Comparative analysis of novel MGISEQ-2000 sequencing platform vs Illumina HiSeq 2500 for whole-genome sequencing

Alexey Gorbachev [1], Nikolay Kulemin [1, 2], Vladimir Naumov [1, 4], Vera Belova [2], Dmitriy Kwon [3], Denis Rebrikov [2, 4], Dmitriy Korostin [2]

Authors' affiliations

[1] Zenome.io, Ltd., Moscow, Russia

[2] Pirogov Russian National Research Medical University, Moscow, Russia

[3] Company Helicon, Ltd., Moscow, Russia

[4] National Medical Research Center for Obstetrics, Gynecology and Perinatology Named After Academician V.I. Kulakov, Moscow, Russia

Abstract

Background: MGISEQ-2000 developed by MGI Tech Co. Ltd. (a subsidiary of the BGI Group) is a new competitor of such next-generation sequencing platforms as NovaSeq and HiSeq (Illumina). Its sequencing principle relies on the DNB and cPAS technologies also used in the previous version of the BGISEQ-500 device, but the reagents for MGISEQ-2000 are refined and the platform utilizes updated software. The cPAS technology has evolved from cPAL previously created by Complete Genomics.

Findings: This article compares the results of the whole-genome sequencing of a DNA sample from a Russian female donor performed on MGISEQ-2000 and Illumina HiSeq 2500 (both PE150). Two platforms were compared in terms of sequencing quality, number of errors and performance. Additionally, we did variant calling using four different software packages: Samtools mpileaup, Strelka2, Sentieon, and GATK.

Conclusions: The accuracy of single nucleotide polymorphism (SNP) detection was similar between the data generated by MGISEQ-2000 and HiSeq 2500, which was used as a reference:
For Samtools mpileaup software package: TPR (Sensitivity) 99.30%, FPR = 0,000498%;
For Strelka2 software package: TPR (Sensitivity) 99,51%, FPR = 0,000254%;
For Sentieon software package: TPR (Sensitivity) 99,57%, FPR = 0,000285%.
For GATK software package: TPR (Sensitivity) 98,70%, FPR = 0,000240%.
At the same time, a separate indel analysis of the overall error rate revealed similar FPR values and lower sensitivity:

For Samtools mpileup: TPR (Sensitivity) 93,62%, FPR = 0,000698%;
For Strelka2: TPR (Sensitivity) 98,84%, FPR = 0,000127%;
For Sentieon: TPR (Sensitivity) 98,68%, FPR = 0,000285%.
For GATK: TPR (Sensitivity) - 98,70%, FPR = 0,000240%.
The method of statistical analysis we use does not allow us to conclusively establish which of the two instruments is the most accurate. However, it can be said with confidence that the data generated by the analyzed sequencing systems are characterized by the comparable magnitude of error and that MGISEQ-2000 can be used for a wide range of research tasks on a par with HiSeq 2500.

Background

The cPAL sequencing technology developed by Complete Genomics first came to light in 2009 [1]. In 2013, Complete Genomics was acquired by BGI (the Beijing Genomic Institute), and the technology was subsequently refined [2]. In 2015, a new commercially available second-generation genome analyzer BGISEQ-500 was first announced [3]. Since then, the cPAL technology has undergone serious modifications.

The cPAS method was an important milestone in the evolution of this technology. The method exploits fluorescently labeled terminated substrates. In cPAS, sequencing occurs as DNA polymerase starts its work using a primer (anchor) complementary to single DNA strand [4]. DNA nanoballs (DNB) are 160,000 to 200,000-bp-long single-stranded DNA fragments used for signal amplification, the replicated butt-joined copies of one of the original DNA library molecules. The copies are created in the process of rolling circle amplification of DNA circles constituting the library. Each DNB rests in a separate section of patterned flow cell, which is ensured by its noncovalent binding to a charged substrate. The flow cell is a silicon wafer coated with silicon dioxide, titanium, hexamethyldisilazane and a photoresist material. DNBs are added to the flow cell and selectively bind to positively-charged aminosilanes in a highly ordered pattern, allowing a very high density of DNA nanoballs to be sequenced [1], [5].

The sequencing process itself consists of a few steps, including the addition of a fluorescently labeled terminated nucleotide (sequencing by synthesis), the cleavage of a terminator during the synthesis process and the detection of the produced fluorescent signal [6], [7], [8]. We would like to emphasize that we were unable to find a detailed description of cPAS-based sequencing in the literature, nor figured out how it is implemented in MGISEQ-2000. However, there is a patent in the public domain that describes the application of the cPAS approach, in which the sequencing process is carried out using fluorescently labeled monoclonal antibodies that recognize a unique chemical modification of one of four terminated dNTPs [9]. Anyway, it is not currently possible to obtain full information about sequencing by MGISEQ-2000.

A couple of years ago, a paper was published demonstrating a similar accuracy of SNP detection and slightly lower accuracy of indel detection for the BGISEQ-500 platform, as compared to HiSeq 2500, using a reference genomic dataset from GIAB [3]. A few recent studies have compared the performance of these two platforms in sequencing ancient DNA [10], metagenome [11] and microRNA [4]. In general, the quality of data generated by BGISEQ-500 has proved to be good, although some of its characteristics are somewhat worse than those of Illumina HiSeq 2500.

The Genome in a Bottle Consortium provides reference genomes for benchmarking [12]. By comparing the obtained genomic variants to a reference sequence, one can assess the accuracy/sensitivity of a tested instrument and the corresponding bioinformatics pipeline for data analysis. In our study, we somewhat stepped back from the conventional methods of analysis as we were pressed for time and material resources. Our intention was to test how suitable is the MGISEQ-2000 platform for assessing the mutational variability of embryonic cells. So, we took the genome of a Russian female egg donor and conducted a genome-wide analysis using two platforms: Illumina HiSeq 2500 and MGISEQ-2000. Since HiSeq 2500 is a well-characterized and popular platform for genomic research, we decided to evaluate the overall error rate in order to understand whether we can use MGISEQ-2000 for our utilitarian tasks.

CRISPR-CAS9-based genome editing technologies are an effective tool for altering the nucleotide sequence of target regions. The application of genome-editing technologies to in vitro fertilization (IVF) at the zygote stage holds clinical promise and allows almost complete elimination of the original DNA sequence in embryonic cells [13]. However, PCR used to assess the efficacy of targeted genome editing provides no information about the nonspecific activity of CRISPR-CAS9 systems, which can potentially affect any part of the genome. In this case, WGS (whole-genome sequencing) of the embryonic cell is needed. We decided to compare the performance quality of two massively parallel sequencing (MPS) platforms by Illumina (HiSeq 2500) and MGI (MGISEQ-2000) using the biological samples provided by one of the egg donors for embryonic genome editing.

Materials and Methods

Ethics

The research was carried out according to The Code of Ethics of the World Medical Association (Declaration of Helsinki). Written informed consent was obtained from the patient, and the study was approved by the Ethical Committee from National Medical Research Center for Obstetrics, Gynecology and Perinatology Named After Academician V.I. Kulakov, Moscow, Russia.

DNA preparation

A sample of genomic DNA was isolated from WBC (white-body cells) by phenol-chloroform extraction. Quality control was done with agarose gel electrophoresis (degradation level) and the Qubit dsDNA BR Assay Kit (concentration measurement). The donor was a female resident of the Russian Federation.

Library preparation for sequencing

MGISEQ-2000

The circularization procedure is essentially the denaturation and renaturation of a DNA library in the presence of excess amounts of a splint oligo (dephosphorylated at the 5'-end) and consisting of inverted complementary sequences of adapters ligated to the library. In the process of renaturation with the splint oligo, an annular molecule is formed with double-stranded structure in the adapter region containing a nick. The nick is sealed by DNA ligase. Linear DNA library molecules are disposed of at the digestion stage using a mixture of nucleases that cleave linear molecules. Good scheme is prepared by MGI’s team [28].

The isothermal synthesis of nanoballs is carried out using the rolling circle amplification (RCA) mechanism and is initiated by the splint oligo. As a result, RCA forms a linear single-stranded DNA consisting of 300-500 repeats. A nanoball is a molecule compactly packed into a coil-like form 200-220 nm in diameter.

The procedure of nanoball loading on the flow cell is simplified and automated: the flow cell has a patterned array structure promoting efficient loading (85.5% in our case) which does not depend on the accuracy of library dilution in the case of unordered cells, for example for Illumina MiSeq or HiSeq 2500. The nanoballs are loaded using a DNB Loader, a device similar to cBot (Illumina); alternatively, the loading procedure can be carried out manually using a plastic DNB manual adapter loader. The instrument and the reagents are prepared for sequencing in the way similar to that offered by Illumina. With MGISEQ-2000, water and maintenance washes must be performed. The ready-to-use reagents are delivered in a cartridge that needs to be pre-thawed. A flow cell for MGISEQ-2000 has four separate lanes and one surface on which DNBs are immobilized.

1000 ng of genomic DNA was fragmented using a Covaris ultrasonicator to achieve a length distribution of 100-700 bp with a peak of 350 bp. Size selection was performed with Ampure XP (Beckman). Library concentrations were measured using Qubit; the amount of DNA used was 289 ng (procedure efficiency 29%). Then, an aliquot of 50 ng of the fragmentation product was transferred to a separate tube for end-repair and A-tailing. For ligation, the equimolarly mixed set of Barcode Adapters 501-508 was used. The ligation product was washed with Ampure XP, then 7 PCR cycles were performed using primers complementary to the ligated adapters After washing the library with Ampure XP, its concentration was measured by Qubit. Before the annealing and circularization with splint oligo, the library was normalized to the amount of 330 ng in a volume of 60 µl. After linear DNA was digested, the concentration of circulated DNA (0.997 ng / uL) was measured by Qubit using the ssDNA kit.

After RCA and formation of DNBs, the end product was measured by Qubit using the ssDNA kit. The typical range of nanoball concentrations suitable for loading is 8-40 ng / uL. In our case, the concentration was 20 ng / uL. Nanoball loading was assisted by a DNB manual loader.

Illumina 2500

500 ng of genomic DNA was enzymatically fragmented by dsDNA Fragmentase (NEB). The library was prepared using the NEBNext Ultra II kit and indexes from the Dual Index Primers Set 2 (all New England Biolabs) according to the manufacturer's instructions; amplification at the last sample preparation stage was done in 3 PCR cycles.

MPS was carried out on the Illumina HiSeq 2500 in the Rapid Run mode (paired-end 150 bp dual indexing) using the 500-cycle v2 reagent kit according to the manufacturer's instructions.

Sequencing

Preparation of genomic libraries and sequencing on MGISEQ-2000 were carried out by our research group at the facilities of MGI Tech. in Shenzhen. Fastq files were generated as described previously using the zebracallV2 software provided by the manufacturer [3].

Library preparation and sequencing on HiSeq 2500 were carried out at the Center for Genome Technologies of Russian National Research Medical University. Fastq files were generated using the Basespace cloud software offered by the manufacturer (https://basespace.illumina.com/analyses/140691740/files/logs).

Raw Data

Fastq files with WGS of E704 sample obtained from HiSeq 2500 and MGISEQ-2000 are avaliable in SRA database (BioProject: PRJNA530191, direct link https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA530191

Data analysis

The detailed description of the sequencing process and the scripts are provided in a Supplementary file 1

Results

Sequencing data summary

In this research, we analyzed two whole-genome datasets yielded by the sequencing of a Russian female donor's gDNA (hereinafter, we will call the sample E704). Her genome was sequenced using two platforms: HiSeq 2500 by Illumina and new MGISEQ-2000 by BGI Complete Genomics that have similar performance characteristics. In the case of MGISEQ-2000, DNA was applied onto a separate lane of the flow cell. Sequencing was performed in a paired-end 150 bp mode. We noted the amount of data generated by MGISEQ-2000 and calculated the average coverage. After that, we sequenced the donor's genome using Illumina HiSeq 2500 in order to obtain a similar amount of data. General sequencing characteristics are presented in Table 1. The detailed description of library preparation is provided in Materials and Methods. We would like to note that we used different methods of DNA fragmentation for library preparation: fragmentation by ultrasound (E704-M) and enzymatic fragmentation (dsDNA fragmentase; E704-I). This fact is important for the discussion of our results.

Table 1: Summary of the dataset*

Platform	DNA Fragmentation method	Reagents/Type	Read (×10^6)	Bases (Gbp)	GC Content	>Q20	>Q30
MGISEQ-2000 E704-M	UltraSound	PE150	780	101.37	40%	99.92 %	95.03 %
HiSeq 2500 E704-I	Enzymatic	PE150	726	94.37	40%	99.99 %	97.18 %

As shown in Table 1, the size of the obtained data, as well as the characteristics of sequencing quality, indicate that the datasets can be analyzed and compared. The use of different fragmentation methods is unlikely to skew the comparison of the two datasets [14].

The average coverage is an important characteristic of whole-genome sequencing, just like its distribution and variability. Figure 1 compares the average coverage distribution for MGISEQ-2000 and HiSeq 2500. The figure shows a slightly higher average coverage for MGISEQ-2000 (32.75X for MGISEQ-2000 vs 30.48X for HiSeq 2500). At the same time, the overall coverage distribution is highly uniform for both datasets (Inter-Quartile Range (IQR = 6)), suggesting good sequencing quality [15].

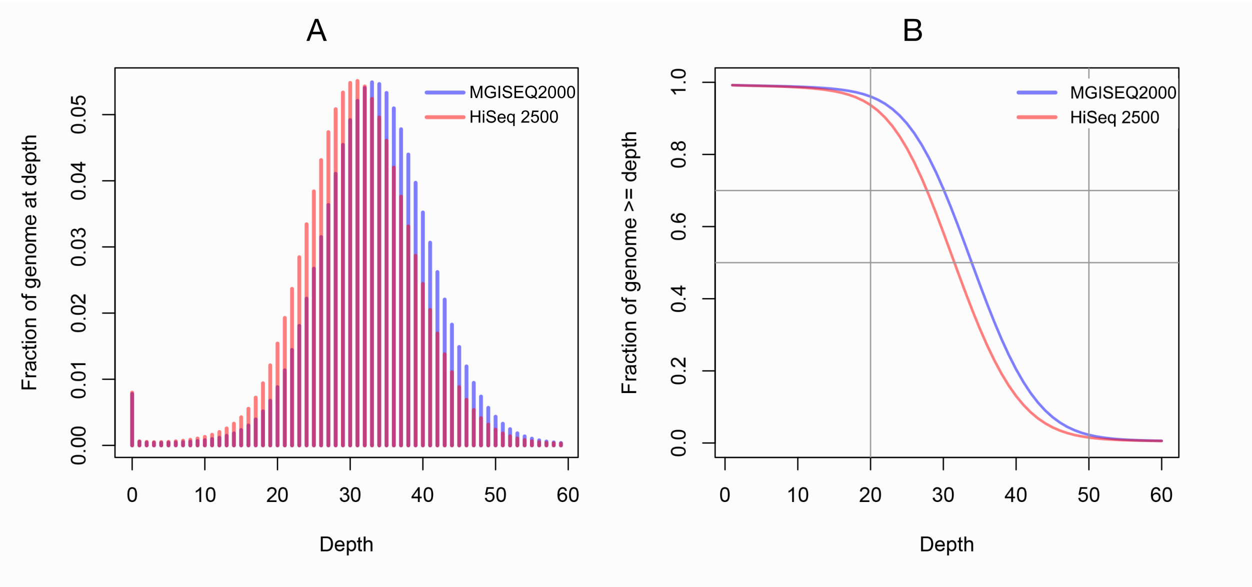

Figure 1. Analysis of the coverage distribution achieved by MGISEQ-2000 and HiSeq 2500 for the E704 sample. A: fraction of genome covered appropriate number of times. B: fraction of genome covered not less than the corresponding number of times. The analysis was performed using the R [16] and BEDtools [17] software packages.

The data presented in Figure 1 were obtained after the FastQC had been carried out during the reads alignment. We specifically mention it at the beginning of the results section so that it is clear that the input data in terms of the coverage distribution and the total reads number were similar.

FastQC analysis

The next step in the comparison of both datasets was to assess the quality of FastQ files by FastQC [18]. We also analyzed every individual FastQ file generated by paired-end sequencing with different barcodes (see *Materials and Methods*).

Data quality exposed by FastQC source file analysis was acceptable and comparable for both platforms. K-mers were found at the start of the reads in the fastq files generated by MGISEQ-2000-based sequencing and at the end of the reads in the files yielded by HiSeq 2500-based sequencing. In HiSeq 2500 fastq files a deviation from the normal GC-content was observed at the start of the reads. K-mers might be explained by the presence of unremoved adapter sequences in both cases. The abnormal GC content could be a result of enzymatic fragmentation, which apparently causes a deviation from the random distribution pattern. Bearing that in mind, we decided to remove 10 nucleotides from both ends of each read in both MGISEQ-2000 and HiSeq 2500 fastq files. Further manipulations were carried out on 130-nucleotide-long fragmented reads. We also trimmed adapter and other technical sequences (data not provided in this article), which allowed us to save more data and work with a higher average read length. This, however, was not crucial for our purposes, so we proceeded to the next steps of the comparative analysis. We merged all obtained fastq library files containing different barcodes so that each platform was represented by only a couple of fastq files - with forward (R1) and reverse (R2) reads, respectively. After merging the fastq files, we repeated the quality assessment procedure using the FastQC service only to find out that the total data generated by both platforms were of acceptable quality and could be safely compared.

Figure 2 shows quality of sequencing data assessed by the FastQC service [18]. Data quality was acceptable for each of the nucleotide positions within a read for both MGISEQ-2000 and HiSeq 2500. However, the quality of data representing each position in the MGISEQ-2000 fastq file was somewhat lower than in the HiSeq 2500 file and tended to gradually deteriorate towards the end of a read (though it was not lower than Q20). For HiSeq 2500-generated data, a decline in quality below Q20 was observed only towards the very end of a read. For each nucleotide, the quality of MGISEQ-2000-based sequencing gradually decreased after 50-60 cycles. In contrast, the total number of high-quality nucleotides was higher for HiSeq 2500 and maintained through the last cycle. A similar picture is demonstrated by the graphs representing the distribution of reads quality (Fig. 2c). With Illumina, the distribution is more uniform, meaning that the average quality is higher. The quality of reads yielded by MGISEQ-2000-based sequencing is acceptable since 95% of all reads were above Q30. The GC-content is similar for both platforms (Fig. 2d); the distribution graphs completely repeat each other.

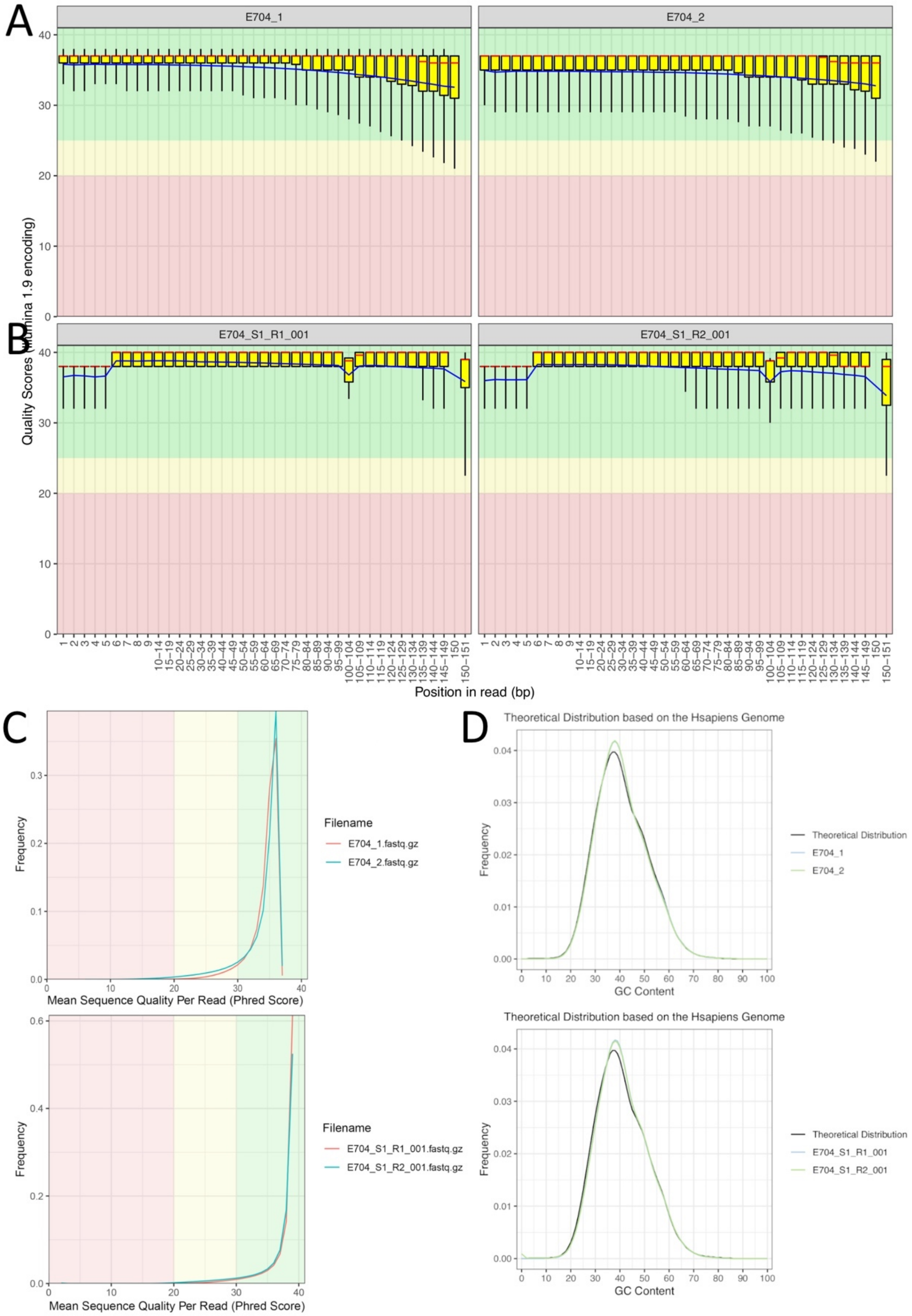
A
E704_1
E704_2
B
E704_S1_R1_001
E704_S1_R2_001
Quality Scores (Illumina 1.9 encoding)
0
10
20
30
40
1 2 3 4 5 6 7 8 9 10–14 15–19 20–24 25–29 30–34 35–39 40–44 45–49 50–54 55–59 60–64 65–69 70–74 75–79 80–84 85–89 90–94 95–99 100–104 105–109 110–114 115–119 120–124 125–129 130–134 135–139 140–144 145–149 150 150–151
Position in read (bp)
C
Frequency
0.0
0.1
0.2
0.3
0
10
20
30
40
Mean Sequence Quality Per Read (Phred Score)
Filename
E704_1.fastq.gz
E704_2.fastq.gz
0.0
0.2
0.4
0.6
Frequency
Mean Sequence Quality Per Read (Phred Score)
Filename
E704_S1_R1_001.fastq.gz
E704_S1_R2_001.fastq.gz
D
Theoretical Distribution based on the Hsapiens Genome
Frequency
0.00
0.01
0.02
0.03
0.04
0 10 20 30 40 50 60 70 80 90 100
GC Content
Theoretical Distribution
E704_1
E704_2
Theoretical Distribution based on the Hsapiens Genome
Frequency
GC Content
Theoretical Distribution
E704_S1_R1_001
E704_S1_R2_001

Figure 2. Post-filtering data quality control. A - Distribution of nucleotide quality parameters across reads. Data are presented for both MGISEQ-2000 (a) and HiSeq 2500 (b) platforms for forward (R1) and reverse (R2) reads, respectively. For each position along the reads, the quality scores of all reads were used to calculate the mean, median, and quantile values; thus the box plot can be shown.
Overall quality score distribution of MGISEQ-2000 and HiSeq 2500 data (c).
Distribution GC-content in the data generated by MGISEQ-2000 and HiSeq 2500 (d). FastQC [18] was used for the analysis.

Reads mapping/alignment and QC

The filtered and trimmed reads were aligned to the reference genome, which was necessary to convert fastq files to BAM files. This was done using Burrows-Wheeler Aligner (BWA-MEM) with default settings recommended for the analysis of genomes sequenced on Illumina systems [19]. The quality of read alignment was assessed using the SAMtools software package and the bamstats software module [20,21].

The quality of read alignment was acceptable for both platforms. The insert size for paired-end libraries corresponds to the theoretical size specified in the manufacturer's protocol: 250 bp for Illumina HiSeq 2500 and 400 bp for MGISEQ-2000. The proportion of aligned reads was 99.9% for both BAM files.

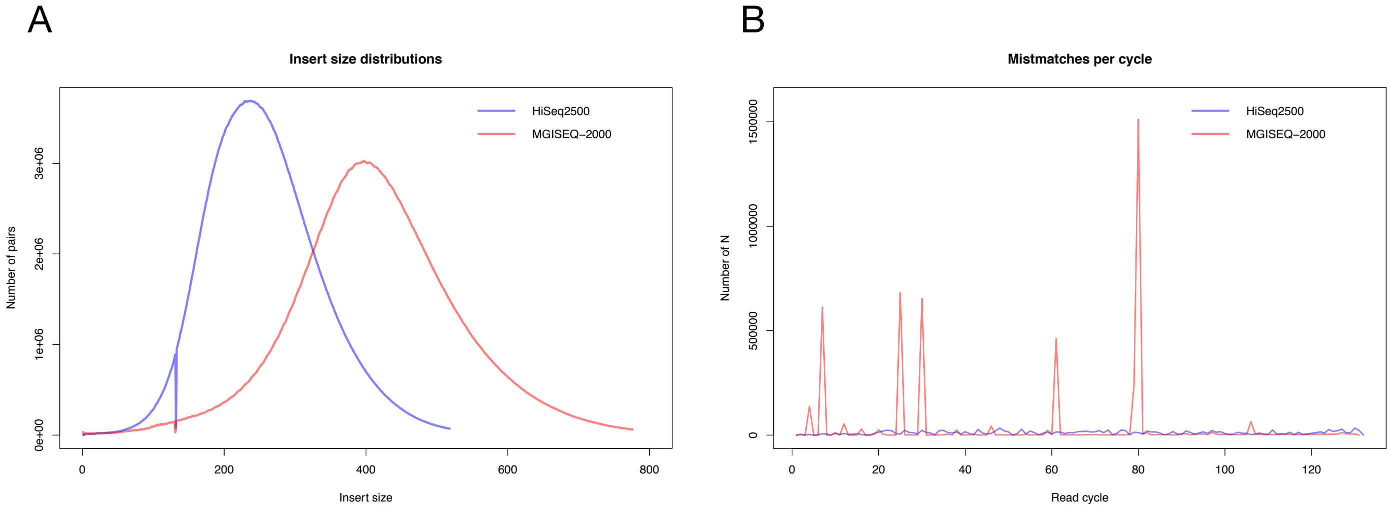

Figure 3. The results of QC analysis of read alignment to the reference genome. A: the distribution of insert lengths between reads of the E704-I library (blue line) and the E704-M library (red line). B: the number of random errors in HiSeq 2500 (blue line) and MGISEQ-2000 (red line). The alignment algorithm used is BWA-MEM [19]. QC analysis was done in bamstats [20,21].

Figure 3 presents the results of the analysis of read alignment to the reference genome. Importantly, the frequency of random sequencing errors is much higher for MGISEQ-2000 and increases with the number of sequencing cycles. Another distinctive feature of MGISEQ-2000 is a shift in the distribution of fragment lengths towards the dominance of shorter fragments, suggesting excessive genome fragmentation. But the distribution of insert lengths in the library is much closer to normal. However, it may be more affected by the process of sample preparation than by the selected sequencing technology (data is not shown).

Variation calling and false positive/negative ratio estimation

In order to further assess the quality of sequencing by MGISEQ-2000, as well as to understand the aspects of its potential use, the generated data were subjected to variant calling. After the data were aligned to the reference genome using BWA-MEM [19], the BAM file was modified using four different pipelines: Samtools [20.21], Strelka2 [22], Sentieon [23], and GATK [24].

The mapping speed, coverage, and sequencing homogeneity were similar for both datasets (Figure 1). All software packages used to process the datasets generated by Illumina and MGI demonstrated similar performance in terms of computation speed, which is consistent with the results obtained for BGISEQ [25].

Alignment results are provided in Table 2; the table shows that both sequencing platforms performed similarly well. The duplication rate for E704-I was higher than for E704-M, amounting to 12.26%. This value, however, was calculated after the fastq files with different barcodes and from different lanes were merged. In each individual fastq file, the duplication rate did not exceed 5-6% for both devices (see Supplementary information). With Illumina HiSeq 2500, 16 separate fastq files (8 for + 8 rev) were generated. The number of fastq files for MGISEQ-2000 was also 16, but they represented a single flow cell, whereas Illumina's files came from two different flow cells. Thus, a higher duplication rate for Illumina results from the use of two cells. Most likely, the probability of getting repeated reads from two independent flow cells is higher than from one cell. As the information contained in fastq files is summed up, it results in an additional 3-4% of duplicates for Illumina-generated data relative to MGISEQ-2000.

Table 2: Mapping statistics for the datasets*

Metrics	**E704-M**	**E704-I**
Clean reads	779784662	725927338
Clean bases	101372006060	94370553940

HG19 length	3095693983	3095693983
Identified bases	2921715981	2919239426
Mapping rate	99,85%	99,93%
Unique rate	90,83%	87,20%
Duplication rate	8,61%	12,26%
Mismatch rate	0,56%	0,54%
Average Depth	32,75	30,48
Coverage at least 4x	99,81%	99,78%
Coverage at least 10x	94,38%	94,30%
Coverage at least 20x	88,87%	84,66%

Since it was not possible to conduct standard benchmarking procedures and determine error values in the reference genomic dataset under this study, we calculated error rates (False Positive, False Negative, etc.) in the E704-M dataset using E704-I as a reference. This approach cannot be used to assess the accuracy of the MGISEQ technology, but it does allow us to conclude that the two compared technologies can be used interchangeably for similar tasks without significant loss of accuracy.

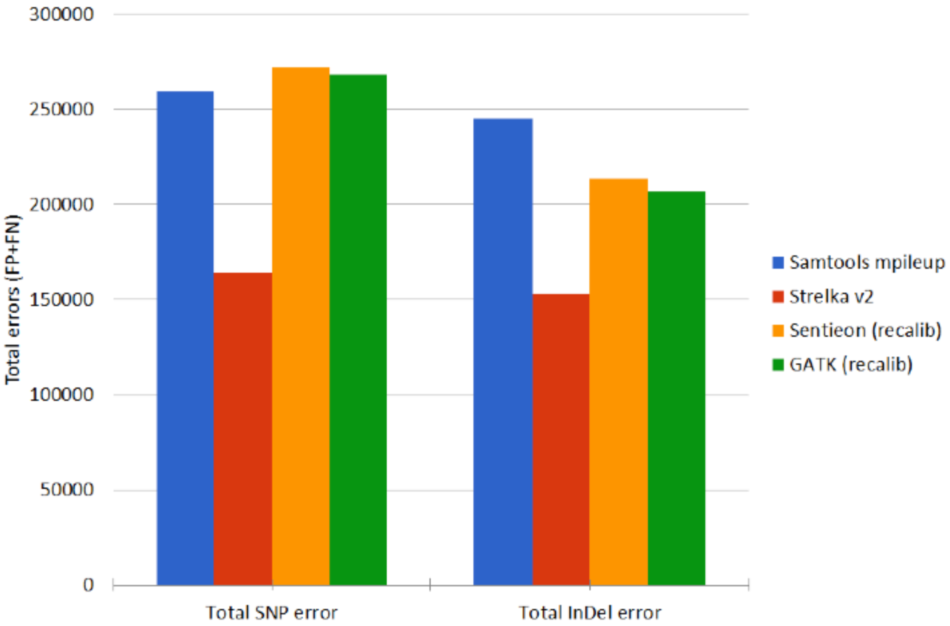

Figure 4. The total number of errors (the sum of FP and FN) in detecting SNPs (total SNP error) and InDels (total InDel Error) occurred in the course of comparing genomic variants in E704-M (A) and E704-I (B). Four software packages were used for variant calling: Samtool, Strelka2, Sentieon, and GATK). Baseline data are shown in the Table contained in the additional file 2.

Figure 4 shows error rates determined by different software packages. The best result was demonstrated by Strelka2 [22]; below we will use the figures yielded by this pipeline. Variant calling results are presented in the Additional file 2. The magnitude of the total error (False Negative + False Positive) between E704-M and E704-I corresponded to the previously obtained results for BGISEQ500 and Illumina [https://blog.dnanexus.com/2018-07-02-comparison-of-bgiseq-500-to-illumina-novaseq-data/].

In total, over 3.7 million SNPs were detected in the datasets by each of the tested platforms. The E704-M sample contained 3,730,684 SNPs; the number of detected SNPs in E704-I was comparable (3,719,768 SNPs). These data are shown in Table 3. In addition, was detected a similar Ti / Tv ratio, which may indirectly indicate the sequencing accuracy.

MGISEQ-2000 was able to detect a little bit more indels (803,736) than HiSeq 2500 (770,193; see table 3). Generally, HiSeq 2500 performance was characterized by a slightly lower average coverage, which partly explains its indel detection rate. However, given that the dbSNP indel rate for HiSeq 2500 was slightly higher (92.1%; E704-I, versus 90.86%; E704-M), this may indicate a lower accuracy of indel detection by the MGISEQ-2000 platform. These observations are consistent with the previous findings for BGISEQ-500 [3].

Table 3: Variant calling statistics for the datasets*

	E704 - MGISEQ-2000	**E704 - Illumina**
SNPs	**3730684**	**3719768**
dbSNP (snp150)	3719888	3696538
dbSNP rate	99,71%	99,38%
Novel	10796	23230
Homozygous	1473069	1463785
Heterzygous	2257468	2255899
Synonymous	13291	13600
Ti/Tv	2,037	2,04
dbSNP Ti/Tv	2,04	2,045
Novel Ti/Tv	1,354	1,308
Indels	**803736**	**770193**
dbSNP (snp150)	730306	709350
dbSNP rate	90,86%	92,10%
Novel	73430	60843
Homozygous	366314	339940
Heterzygous	437422	430253

*The table features data yielded by Strelka2. dbSNP is the total number of SNPs found in the dbSNP database. dbSNP rate is the ratio of SNPs present in dbSNP to all detected SNPs. Ti / Tv is the transition to transversion ratio.

To assess the accuracy of detection of certain genomic variants, we chose the E704-I dataset as a reference for E704-M. We would like to emphasize that we do realize that our approach is not accurate enough to be used for benchmarking. But since such studies had been carried out many times for HiSeq 2500, we decided to determine the level of differences for a single genome. Sequencing by two different tools allowed us to estimate their interchangeability/similarity. We understand that our approach is less accurate and cannot be used to directly measure error rates in detecting various mutations, as proposed by the Genome in a Bottle Consortium [12]. However, we believe that it allows us to compare the tested platforms, using the HiSeq 2500 data as a reference, given that the permissible rate of errors for the latter technology has already been established by the Consortium.

For all the SNPs detected, we estimated the magnitude of various errors and calculated the F1-metric using vcf-compare (vcftools [26]) and snpeff [27]).

Table 4 compares the variants obtained through variant calling by Strelka2; the data generated by other software packages are listed in the Additional file 2.

As a result, using the “accessible genome” matrix, the sensitivity of determining SNPs (Sensitivity) in E704-M was 99.51% compared to E704-I, with an FPR (false positive rate) value — 0.000254% (F1 metrics = 99.65%). For InDels, the sensitivity was 98.84% (F1 metrics = 98.81%). It is worth noting that although we didn’t compare with the reference sequence, the level of convergence of genotypes for the two platforms MGISEQ-2000 and Illumina Hiseq2500 is high enough for both the accessible genome and the complete sequence of the read genome and shows a higher accuracy of the MGISEQ-2000 sequencing relative to Previously obtained data for BGISEQ-500 [3]. This data are shown in Table 4.

Table 4: Variant calling for E704-M vs E704-I*

		MGI vs Illumina
Identified bases (accessible genome)		2182021466
SNPs	REF matches (full genome - VCF)	2179423698
	All features in MGISEQ	2597768
	REF matches (in VCF)	2592230
	ALT matches (in VCF)	2591850
	REF mismatches (in VCF)	0
	ALT mismatches (in VCF)	380
	In MGISEQ	5538
	In reference	12780
	In both	2592230
	True Positive	2592230
	False Positive	5538
	True Negative	2179423698
	False Negative	12780
	TPR (Sensitivity, Recall)	99,51%
	TNR (Specificity)	99,999746%
	FNR	0,49%
	FPR	0,000254%

	PPV (Precision)	99,79%
	FOR	0,00%
	FDR	0,21%
	NPV	100,00%
	F1-Metrics	**99,65%**
InDels	REF matches for INDEL (VCF)	2181793391
	All features in MGISEQ	228212
	REF matches	224595
	ALT matches	223144
	REF mismatches	842
	ALT mismatches	1451
	In MGISEQ	2775
	In reference	2638
	In both	225437
	True Positive	225437
	False Positive	2775
	True Negative	2181793391
	False Negative	2638
	TPR (Sensitivity)	98,84%
	TNR (Specificity)	100,00%
	FNR	1,16%
	FPR	0,000127%
	PPV (Precision)	98,78%
	FOR	0,00%
	FDR	1,22%
	NPV	100,00%
	F1-Metrics	**98,81%**

Discussion

We have compared two genomic datasets generated by Illumina HiSeq 2500 and MGISEQ-2000-based sequencing. As part of our study, we aimed to understand whether MGISEQ-2000 could be used for the whole-genome sequencing of embryos, SNP detection and other tasks faced by our laboratory.

Our study has revealed that MGISEQ-2000 generates datasets possessing similar characteristics, as compared to the data yielded by the “gold standard” of the NGS analysis — the Illumina platform. Given a comparable amount of output data (101.37Gb for MGISEQ and 94.37Gb for Illumina), the average coverage was comparable between the two sets: 32.75X for MGISEQ-2000 vs 30.48X for HiSeq250; the coverage distribution patterns were almost identical (Figure 1).

The analysis demonstrates that sequencing quality is similar for both instruments. The existing differences can be explained by the specifics of the preliminary steps of library preparation and not by the sequencing technique itself.

Four different pipelines were used to perform variant calling. The detection rate of genomic variants was similar between the datasets. The computational time required to process the obtained data was comparable between all software packages and all datasets used. The performance of Strelka2 was characterized by the lowest number of errors (Figure 4).

The quality of data obtained with MGISEQ-2000 is inferior in some respects to that generated by Illumina HiSeq 2500. Specifically, the frequency of random sequencing errors, the percentage of quality reads, and the accuracy of indel detection are higher for HiSeq 2500. However, the magnitude of those differences is small and insignificant for most research tasks. Last but not least, sequencing costs are an important factor for the laboratories in developing countries, including the Russian Federation. To our knowledge, the MGISEQ-2000 platform is comparable to NovaSeq in terms of its costs, but advantageously requires a smaller number of samples per run.

List of abbreviations

bp - base-pair
cPAS - combinatorial Probe-Anchor Synthesis
dATP - deoxyadenosine triphosphate
dTTP - deoxythymidine triphosphate
DNBs - DNA nanoballs
FNR - false negative rate
FPR - false positive rate

FN - false negative

FP - false positive

GIAB - Genome in A Bottle

MPS - Massive Parallel Sequencing

PCR - polymerase chain reaction

PE150 - pair-end 150 bp

SNPs - Single Nucleotide Polymorphisms

indels - insertions and deletions

WGS - Whole Genome Sequencing

WBC - White Blood Cell

Conflicts of interest

DKw - is an employee of OOO “Helicon company”, distributor of MGI Tech LLC on Russian Federation

Authors’ contributions

DR and DK had designed the project. DKw and DK conducted sample preparation and sequencing library construction. VB, DKw and DK conducted sequencing. NK, VN and AG conducted data analysis. DK and AG wrote the manuscript.

DR - Denis Rebrikov

DK - Dmitriy Korostin

VB - Vera Belova

DKw - Dmitry Kwon

NK - Nikolay Kulemin

AG - Alexey Gorbachev

VN - Vladimir Naumov

References

1. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science. 2010;327:78–81.

2. Specter M. The Gene Factory [Internet]. The New Yorker. The New Yorker; 2013 [cited 2018 Dec 25]. Available from: https://www.newyorker.com/magazine/2014/01/06/the-gene-factory

3. Huang J, Liang X, Xuan Y, Geng C, Li Y, Lu H, et al. A reference human genome dataset of the BGISEQ-500 sequencer. Gigascience. 2017;6:1–9.

4. Fehlmann T, Reinheimer S, Geng C, Su X, Drmanac S, Alexeev A, et al. cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. Clin Epigenetics. BioMed Central; 2016;8:123.

5. Chrisey LA, Lee GU, O'Ferrall CE. Covalent attachment of synthetic DNA to self-assembled monolayer films. Nucleic Acids Res. 1996;24:3031–9.

6. Canard B, Sarfati RS. DNA polymerase fluorescent substrates with reversible 3'-tags. Gene. 1994;148:1–6.

7. Mitra RD, Shendure J, Olejnik J, Edyta-Krzymanska-Olejnik, Church GM. Fluorescent in situ sequencing on polymerase colonies. Anal Biochem. 2003;320:55–65.

8. Tsien RY, Ross P, Fahnestock M, Johnston AJ. Dna sequencing [Internet]. Patent. 1991 [cited 2018 Dec 25]. Available from: https://patents.google.com/patent/CA2044616A1/en

9. Drmanac R, Drmanac S, Li H, Xu X, Callow MJ, Eckhardt L, et al. Stepwise sequencing by non-labeled reversible terminators or natural nucleotides [Internet]. US Patent. 2018 [cited 2018 Dec 25]. Available from: https://patentimages.storage.googleapis.com/46/2d/9b/5a6013e915f9b7/US20180223358A1.pdf

10. Mak SST, Gopalakrishnan S, Carøe C, Geng C, Liu S, Sinding M-HS, et al. Comparative performance of the BGISEQ-500 vs Illumina HiSeq 2500 sequencing platforms for palaeogenomic sequencing. Gigascience. 2017;6:1–13.

11. Fang C, Zhong H, Lin Y, Chen B, Han M, Ren H, et al. Assessment of the cPAS-based BGISEQ-500 platform for metagenomic sequencing. Gigascience. 2018;7:1–8.

12. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol. 2014;32:246–51.

13. Kodyleva TA, Kirillova AO, Tyschik EA, Makarov VV, Khromov AV, Guschin VA, et al. The efficacy of CRISPR-Cas9-mediated induction of the CCR5delta32 mutation in the human embryo. Bulletin of RSMU. 2018;70–4.

14. Knierim E, Lucke B, Schwarz JM, Schuelke M, Seelow D. Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. PLoS One. 2011;6:e28240.

15. Sequencing Coverage for NGS Experiments [Internet]. [cited 2019 Jan 28]. Available from: https://www.illumina.com/science/education/sequencing-coverage.html

16. Ripley BD. The R project in statistical computing. MSOR Connections The newsletter of the LTSN Maths, Stats & OR Network. 2001;1:23–5.

17. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

18. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data

[Internet]. [cited 2019 Jan 28]. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

19. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.

20. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27:2987–93.

21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

22. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods. 2018;15:591–4.

23. Weber JA, Aldana R, Gallagher BD, Edwards JS. Sentieon DNA pipeline for variant detection - Software-only solution, over 20× faster than GATK 3.3 with identical results [Internet]. PeerJ PrePrints; 2016 Jan. Report No.: e1672v2. Available from: https://peerj.com/preprints/1672/

24. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.

25. Carroll A. Comparison of BGISEQ 500 to Illumina NovaSeq Data [Internet]. Inside DNAnexus. 2018 [cited 2019 Feb 15]. Available from: https://blog.dnanexus.com/2018-07-02-comparison-of-bgiseq-500-to-illumina-novaseq-data/

26. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.

27. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly . 2012;6:80–92.

28. Oligos and primers for BGISEQ&MGISEQ NGS system [Internet]. [cited 2019 April 03]. Avaliable from: http://en.mgitech.cn/include/upload/kind/file/20181108/20181108161128_5692.pdf