scientific reports

OPEN

Comparative assessment of the Sikun 2000 sequencing platform for whole genome sequencing

Keya Cai[2,3,6], Sibin Li[2,3,6], Meng Pan[1,4,6], Hui Lu[1,3,4], Liang Wang[2], Shengquan Fang[5], Lingzhi Gou[2], Jiang Tang[2], Yan Kong[1,3], Li Zhao[2]✉ & Yongyong Ren[1,3]✉

DNA sequencing technology has significantly advanced over the past five decades, particularly with the introduction of next-generation sequencing (NGS) platforms like Illumina's NovaSeq. In October 2023, Sikun introduced the Sikun 2000, a desktop NGS platform designed for rapid, cost-effective sequencing with up to 200 Gb of data per run. This study is the first to evaluate the performance of the Sikun 2000 in whole genome sequencing (WGS) and compares it with the Illumina NovaSeq 6000 and NovaSeq X platforms using five well-characterized human Genomes in a Bottle dataset. Results show that the Sikun 2000 performs competitively in variant detection, particularly excelling in single nucleotide variant accuracy. It also demonstrated a higher sequencing depth and lower proportion of low-quality reads than the NovaSeq platforms. However, its performance in insertion-deletion (Indel) detection was slightly lower than that of NovaSeq platforms. Overall, the Sikun 2000 is a qualified sequencing platform for WGS, and further evaluation in clinical and research applications is required as data continues to accumulate.

Keywords Whole genome sequencing, Sikun 2000, Variation detection

Over the past five decades, DNA sequencing technology has advanced remarkably, both in technical capabilities and the breadth of its applications across life sciences research[1]. Next-generation sequencing (NGS), in particular, has revolutionized high-throughput sequencing, surpassing the earlier Sanger sequencing method[2] by offering greater scalability, efficiency, and cost-effectiveness. Taking these advantages, researchers could obtain the information of all genes in one whole genome sequencing (WGS), which has greatly accelerated the implementation of several global genomic initiatives, such as the 1000 Genomes Project[3] and the International HapMap Project[4].

Following the completion of the Human Genome Project[5,6], the first second-generation sequencing platform for short-read sequencing, Roche 454, was introduced in 2015[7]. Since then, three major short-read sequencing platforms have emerged: Illumina (NextSeq, NovaSeq, etc.)[8], MGI (BGISEQ-500, MGISEQ-2000, etc.)[9], and Thermo Fisher (SOLiD, Ion Torrent, etc.)[10]. Among these, Illumina has maintained a leading market position[11]. In 2017 and again in 2022, Illumina introduced the NovaSeq 6000 and the NovaSeq X, respectively. These models, integrating cutting-edge two-color optics and patterned flow cell technology, quickly became the widely utilized sequencing platforms. As the scope of NGS applications continues to expand, developers are consistently introducing new platforms to meet the evolving demands of diverse research scenarios.

In October 2023, Sikun introduced its latest desktop NGS platform for short-read sequencing, the Sikun 2000, based on sequencing by synthesis (SBS) technology and reversible terminator chemistry. In addition to significant enhancements to its optical system, which incorporates a microlens array to minimize light energy dissipation and reduce optical noise[12], the Sikun 2000 also features advances in reagent chemistry. Specifically, Sikun developed and applied a modified nucleotide in which the fluorescent compound is linked via an ether-containing heteroalkyl structure $-(CH_2)_m-O-(CH_2)_n-COR$, enhancing DNA polymerase affinity and thereby improving sequencing speed and accuracy[13]. The system also integrates a fluorescence image reconstruction

[1]SJTU-Yale Joint Center for Biostatistics and Data Science, Technical Center for Digital Medicine, National Center for Translational Medicine, Shanghai Jiao Tong University, Shanghai 200240, China. [2]Sikun Life Science Company Limited, ZhengZhou 450016, China. [3]Institute of Bioinformatics, Shanghai Academy of Experimental Medicine, Shanghai 201203, China. [4]Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China. [5]Department of Gastroenterology, Yueyang Hospital of Integrated Traditional Chinese and Western Medicine, Shanghai University of Traditional Chinese Medicine, Shanghai 200437, China. [6]These authors contributed equally: Keya Cai, Sibin Li and Meng Pan. ✉email: zhaoli@sikun.com; yongyong.ren@sjtu.edu.cn

Platform	Sikun 2000	Novaseq 6000	Novaseq X
Parameter	**Average (std)**	**Average (std)**	**Average (std)**
Reads (M)	610(0)	610(0)	610(0)
Bases (G)	91.5(0)	91.5(0)	91.5(0)
GC (%)	41.1(0)	41.9(0.04)	42.1(0.09)
Q20 (%)	98.52(0.06)	98.25(0.08)	99.15(0.09)
Q30 (%)	93.36(0.21)	94.89(0.20)	97.36(0.27)
Low-quality reads (%)	0.0088(0.002)	0.8338(0.023)	0.9780(0.088)

Table 1. Data production.

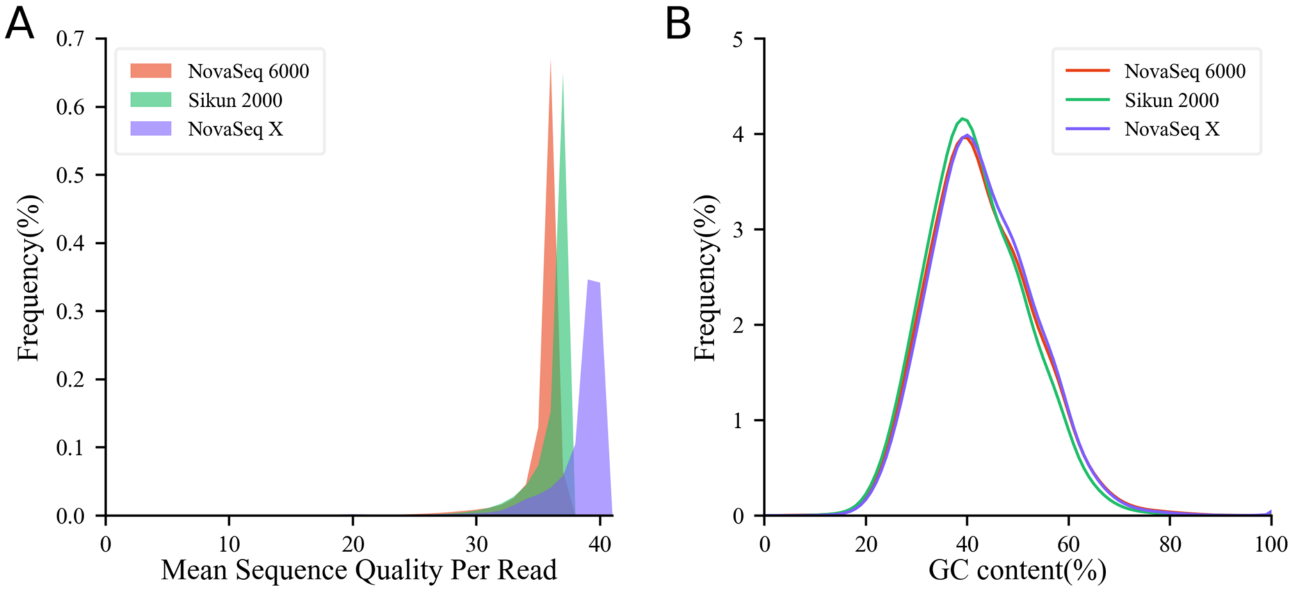

Fig. 1. Quality control of the dataset from Sikun 2000, Novaseq 6000 and Novaseq X. (**A**) The overall quality score distribution of Sikun 2000, Novaseq 6000 and Novaseq X data. (**B**) The GC content distribution of Sikun 2000, Novaseq 6000 and Novaseq X data.

algorithm[14] that corrects for pixel shifts and suppresses noise, further improving base-calling precision. The Sikun 2000 is capable of generating up to 200 Gb of data per run within 22 h, ensuring rapid, low-cost sequencing even with small sample volumes. However, its performance in WGS applications has yet to be evaluated within the scientific community.

In this study, we conducted a performance evaluation of the Sikun 2000 for WGS by comparing it with the NovaSeq 6000 and NovaSeq X platforms in short-read WGS. Using five well-characterized human samples (NA12878, NA24385, NA24149, NA24143, and NA24631), we assessed sequencing quality and variant detection accuracy across the three platforms. Our results demonstrate that the Sikun 2000 performs competitively in variant detection accuracy when compared to the NovaSeq 6000 and NovaSeq X.

Results

Comparison of data production

The DNA from five samples (HG001-HG005) was used in this study and was sequenced to >30× on each platform. The data produced by three sequencing instruments were assessed using six evaluation metrics and identified distinct advantages for each platform. As illustrated in Table 1, for the base quality when measured by Q20, the Sikun 2000 showed a slightly higher percentage than NovaSeq 6000 (98.52% vs. 98.25%, $p = 0.03$, Wilcoxon signed-rank test, alternative="greater") and a slightly lower than NovaSeq X (98.52% vs. 99.14%, $p = 0.03$, Wilcoxon signed-rank test, alternative="less"). When assessed with Q30, the Sikun 2000 had a lower percentage than NovaSeq 6000 (93.36% vs. 94.89%, $p = 0.03$, Wilcoxon signed-rank test, alternative="less") and NovaSeq X (93.36% vs. 97.37%, $p = 0.03$, Wilcoxon signed-rank test, alternative="less"). However, Sikun 2000 has a much lower proportion of low-quality reads compared to NovaSeq 6000 (0.0088% vs. 0.8338%, $p = 0.03$, Wilcoxon signed-rank test, alternative="less") and NovaSeq X (0.0088% and 0.9780%, $p = 0.03$, Wilcoxon signed-rank test, alternative="less"). Additionally, the distribution of reads quality and GC content were also measured using fastQC. Figure 1A shows that the mean sequence quality per read for all platforms was concentrated between Q30 and Q40. Specifically, Sikun 2000 exhibited a peak lower than that of NovaSeq X but superior to

Platform	Sikun 2000	Novaseq 6000	Novaseq X
Parameter	**Average(std)**	**Average(std)**	**Average(std)**
Clean reads(M)	610(0)	604(0)	604(0.4)
Mapped rate(%)	99.62(0.014)	99.90(0.005)	99.67(0.039)
Coverage (%)	92.01(0.31)	92.018(0.31)	92.01(0.31)
4x coverage (%)	91.64(0.26)	91.46(0.17)	91.51(0.19)
10x coverage (%)	88.13(0.45)	86.53(0.75)	87.49(0.86)
Average depth (×)	24.48(0.15)	20.41(0.15)	21.85(0.57)
Dup rate (%)	1.93(0.15)	18.53(1.06)	8.23(2.02)

Table 2. Read alignment results.

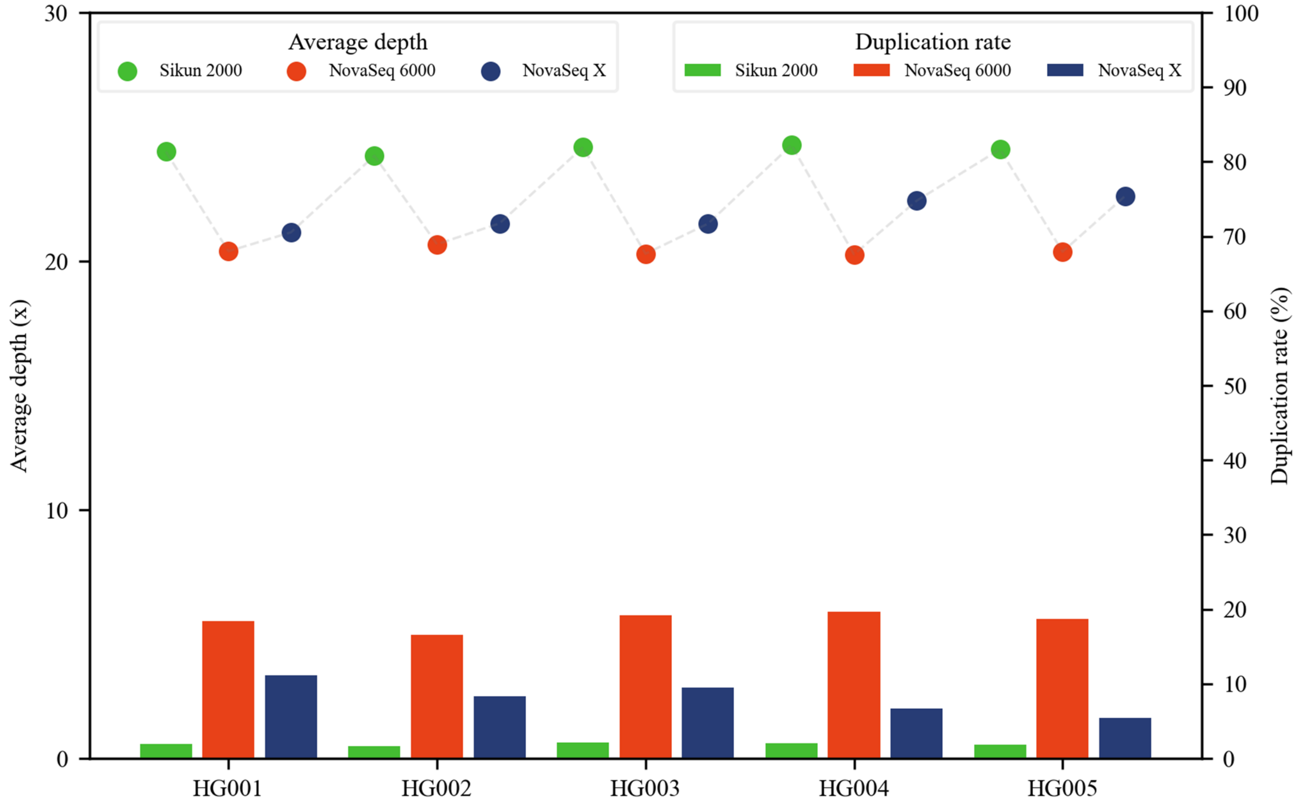

Fig. 2. Comparison of average sequencing depth and duplication rate across different platforms.

NovaSeq 6000. The GC content was similar for the three platforms (Fig. 1B), and the distribution curves are almost identical.

The DNA from five samples (HG001-HG005) was sequenced and downsampled into the same number of reads.

Q20: the accuracy of bases is 99%.

Q30: the accuracy of bases is 99.9%.

Low-quality Reads: The reads with more than 40% of bases having a quality score below 15 or reads containing more than 5% ambiguous bases (Ns).

Comparison of reads alignment

The reads from three platforms were aligned to the human reference genome hg19 using BWA. As shown in Table 2, approximately 92% of bases were covered by at least one read, and more than 86% were covered by at least 10 reads, indicating comprehensive and uniform sequencing across the genome on all platforms. However, Sikun 2000 outperformed the other platforms in both average depth and duplication rate (Table 2; Fig. 2). Its average depth (24.48× ± 0.15) was significantly higher than both NovaSeq 6000 (20.41× ± 0.15) and NovaSeq X (21.85× ± 0.57) ($P<0.05$, Wilcoxon signed-rank test, alternative="greater"). Additionally, Sikun 2000 had a significantly lower duplication rate (1.93% ± 0.15) compared to NovaSeq 6000 (18.53% ± 1.06) and NovaSeq X

(8.23% ± 2.02) ($P < 0.05$, Wilcoxon signed-rank test, alternative="less"). These results highlight that Sikun 2000 provides deeper coverage and generates fewer redundant reads, enhancing data quality and variant detection.

Comparison of variant detection

The variant detection was performed following the guidelines from GATK HaplotypeCaller[15]. The Jaccard similarity was used to measure the concordance between datasets from different sequencing platforms. As shown in Fig. 3, the mean concordance of common variants between Sikun 2000 and NovaSeq 6000 was approximately 92.42%, which was similar to the mean concordance of common variants between Sikun 2000 and NovaSeq X (92.13%) for SNVs. The mean proportion of common Indels between Sikun 2000 and NovaSeq 6000 was approximately 66.63%, nearly identical to that of Sikun 2000 and NovaSeq X (65.22%). The average concordance for SNV and Indel detection between NovaSeq 6000 and NovaSeq X was 92.06% and 70.62%, respectively. For SNV detection, the inter-platform concordance between Sikun 2000 and the NovaSeq platforms was higher than the intra-platform concordance between the two NovaSeq platforms. However, for Indel detection, the intra-platform concordance between the two NovaSeq models was slightly higher than the inter-platform concordance between Sikun 2000 and the NovaSeq platforms.

To further evaluate the detection of genomic variants, we compared the F-score, Recall, and Precision for SNP and Indel calling between Sikun 2000, NovaSeq 6000, and NovaSeq X using the GIAB dataset. As shown in Fig. 4A and B, the average Recall, Precision, and F1-score for SNPs from Sikun 2000 were slightly higher than those of NovaSeq 6000 (97.24% vs. 97.02%; 98.48% vs. 98.30%; and 97.86% vs. 97.64%; $P < 0.05$ for all, Wilcoxon signed-rank test, alternative="greater") and NovaSeq X (97.24% vs. 96.84%; 98.48% vs. 98.02%; and 97.86% vs. 97.44%; $P < 0.05$ for all, Wilcoxon signed-rank test, alternative = "greater"). In terms of Indels, the performance of Sikun 2000 was not as strong as that of NovaSeq 6000 (83.08% vs. 87.08%; 84.46% vs. 86.46%; $P < 0.05$ for all, Wilcoxon signed-rank test, alternative="less"), but comparable to NovaSeq X (83.08% vs. 86.74%, $P = 0.06$; 84.46% vs. 85.68%, $P = 0.31$, Wilcoxon signed-rank test, alternative="two-sided") in Recall and F1-score. For Precision, Sikun 2000 slightly outperformed NovaSeq X (85.98% vs. 84.68%, $P = 0.03$, Wilcoxon signed-rank test, alternative="greater") and performed similarly to NovaSeq 6000 (85.98% vs 85.80%, $P = 0.19$, Wilcoxon signed-rank test, alternative="two-sided").

In addition, we analyzed the categories of variants identified by the three sequencing platforms to assess their potential impact on downstream data analysis. A total of 23 variant-related features were evaluated for both true positive and false positive SNVs and Indels, including genomic location (e.g., exonic, intronic, splice sites), mutation type (e.g., synonymous, missense, nonsense), and predicted functional impact (e.g., high or low). Compared to the NovaSeq 6000 and NovaSeq X, the Sikun 2000 showed no statistically significant differences in recall of true positive SNVs and Indels across 10 evaluated categories, such as variants located in gene transcript regions, splice regions, and those predicted to have high functional consequences, including nonsense and stop-gain mutations. Overall, the NovaSeq series (NovaSeq 6000 and NovaSeq X) detected a higher number of true positive variants than the Sikun 2000 in about 10 categories, including 5' UTR regions, exonic regions, missense mutations, and predicted to be low-impact variants. Conversely, the NovaSeq platforms also generated more false positives than the Sikun 2000 in approximately 14 categories, notably in exonic and intronic regions, as well as in missense and stop-gain variants. Detailed results are presented in Supplementary Data 1.

Discussion

Comparing the newly developed Sikun 2000 platform with established sequencing platforms is crucial for assessing its potential in both academic and clinical applications. Such comparisons allow us to evaluate the Sikun 2000's performance in terms of data quality, accuracy, and overall effectiveness. These evaluations are essential for determining whether the Sikun 2000 meets the rigorous standards required for genomic research and clinical diagnostics, where consistency and reliability are paramount.

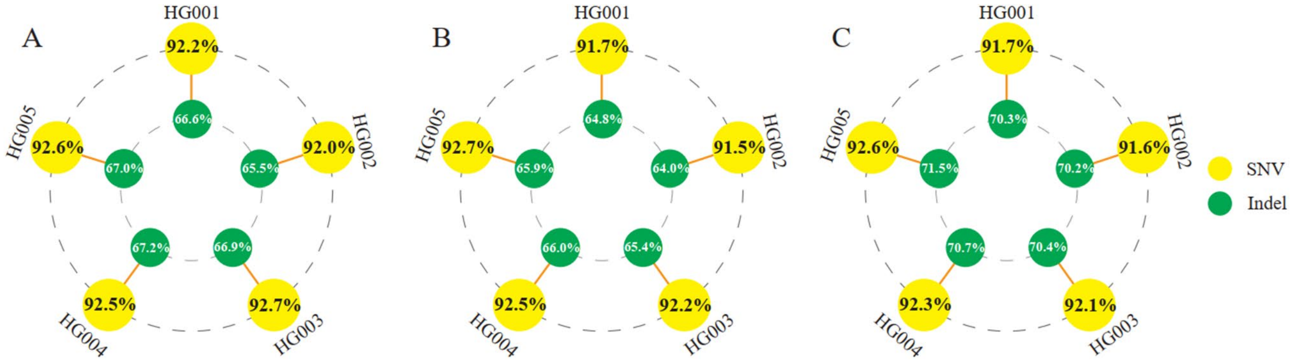

Fig. 3. Concordance of variation detection from Sikun 2000, Novaseq 6000 and Novaseq X. (**A**) The Jaccard similarity of SNVs and Indels detected by the Sikun 2000 and Novaseq 6000 platforms in five samples. (**B**) The Jaccard similarity of SNVs and Indels detected by the Sikun 2000 and Novaseq X platforms in five samples. (**C**) The Jaccard similarity of SNVs and Indels detected by the NovaSeq 6000 and NovaSeq X platforms in five samples.

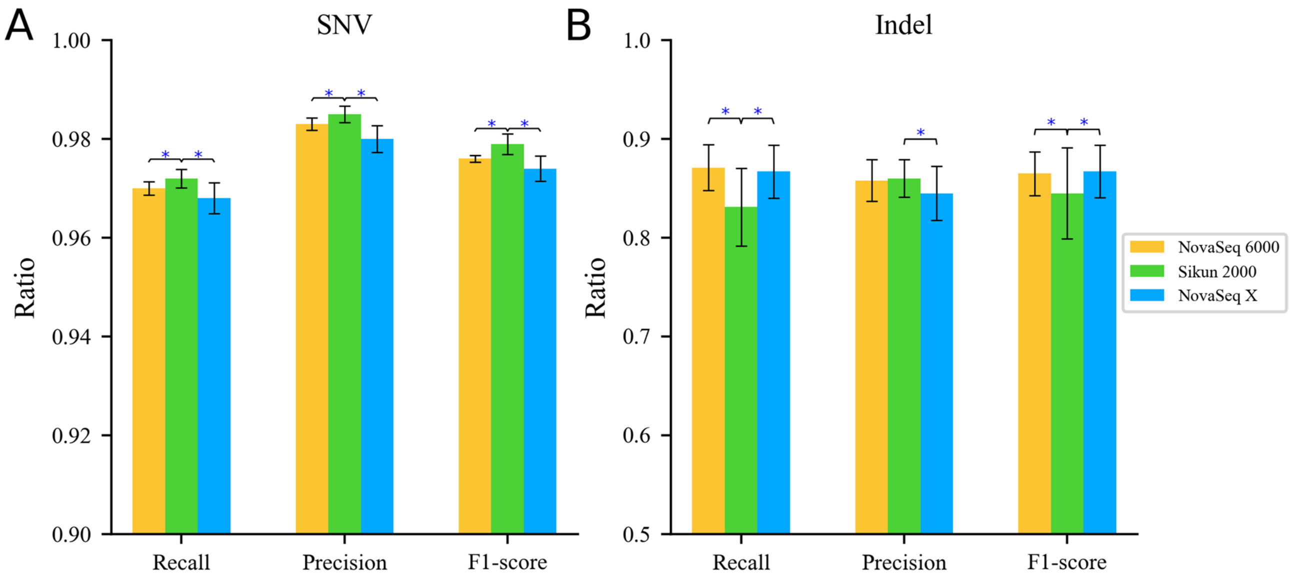

Fig. 4. Comparison of variation accuracy from Sikun 2000, Novaseq 6000 and Novaseq X. (**A**) Comparison of average Recall, Precision, and F-score for SNP detection. (**B**) Comparison of average Recall, Precision, and F1-score for Indel detection. Error bars represent 95% confidence intervals ($n=5$ biologically independent samples). Asterisks denote statistical significance based on a one-sided paired Wilcoxon signed-rank test (*$p<0.05$).

This study is the first to directly compare Sikun 2000 with the widely used Illumina NovaSeq 6000 and NovaSeq X platforms in the context of whole-genome sequencing. By evaluating all three platforms under identical conditions, we provide a comprehensive assessment of how Sikun 2000 compares to these established systems. The results offer valuable insights into the relative advantages and limitations of the new platform, providing a foundation for its future use in diverse genomic applications.

The sequencing quality across all three platforms demonstrated distinct strengths. NovaSeq X achieved the highest Q20 and Q30 values, while Sikun 2000 outperformed NovaSeq 6000 with a better Q20 score. Notably, the Sikun 2000 exhibited a significantly lower proportion of low-quality reads, representing a key advantage in generating high-quality sequencing data. Due to the limited availability of detailed technical information on the NovaSeq 6000 and NovaSeq X platforms, it is not possible to perform a direct head-to-head comparison to determine the exact cause of this difference. However, one potential contributing factor may be the microlens array-based lighting system[12] employed by the Sikun 2000, which is designed to provide superior illumination uniformity. This may help improve imaging quality, particularly at the edges of the field of view, leading to better read accuracy and a reduction in low-quality sequences.

In terms of genome coverage uniformity, all platforms performed similarly, with over 86% of bases covered by at least 10 reads. However, the Sikun 2000 demonstrated a notably lower duplicate rate, which may further, contribute to its improved data quality. One possible explanation is that the Sikun 2000 performs secondary processing of read images using an image reconstruction algorithm[14]. This algorithm enhances sequence recognition by improving image resolution specifically in overlapping signal regions, enabling more accurate distinction between closely spaced reads in high-density areas. However, whether this is the decisive factor contributing to the lower duplicate rate observed in comparison with the NovaSeq platforms remains uncertain and requires further technical information for verification. In combination, the lower proportion of low-quality reads and the reduced duplication rate may contribute to the observed increase in average sequencing depth on the Sikun 2000.

In variant detection using the BWA-GATK pipeline, the NovaSeq X and NovaSeq 6000 platforms demonstrated superior accuracy in Indel detection, while the Sikun 2000 showed better performance in SNV detection. One possible reason for the relatively lower performance of the Sikun 2000 in Indel detection is the presence of overlapping fluorescence signals in adjacent areas of the flow cell during imaging. To avoid interference, the affected regions are directly excluded during image processing, rather than being retained and filtered later as low-quality reads. Given that Indels are typically less abundant and more sensitive to coverage loss than SNVs, the removal of reads from these overlapping regions may have a disproportionate impact on Indel detection. Further improvements in distinguishing and recovering usable reads from these regions may help enhance Indel calling performance and offer deeper insights into their impact on variant detection.

The observed differences in variant detection across the three sequencing platforms lead to distinct impacts on downstream analyses. The NovaSeq 6000 and NovaSeq X platforms demonstrated higher sensitivity by detecting more true positive variants in multiple categories, such as exonic, synonymous, and missense gene variants. However, they also exhibited a higher number of false positives compared to the Sikun 2000, particularly in genic and functionally relevant regions. This elevated false positive rate may affect the reliability of downstream analyses, but can be effectively mitigated using variant filtering tools such as FVC, which can remove most false positives while retaining the majority of true variants[16]. In contrast, the Sikun 2000 showed higher specificity, generating fewer false positives across variant categories. This makes it advantageous in applications requiring high precision, such as clinical diagnostics. However, its sensitivity was relatively lower in certain categories, indicating room for improvement. Enhancing the detection of true variants through further optimization of signal capture or variant calling processes may help improve the overall performance of the Sikun 2000.

It is important to recognize that the Sikun 2000 platform is still in its rapid development phase. Current efforts are focused on recalibrating the optical system's imaging area to minimize regions of repeated imaging and reduce read loss. At the same time, light-tolerant modified nucleotides are being developed to improve the quality of reads in regions subject to repeated exposure. Additionally, the performance of the polymerase and the speed of image capture are being continuously optimized, with the goal of reducing sequencing time to under 20 h per run in PE150 mode. As the technology continues to evolve and more data become available, large-scale analyses will become more feasible in the future, enabling a more comprehensive understanding of its performance in both research and clinical applications.

Methods
Library Preparation and sequencing
Genomic DNA from HG001 (NA12878), HG002 (NA24385), HG003 (NA24149), HG004 (NA24143), and HG005 (NA24631) was obtained from the Coriell Institute. Quantification of the genomic DNA was performed using a Qubit 4.0 fluorometer (Life Technologies, Paisley, UK). Whole genome sequencing (WGS) libraries for these five samples were prepared using the IGT® Enzyme Plus Library Prep Kit V3 and IGT® Adapter & UDI Primer 1–16 (for Illumina, tube). The Sikun 2000 is compatible with Illumina libraries and adapters. The genomic DNA was enzymatically fragmented and the 3' ends were modified with a dATP sticky end using the Fragment & ERA Enzyme Mix. A dTTP-tailed adapter sequence was ligated to both ends of the DNA fragments. After ligation, each sample was amplified for four cycles and purified. The concentration and integrity of the purified products were assessed using the Qubit 4.0 fluorometer and the High Sensitivity DNA Kit on the Agilent Bioanalyzer 2100 system (Agilent Technologies, CA, USA), respectively. Given that 150 bp paired-end (PE150) sequencing is currently the most widely used strategy on second-generation sequencing platforms particularly in studies of cancer[17–19] and genetic diseases[20–22]. The libraries were constructed with a primary fragment size peak of approximately 350 bp to ensure optimal read overlap and sequencing data quality. These same WGS libraries were used across platforms to eliminate potential bias from sample preparation and library construction.

Sequencing quality check, mapping, and data analysis
To minimize potential biases introduced by varying sequencing depths across platforms, the raw sequencing data were downsampled to the same number of reads by fastp (version 0.19.6)[23] without applying quality control procedures. This normalization step was performed prior to the comparison of sequencing output metrics presented in Table 1. The quality of the downsampled sequencing data was first assessed using FastQC[24] and fastp. Low-quality reads were defined as those with more than 40% of bases having a sequencing quality score below 15 or reads containing more than 5% ambiguous bases (Ns). Prior to read alignment, the raw sequencing data were processed using fastp with default parameters to perform quality control, including adaptor trimming and removal of low-quality bases. The reads were then mapped to the human reference genome (hg19 version) using bwa-mem (version 2.2.1)[25]. The resulting SAM files were converted to BAM format and sorted using Samtools (version 1.19)[26].

The sorted BAM files underwent deduplication and base quality score recalibration (BQSR) were performed following the GATK best practice guidelines (version 4.0.12.0)[27]. SNP and Indel mutations were identified using GATK HaplotypeCaller[15] and stored in variant call format (VCF). These genetic variants in high-confidence regions were then compared to the gold standard variants from the GIAB dataset[28]. To evaluate the accuracy of variant calls, we used hap.py (version 0.3.14, https://github.com/Illumina/hap.py)[29] to compute precision, recall, and F1-score metrics. All data visualizations were conducted using Python (version 3.9). The genetic variants were annotated by using SnpEff (version 5.2c). All statistical comparisons across the five samples were conducted using the Wilcoxon signed-rank test, implemented via the scipy.ststs module in the SciPy library (version 1.13.1) in python (version 3.9).

Data availability
Sequence data that support the findings of this study is publicly accessible in the BMAP[30] data repository with the primary accession code DRS1089965008844238848 (https://bmap.sjtu.edu.cn/datastorage/main/62).

Received: 21 February 2025; Accepted: 26 May 2025
Published online: 30 May 2025

References
1. Liu, L. et al. Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology* 1–11, (2012). https://doi.org/10.1155/2012/251364 (2012).
2. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U S A.* **74**, 5463–5467. https://doi.org/10.1073/pnas.74.12.5463 (1977).
3. McVean, G. A. et al. An integrated map of genetic variation from 1092 human genomes. *Nature* **491**, 56–65. https://doi.org/10.1038/nature11632 (2012).
4. Frazer, K. A. et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861. https://doi.org/10.1038/nature06258 (2007).
5. Venter, J. C. et al. The sequence of the human genome. *Science* **291**, 1304–1351. https://doi.org/10.1126/science.1058040 (2001).

6. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921. https://doi.org/10.1038/35057062 (2001).
7. Margulies, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380. https://doi.org/10.1038/nature03959 (2005).
8. Bentley, D. R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59. https://doi.org/10.1038/nature07517 (2008).
9. Kumar, K. R., Cowley, M. J. & Davis, R. L. Next-generation sequencing and emerging technologies. *Semin. Thromb. Hemost.* **45**, 661–673. https://doi.org/10.1055/s-0039-1688446 (2019).
10. Rothberg, J. M. et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348–352. https://doi.org/10.1038/nature10242 (2011).
11. Lang, J. et al. Evaluation of the MGISEQ-2000 sequencing platform for illumina target capture sequencing libraries. *Front. Genet.*https://doi.org/10.3389/fgene.2021.730519 (2021).
12. Qiao, S. Q., Cheng, Y. D., Tang, X. M. & Wang, J. JM. The lighting system and microscope equipment. State Intellectual Property Office of China. Invention Patent No. CN114671843 (2022).
13. Long, H. Y., Wang, Z. D., Wang, Z. Y. & Zeng, Z. G. The fluorescent compound, fluorescent modified nucleotide, and reagent kit. State Intellectual Property Office of China. Invention Patent No. CN112233445A (2022).
14. Song, Z. Q. & Wang, Y. J. DY. The image reconstruction method, equipment, and storage medium for fluorescent images used in nucleic acid sequencing. *State Intellectual Property Office of China.* Invention Patent No. CN 117934282 A (2024).
15. Poplin, R. et al. Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv* **291**, 1304–1351. https://doi.org/10.1101/201178 (2018).
16. Ren, Y. et al. FVC as an adaptive and accurate method for filtering variants from popular NGS analysis pipelines. *Commun. Biology*. https://doi.org/10.1038/s42003-022-03397-7 (2022).
17. Tang, J. et al. Enhancing transcription–replication conflict targets ecDNA-positive cancers. *Nature* **635**, 210–218. https://doi.org/10.1038/s41586-024-07802-5 (2024).
18. Nguyen, D. D. et al. The interplay of mutagenesis and EcDNA shapes urothelial cancer evolution. *Nature* **635**, 219–228. https://doi.org/10.1038/s41586-024-07955-3 (2024).
19. Kamizela, A. E. et al. Timing and trajectory of BCR::ABL1-driven chronic myeloid leukaemia. *Nature* https://doi.org/10.1038/s41586-025-08817-2 (2025).
20. Huang, Q. Q. et al. Examining the role of common variants in rare neurodevelopmental conditions. *Nature* **636**, 404–411. https://doi.org/10.1038/s41586-024-08217-y (2024).
21. Green, G. S. et al. Cellular communities reveal trajectories of brain ageing and Alzheimer's disease. *Nature* **633**, 634–645. https://doi.org/10.1038/s41586-024-07871-6 (2024).
22. Stæger, F. F. et al. Genetic architecture in Greenland is shaped by demography, structure and selection. *Nature* **639**, 404–410. https://doi.org/10.1038/s41586-024-08516-4 (2025).
23. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890. https://doi.org/10.1101/274100 (2018).
24. Thrash, A., Arick, M., Peterson, D. G. & Quack A quality assurance tool for high throughput sequence data. *Anal. Biochem.* **548**, 38–43. https://doi.org/10.1016/j.ab.2018.01.028 (2018).
25. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324 (2009).
26. Li, H. et al. The sequence alignment/map format and samtools. *Bioinformatics* **25**, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352 (2009).
27. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498. https://doi.org/10.1038/ng.806 (2011).
28. Zook, J. M. et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32**, 246–251. https://doi.org/10.1038/nbt.2835 (2014).
29. Krusche, P. et al. Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* **37**, 555–560. https://doi.org/10.1038/s41587-019-0054-x (2019).
30. Ren, Y. et al. BMAP: a comprehensive and reproducible biomedical data analysis platform. *bioRxiv* (2024). 2007.2015.603507. https://doi.org/10.1101/2024.07.15.603507 (2024).

Acknowledgements

This work was supported by the National Key R&D Program of China (Grant Nos. 2024YFC3505900 and 2024YFC3505901); Science and Technology Commission of Shanghai Municipality (Grant Nos. 23JS1400700 and 24JS2840200); the Science and Technology Innovation Key R&D Program of Chongqing (Grant No. CSTB2024TIAD-STX0006); the Key Science and Technology Project of Zhengzhou (Grant No. 2021KJZX0063); Neil Shen's SJTU Medical Research Fund (LX, LL, HL).

Author contributions

Y.R. and L.Z. conceived the concept of the study. K.C., S.L., M.P., and Y.R. performed the sequencing data analysis, statistical analysis, and drafted the manuscript. H.L. participated in the study design. Y.K. designed the visualization of the pictures. S.F supported the data interpretation. L.W, L.G, J.T, and L.Z contributed to the data sequencing. Y.R. and L.Z. supervised all aspects of the study. All authors reviewed and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-04170-6.

Correspondence and requests for materials should be addressed to L.Z. or Y.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.