the FFPE sample presents relatively low sequencing depth compared to fresh-frozen tissue, which hinders accurate variant calling[18]. Therefore, an auxiliary detection method is recommended for clinical sample sequencing to confirm the low-frequency variants. Through comparative analysis, we noticed that GenoLab M showed a slightly higher detection rate than NovaSeq 6000 in one *EGFR* SNP mutation (expected VAF = 1%), one *EGFR* deletion mutation (expected VAF = 1.5%) and one *CTNNB1* deletion mutation (expected VAF = 9%). For NextSeq 550 platform, six variants with expected VAF<9% could not be detected compared with GenoLab M. These findings indicate that the GenoLab M platform has consistent, even slightly higher advantages in variant site detection compared with NovaSeq 6000 and NextSeq 550. Due to the lack of repeat sequencing samples on the NextSeq 550 platform, further confirmation is required to draw objective comparative conclusions.

For the novel GenoLab M sequencer, as well as recognized NovaSeq 6000 and NextSeq 550 platforms, it is the first HD832 reference dataset of target sequencing enriched by TSO500 panel, from which we provide a performance evaluation and data reference for future research. Combined with the published data focused on transcriptome sequencing and submitting data for WES, the GenoLab M platform shows excellent performance in various NGS sequencing fields, comparable to Illumina's platforms. However, the performances of GenoLab M in other sequencing areas need to be further confirmed to prove its applicability, such as WGS, metagenomic sequencing, small RNA sequencing, etc. The sequencing dataset in this study can also be used as a reference for future development of sequencing technology and optimization of analytical tools.

In summary, the GenoLab M sequencer demonstrated comparable performance concerning data quality and variant allele frequency calling in tumor target genome sequencing compared to NovaSeq 6000 and NextSeq 550 instruments. Notably, the GenoLab M platform showed better performance in detecting variants with very low allele frequency.

Materials & Methods

Sample

The DNA sample used in this study is OncoSpan FFPE (Catalog ID: HD832, Horizon Discovery, USA), which is a well-characterized, cell line-derived Reference Standard containing 386 variants in 152 cancer genes. In this study, the mutation cohort captured by the TSO500 panel theoretically includes 212 variants, with 25 confirmed by droplet digital PCR (ddPCR). The allelic frequency (AF) of these mutations varies between 1-100%.

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 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.

Library preparation and target enrichment

The extracted DNA passing quality control (QC) was sheared on the Covaris E220 evolution (Covaris Ltd, USA) 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, USA) 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. UMI1 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 add 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, USA), and bead-based normalized for sequencing.

Sequencing

Before sequencing, the size distribution of the sequencing library HD832 was characterized using Agilent Tapestation 4200. WuXi Nextcode LTD. carried out HD832 library preparation and sequencing on Illumina NextSeq550 and NovaSeq 6000 platforms in Shanghai. Meanwhile, sequencing on the GenoLab M platform was performed at the lab of GeneMind LTD. in Shenzhen. The sequencing length is 2X150 bp.

Bioinformatics analysis and variants selection

Raw data was analyzed with FastQC for assessing sequencing performance metrics, including read count, base quality across reads (Q20&Q30), and GC content. The sequencing reads were mapped using Minimap2 to human reference genome (hg19) with default algorithms. Then, initial read mapping was further optimized by removing InDels and PCR duplicates using SAMBLASTER. Somatic tumor mutations were called with the HaplotypeCaller-GATK, Mutect2-GATK and SAMtools. The variants were compared for concordance among different platforms.

Declarations

Acknowledgements

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

Author contributions

Lei Sun conceived and designed the research, reviewed and revised the manuscript. Juan Lai and Yongfeng Liu wrote the manuscript. Ping Wu, Qin Yan and Zhiliang Zhou reviewed and revised the manuscript. Baosheng Feng, Xue Fan and Xin Guo performed sample prepared and sequencing. Miao Wang supported data mining and figure drawing. All authors read and approved the final version of the manuscript.

Data availability

The target genome sequencing data are available in CNGB Sequence Archive (https://db.cngb.org/cnsa/) under project accession number CNP0002709.

Ethics approval and consent to participate

This study was approved by the Ethics Committee of GeneMind Biosciences Company Limited. All methods were carried out in accordance with relevant guidelines and regulations.

Competing interests

The authors from GeneMind declare that they have no competing interests.

Funding

Not applicable.

References

1. Bewicke-Copley, F., Kumar, E. A., Palladino, G., Korfi, K. & Wang, J. Applications and analysis of targeted genomic sequencing in cancer studies. *Computational structural biotechnology journal* **17**, 1348-1359 (2019).
2. Tan, A. C. *et al.* 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 (2020).
3. Nagahashi, M. *et al.* Next generation sequencing-based gene panel tests for the management of solid tumors. *Cancer science* **110**, 6-15 (2019).
4. Zhao, C. *et al.* TruSight oncology 500: enabling comprehensive genomic profiling and biomarker reporting with targeted sequencing. *BioRxiv* (2020).
5. Buchhalter, I. *et al.* Size matters: dissecting key parameters for panel-based tumor mutational burden analysis. *International journal of cancer* **144**, 848-858 (2019).
6. Li, H. *et al.* Genetic alteration of Chinese patients with rectal mucosal melanoma. *BMC cancer* **21**, 1-9 (2021).
7. Sahajpal, N. S. *et al.* 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 (2020).
8. Yalçıntepe, S. *et al.* The Application of Next Generation Sequencing Maturity Onset Diabetes of the Young Gene Panel in Turkish Patients from Trakya Region. *Journal of Clinical Research in Pediatric Endocrinology* **13**, 320 (2021).
9. Wang, C.-Y. *et al.* Development and validation of an expanded targeted sequencing panel for non-invasive prenatal diagnosis of sporadic skeletal dysplasia. *BMC Medical Genomics* **14**, 1-13 (2021).
10. Sahajpal, N. S. *et al.* 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 (2020).

11. Ramos-Paradas, J. *et al.* Tumor mutational burden assessment in non-small-cell lung cancer samples: results from the TMB(2) harmonization project comparing three NGS panels. *J Immunother Cancer* **9**, doi:10.1136/jitc-2020-001904 (2021).
12. Pestinger, V. *et al.* Use of an integrated pan-cancer oncology enrichment next-generation sequencing assay to measure tumour mutational burden and detect clinically actionable variants. *Molecular diagnosis therapy* **24**, 339-349 (2020).
13. Conroy, J. M. *et al.* A scalable high-throughput targeted next-generation sequencing assay for comprehensive genomic profiling of solid tumors. *Plos one* **16**, e0260089 (2021).
14. Liu, Y. *et al.* Comparative performance of the GenoLab M and NovaSeq 6000 sequencing platforms for transcriptome and LncRNA analysis. *BMC genomics* **22**, 1-12 (2021).
15. Smith, E. N. *et al.* Biased estimates of clonal evolution and subclonal heterogeneity can arise from PCR duplicates in deep sequencing experiments. *Genome biology* **15**, 1-10 (2014).
16. Petrackova, A. *et al.* Standardization of sequencing coverage depth in NGS: recommendation for detection of clonal and subclonal mutations in cancer diagnostics. *Frontiers in oncology*, 851 (2019).
17. Wang, M. *et al.* SomaticCombiner: improving the performance of somatic variant calling based on evaluation tests and a consensus approach. *Scientific reports* **10**, 1-16 (2020).
18. Yan, Y. H. *et al.* Confirming putative variants at≤ 5% allele frequency using allele enrichment and Sanger sequencing. *Scientific reports* **11**, 1-9 (2021).

Figures

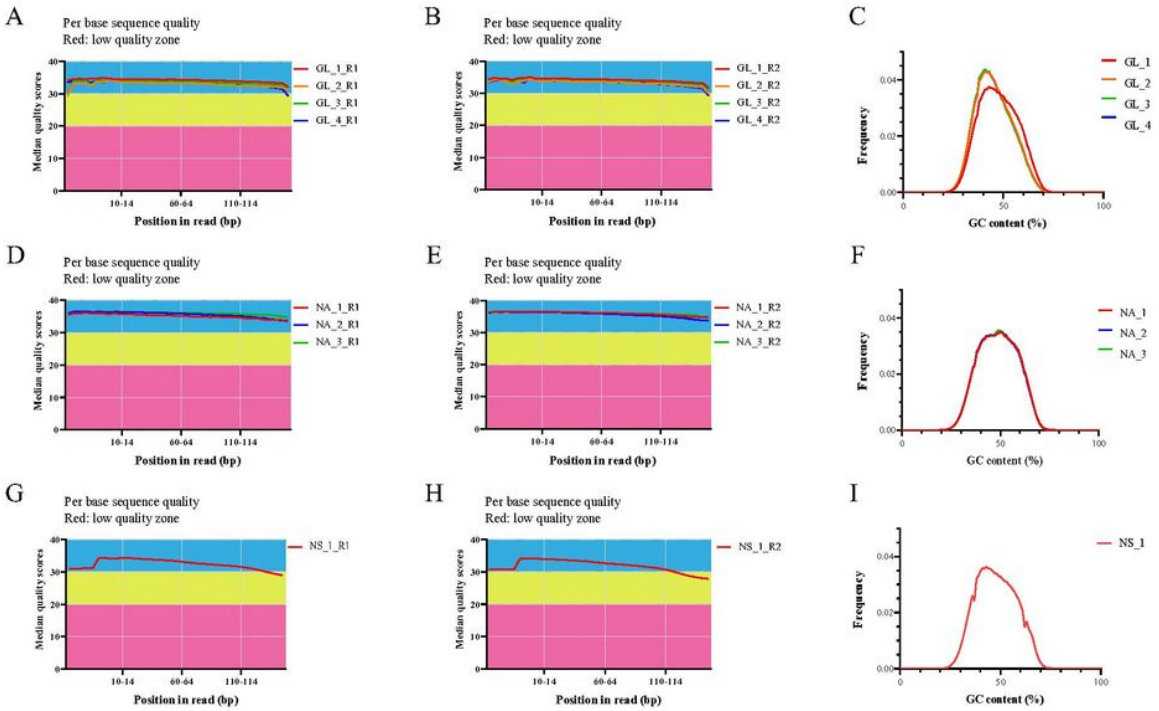

Figure 1

Data quality for three sequencing platforms. (A, B, C) GenoLab M platform, (D, E, F) NovaSeq 6000 platform, (G, H, I) NextSeq 550 platform. A, D, and G represent the overall per base sequence quality of forward (R1) reads; B, E, and H represent the overall per base sequence quality of reverse (R2) reads; C, F, and I represent the distribution of per sequence GC content. GL, GenoLab M; NA, NovaSeq 6000; NS, NextSeq 550.

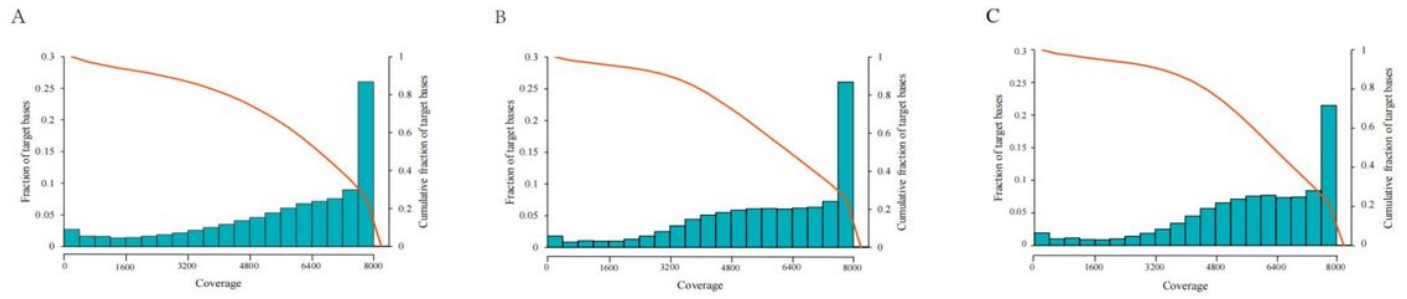

Figure 2

The coverage distribution of the target genes. (A) GenoLab M, (B) NovaSeq 6000, (C) NextSeq 550 platforms. Representative data from sample HD832 showing the fraction of bases covered at incremental depth windows (blue bars and Left Y axis) and the cumulative fraction of bases covered at the specified coverage (red line and Right Y axis).

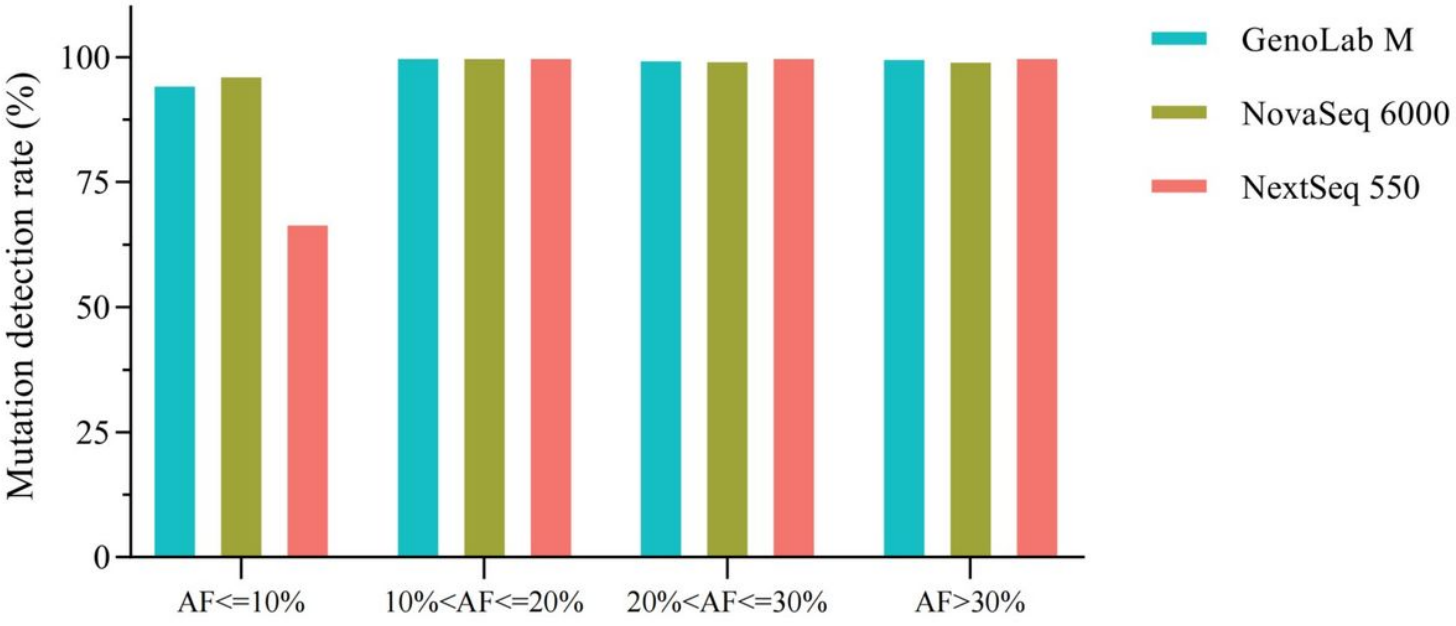

Figure 3

Analysis of mutation detection rate for GenoLab M, NovaSeq 6000, and NextSeq 550 platforms at different allele frequencies.

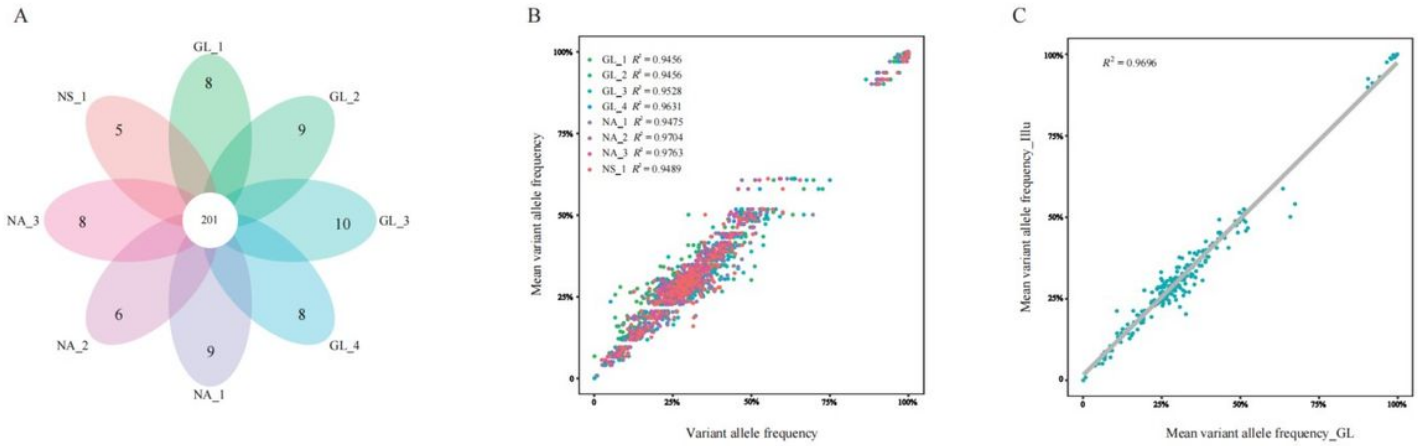

Figure 4

Comparison of concordant variants in HD832 sample captured by TSO500 panel cross-platform. (A) Venn-Diagram of target-region-mapping variants. The number outside showed the individual number of mutants identified from each dataset generated on different platforms. The core of the Venn-Diagram listed the number of the overlapped variants detected by all three platforms. (B) Linear models calculated the adjusted R^2 of the detected variant allele frequency (actual) compared with mean variant allele frequency (expected) of three platforms. (C) Linear models calculated the adjusted R^2 of the allele frequency measurements in concordant variant callings made by the GenoLab M and Illumina's platforms. GL, GenoLab M; NA, NovaSeq 6000; NS, NextSeq 550; Illu, Illumina's platforms (NovaSeq 6000 and NextSeq 550).

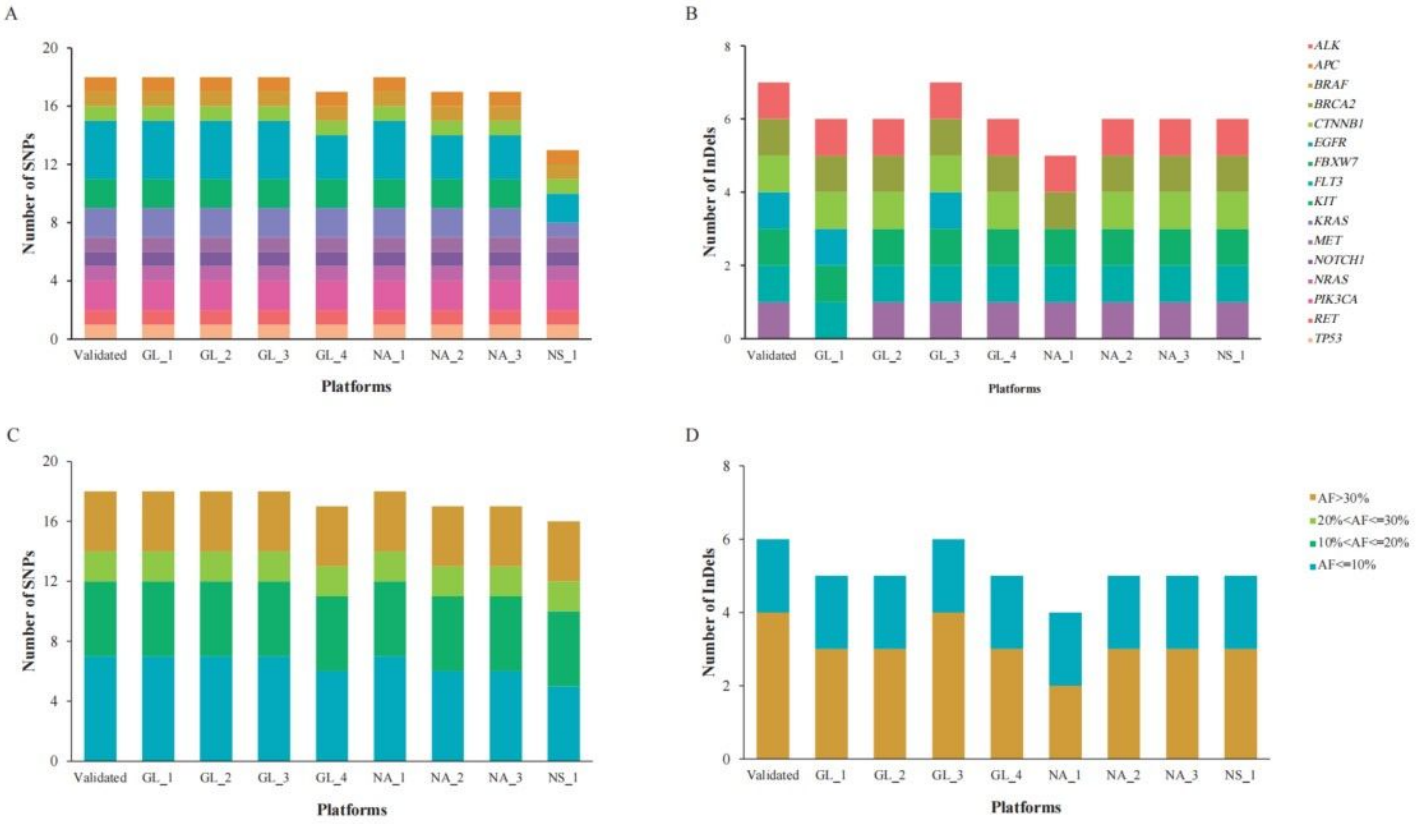

Figure 5

Distribution of the ddPCR-confirmed variants detected by three platforms. The variants included SNPs and InDels, which were distributed across a range of genes (A and B) and variant allele frequencies (C and D). GL, GenoLab M; NA, NovaSeq 6000; NS, NextSeq 550.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryinformation.pdf