frontiers in Genetics

ORIGINAL RESEARCH
published: 27 October 2021
doi: 10.3389/fgene.2021.730519

Evaluation of the MGISEQ-2000 Sequencing Platform for Illumina Target Capture Sequencing Libraries

Jidong Lang[1,2,3*†], Rongrong Zhu[4†], Xue Sun[1], Siyu Zhu[5], Tianbao Li[1,2], Xiaoli Shi[1], Yanqi Sun[1], Zhou Yang[1], Weiwei Wang[1,2], Pingping Bing[3], Binsheng He[3*] and Geng Tian[1,2*]

[1]Bioinformatics and R and D Department, Geneis (Beijing) Co. Ltd., Beijing, China, [2]Qingdao Geneis Institute of Big Data Mining and Precision Medicine, Qingdao, China, [3]Academician Workstation, Changsha Medical University, Changsha, China, [4]Vascular Surgery Department, Tsinghua University Affiliated Beijing Tsinghua Changgung Hospital, Beijing, China, [5]Department of Medicine, School of Medicine, University of California at San Diego, La Jolla, CA, United States

OPEN ACCESS

Edited by:
Liqian Zhou,
Hunan University of Technology,
China

Reviewed by:
Bing Wang,
Anhui University of Technology, China
Attila Patocs,
Semmelweis University, Hungary
Li Peng,
Hunan University of Science and
Technology, China

***Correspondence:**
Jidong Lang
langjd@geneis.cn
Binsheng He
hbscsmu@163.com
Geng Tian
tiang@geneis.cn

†These authors have contributed equally to this work.

Specialty section:
This article was submitted to
RNA,
a section of the journal
Frontiers in Genetics

Received: 25 June 2021
Accepted: 24 September 2021
Published: 27 October 2021

Citation:
Lang J, Zhu R, Sun X, Zhu S, Li T, Shi X, Sun Y, Yang Z, Wang W, Bing P, He B and Tian G (2021) Evaluation of the MGISEQ-2000 Sequencing Platform for Illumina Target Capture Sequencing Libraries. Front. Genet. 12:730519. doi: 10.3389/fgene.2021.730519

Illumina is the leading sequencing platform in the next-generation sequencing (NGS) market globally. In recent years, MGI Tech has presented a series of new sequencers, including DNBSEQ-T7, MGISEQ-2000 and MGISEQ-200. As a complex application of NGS, cancer-detecting panels pose increasing demands for the high accuracy and sensitivity of sequencing and data analysis. In this study, we used the same capture DNA libraries constructed based on the Illumina protocol to evaluate the performance of the Illumina Nextseq500 and MGISEQ-2000 sequencing platforms. We found that the two platforms had high consistency in the results of hotspot mutation analysis; more importantly, we found that there was a significant loss of fragments in the 101–133 bp size range on the MGISEQ-2000 sequencing platform for Illumina libraries, but not for the capture DNA libraries prepared based on the MGISEQ protocol. This phenomenon may indicate fragment selection or low fragment ligation efficiency during the DNA circularization step, which is a unique step of the MGISEQ-2000 sequence platform. In conclusion, these different sequencing libraries and corresponding sequencing platforms are compatible with each other, but protocol and platform selection need to be carefully evaluated in combination with research purpose.

Keywords: illumina sequencing platform, MGISEQ-2000 sequencing platform, next generation sequencing, DNA nanoball, target capture library

INTRODUCTION

With the launch of the Human Genome Project, next-generation sequencing (NGS) technology has had a huge impact on the biological field in the past 20 years (Consortium, 2015; Yang et al., 2015; Goodwin et al., 2016). Different companies and research institutions have developed various sequencing approaches and platforms, such as Roche's 454 sequencing platform, Illumina's sequencing by synthesis (SBS) technology, and PacBio's single-molecule nanopore sequencing technology (Rivas et al., 2015; Goodwin et al., 2016). Among them, the sequencers or sequencing platforms developed by the Illumina Company have a dominant position in the sequencing market due to their high throughput and high sequencing accuracy. Over time, the development of machine hardware and the diversification of bioinformatics analysis software tools have led to drastic reductions in sequencing costs and increases in convenience and usability, even for new developed techniques like single cell sequencing (Yang et al., 2020a; Xu et al., 2020). For example, NGS technology plays a vital role in analyzing somatic mutations that occur in multiple tumor types. The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013) and International Cancer Genome Consortium (ICGC) (Hudson et al., 2010) have sequenced thousands of tumors from more than 50 cancer types and summarized the significant genetic somatic mutations that occur during the process of tumorigenesis (Alexandrov et al., 2013). These data have played an extremely important role in promoting cancer genome research and development (He et al., 2020a; He et al., 2020b; Liu et al., 2021).

Recently, MGI Tech Co., Ltd (referred to MGI) launched a series of NGS sequencers and platforms based on DNA nanoball (DNB) and probe-anchor synthesis (cPAS) technology, such as MGISEQ-200, MGISEQ-2000, and DNBSEQ-T7 (Fehlmann et al., 2016). They have gradually achieved a certain sales volume and have become another option for high-throughput sequencing. For example, MGISEQ-2000 can generate approximately 1.44 TB sequencing data per run with a running cost of only 10 USD/GB. Several studies have compared the performance between MGI and the Illumina sequencing platform, and the results showed that they were highly consistent for different types of sequencing libraries, including whole-exome sequencing (WES) (Xu et al., 2019), whole-genome sequencing (WGS) (Patch et al., 2018), transcriptome sequencing (Zhu et al., 2018; Jeon et al., 2019; Patterson et al., 2019; Zeng et al., 2020), single-cell transcriptome sequencing (Natarajan et al., 2019; Peng et al., 2020a; Senabouth et al., 2020; Zhuang et al., 2021), metagenome sequencing (Fang et al., 2018) and small RNA sequencing (Huang et al., 2017) libraries.

When MGI launched their sequencers, they indicated that they were compatible with the sequencing libraries constructed based on Illumina protocols, that is, that the MGISEQ platform could sequence the Illumina libraries. In our study, we used the same capture DNA libraries constructed based on the Illumina protocol for sequencing with the Illumina NextSeq 500 and MGISEQ-2000 sequencing platforms. We found that the two platforms had high consistency in the hotspot mutation analysis and that there was a significant loss of the 101–133 bp fragments on the MGISEQ-2000 sequencing platform but not in the capture DNA libraries based on the MGISEQ protocol. We hypothesized that this might be related to fragment selection or low ligation efficiency during the DNA circularization step, a step that is unique to the MGISEQ-2000 sequence platform. Hence, although the selection of sequencers and platforms is becoming increasingly diversified and all theoretically compatible and applicable to each other, the choice of platform for practical applications may need to be further evaluated according to the research purpose and library characteristics.

MATERIALS AND METHODS

Sample Collection and Experimental Groups

Our research was approved by the Qingdao Geneis Institute of Big Data Mining and Precision Medicine in November 2019, and the research ID was Ethics-QD-[2020] No. 001. A total of 272 samples (patient age: 29–91 years old) were collected at Qingdao Geneis Institute of Big Data Mining and Precision Medicine from December 2019 to March 2020, including 79 plasma samples, 21 white blood cell samples and 172 formalin-fixed and paraffin-embedded (FFPE) samples. Informed written consent forms were obtained from patients, and identifying information was removed. The clinical information of the samples is shown in **Table 1**.

TABLE 1 | Clinical information for collected samples.

Clinical characteristics		All samples (n = 272)
Unknown		46
Age, Median (Range)-yrs		62.5 (29.0–91.0)
Age groups-No.%	15–49 years	24/226 (10.62)
	50–64 years	97/226 (42.92)
	≥65 years	105/226 (46.46)
Sex-No.%	Female	103/226 (45.58)
	Male	123/226 (54.42)
Disease-No.%	Lung cancer	166/226 (73.45)
	Colon cancer	13/226 (5.75)
	Rectal cancer	11/226 (4.87)
	Gastric cancer	6/226 (2.65)
	Breast cancer	5/226 (2.21)
	Esophageal cancer	5/226 (2.21)
	Colorectal cancer	4/226 (1.77)
	Nasopharyngeal carcinoma	2/226 (0.88)
	Liver cancer	1/226 (0.44)
	Ovarian cancer	1/226 (0.44)
	Tongue cancer	1/226 (0.44)
	Unknown	11/226 (4.87)

We randomly selected 204 (75%: 204/272) samples to construct capture libraries based on the Illumina protocol and performed data analysis. The remaining samples were divided into two groups of 34 samples (12.5%: 34/272) using different capture panels and constructing capture libraries based on the MGISEQ protocol for sequencing and data analysis, respectively.

Library Preparation Based on Illumina Platform and Sequencing

DNA for NGS-based analysis was extracted using the GeneRead Kit (Qiagen, Hilden, Germany) for FFPE tissue and the QIAamp DNA Blood Mini Kit (Qiagen, Hilden, Germany) for white blood cell samples. DNA (200 ng) was used to build the library by using the NEBNext Ultra II DNA library Prep Kit for Illumina (96 reactions) (NEB, Ipswich, MA, United States). Cell-free DNA was extracted using a QIAamp Circulating Nucleic Acid Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. The extracted DNA (20 ng/sample) was then used to build libraries using Accel-NGS® 2S Plus DNA Library Kits (96 reactions; Swift BioSciences, Ann Arbor, MI, United States). Integrated DNA Technologies (IDT, Skokie, IL, United States) or Agilent Technologies (Santa Clara, CA, United States) custom probes were used for hybridization capture. We used the IDT 38-hotspot gene panel or Agilent 519 gene panel (**Supplementary Table S5**) for all 272 libraries.

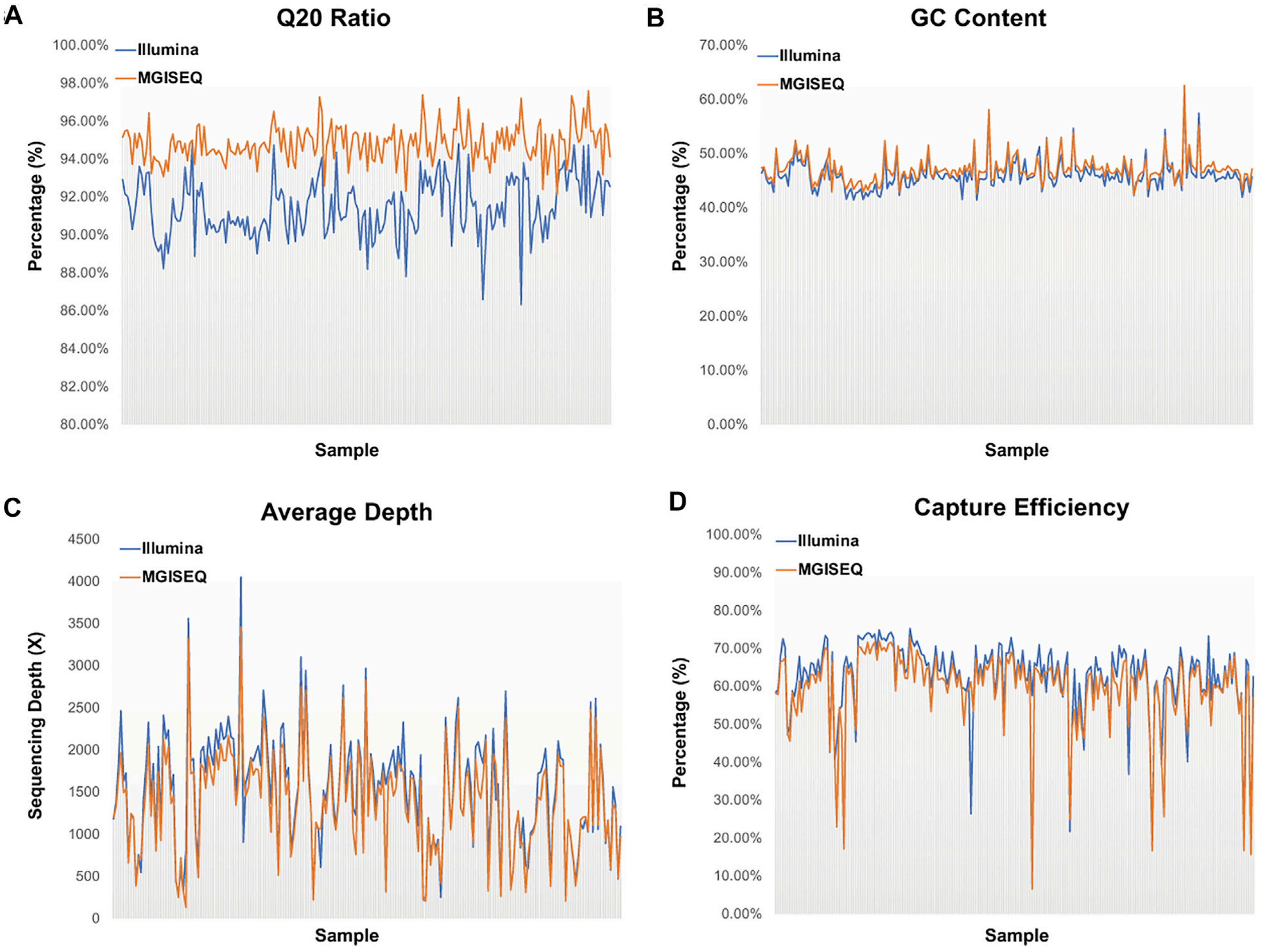


FIGURE 1 | Comparison of Sequencing Data Quality Control Parameters between Illumina and MGISEQ-2000 Platforms. **(A)** Distribution of Q20 ratio by each sample. **(B)** Distribution of GC content by each sample. **(C)** Distribution of average depth by each sample. **(D)** Distribution of probe capture efficiency by each sample.

Quantification was performed with an Illumina/Universal Library Quantification Kit (Kapa Biosystems, Wilmington, MA, United States) on an ABI 7500 Real Time Polymerase Chain Reaction (PCR) System (Applied Biosystems, Waltham, MA, United States). The quality control for Agilent 2,100 Bioanalyzer used a High-Sensitivity DNA Kit (Agilent Technologies, Santa Clara, CA, United States). Next-generation sequencing-based analysis was performed on a NextSeq500 or MiSeqDX instrument according to the manufacturer's instructions (Illumina, San Diego, CA, United States). With the NextSeq500/550 High Output V2 Kit or MiSeqTMDX Reagent V3 Kit, Illumina NextSeq500 or MiSeqDX (Illumina, San Diego, CA, United States) was used for DNA sequencing in 302 cycles for 151 bp paired-end sequencing. All 272 libraries were also analyzed on a MGISEQ2000 instrument according to the manufacturer's instructions (BGI, Shenzhen, Guangdong, China). With the MGISEQ-2000RS High Output kit (BGI, Shenzhen, Guangdong, China), MGISEQ-2000 (BGI, Shenzhen, Guangdong, China) was used for DNA sequencing in 200 cycles and 300 cycles for 100 bp and 150 bp paired-end sequencing, respectively.

Library Preparation Based on the MGISEQ Platform and Sequencing

DNA libraries were prepared with the MGIEasy FS DNA Library Prep Set (BGI, Shenzhen, Guangdong, China). DNA (50–200 ng) was fragmented physically with a Covaris S220 instrument (Covaris, Woburn, MA, United States), followed by A-tailing, adapter ligation and PCR amplification. DNA library quality was assessed using a Qubit and Agilent 2,100 Bioanalyzer with a High Sensitivity DNA Kit. Cot-1 DNA blocking reagent (Thermo Fisher Scientific, Waltham, MA, United States), IDT universal blocking oligonucleotides and IDT adapter-specific blocking oligonucleotides were added to the pooled libraries and dried in a SpeedVac. The dried mixture was redissolved in mixed liquids of IDT hybridization buffer, IDT hybridization enhancer and BOKE capture probes (BOKE bioscience, Bejing, China). After hybridization at 65°C for 4 h, the target regions were captured with M270 streptavidin

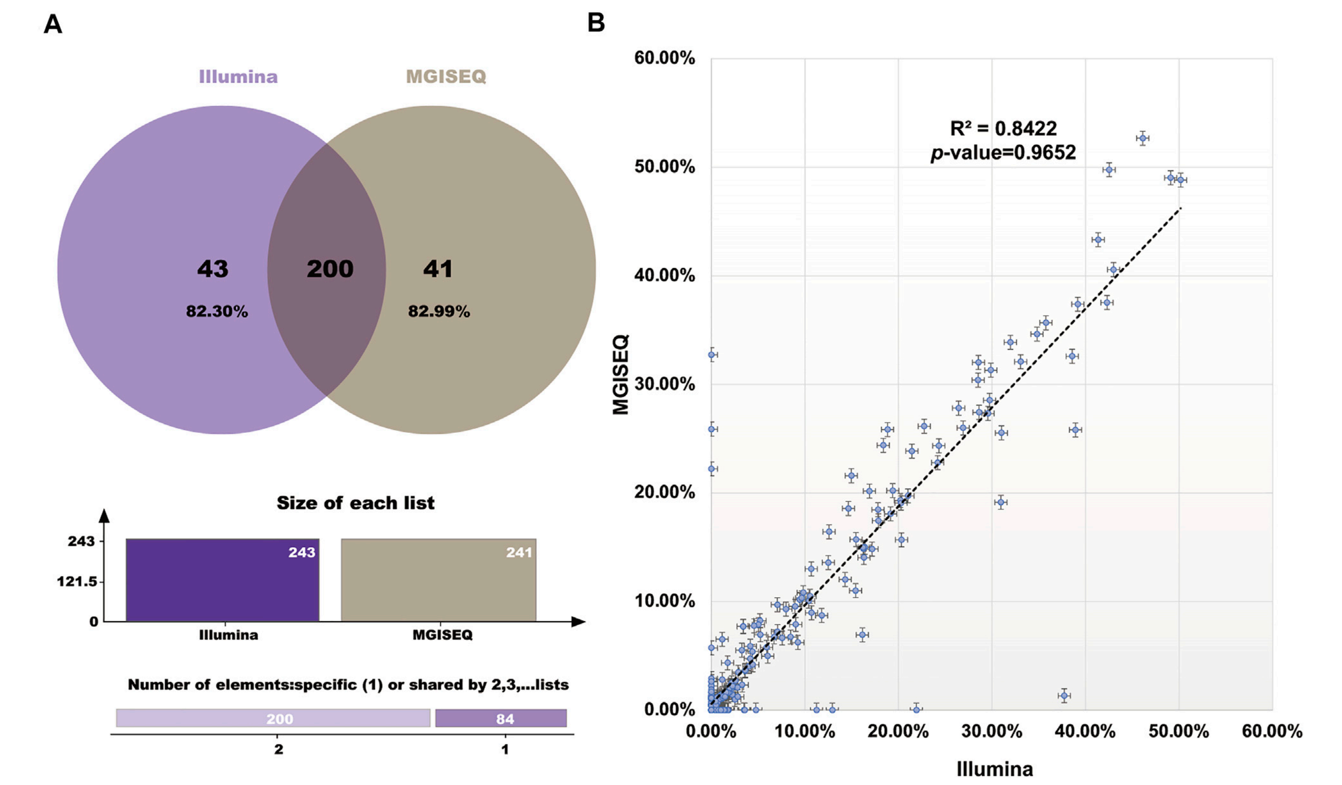


FIGURE 2 | Hotspot Mutation Results Comparison between the Illumina and MGISEQ-2000 Platforms. **(A)** The Venn digram of the detected hotspot mutations comparision. **(B)** The correlation comparison of the detected hotspot mutation frequency values.

beads by incubation at 65°C for 45 min and then washed 3 times at 65°C and another 3 times at room temperature with IDT xGen lockdown reagents. Then, 15 postcapture amplification cycles were performed to obtain the captured libraries. Final libraries were pooled and sequenced using the MGISEQ-2000 sequencing platform with a 150 bp paired-end cycle kit.

Data Normalization and Statistics

As the volume of sequencing data and read length of the Illumina and MGISEQ-2000 platforms were different (**Supplementary Table S1**), we "normalized" all 272 sample sequencing datasets, that is, each sample had the same read length and read number. We used seqtk (version: 1.0-r73-dirty) (https://github.com/lh3/seqtk) to "normalize" the raw sequencing data. We used a in-house perl program to caculate the number of reads, Q20 ratio and GC content (**Supplementary Table S2**).

Data Preprocessing and Analysis

The normalized data were cleaned by Trimmomatic (version: 0.39) (Bolger et al., 2014), which filtered out the adapter contamination reads and low-quality reads and the parameter's setting was ILLUMINACLIP:adapter sequence:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 (adapter sequences for Illumina Nextseq 500 and MGISEQ-2000 were AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC/AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA and AAGTCGGAGGCCAAGCGGTCTTAGGAAGACAA/AAGTCGGATCGTAGCCATGTCGTTCTGTGAGCCAAGGAGTTG, respectively). BWA-ALN algorithm (version: 0.7.12) (Li and Durbin, 2009) was applied for alignment with the reference genome hg19 (parameters: -o 1 -e 50 -t 4 -i 15 -q 10). The output SAM file was sorted and deduplicated with Samtools (version: 0.1.19) (Li et al., 2009), and the BAM format file was obtained. We used FreeBayes (version: 1.0.2) (Garrison and Marth, 2012) to detect SNP/InDel mutations (parameters: -j -m 10 -q 20 -F 0.001 -C 1). The mutations were annotated from the ANNOVAR database (Wang et al., 2010). Fragment size distribution was summarized from the paired-end alignment information (column ninth) in the BAM format file. Statistical analysis used the statistical functions in Microsoft Excel 2019 and R software (version 3.2.5).

RESULTS

Data Quality Control Parameters Were Significantly Different Between the Illumina and MGISEQ-2000 Sequencing Platforms

We compared the Q20 rate, GC content, mean depth and capture efficiency of 204 samples generated based on the Illumina library protocol, which were captured by the IDT 38-hotspot gene panel and sequenced on the Illumina and MGISEQ-2000 sequencing

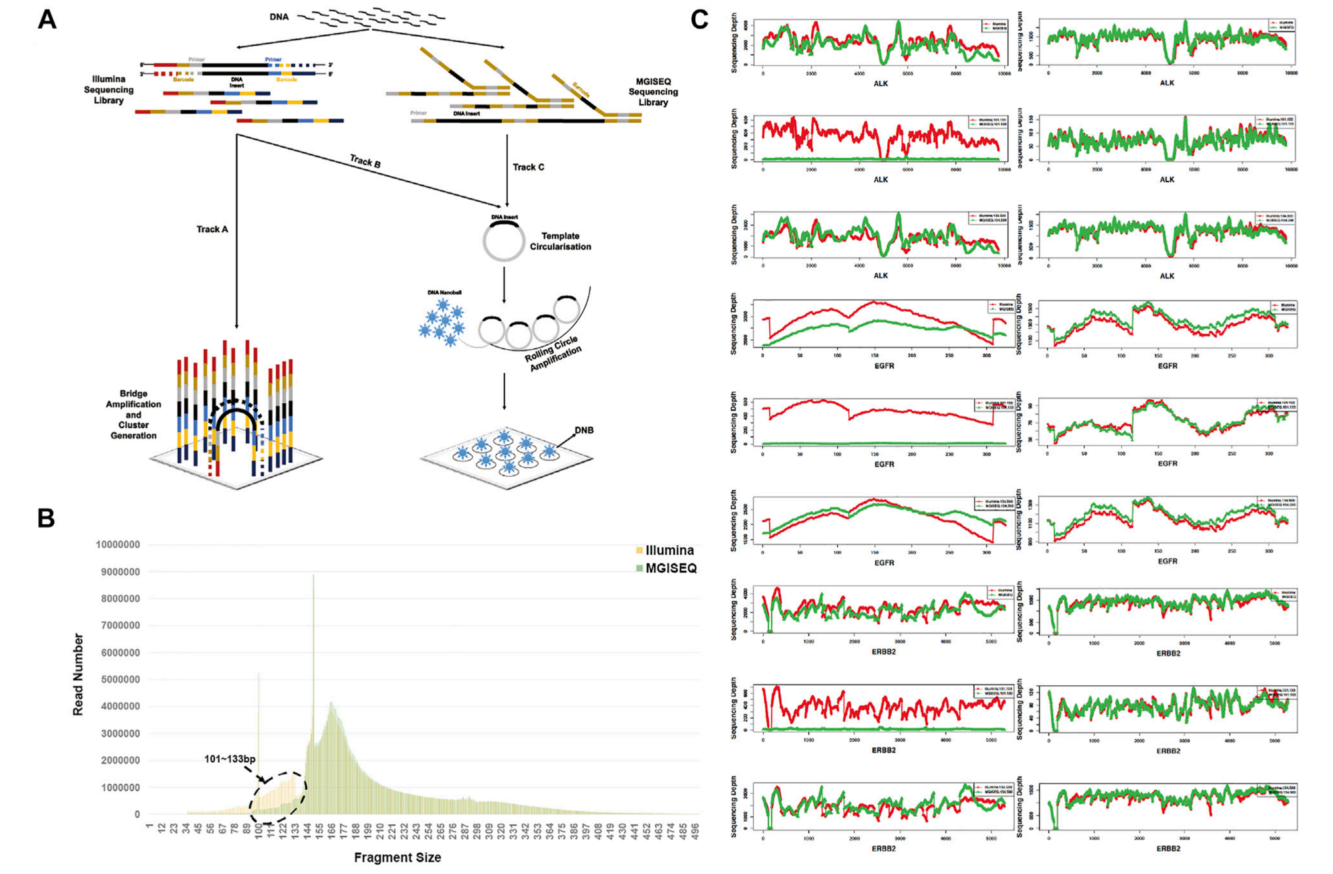


FIGURE 3 | The Insert Fragment Size Analysis of Illumina-based Capture Library On the Illumina and MGISEQ-2000 Platforms. **(A)** The principles of Illumina and MGISEQ sequencers. Track A represents the Illumina library combined with bridge PCR amplification of the Illumina platform. Track B represents the Illumina library combined with DNA circularization for DNB of the MGISEQ-2000 platform. Track C represents the MGISEQ library combined with DNA circularization for DNB of the MGISEQ-2000 platform. **(B)** Compared with the Illumina platform data, the MGISEQ-2000 platform data had a significant loss of 101–133 bp fragments. **(C)** Statistical analysis of sequencing depth distribution in ALK, EGFR and ERBB2 with one sample. Sequencing depth distribution of 101–133 bp (left panel) and 134–500 bp (right panel) in ALK (top), EGFR (middle), and ERRBB2 (bottom). In each figure panel, the top panel shows the sample total sequencing depth distribution, the middle panel shows the sequencing depth of 101–133 bp fragment size, and the bottom panel shows the 134–500 bp fragment size sequencing depth distribution.

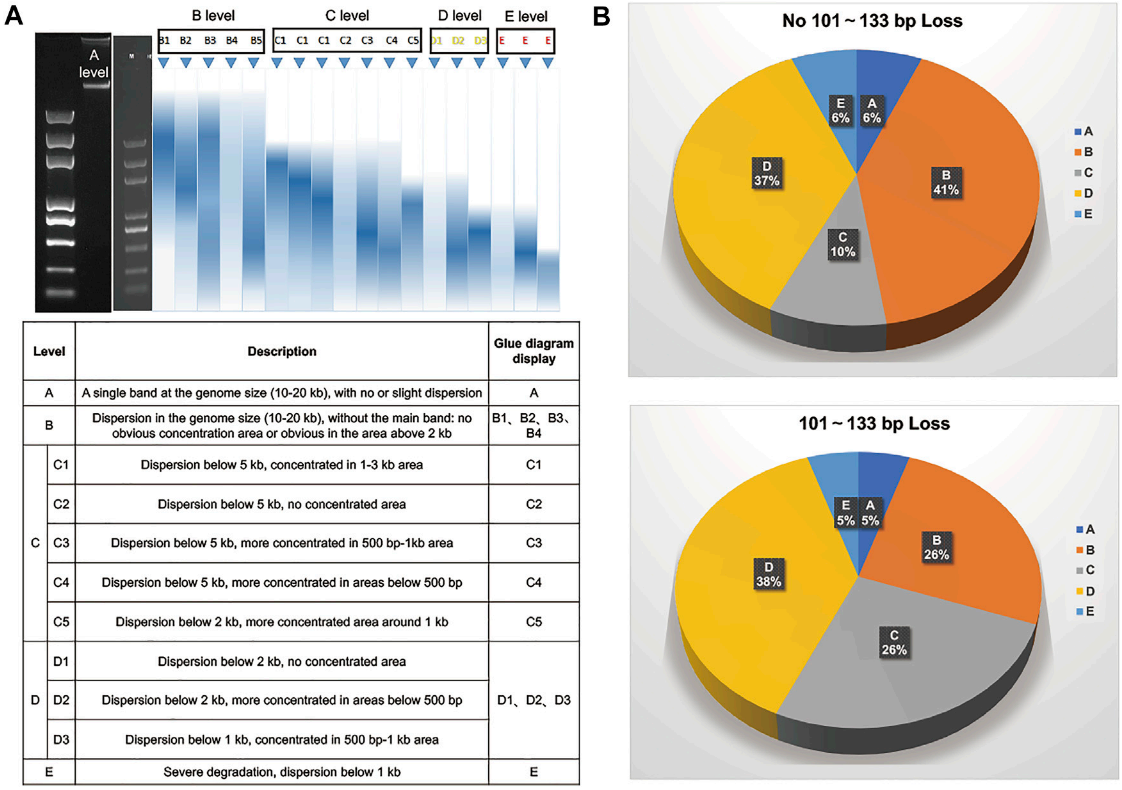
platforms (**Figure 1**, details in **Supplementary Table S3**), respectively. We found that all of the quality control parameters had significant differences, with *p*-values of 4.87e-85, 1.15e-4, 0.0326 and 0.0035, respectively, in the two-tailed heteroscedasticity *t*-test analysis. We thought that these differences could be due to the sequencing principles, the algorithm used for base recognition or the sequencing platform characteristics. For example, the Nextseq500 platform treated all unrecognized bases as G, while HiSeq-2000, MGISEQ-2000 and other previous four-color imaging sequencers treated these bases as N. Therefore, the GC content tended to be higher in the Illumina NextSeq500 results than in the others.

Hotspot Mutations Showed High Consistency Between the Illumina and MGISEQ-2000 Sequencing Platforms.

The hotspot mutations (SNPs and InDels) detected in 204 sample datasets were compared between the Illumina and MGISEQ-2000 platforms (**Supplementary Table S4**). We defined a positive detection filter condition as mutation frequency ≥ 0.4% for plasma samples and mutation frequency ≥ 1% for FFPE samples. We found that the hotspot mutation detection results had high consistency rates of 82.30% (Illumina: 200/243) and 82.99% (MGISEQ-2000: 200/241) (**Figure 2A**). Furthermore, no significant difference (R^2 = 0.8422, *p*-value = 0.9652) in mutation frequency was observed between the Illumina and MGISEQ-2000 platform data. (**Figure 2B**).

MGISEQ-2000 sequencing platform data based on Illumina libraries showed a significant loss of the 101–133 bp fragment.

Insert fragment size and distribution were evaluated and analyzed for all 204 samples. As we used the same sample library for sequencing, the theoretical difference only existed in Illumina's bridge PCR amplification and MGISEQ-2000s DNB circularization. (**Figure 3A**) (Goodwin et al., 2016; Chen et al., 2019; Korostin et al., 2020). Combining all 204 sample data for fragment size analysis, our results revealed a significant loss of 101–133 bp fragments in the MGISEQ-2000 platform data, with a



Level		Description	Glue diagram display
A		A single band at the genome size (10-20 kb), with no or slight dispersion	A
B		Dispersion in the genome size (10-20 kb), without the main band: no obvious concentration area or obvious in the area above 2 kb	B1、B2、B3、B4
C	C1	Dispersion below 5 kb, concentrated in 1-3 kb area	C1
	C2	Dispersion below 5 kb, no concentrated area	C2
	C3	Dispersion below 5 kb, more concentrated in 500 bp-1kb area	C3
	C4	Dispersion below 5 kb, more concentrated in areas below 500 bp	C4
	C5	Dispersion below 2 kb, more concentrated area around 1 kb	C5
D	D1	Dispersion below 2 kb, no concentrated area	D1、D2、D3
	D2	Dispersion below 2 kb, more concentrated in areas below 500 bp	
	D3	Dispersion below 1 kb, concentrated in 500 bp-1 kb area	
E		Severe degradation, dispersion below 1 kb	E

FIGURE 4 | Statistical Analysis On The Quality of 204 Samples. **(A)** Sample quality grading table of gDNA agarose gel electrophoresis. **(B)** The distribution of different sample quality levels in samples with and without loss of 101–133 bp fragment size. The top figure represented sample quality grade distribution of samples without 101–133 bp fragment size loss. The bottom figure represented sample quality grade distribution of samples with 101–133 bp fragment size loss.

t-test *p*-value of 3.3072e-17 (**Figure 3B**), while other fragment sizes, such as 134–500 bp (*t*-test *p*-value = 0.7264), did not show a difference. Although significant differences were found in the Q20 rate, GC content and other quality control statistics, these should be attributable to the sequencer system characteristics and should not have a great impact on the fragment size distribution. Therefore, the loss of the 101–133 bp fragment size may be related to the DNA cyclization step, that is, there may be fragment size selection in the circularization step or enrichment bias for longer DNA molecules and low ligation efficiency for shorter DNA molecules.

Then, we extracted 101–133 bp and 134–500 bp fragment size information from BAM files for each sample and analyzed the sequencing depth distribution of three common cancer genes, ALK receptor tyrosine kinase (*ALK*), epidermal growth factor receptor (*EGFR*) and erb-b2 receptor tyrosine kinase 2 (*ERBB2*). The results showed that 69.12% (141/204) of samples had 101–133 bp fragment size loss, while the sequencing depth distribution of 134–500 bp fragments was consistent with the overall total sequencing depth, indicating that the phenomenon was not due to stochasticity in specific genes (**Figure 3C**). The sequencing depth distribution of all samples was in the Supplementary Figures by each sample.

As we know, the use of FFPE or hemolyzed samples may have a great influence on the distribution of DNA fragment size. Therefore, we performed statistical analysis on the quality of 204 samples with and without 101–133 bp loss. First, we defined the sample quality levels with DNA agarose gel electrophoresis as A, B, C, D or E (**Figure 4A**). Then, all samples in each grade were subgrouped according to whether the 101–133 bp fragment size was lost. We found that the sample proportions of A, D and E levels were consistent in the two groups, while B and C levels were quite different. The proportions of B [C] level samples in the 101–133 bp loss group and 101–133 bp nonloss group were 25.53% (36/141) [26.24% (37/141)] and 41.27% (26/63: 6) [9.52% (6/63)], respectively (**Figure 4B**). Therefore, our results showed that the circularization step of MGISEQ-2000 not only biased the selection of DNA fragment size but also may have a greater impact on samples with quality grade B or C.

Fragment Size Loss had no Probe Preference and was not Obvious in the Database of MGISEQ-2000 Libraries.

To verify whether the phenomenon was related to capture-probe preference, we analyzed the fragment size distribution of the sequencing data from 34 samples that were captured with an Agilent 519 gene panel and sequenced separately by Illumina Nextseq500 and MGISEQ-2000. As shown in **Figure 5A**, the same 101–133 bp fragment size loss was found. In addition, we

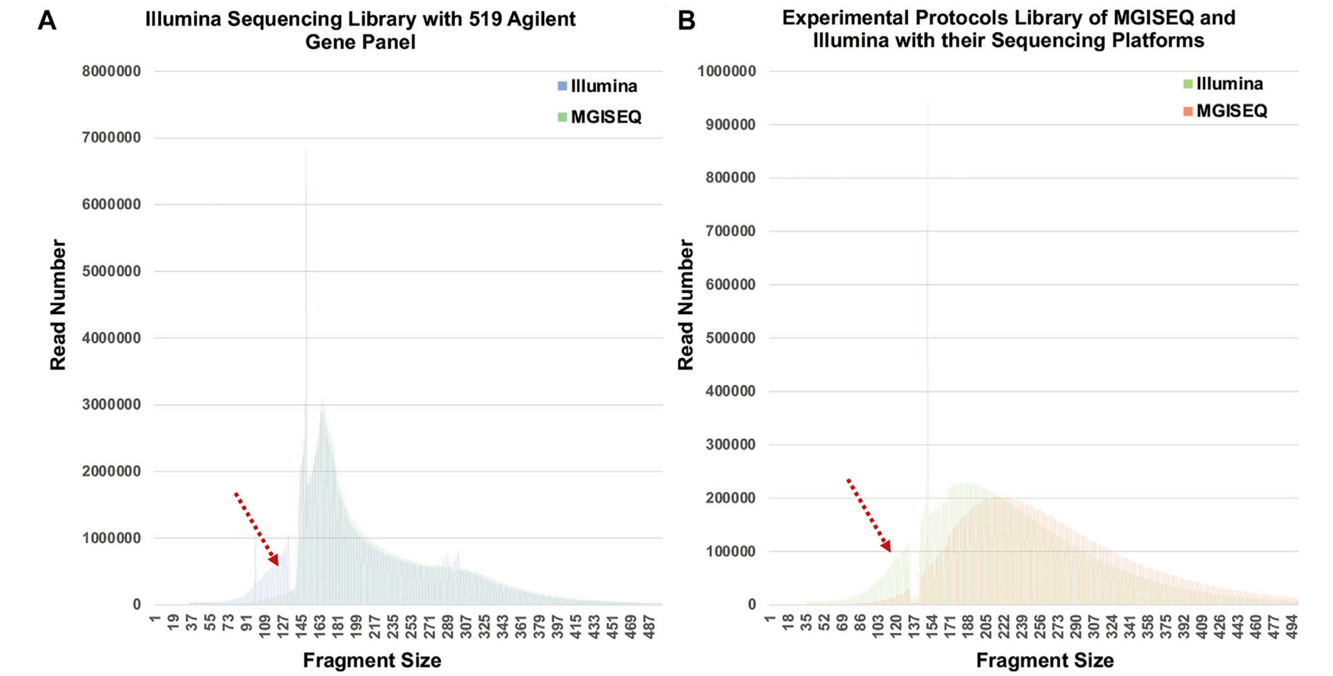


FIGURE 5 | Fragment Distribution in Illumina Nextseq500 and MGISEQ-2000. **(A)** The libraries were constructed following the instructions of Illumina and captured with an Agilent 519 gene panel. **(B)** Fragment distribution when experiments were performed according to the experimental systems and kits recommended by Illumina and MGISEQ, respectively.

constructed 34 other libraries according to the experimental protocols of MGISEQ and Illumina and generated data on their sequencing platforms. We also analyzed the fragment size distribution and found that the fragment size (peak 183 bp) distribution on the Illumina platform had a “left offset” compared to that (peak 214 bp) on the MGISEQ-2000 platform. The fragment size distribution curve of the MGISEQ data was smooth, and there was no obvious 101–133 bp fragment size loss (**Figure 5B**).

DISCUSSION

In recent decades, next-generation sequencing technology has undergone rapid development. With the greatly reduced sequencing cost, increasing scientific research and technical product development are being applied to NGS. In particular, to meet the needs of precision medicine and big data mining, the number and scale of cancer omics research and clinical projects are constantly increasing (Yang et al., 2020b; Zeng et al., 2020). For a large number of samples, the expenses and costs borne are unaffordable; thus, sequencing costs are still the bottleneck for large-scale NGS applications. At present, Illumina sequencers dominate the high-throughput sequencing market, but MGI sequencers based on DNB technology have gradually become more popular worldwide. Recently, several studies have compared the performance of BGI-500 and the Illumina HiSeq machine and showed that both of them could produce high-quality data in various applications. However, a comparison of their quality for capture panel sequencing (except WES), which is widely used in tumor research, has not been published.

In this study, we compared the data produced from the same library by different sequencing platforms. For the library preparation step, Illumina used bridge PCR technology, while MGI achieved single-molecule template amplification by DNB circularization amplification. We applied both the Illumina (Nextseq500 and MiSeqDx) platform and MGISEQ (MGISEQ-2000) platform to the same library constructed by the Illumina protocol. Theoretically, any difference in sequencing data should have been caused by the differences between bridge PCR and circularization amplification or the consequent sequencing system differences. Comparison of the data analysis results revealed the disadvantage of fragment size selection and short fragment size ligation efficiency in the circularization step. These results suggest that the sequencing data based on Illumina library preparations and in which sample types with shorter fragment sizes (such as hemolyzed plasma samples) or a more complex distribution of DNA fragment sizes (such as FFPE samples with longer storage times) are used may encounter short DNA fragment size loss on the MGISEQ sequencing platform. Therefore, we should evaluate the compatibility of sequencing libraries and sequencing platforms for scientific research that focuses on the distribution of fragment size, especially for small RNA (Fehlmann et al., 2016), cell-free DNA (cfDNA) and circulating tumor DNA (ctDNA) research (Underhill et al., 2016; Liu et al., 2020). Although the sequencing library is basically compatible with different sequencing platforms, appropriate experimental systems and sequencing platforms should be selected based on the research purpose and sample type. Otherwise, there may be an unexpected impact on the sequencing results. Our data showed the results of only target capture panel sequencing; the assessment of other sequencing applications requires further investigation.

Considering that the alignment algorithm may also have an impact on the fragment size distribution analysis, we replaced the BWA "aln" algorithm mentioned in the article with the BWA "mem" algorithm. The "mem" algorithm is much looser than the "aln" algorithm, and it can perform local alignment and splicing. The "mem" algorithm allows multiple different parts of the sequencing reads to have their own optimal matches, resulting in multiple optimal alignment positions for the reads and greatly improving the alignment rate. After comparing and analyzing the combined data with 204 samples of the IDT 38-hotspot gene panel and 34 samples of the Agilent 519 gene panel by using the "mem" algorithm, we found that the number of reads in the 101–133 bp fragment size from the MGISEQ-2000 platform data was significantly improved (**Supplementary Figure S1**), but there were still significant differences, with *t*-test *p*-values of 0.0277 and 0.0252, respectively. The conclusion was consistent with that based on the "aln" algorithm.

We also found that the data without the 101–133 bp fragment size loss were derived from different sequencing read lengths of the Illumina Nextseq500 and MGISEQ-2000 platforms, while the data with the same sequencing read length showed the 101–133 bp fragment size loss. To investigate whether the data with or without the phenomenon were related to the sequencing read length, we reanalyzed and compared data with the same number of sequencing reads but not read length, and found that the results were consistent with the previous conclusion. Since the 101–133 bp fragment size loss was concentrated in the data with long read length (150 bp) but not in the data with short read length (100 bp), we hypothesized that the phenomenon may also be related to the sequencing read length. We will conduct more in-depth research on this point in our future work.

In summary, the MGISEQ-2000 platform has good compatibility with Illumina sequencing libraries, but the DNB circularization step may cause fragment size selection or have low ligation efficiency for short DNA fragment sizes. For the accuracy of downstream data analysis, we recommend that different sequencing platforms should be used with their official experimental systems and kits. If the experiment needs to change between different platforms, for cost considerations or other reasons, the selected platform should be evaluated carefully with respect to the purpose of the research or actual needs, as it may have a significant impact on outcomes. In the future, it would be interesting to compare the performances of two platforms in specific applications like cancer diagnosis (He et al., 2020b; Peng L.-H. et al., 2020), prognosis (Peng et al., 2020c; Song et al., 2020; Zhou et al., 2020), evolution inference (Yang et al., 2013; Yang et al., 2014), drug repositioning (Peng et al., 2015; Zhou et al., 2019; Liu et al., 2020), and so on. However, it is out of the scope of this study.

DATA AVAILABILITY STATEMENT

The data has been uploaded to NCBI - BioProject 744584.

AUTHOR CONTRIBUTIONS

GT, JL and BH designed the study, collected, analyzed and interpreted the data, and wrote the article. XuS and ZY performed the experiment. RZ, SZ, TL, XiS, YS, WW and PB reviewed and modified the article. All authors approved the final version of the article.

ACKNOWLEDGMENTS

We thank Tingting Hui in Geneis (Beijing) Co. Ltd. for modifying and adjusting the figures.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.730519/full#supplementary-material

REFERENCES

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J., and Stratton, M. R. (2013). Deciphering signatures of mutational processes operative in human cancer. *Cel Rep.* 3, 246–259. doi:10.1016/j.celrep.2012.12.008

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi:10.1093/bioinformatics/btu170

Chen, J., Li, X., Zhong, H., Meng, Y., and Du, H. (2019). Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers. *Sci. Rep.* 9, 9345. doi:10.1038/s41598-019-45835-3

Consortium, G. T. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348, 648–660. doi:10.1126/science.1262110

Fang, C., Zhong, H., Lin, Y., Chen, B., Han, M., Ren, H., et al. (2018). Assessment of the cPAS-based BGISEQ-500 platform for metagenomic sequencing. *Gigascience* 7, 1–8. doi:10.1093/gigascience/gix133

Fehlmann, T., Reinheimer, S., Geng, C., Su, X., Drmanac, S., Alexeev, A., et al. (2016). cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. *Clin. Epigenet* 8, 123. doi:10.1186/s13148-016-0287-1

Garrison, E., and Marth, G. (2012). Haplotype-Based Variant Detection from Short-Read Sequencing. *Quantitative Biol.* arXiv:1207.3907v2.

Goodwin, S., Mcpherson, J. D., and Mccombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351. doi:10.1038/nrg.2016.49

He, B., Dai, C., Lang, J., Bing, P., Tian, G., Wang, B., et al. (2020a). A machine learning framework to trace tumor tissue-of-origin of 13 types of cancer based on DNA somatic mutation. *Biochim. Biophys. Acta (Bba) - Mol. Basis Dis.* 1866, 165916. doi:10.1016/j.bbadis.2020.165916

He, B., Lang, J., Wang, B., Liu, X., Lu, Q., He, J., et al. (2020b). TOOme: A Novel Computational Framework to Infer Cancer Tissue-of-Origin by Integrating Both Gene Mutation and Expression. *Front. Bioeng. Biotechnol.* 8, 394. doi:10.3389/fbioe.2020.00394

Huang, J., Liang, X., Xuan, Y., Geng, C., Li, Y., Lu, H., et al. (2017). A reference human genome dataset of the BGISEQ-500 sequencer. *Gigascience* 6, 1–9. doi:10.1093/gigascience/gix024

Hudson, T. J., Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., et al. (2010). International network of cancer genome projects. *Nature* 464, 993–998. doi:10.1038/nature08987

Jeon, S. A., Park, J. L., Kim, J.-H., Kim, J. H., Kim, Y. S., Kim, J. C., et al. (2019). Comparison of the MGISEQ-2000 and Illumina HiSeq 4000 sequencing platforms for RNA sequencing. *Genomics Inform.* 17, e32. doi:10.5808/gi.2019.17.3.e32

Korostin, D., Kulemin, N., Naumov, V., Belova, V., Kwon, D., and Gorbachev, A. (2020). Comparative analysis of novel MGISEQ-2000 sequencing platform vs Illumina HiSeq 2500 for whole-genome sequencing. *PLoS One* 15, e0230301. doi:10.1371/journal.pone.0230301

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Liu, F., Peng, L., Tian, G., Yang, J., Chen, H., Hu, Q., et al. (2020). Identifying small molecule-miRNA associations based on credible negative sample selection and random walk. *Front. Bioeng. Biotechnol.* 8, 131. doi:10.3389/fbioe.2020.00131

Liu, H., Qiu, C., Wang, B., Bing, P., Tian, G., Zhang, X., et al. (2021). Evaluating DNA Methylation, Gene Expression, Somatic Mutation, and Their Combinations in Inferring Tumor Tissue-of-Origin. *Front. Cel Dev. Biol.* 9, 619330. doi:10.3389/fcell.2021.619330

Liu, X., Lang, J., Li, S., Wang, Y., Peng, L., Wang, W., et al. (2020). Fragment Enrichment of Circulating Tumor DNA With Low-Frequency Mutations. *Front. Genet.* 11, 147. doi:10.3389/fgene.2020.00147

Natarajan, K. N., Miao, Z., Jiang, M., Huang, X., Zhou, H., Xie, J., et al. (2019). Comparative analysis of sequencing technologies for single-cell transcriptomics. *Genome Biol.* 20, 70. doi:10.1186/s13059-019-1676-5

Patch, A.-M., Nones, K., Kazakoff, S. H., Newell, F., Wood, S., Leonard, C., et al. (2018). Germline and somatic variant identification using BGISEQ-500 and HiSeq X Ten whole genome sequencing. *PLoS One* 13, e0190264. doi:10.1371/journal.pone.0190264

Patterson, J., Carpenter, E. J., Zhu, Z., An, D., Liang, X., Geng, C., et al. (2019). Impact of sequencing depth and technology on de novo RNA-Seq assembly. *BMC Genomics* 20, 604. doi:10.1186/s12864-019-5965-x

Peng, L., Liao, B., Zhu, W., Li, Z., and Li, K. (2017). Predicting Drug-Target Interactions With Multi-Information Fusion. *IEEE J. Biomed. Health Inform.* 21 (2), 561–572. doi:10.1109/JBHI.2015.2513200

Peng, L.-H., Zhou, L.-Q., Chen, X., and Piao, X. (2020b). A computational study of potential miRNA-disease association inference based on ensemble learning and kernel ridge regression. *Front. Bioeng. Biotechnol.* 8, 40. doi:10.3389/fbioe.2020.00040

Peng, L., Tian, X., and Shen, L. (2020c). Identifying effective antiviral drugs against SARS-CoV-2 by drug repositioning through virus-drug association prediction. *Front. Genet.* 11, 1072. doi:10.3389/fgene.2020.577387

Peng, L., Tian, X., Tian, G., Xu, J., Huang, X., Weng, Y., et al. (2020a). Single-cell RNA-seq clustering: datasets, models, and algorithms. *RNA Biol.* 17 (6), 765–783. doi:10.1080/15476286.2020.1728961

Rivas, M. A., Pirinen, M., Conrad, D. F., Lek, M., Tsang, E. K., Karczewski, K. J., et al. (2015). Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* 348, 666–669. doi:10.1126/science.1261877

Senabouth, A., Andersen, S., Shi, Q., Shi, L., Jiang, F., Zhang, W., et al. (2020). Comparative performance of the BGI and Illumina sequencing technology for single-cell RNA-sequencing. *NAR Genom Bioinform* 2, lqaa034, 2020 . lqaa034. doi:10.1093/nargab/lqaa034

Song, Z., Chen, X., Shi, Y., Huang, R., Wang, W., Zhu, K., et al. (2020). Evaluating the Potential of T Cell Receptor Repertoires in Predicting the Prognosis of Resectable Non-Small Cell Lung Cancers. *Mol. Ther. - Methods Clin. Dev.* 18, 73–83. doi:10.1016/j.omtm.2020.05.020

Underhill, H. R., Kitzman, J. O., Hellwig, S., Welker, N. C., Daza, R., Baker, D. N., et al. (2016). Fragment Length of Circulating Tumor DNA. *Plos Genet.* 12, e1006162. doi:10.1371/journal.pgen.1006162

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38, e164. doi:10.1093/nar/gkq603

Weinstein, J. N., Collisson, E. A., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., et al. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* 45, 1113–1120. doi:10.1038/ng.2764

Xu, J., Cai, L., Liao, B., Zhu, W., and Yang, J. (2020). CMF-Impute: an accurate imputation tool for single-cell RNA-seq data. *Bioinformatics* 36, 3139–3147. doi:10.1093/bioinformatics/btaa109

Xu, Y., Lin, Z., Tang, C., Tang, Y., Cai, Y., Zhong, H., et al. (2019). A new massively parallel nanoball sequencing platform for whole exome research. *BMC Bioinformatics* 20, 153. doi:10.1186/s12859-019-2751-3

Yang, J., Grünewald, S., and Wan, X.-F. (2013). Quartet-net: a quartet-based method to reconstruct phylogenetic networks. *Mol. Biol. Evol.* 30, 1206–1217. doi:10.1093/molbev/mst040

Yang, J., Grünewald, S., Xu, Y., and Wan, X.-F. (2014). Quartet-based methods to reconstruct phylogenetic networks. *BMC Syst. Biol.* 8, 21. doi:10.1186/1752-0509-8-21

Yang, J., Huang, T., Huang, T., Petralia, F., Long, Q., Zhang, B., et al. (2015). Synchronized age-related gene expression changes across multiple tissues in human and the link to complex diseases. *Sci. Rep.* 5, 15145. doi:10.1038/srep15145

Yang, J., Liao, B., Zhang, T., and Xu, Y. (2020a). Editorial: Bioinformatics Analysis of Single Cell Sequencing Data and Applications in Precision Medicine. *Front. Genet.* 10, 1358. doi:10.3389/fgene.2019.01358

Yang, J., Peng, S., Zhang, B., Houten, S., Schadt, E., Zhu, J., et al. (2020b). Human geroprotector discovery by targeting the converging subnetworks of aging and age-related diseases. *Geroscience* 42, 353–372. doi:10.1007/s11357-019-00106-x

Zeng, L., Yang, J., Peng, S., Zhu, J., Zhang, B., Suh, Y., et al. (2020). Transcriptome analysis reveals the difference between "healthy" and "common" aging and their connection with age-related diseases. *Aging Cell* 19, e13121. doi:10.1111/acel.13121

Zhou, L., Li, Z., Yang, J., Tian, G., Liu, F., Wen, H., et al. (2019). Revealing drug-target interactions with computational models and algorithms. *Molecules* 24 (9), 1714. doi:10.3390/molecules24091714

Zhou, L., Wang, J., Liu, G., Lu, Q., Dong, R., Tian, G., et al. (2020). Probing antiviral drugs against SARS-CoV-2 through virus-drug association prediction based on the KATZ method. *Genomics* 112 (6), 4427–4434. doi:10.1016/j.ygeno.2020.07.044

Zhu, F.-Y., Chen, M.-X., Ye, N.-H., Qiao, W.-M., Gao, B., Law, W.-K., et al. (2018). Comparative performance of the BGISEQ-500 and Illumina HiSeq4000 sequencing platforms for transcriptome analysis in plants. *Plant Methods* 14, 69. doi:10.1186/s13007-018-0337-0

Zhuang, J., Cui, L., Qu, T., Ren, C., and Yang, J. (2021). A streamlined scRNA-Seq data analysis framework based on improved sparse subspace clustering. *IEEE Access*, 1. doi:10.1109/access.2021.3049807

Conflict of Interest: JL, XuS, TL, XiS, YS, ZY, WW, and GT were employed by Geneis (Beijing) Co. Ltd.

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.

The handling Editor declared a past co-authorship/collaboration with several of the authors.

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.