Accuracy benchmark of the GeneMind GenoLab M sequencing platform for WGS and WES analysis

Chaoyang Li
GeneMind Biosciences Company Limited
Xue Fan
Longgang District Maternity&Child Healthcare Hospital of Shenzhen City
Xin Guo
Longgang District Maternity&Child Healthcare Hospital of Shenzhen City
Yongfeng Liu
GeneMind Biosciences Company Limited
Miao Wang
GeneMind Biosciences Company Limited
Xiaochao Zhao
GeneMind Biosciences Company Limited
Ping Wu
GeneMind Biosciences Company Limited
Qin Yan
GeneMind Biosciences Company Limited
Lei Sun (sunlei@genemind.com)
GeneMind Biosciences Company Limited

Research Article

Keywords: GenoLab M, NovaSeq 6000, Nextseq 550, WGS, WES, NA12878

Posted Date: March 7th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1402182/v1

Abstract

Background: GenoLab M is a recently developed next-generation sequencing (NGS) platform from GeneMind Biosciences. To establish the performance of GenoLab M, we present the first report to benchmark and compare the WGS and WES sequencing data of the GenoLab M sequencer to NovaSeq 6000 and NextSeq 550 platform in various types of analysis. 30-fold sequencing from Illumina NovaSeq platform and processed by GATK pipeline is currently considered as the golden standard of WGS. Thus this dataset is generated as a benchmark reference in this study.

Results: GenoLab M showed an average of 94.62% of Q20 percentage for base quality, while the NovaSeq was slightly higher at 96.97%. However, GenoLab M outperformed NovaSeq or NextSeq at a duplication rate, suggesting more usable data after deduplication. For WGS short variant calling, GenoLab M showed significant accuracy improvement over the same depth dataset from NovaSeq, and reached similar accuracy to NovaSeq 33X dataset with 22x depth. For 100X WES, the F-score and Precision in GenoLab M were higher than NovaSeq or NextSeq, especially for InDel calling.

Conclusions: GenoLab M is a promising NGS platform for high-performance WGS and WES applications. For WGS, 22X depth in the GenoLab M sequencing platform offers a cost-effective alternative to the current mainstream 33X depth on Illumina.

Background

The past fifteen years have witnessed a new era in DNA sequencing technologies[1], starting from the release of the Roche 454 sequencer, which unveiled the curtain of next-generation sequencing (NGS)[2]. Compared to Sanger sequencing technology[3], NGS has remarkably higher throughput and reduced costs [1]. As technology upgrades and iterates, NGS technologies have dramatically decreased the cost of human whole genome sequencing (WGS) and whole-exome sequencing (WES). As a result, the rapid development of technology leads to brilliant achievements in WGS projects such as the 1000 genome project [4], the HapMap project[5], and extensive cohort studies worldwide. WGS and WES have been and are being widely performed to discover genetic disease-associated genes and identify driver mutations in hereditary tumors [6–8]. It lays the foundations for the prior understanding of how mutated genes affect disease phenotype and the further interpretation of pathogenic mechanisms [6–8].

Since the completion of the Human Genome Project in 2003, various sequencing platforms have been developed: Roche 454, Illumina series (GA, HiSeq, Miseq, NextSeq, NovaSeq, etc.) [9], MGI (BGISEQ-500, MGISEQ2000, DNBSEQ-T7) [10], Ion Torrent [11], and GenapSys [12]. Benefiting from continued technology development and product commercialization, Illumina’s sequencing by synthesis (SBS) based sequencers have dominated the sequencing market for a long time. In 2016, NextSeq 550 was released as mid-throughput desktop sequencing instrument, which can be applied in many fields, including transcriptome sequencing, targeted sequencing, WES, metagenomics sequencing, and genotyping. In June 2017, NovaSeq 6000 was launched, which incorporates Illumina’s SBS chemistry and two-color

optics. Combined with patterned flow cell technology and reversible terminator-based method [10], it can produce 6 TB of sequencing data in a single run at a cost of approximately 10 USD/GB [13]. As NGS applications expand in various research areas and clinical settings, there is an unmet demand to develop a novel NGS platform that is accurate, flexible, and cost-efficient for applications.

In October 2020, GeneMind Biosciences Company Limited (GeneMind) launched a new sequencing instrument (GenoLab M™) based on their previous work on single molecule sequencer GenoCare™.[14] The GenoLab M sequencer employs SBS techniques and reversible termination approaches [15]. In 2021, the first study using GenoLab M was published [15], revealing that the GenoLab M is a p promising sequencing platform for transcriptomics and LncRNA studies in animal, plant, and human with comparable performance but a lower cost compared to NovaSeq 6000. However, the performance of the GenoLab M platform in other application areas has not yet been released, especially in WGS and WES.

In 2014, Genome in a Bottle (GIAB) published A golden standard genotype dataset (including reference sample NA12878), providing a resource for comparison of variants calling pipelines [16]. Recently, several studies used the GIAB variant dataset for comparisons among different variants callers or sequencing platforms [17–20]. Generally, data depth of WGS and WES were above 30 fold and 100 fold [13, 18, 21–23]. Early in the history of WGS, the field converged around the concept that 30-fold represents a “high quality” genome with the ideal trade-off of accuracy and cost. Together with Genome Analysis Tool kit (GATK) [24] as the best practice analysis pipeline [25], this depth concept has become deeply ingrained in the community mindset, even when the sequencing and analysis fields have evolved rapidly. It is well recognized that GATK works well with dominated Illumia data, but is not yet proven on other sequencing platforms. Also, 30-fold data in WGS is potentially redundant, not only on the cost of sequencing but also the analysis computation and storage costs. There are quite a few previously published lower depth WGS studies, such as a large group WGS project of Icelanders in 2015 with a median sequencing depth was 20X [26]. In 2018, Anna Supernat et al., have compared three variant callers (DeepVariant [27], GATK, and SpeedSeq [27]) for WGS reference sample sequenced at different depths (10X, 15X, and 30X). It was observed that the F-Scores obtained by DeepVariant at 15X were comparable to SpeedSeq and GATK at 30X. Yifan Jiang et al, found that the optimal sequencing depth for whole genome re-sequencing in pigs was 10X, an ideal practical depth for achieving plateau coverage and discovering accurate variants with greater than 99% genome coverage [28]. With all these preliminary supporting studies and the emerging sequencing and analysis technologies with improved accuracy, a lower sequencing depth than 30X may be considered as the current best practice.

This study obtained both WES and WGS datasets of the NA12878 standard sample generated from multiple sequencing platforms, including NextSeq 550, NovaSeq 6000, and GenoLab M. On the analysis part, two pipelines were chosen: Sentieon DNAscope pipeline, a machine learning based variant calling workflow (https://github.com/Sentieon/sentieon-dnascope-ml), and DNAseq workflow, which is an accelerated GATK re-implementation [29]. We compared WGS performance in GenoLab M with 22X data and NovaSeq 6000 with 33X data.

Method

Samples preparation and sequencing

We ordered 50 ug NA12878 cell line genomic DNA from Sequanta Technologies Co., Ltd. After quality control, in brief, the genomic DNA was constructed as Illumina WES via SureSelect Human All Exon V8 kit (Agilent Technologies Inc.) and WGS library via TruSeq Nano DNA library kit (Illumina, Inc.). 1 ug DNA to was fragmented by Covaris E220 to 100-250 bp for WES, and to 350-450 bp for WGS. Then, end of each DNA fragment was repaired and an A base was added to the 3'end to form a sticky end, and then the Illumina adapter was ligated to both ends of DNA fragments. PCR amplification was applied to each sample after ligation. While WGS libraries were completed, the WES libraries went through additional steps, including SureSelect Human All Exon V8 capture, PCR amplification and purification.

WES library was split and loaded into GenoLab M and NextSeq 550 or NovaSeq 6000 for 150bp paired-end sequencing. And WGS library was sequenced on GenoLab M and Novaseq.

Reads mapping and bam processing

Secondary analysis was performed via Sentieon software v 202112.01 [30], a complete suite of tools that can be used to process raw reads to variant calling result. Raw reads were aligned to the hg38 (https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/references/GRCh38/) by "Sentieon BWA" and sorting was done by the "sort" utility tool. BAM files were then adjusted by Samtools v1.10 to the desired depth for later analysis and comparison, specifically 22X and 33X for the WGS dataset, and 100X for the WES dataset. Quality metrics were generated from these BAM files by Sentieon QC tools. Next, "LocusCollector" and "Dedup" tools were used to mark duplicate reads, to prepare the BAM files for variant calling step.

Running DNAseq (GATK re-implementation) and DNAscope

The Sentieon DNAseq pipeline is a re-implementation of the GATK best practice pipeline, returning identical results at a much higher speed [29]. DNAseq is typically 5-10X faster than GATK pipeline on the same generic CPU platform. Therefore here in this study, we ran DNAseq pipeline and treated the result the same as the data from GATK pipeline. Deduped BAM files were firstly processed by "QualCal" tool to conduct base quality score recalibration, and variants were called by "Haplotyper" tool to provide the matching result of GATK. VQSR was not performed because we don't believe this extra step will improve overall variant calling accuracy [31].

Deduped BAM files were directly input into DNAscope pipeline, as BQSR step is not needed here. DNAscope variant caller first generated candidate variants, filtered in the next step. GenoLab M machine learning model was applied on both variant generation and filtering steps. DNAscope is designed as a successor to GATK HaplotypeCaller, as it uniquely combines the well-validated methods from haplotype-based variant callers with machine learning to achieve improved accuracy. The candidate variants calling comprises three parts: active region detection, local haplotype assembly, and read-likelihood calculation

(Pair-HMM). Later the variant candidates with rich annotations are passed to a machine learning model for variant genotyping, leading to improvements in both variant calling and genotyping accuracy.

The GenoLab model for DNAscope was constructed during this project using several WGS and WES datasets sequenced from reference samples. Due to the limited training dataset, separated WGS and WES models were trained. The training was performed across all chromosomes with the exception of chromosome 20. It should be noted that none of the evaluated datasets was used during training.

Variant accuracy evaluation

All VCF files generated from DNAseq or DNAscope pipelines were taken as input for accuracy evaluation. They were compared against the NIST truth set v4.2.1 using hap.py v0.3.14 with RTGtools vcfeval v3.10.1 as the variant comparison engine [32] to calculate an F-score as a representation of accuracy. Stratification region files v2.0 were downloaded from GIAB project and used for stratification analysis [33].

Results

NGS datasets summary

To avoid biased results by different sample prep and library construction processes, we used the same WGS or WES library. In total, there are 3 WES and 2 WGS datasets obtained from GenoLab M, and NovaSeq 6000 or NextSeq 550 (Figure 1), and the dataset were subsampled to 100x for WES and 22x for WGS to generate additional datasets for comparison. FASTQ and BAM quality statistics were calculated, as shown in Table 1. For the base quality (over Q20) base percentages, the GenoLab M showed an average of 94.62%, slightly lower than NovaSeq's performance at 96.97%. While the duplication rate of GenoLab M outperformed NovaSeq or NextSeq, which was only half of NovaSeq's duplication rate at the same sequencing depth. A lower duplication rate usually leads to higher data usage and less waste.

The performance of 22× WGS data in GenoLab M

Subsequently, we compared the WGS SNP&InDel calling accuracy of GenoLab M and NovaSeq with analysis algorithms adapted to each sequencer at 22X and 33X depth. As shown in Figure 2A&B, the F-score, Recall, and Precision of SNP and InDel from 33X WGS were higher than 22X WGS from the same sequencing platform. At the same depth, GenoLab M showed higher recall and precision in SNP and InDel calling than NovaSeq. Interestingly, 22X WGS from GenoLab M had similar performance in SNP, and a slight advantage in InDel, compared to 33X WGS from NovaSeq. GenoLab M's analysis ML model could be part of the reason. The characteristics of the sequencing data are also likely to contribute to the difference. In addition, stratification comparison was performed including Chromosome 20 (chr20), which was not included in any of DNAscope's model training dataset; Segmental duplications region (SDR); and "Not in all Difficult Regions" (NIADR). As displayed in Figure 2C&D, stratification comparison was similar to the whole genome, especially in SDR, 22X GenoLab M dataset reached better performance (F-score of

0.941 and 0.923, respectively) in SNP and InDel calling compared to 33X NovaSeq dataset (F-scores 0.884 and 0.870, respectively).

The variant calling results of two platforms at 22X or 33X depth were filtered using GIAB NA12878 truth vcf file. The distribution of the after-filter variants representing concordance of each dataset was shown in Venn diagrams (SNP, Figure 3A and InDel, Figure 3B). All four datasets jointly identified 3,242,150 SNPs and 463,305 InDels, which were more than 99% of the truth variants. For common sets of variants, the proportion of SNP (96.27%, 3,133,010) was significantly higher than that of InDel (85.45%, 399,648). Besides, 22X WGS from GenoLab M (98.24% and 92.75%) showed indistinguishable SNP detection and slightly inferior InDel, compared with 33X data from NovaSeq (98.70% and 95.15%).

Variants calling performance in WES datasets

Three WES datasets at their raw sequencing depth and three more datasets subsampled to 100X were generated for WES performance assessment. As expected, SNP and InDel F-score, Recall, and Precision of the subsampled datasets dropped from their original depth (Figure 4). At 100X, the F-score and Precision in GenoLab M were higher than NovaSeq or NextSeq, while the Recall in GenoLab M was slightly lower.

Same as with WGS concordance analysis, the variant calling results of six WES datasets were filtered by reference truth, and concordance was shown in Figure 5. All six datasets jointly identified 20,707 SNPs and 425 InDels, which were more than 97% of the truth variants' amount, with the majority shared among all six datasets. For InDel, 100X depth in all platforms has no specific number, compared with raw data, while, for SNP, GenoLab M and NovaSeq have a small number of mutation detection. Overall, at 100X depth, GenoLab M (20,371) displayed comparable recall in SNP detection compared with NovaSeq (20,490) or NextSeq (20,388), and slightly inferior in InDel detection.

Discussion

In the past ten years, with the development of NGS sequencers by companies such as Illumina, MGI, and Ion Torrent, the application of WES or WGS to identify variants of the human genome became accessible for the public and even individuals. To further expand the accessibility, various variants calling pipelines have been developed to adapt each of these sequencing platforms, introduced by published benchmark studies. For WGS, 30-fold represents a "high quality" genome, and GATK is one popular bioinformatics analysis tool.

In this study, WES and WGS datasets of the NA12878 standard sample were generated from NextSeq 550, NovaSeq 6000, and GenoLab M. We measured the base quality (Q20&Q30), duplication rate, and the average sequencing depth of each dataset. Since GenoLab M is a new sequencing platform, GenoLab ML model for DNAscope was constructed using several WGS and WES datasets generated from reference samples. For Illumina platforms, GATK pipeline analysis was performed. For Q20 percentages, the GenoLab M showed an average of 94.62%, and the NovaSeq 6000 was 96.97%, with a slight

preponderance. At the same time, the duplication rate of GenoLab M was only half of NovaSeq 6000 under the same sequencing depth (Table 1).

Analysis observed that 22X GenoLab M WGS showed higher accuracy than 22X NovaSeq accuracy and reached a similar performance of 33X NovaSeq (Fig. 2A&B). Both low duplication sequencing and Genolab analysis ML model contribute to the variant calling accuracy. Here we believe GenoLab M offers a cost-effective alternative to the NovaSeq 6000 platform with less depth (22X) and similar data quality for human resequencing applications. GenoLab's lower duplication rate may lead to better data efficiency. The human genome shows a complex pattern of highly identical, interspersed segmental duplication, also known as SDR [34, 35].This region poses particular challenges for gene annotation because:

1. Enriched in assembly gaps [36];
2. More prone to copy number polymorphism among individuals[37];
3. Different paralogs are difficult to distinguish because of their high sequence identity [38].

- The existence of SDR predisposes humans to large-scale rearrangements due to unequal crossing-over leading to genomic instability associated with neurodevelopmental delay and autism [39]. The demonstrated accuracy advantages of GenoLab M sequencing platform in the SDR of the human genome may be suitable to NGS projects on neurodegeneration disease and autism.

In WES analysis, recall of GenoLab M was still lower than NovaSeq or NextSeq at te same sequencing depth, which serves as a development target for us. To improve overall variant calling accuracy, more GenoLab M reference datasets are required to assemble a larger training set for future DNAscope model training. Also, the collection and sequencing of more clinical or scientific samples will further help GeneMind R&D to improve sequencing instruments' performance, such as increasing the Quality value (Q20&Q30) and throughput.

Conclusions

22X WGS in GeneMind sequencing platform showed a similar performance to 33X depth in Illumina NovaSeq 6000, which offers an effective alternative. And 100X WES of GenoLab M showed similar or superior performance to Illumina platforms at the same depth, which also has application prospects in WES.

Declarations

Ethics approval and consent to participate

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

Consent for publication

Not applicable.

Availability of data and materials

The bam files of WGS and WES are available in CNGB Sequence Archive (https://db.cngb.org/cnsa/) under project accession number CNP0002694.

Competing interests

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

Funding

Not applicable.

Authors' contributions

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

Acknowledgements

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

References

1. Zheng J, Zhang H, Banerjee S, Li Y, Zhou J, Yang Q, Tan X, Han P, Fu Q, Cui X: **A comprehensive assessment of Next-Generation Sequencing variants validation using a secondary technology**. Molecular genetics & genomic medicine 2019, **7**(7):e00748.
2. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z: **Genome sequencing in microfabricated high-density picolitre reactors**. Nature 2005, **437**(7057):376–380.
3. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors**. *Proceedings of the national academy of sciences* 1977, **74**(12):5463–5467.
4. Consortium GP: **An integrated map of genetic variation from 1,092 human genomes**. Nature 2012, **491**(7422):56–65.
5. Consortium IH: **A second generation human haplotype map of over 3.1 million SNPs**. Nature 2007, **449**(7164):851.
6. Saunders CJ, Miller NA, Soden SE, Dinwiddie DL, Noll A, Alnadi NA, Andraws N, Patterson ML, Krivohlavek LA, Fellis J: **Rapid whole-genome sequencing for genetic disease diagnosis in neonatal**

intensive care units. Science translational medicine 2012, **4**(154):154ra135-154ra135.

7. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J: **Exome sequencing as a tool for Mendelian disease gene discovery**. Nature Reviews Genetics 2011, **12**(11):745–755.
8. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER: **The next-generation sequencing revolution and its impact on genomics**. Cell 2013, **155**(1):27–38.
9. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR: **Accurate whole human genome sequencing using reversible terminator chemistry**. nature 2008, **456**(7218):53–59.
10. Kumar KR, Cowley MJ, Davis RL: **Next-generation sequencing and emerging technologies**. In: *Seminars in thrombosis and hemostasis: 2019*: Thieme Medical Publishers; 2019: 661–673.
11. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M: **An integrated semiconductor device enabling non-optical genome sequencing**. Nature 2011, **475**(7356):348–352.
12. Esfandyarpour H, Parizi KB, Barmi MR, Rategh H, Wang L, Paliwal S, Golnabi H, Kenney P, Reel R, Lee F: **High accuracy DNA sequencing on a small, scalable platform via electrical detection of single base incorporations**. Biorxiv 2020:604553.
13. Jeon SA, Park JL, Park S-J, Kim JH, Goh S-H, Han J-Y, Kim S-Y: **Comparison between MGI and Illumina sequencing platforms for whole genome sequencing**. Genes & Genomics 2021, **43**(7):713–724.
14. Zhao L, Deng L, Li G, Jin H, Cai J, Shang H, Li Y, Wu H, Xu W, Zeng L: **Single molecule sequencing of the M13 virus genome without amplification**. PLoS One 2017, **12**(12):e0188181.
15. Liu Y, Han R, Zhou L, Luo M, Zeng L, Zhao X, Ma Y, Zhou Z, Sun L: **Comparative performance of the GenoLab M and NovaSeq 6000 sequencing platforms for transcriptome and LncRNA analysis**. BMC genomics 2021, **22**(1):1–12.
16. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M: **Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls**. Nature biotechnology 2014, **32**(3):246–251.
17. Hwang S, Kim E, Lee I, Marcotte EM: **Systematic comparison of variant calling pipelines using gold standard personal exome variants**. Scientific reports 2015, **5**(1):1–8.
18. Chen J, Li X, Zhong H, Meng Y, Du H: **Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers**. Scientific reports 2019, **9**(1):1–13.
19. Cornish A, Guda C: **A comparison of variant calling pipelines using genome in a bottle as a reference**. *BioMed research international* 2015, **2015**.
20. Yu X, Sun S: **Comparing a few SNP calling algorithms using low-coverage sequencing data**. BMC bioinformatics 2013, **14**(1):1–15.

21. Korostin D, Kulemin N, Naumov V, Belova V, Kwon D, Gorbachev A: **Comparative analysis of novel MGISEQ-2000 sequencing platform vs Illumina HiSeq 2500 for whole-genome sequencing**. Plos one 2020, **15**(3):e0230301.
22. Kim H-M, Jeon S, Chung O, Jun JH, Kim H-S, Blazyte A, Lee H-Y, Yu Y, Cho YS, Bolser DM: **Comparative analysis of 7 short-read sequencing platforms using the Korean reference genome: MGI and Illumina sequencing benchmark for whole-genome sequencing**. GigaScience 2021, **10**(3):giab014.
23. Foox J, Tighe SW, Nicolet CM, Zook JM, Byrska-Bishop M, Clarke WE, Khayat MM, Mahmoud M, Laaguiby PK, Herbert ZT: **Performance assessment of DNA sequencing platforms in the ABRF Next-Generation Sequencing Study**. Nature Biotechnology 2021, **39**(9):1129–1140.
24. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data**. Genome research 2010, **20**(9):1297–1303.
25. Franke KR, Crowgey EL: **Accelerating next generation sequencing data analysis: an evaluation of optimized best practices for Genome Analysis Toolkit algorithms**. Genomics & informatics 2020, **18**(1).
26. Gudbjartsson DF, Sulem P, Helgason H, Gylfason A, Gudjonsson SA, Zink F, Oddson A, Magnusson G, Halldorsson BV, Hjartarson E: **Sequence variants from whole genome sequencing a large group of Icelanders**. Scientific data 2015, **2**(1):1–11.
27. Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM: **SpeedSeq: ultra-fast personal genome analysis and interpretation**. Nature methods 2015, **12**(10):966–968.
28. Jiang Y, Jiang Y, Wang S, Zhang Q, Ding X: **Optimal sequencing depth design for whole genome re-sequencing in pigs**. Bmc Bioinformatics 2019, **20**(1):1–12.
29. Kendig KI, Baheti S, Bockol MA, Drucker TM, Hart SN, Heldenbrand JR, Hernaez M, Hudson ME, Kalmbach MT, Klee EW: **Sentieon DNASeq variant calling workflow demonstrates strong computational performance and accuracy**. Frontiers in genetics 2019:736.
30. Freed D, Aldana R, Weber JA, Edwards JS: **The Sentieon Genomics Tools-A fast and accurate solution to variant calling from next-generation sequence data**. BioRxiv 2017:115717.
31. Zhao S, Agafonov O, Azab A, Stokowy T, Hovig E: **Accuracy and efficiency of germline variant calling pipelines for human genome data**. Scientific reports 2020, **10**(1):1–12.
32. Cleary JG, Braithwaite R, Gaastra K, Hilbush BS, Inglis S, Irvine SA, Jackson A, Littin R, Rathod M, Ware D: **Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines**. BioRxiv 2015:023754.
33. Krusche P, Trigg L, Boutros PC, Mason CE, De La Vega FM, Moore BL, Gonzalez-Porta M, Eberle MA, Tezak Z, Lababidi S: **Best practices for benchmarking germline small-variant calls in human genomes**. Nature biotechnology 2019, **37**(5):555–560.

34. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE: **Recent segmental duplications in the human genome**. Science 2002, **297**(5583):1003–1007.
35. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE: **Segmental duplications: organization and impact within the current human genome project assembly**. Genome research 2001, **11**(6):1005–1017.
36. Alkan C, Sajjadian S, Eichler EE: **Limitations of next-generation genome sequence assembly**. Nature methods 2011, **8**(1):61–65.
37. Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M: **Global diversity, population stratification, and selection of human copy-number variation**. Science 2015, **349**(6253):aab3761.
38. Dougherty ML, Underwood JG, Nelson BJ, Tseng E, Munson KM, Penn O, Nowakowski TJ, Pollen AA, Eichler EE: **Transcriptional fates of human-specific segmental duplications in brain**. Genome research 2018, **28**(10):1566–1576.
39. Cantsilieris S, Sunkin SM, Johnson ME, Anaclerio F, Huddleston J, Baker C, Dougherty ML, Underwood JG, Sulovari A, Hsieh P: **An evolutionary driver of interspersed segmental duplications in primates**. Genome biology 2020, **21**(1):1–35.

Tables

Table 1 is available in the Supplemental Files section.

Figures

Figure 1

The flowchart of combinations using three sequencers and two variant calling pipelines for germline variants. Key process for NGS data generation and analysis were shown on the left. Squares in the flowchart represent data files, and rhombus indicate processes. NovaSeq means NovaSeq 6000, NextSeq means NextSeq 550.

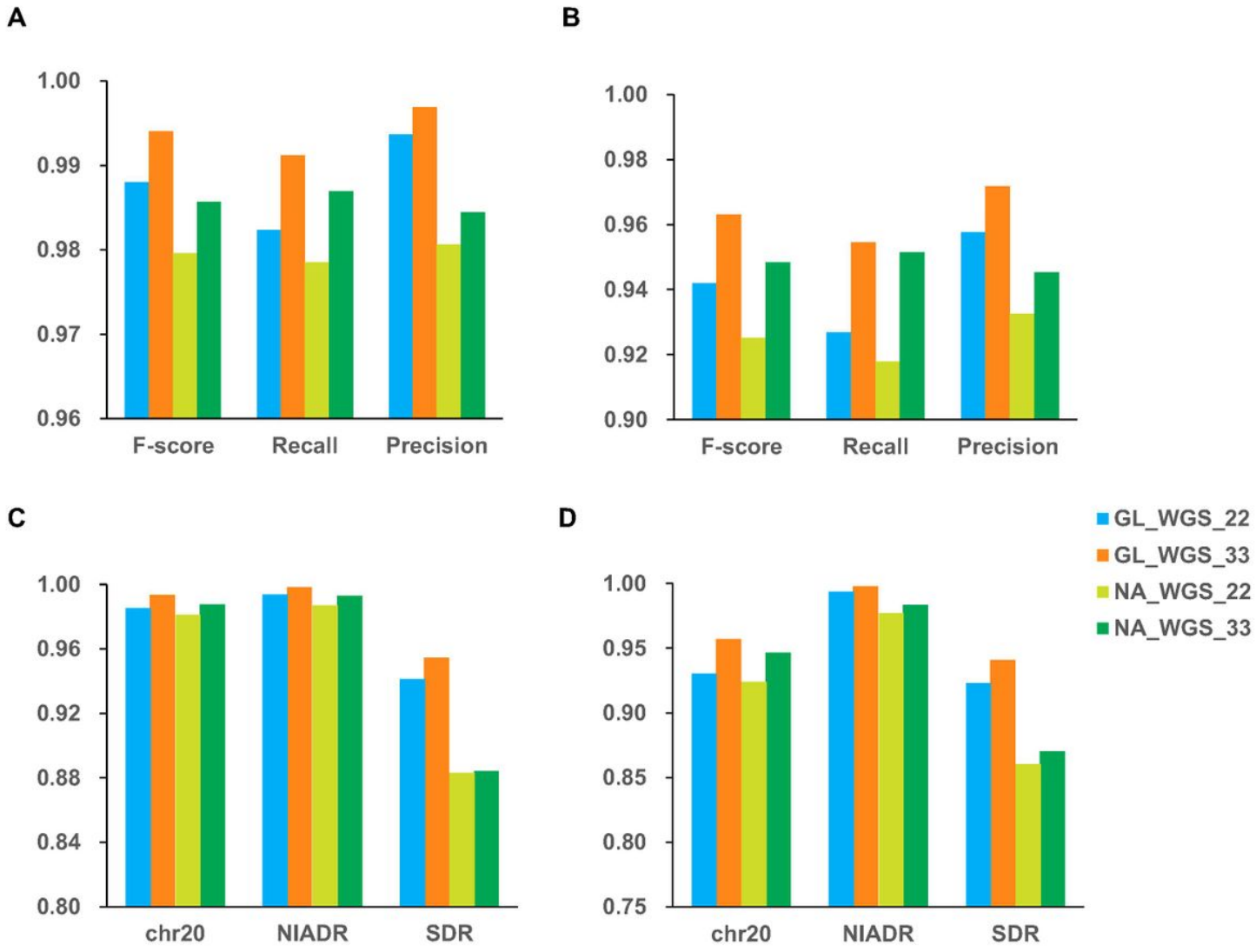
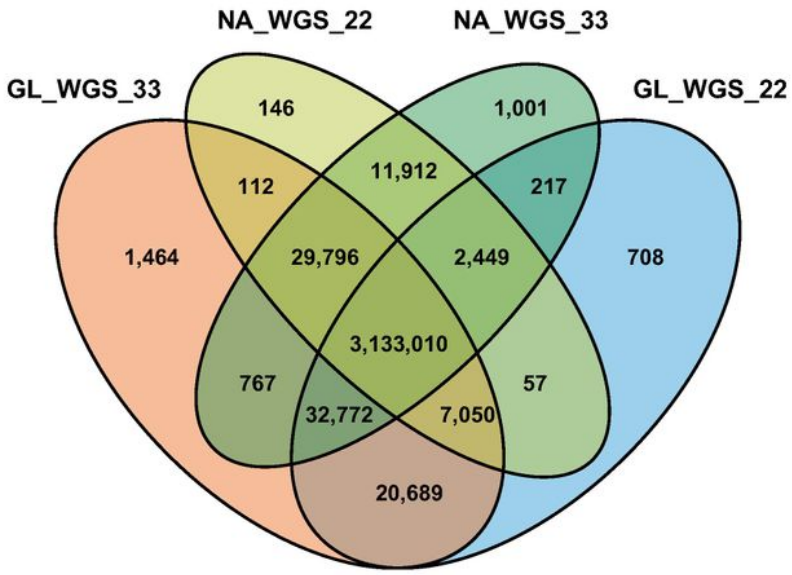


Figure 2

Comparison of variants calling performances in GenoLab M and NovaSeq 6000 from 33X and 22X coverage of the NA12878 sample. A SNP and B InDel on whole genome, C SNP and D InDel F-score on stratification region. GL_ and NA_ means GenoLab M and NovaSeq 6000, chr 20 means chromosome 20, NIADR means Not in all Difficult Regions, SDR means Segmental Duplications Regions.

A



B

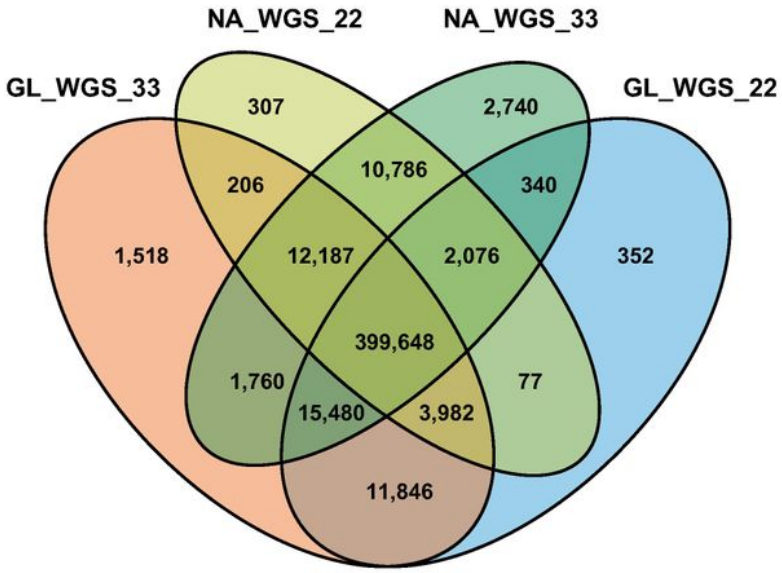


Figure 3

Venn diagram of variants calling performances in WGS datasets. A SNP and B InDel.

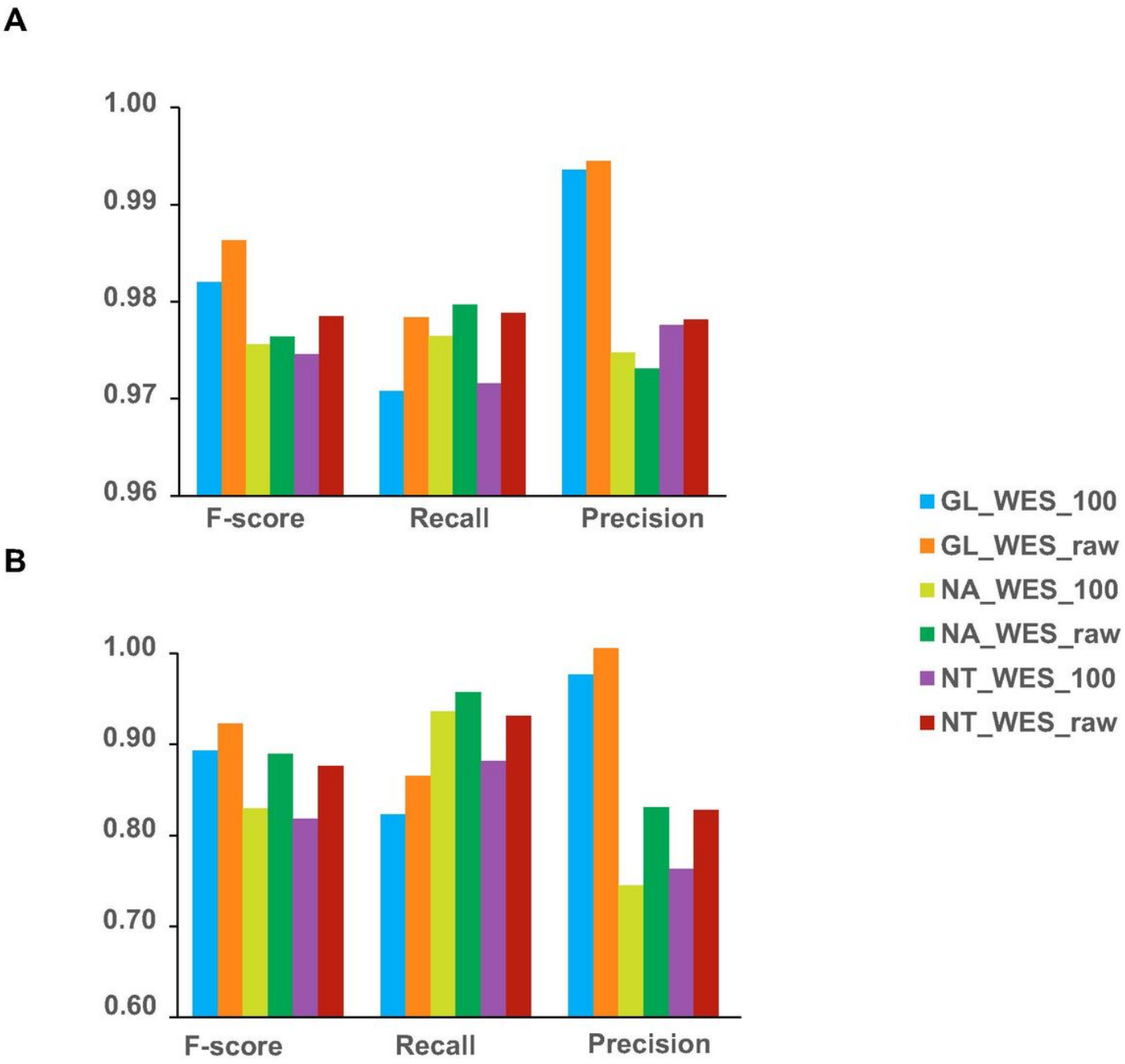


Figure 4

Comparison of variants calling performances in six WES datasets. GL_, NA_ and NT_ means GenoLab M, NovaSeq 6000 and NextSeq 550, respectively. A SNP and B InDel.

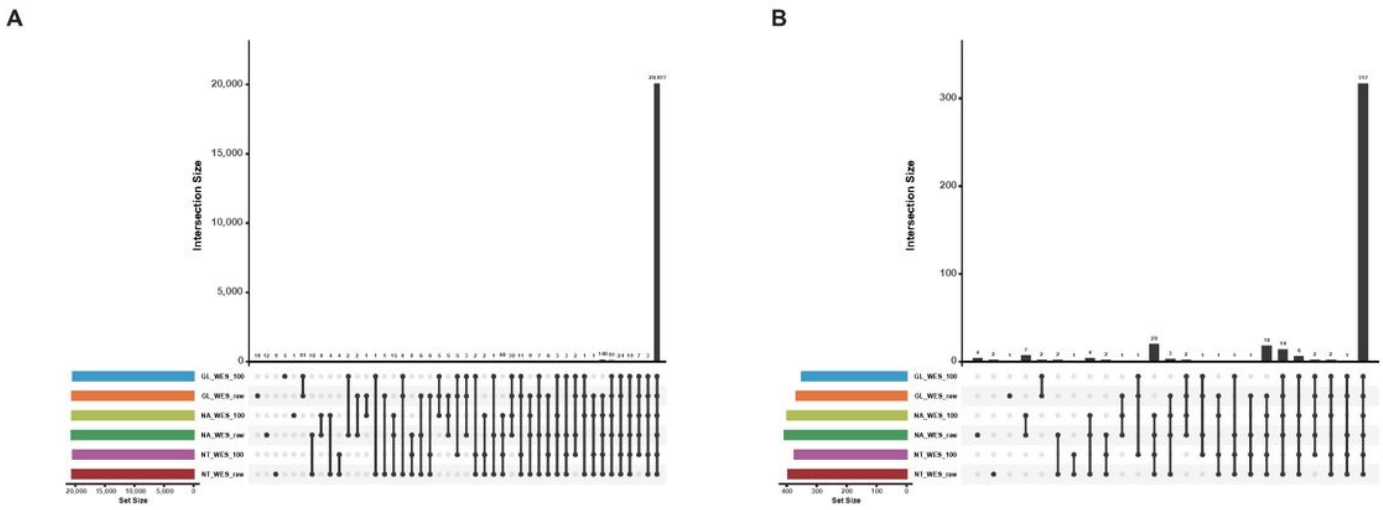


Figure 5

Upset diagram of variant Calling results of all combinations in WES datasets. A SNP and B InDel.

Supplementary Files

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

- Table1.xlsx