https://doi.org/10.1093/nargab/lqac033

Nano2NGS-Muta: a framework for converting nanopore sequencing data to NGS-liked sequencing data for hotspot mutation detection

Jidong Lang*, Jiguo Sun, Zhi Yang, Lei He, Yu He, Yanmei Chen, Lei Huang, Ping Li, Jialin Li and Liu Qin

Bioinformatics and Product Development Department, Qitan Technology (Beijing) Co., Ltd, Beijing 100192, China

Received November 14, 2021; Revised March 30, 2022; Editorial Decision April 11, 2022; Accepted April 13, 2022

ABSTRACT

Nanopore sequencing, also known as single-molecule real-time sequencing, is a third/fourth generation sequencing technology that enables deciphering single DNA/RNA molecules without the polymerase chain reaction. Although nanopore sequencing has made significant progress in scientific research and clinical practice, its application has been limited compared with next-generation sequencing (NGS) due to specific design principle and data characteristics, especially in hotspot mutation detection. Therefore, we developed Nano2NGS-Muta as a data analysis framework for hotspot mutation detection based on long reads from nanopore sequencing. Nano2NGS-Muta is characterized by applying nanopore sequencing data to NGS-liked data analysis pipelines. Long reads can be converted into short reads and then processed through existing NGS analysis pipelines in combination with statistical methods for hotspot mutation detection. Nano2NGS-Muta not only effectively avoids false positive/negative results caused by non-random errors and unexpected insertions-deletions (indels) of nanopore sequencing data, improves the detection accuracy of hotspot mutations compared to conventional nanopore sequencing data analysis algorithms but also breaks the barriers of data analysis methods between short-read sequencing and long-read sequencing. We hope Nano2NGS-Muta can serves as a reference method for nanopore sequencing data and promotes higher application scope of nanopore sequencing technology in scientific research and clinical practice.

INTRODUCTION

Nanopore sequencing, also known as third-generation sequencing (TGS) or single-molecule real-time DNA sequencing, enables the identification of single DNA molecules without requiring the polymerase chain reaction (PCR). TGS is dominated by two technologies: Pacific Biosciences' (Pacbio) single molecule fluorescent sequencing through their single molecule real time (SMRT) technique (1) and Oxford Nanopore Technologies (ONT) or Qitan Technology's nanopore sequencing by electrophoresis (2–5). The former technique harnesses the intrinsic speed of DNA polymerase. Ten bases can be detected per second, which is 20 000 times the speed of chemical sequencing. This technique also exploits the processing capacity of DNA polymerase, and long DNA molecules can be sequenced in a single reaction. The latter technique is marked by the absence of PCR amplification or chemical labeling during real-time sequencing of DNA or RNA molecules, thus avoiding the introduction of false mutations during operation and ensuring high fidelity. The sequencing speed can reach 450 bp/s for DNA and 70 nt/s for RNA. NGS can generate reads of hundreds of bases, whereas TGS can produce reads of several kilobases, or even ultra-long reads (several megabases) (6,7).

In recent years, nanopore sequencing technology has made great achievements in the application of genome (especially bacterial genome) assembly and metagenomics. Long reads from nanopore sequencing platforms such as ONT are widely used in the study of bacterial genomes (8–10). Compared with short reads from next generation sequencing, long reads can span larger genomic repeats and complex genomic structures, thus facilitating downstream genome assembly and analysis (11–13). Meanwhile, most metagenomic studies are based on the NGS platform (Illumina), whose sequencing time is >16 h, and the overall sample-to-answer turnaround time is 48–72 h, although there are some studies into developing real-

*To whom correspondence should be addressed. Tel: +86 13811153846; Email: langjidong@hotmail.com

© The Author(s) 2022. Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

time analysis methods for NGS sequencing platforms, aiming to shorten the overall run time (14–16). In contrast, nanopore sequencing identifies pathogen sequences with real-time computational analysis within 50 min if the host DNA background can be effectively removed (e.g. using saponin (17)), and the detection can be completed within 6 h (18). Nanopore sequencing can be used to detect the genomes of a wide range of pathogenic bacteria and emerging viruses and holds great promise for clinical applications, such as real-time surveillance of epidemics at specific sites.

Currently, the application of mutation detection based on nanopore sequencing data is far less extensive than that of genome assembly and metagenomic detection. Although methods for detecting single nucleotide mutations (including germline mutations and somatic mutations) based on TGS data have not yet been well established, several research groups worldwide have been developing algorithms to accurately identify mutations such as single-nucleotide variants (SNVs) and insertions-deletions (indels) for TGS data. These algorithms include the Longshot with a pair-hidden Markov model (19), the Clair with a deep neural network model (20), and the PEPPER-Margin-DeepVariant developed and optimized via DeepVariant (21). However, there are challenges in the accurate detection of single bases from nanopore sequencing data, such as low sample quality, low stability of current passing through nanopores, and low accuracy of the base calling model. These challenges lead to poor sequencing quality, errors in base calling, and non-random systematic errors (7,22), which greatly diminish the algorithm's accuracy. Despite the advances in methods for detecting single base mutations based on nanopore sequencing data, these methods have obvious shortcomings. Most notably, they are limited by sequencing quality, the alignment algorithm, distribution of training data in deep learning, narrow applications scenarios, and low robustness.

Thus, we developed the Nano2NGS-Muta framework to address the above problems. The main idea is to convert long reads into NGS-liked short reads for downstream analysis, to some extent effectively avoiding the problems caused by non-random systematic errors or low alignment rate resulting from a high sequencing error rate, and improving the reads effective utilization. Nano2NGS-Muta was designed to evaluate and correct the detection results based on the 'common sense' that there is higher base quality in the middle than in the ends of the reads, the unique identifiers (UIDs)/unique molecular identifiers (UMIs) used in data analysis, and the weight used in statistical analysis. Nano2NGS-Muta can effectively control the sensitivity and specificity, and is theoretically particularly suitable for detection of mutations in a hotspot panel for clinical diagnosis. Nano2NGS-Muta is highly compatible with the conventional methods used for NGS data analysis, such as GATK Best Practice pipeline, thus expanding the tools of nanopore sequencing for hotspot mutation detection. Theoretically, Nano2NGS-Muta can break the barriers of data analysis methods between short-read sequencing and long-read sequencing, thus increasing the potential applications of nanopore sequencing technology in scientific research and clinical diagnosis.

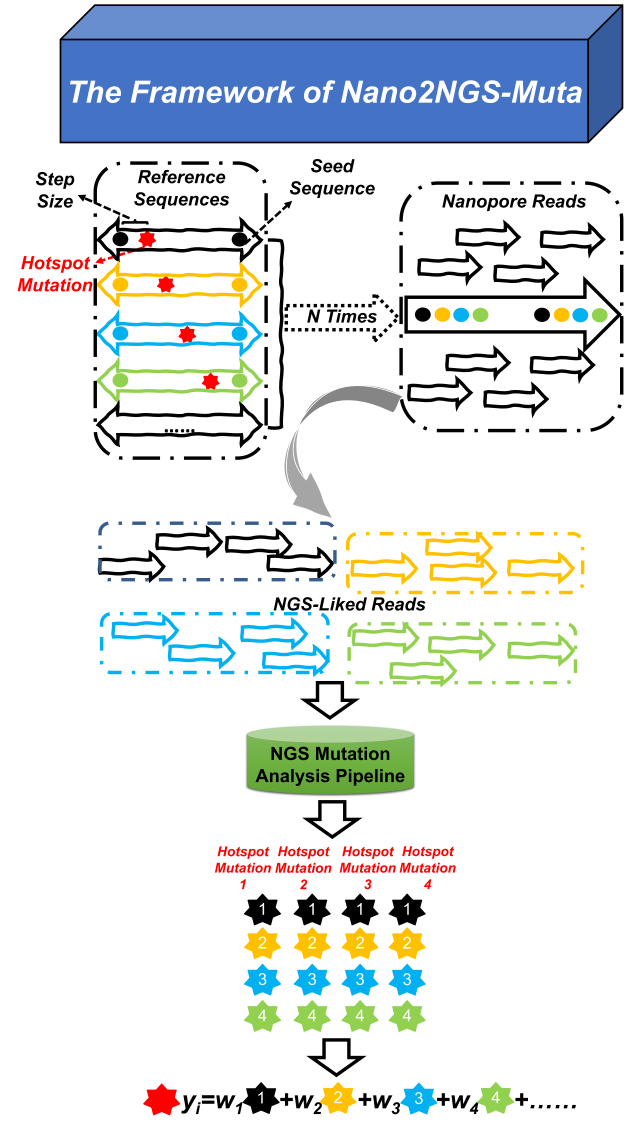

Figure 1. Schematic diagram of the Nano2NGS-Muta framework.

MATERIALS AND METHODS

Principles of Nano2NGS-Muta framework

Nano2NGS-Muta is used for hotspot mutation detection with long reads from nanopore sequencing (Figure 1) based on the following three steps. (i) Long reads are converted to short reads for subsequent analyses. First, we define the length of the short read is L, the number of times to extracted short read is N, the step size of the hotspot to be detected between each extraction on the short read is D, and N equals integer of L divided by D and subtract 1. To do this, the hotspot should be scattered to various positions at the short read, such as the 5′ end, the middle, and the 3′ end. The L-length target reference sequences containing different hotspot positions are extracted from the reference genome. Second, m bases at the top and end of each extracted reference sequence are obtained as paired seed sequences, and the seed sequences should not contain the hotspots. That is, the paired seed sequences are located outside the hotspots. Finally, *N* paired seed sequences are used to extract short reads from the original nanopore sequencing reads with zero mismatch to obtain *N* NGS-liked short read sets. (ii) The *N* short read sets were analyzed by NGS mutation analysis pipeline, respectively. (iii) The mutation detection results in *N* short read sets are integrated and filtered. Here, we use a 'common sense' to identify false-positive variants in NGS: it is generally accepted that NGS sequencing is more accurate for bases at the middle of reads, and the accuracy of bases at both ends decreases due to sequencing fluctuation at the 5′ end or poor sequencing quality at the 3′ end. The position of the mutation in the read also affects the score of the quality during sequence alignment, though the influence is insignificant due to random sequencing. On this basis, we assign weights to the *N* analysis results. High weights are assigned to hotspots in the middle; low and identical weights are assigned to hotspots at both ends; and the total weight is 1. For instance, hotspots located at the 5′ end, middle, and the 3′ end of extracted reads ($N = 3$) are assigned weights of 0.3, 0.4, and 0.3, respectively. Finally, the *N* analysis results and their weights are used to calculate the weighted sum according to the following equation, which are the final hotspot mutation detection results.

$$y = w_1x_1 + w_2x_2 + \ldots + w_{n-1}x_{n-1} + w_nx_n,$$

where w_n is the weight of hotspot mutation detection result in the nth short read set, and $w_{(n+1)/2} > w_n = w_1 > w_{n-1} = w_2 > \ldots\ldots$, if n is odd; or $w_{n/2} = w_{\frac{n}{2}+1} > w_n = w_1 > w_{n-1} = w_2 > \ldots\ldots$, if n is even; $\sum_1^n w_n = 1$; x_n is the hotspot mutation detection result in the nth short read set.

Simulation of nanopore sequencing data containing hotspot mutations

Data simulation was performed on 10 single-base mutations (*BRAF-p.V600E*, *EGFR-p.L858R*, *EGFR-p.T790M*, *EGFR-p.G719A*, *KRAS-p.G12C*, *KRAS-p.G13D*, *PIK3CA-p.H1047R*, *PIK3CA-p.E545K*, *NRAS-p.Q61R*, *FGFR3-p.Y375C*) and two indel mutations (*EGFR-p.E746A750del* and *EGFR-p.A767V769dup*). For each hotspot, 9000 wild-type reads and 1000 mutant reads were randomly generated (read lengths: 1020–3968 bp; mean = 2467 bp; sd = 783 bp). For each hotspot, data were mixed at theoretical mutation frequencies of 10%, 5%, 1%, 0.5% and 0.1%, with three replicates for each frequency. Negative control samples were randomly generated from 9000 wild-type reads with three replicates. Read duplication was removed from all data. Base quality was randomly selected between the ASCII code '+' and 'K' (phred33: Q10–Q42) (Supplementary Sheet Table S2). Since some hotspots were close to each other, the simulated reads might contain neighbor hotspots, which led to increased wild-type supporting read count and slight fluctuation in the frequency of such hotspots.

Experimental procedures for standard samples containing hotspot mutations

A genomic DNA (gDNA) standard product (GeneWell Biotechnology Co., Ltd, Shenzhen, Guangdong, China) with a tumor fraction of about 5% was used in this study, and it contained 7 hotspot mutations, *BRAF-p.V600E* (8.00%), *EGFR-p.L858R* (5.00%), *EGFR-p.T790M* (5.00%), *KRAS-p.G12C* (5.00%), *KRAS-p.G13D* (5.00%), *EGFR-p.G719A* (5.00%) and *EGFR-p. E746A750del* (5.00%). The original gDNA standard product, its 5-diluted and 10-diluted samples with the GM12878 cell line (Coriell Institute, Camden, New Jersey, USA) and the GM12878 cell line gDNA samples have also three replicates. We designed 6 forward and reverse amplification primers for these 7 hotspot mutations, which were *5′-CAGCTTGCTGCAATGCACACAAGTT-TTCTGTAGATTTCGAGGCCAGAGTCCTT-3′*, *5′-AGTTGGGCTCAGCAAGGTAGGCATC-TGATTCCAATGCCATCCACTTGATAGG-3′*, *5′-GCCTGACTCAGTGCAGCATGGATTTC-GAGAGATGACGGGCAACGGCGTAT-3′*, *5′-TGCTTGGGATGGAAGTTCTACTC-CATATTGACTTCTAACACTTAGAGGTGG-3′*, *5′-TGGTGACATGTTGGTACATCCATCCG-GCCTGAGGTTCAGAGCCATGGACC-3′ and 5′-TGCGTTCGGCACGGTGTATAAGGTA-TCGATTCTGCTTCCCTAGTCCGCTG-3′*, respectively. Then we performed PCR amplification, end repaired, and ligated nanopore sequencing adapters to build sequencing libraries. Finally, the nanopore sequencing was performed on the QNome-9604 instrument according to the manufacturer's instructions (Qitan Technology (Beijing) Co., Ltd, Beijing, China), which is a new nanopore sequencing platform.

Analysis of simulated data and standard data

Nano2NGS-Muta was used for data conversion. Read length (*L*) was set to 101 bp, step size (*D*) was 10 and seed length (*m*) was 10 bp. The positions of hotspots in the extracted sequence were 11, 21, 31, 41, 51, 61, 71, 81 and 91 ($N = 9$). Paired seed sequences were used to select the target sequences from 18 simulated datasets with zero mismatches. SAM files were generated from alignment using the BWA-MEM algorithm (version: 0.7.17-r1188) (23). SAM files were sorted to generate BAM files using Samtools (version: 1.12) (24). Mutations were detected using Freebayes (version: v1.0.2) (25) and recorded in VCF files. The data in VCF files were annotated using ANNOVAR (26). Weights were assigned to hotspots according to their positions (0.05, 0.075, 0.10, 0.15, 0.25, 0.15, 0.10, 0.075 and 0.05). Finally, the weighted sum was calculated for all hotspot mutations to obtain the final analysis results. Mutation analysis and comparison were performed on the simulated data and standard data using the Longshot (version 0.4.1), PEPPER-Margin-DeepVariant (r0.4.1), and iGDA (27) algorithms. Minimap2 (version 2.21-r1071) (28) and Sambamba (version 0.8.0) (29) were used for alignment and sorting, respectively.

RESULTS

Nano2NGS-Muta accurately detected hotspot mutations in simulated data

We compared the detection performance of hotspot mutations by Nano2NGS-Muta + Freebayes, Longshot, PEPPER-Margin-DeepVariant and iGDA on simulated data. We found that the results of Nano2NGS-Muta + Freebayes were exactly as expected (Supplementary Sheet Table S2). Neither false-positive nor false-negative results were observed, resulting in 100% sensitivity and specificity (Figure 2A). Longshot detected only six mutations at 10% frequency (*BRAF-p.V600E*, *EGFR-p.L858R*, *PIK3CA-p.H1047R*, *PIK3CA-p.E545K*, *NRAS-p.Q61R* and *FGFR3-p.Y375C*). No mutations were detected by PEPPER-Margin-DeepVariant and iGDA. For PEPPER-Margin-DeepVariant, the result may be explained by the fact that the lower limit of mutation detection was set to the default value (may be) of 20%. iGDA used the information of multiple loci to detect low-frequency SNVs and showed poor performance on detection of low-frequency hotspot mutations due to the intrinsic limitations of the algorithm's design. This result highlights the limitations of mutation detection by PEPPER-Margin-DeepVariant and iGDA. We compared the mutation frequency detected by Nano2NGS-Muta + Freebayes between repetitions and found no significant differences (Figure 2B).

Nano2NGS-Muta had better performance in standard samples

We also compared the detection performance of hotspot mutations by Nano2NGS-Muta + Freebayes, Longshot, PEPPER-Margin-DeepVariant and iGDA algorithms on sequencing data for standard samples (Supplementary Sheet Tables S3 and S4). We found that except for Nano2NGS-Muta + Freebayes, none of the other three software detected the seven known hotspot mutations. For Nano2NGS-Muta + Freebayes (Figure 2C), if we set 0.1% as the threshold for all sites in the data of repeated experiments, the detection sensitivity of this method was 95.24%, 100.00% and 85.71%, and the specificity was 28.57%, 28.57% and 42.86%, respectively. Although there were false positive results, there was no false-negative result in which none of the other three methods have detected the target mutations. While we also found that in the 0.5% and 1% mutation frequency analysis results, the mutation frequency values of replicated samples have more obvious fluctuations, and the NM12878 data also had 0.19% to 1.86% of positive results detected. This may be related to the base errors introduced during the experiment, sequencing process, basecalling algorithm and/or the background noise of the system. Therefore, we inferred that for the Nano2NGS-Muta + Freebayes mode for nanopore sequencing data, the mutation frequency limit of detection (LoD) may be 2% or even 5%.

DISCUSSION

Nanopore sequencing, or long-read sequencing, provides many advantages over short-read sequencing (30). Compared with the commercial short-read sequencers, such as Illumina's HiSeq, NextSeq and MiSeq instruments (31,32), BGI's MGISEQ and BGISEQ instruments (33,34), and Thermo Fisher's Ion Torrent instrument (35), which generate reads of up to 600 bases, long-read sequencing technologies can produce >10 kb reads (30). Short-read sequencing has evolved rapidly over the last decade and is very economical and efficient; its sequencing data are highly accurate; and there are various well-developed data analysis tools and workflows (32,36). These features are lacking in long-read sequencing technologies, and the developed tools have limited application (6,37). Most typically, the clinical application of liquid biopsy or early cancer screening, nanopore sequencing platform is helpless so far. Although there are some research methods that can be applied to cell-free DNA (cfDNA) and/or circulating tumor-derived DNA (ctDNA), they are limited to the direction of experimental development and optimization or copy number variation detection (38,39). There are still no more breakthroughs in these researches and/or applications, so NGS technology is still the main choice. But more and more studies is devoted to the development of new methods, such as Gorzynski *et al.* developed a rapid whole-genome sequencing method, which using nanopore DNA sequencing technology to improve the prognosis of critically ill patients, and the time to identify disease-causing genetic variants reduced to <8 h (40).

The main idea of Nano2NGS-Muta framework is to convert long reads into NGS-liked short reads for data analysis. The converted data are compliable with the analysis algorithms and software used for NGS short reads. The framework takes advantage and avoids the shortcomings of nanopore sequencing data while improving the reads effective utilization, thus accelerating the commercialization of nanopore sequencing technology. We also tried to apply this concept to the metagenomic taxonomical classification detection, named Nano2NGS-Meta (Supplementary Figure S1), and the satisfactory results were obtained. By analyzing the simulated data and standard sequencing data with Nano2NGS-Meta, and comparing with the results analyzed by commonly used software for nanopore sequencing data, we found that the performance of Nano2NGS-Meta can be on a par with that of the others, even slightly better in some places (Supplementary Figures S2 and S3, Sheet Tables S5–S8). For example, sample P3 contained culturable *Pseudomonas aeruginosa* and *Streptococcus pneumoniae*. Only *Pseudomonas aeruginosa* was identified in the Charalampous *et al*. paper, whereas both species were detected by Nano2NGS-Meta. Similarly, sample P34 contained *Staphylococcus aureus*, which was also identified by Nano2NGS-Meta but missed by the paper's method (Supplementary Table S1).

However, there are still some challenges in the current version. The problems in Nano2NGS-Muta include the selection of read length, the position of hotspots, and the efficiency of software, which may affect the stability and accuracy of data analysis. The algorithms and processes used for subsequent NGS data analysis also need to be considered. Although a higher number of sampling dataset is associated with a higher accuracy of results, each sampling dataset should be analyzed, which increases not only the analysis time but also the consumption of computing resources. GATK analysis pipeline, for example, consumes large amount of memory and storage. All these problems need to be addressed considering data volume and practical application. It is necessary to balance analysis accuracy with the time and consumption of computational resources. Another is the LoD to this method, that is, when

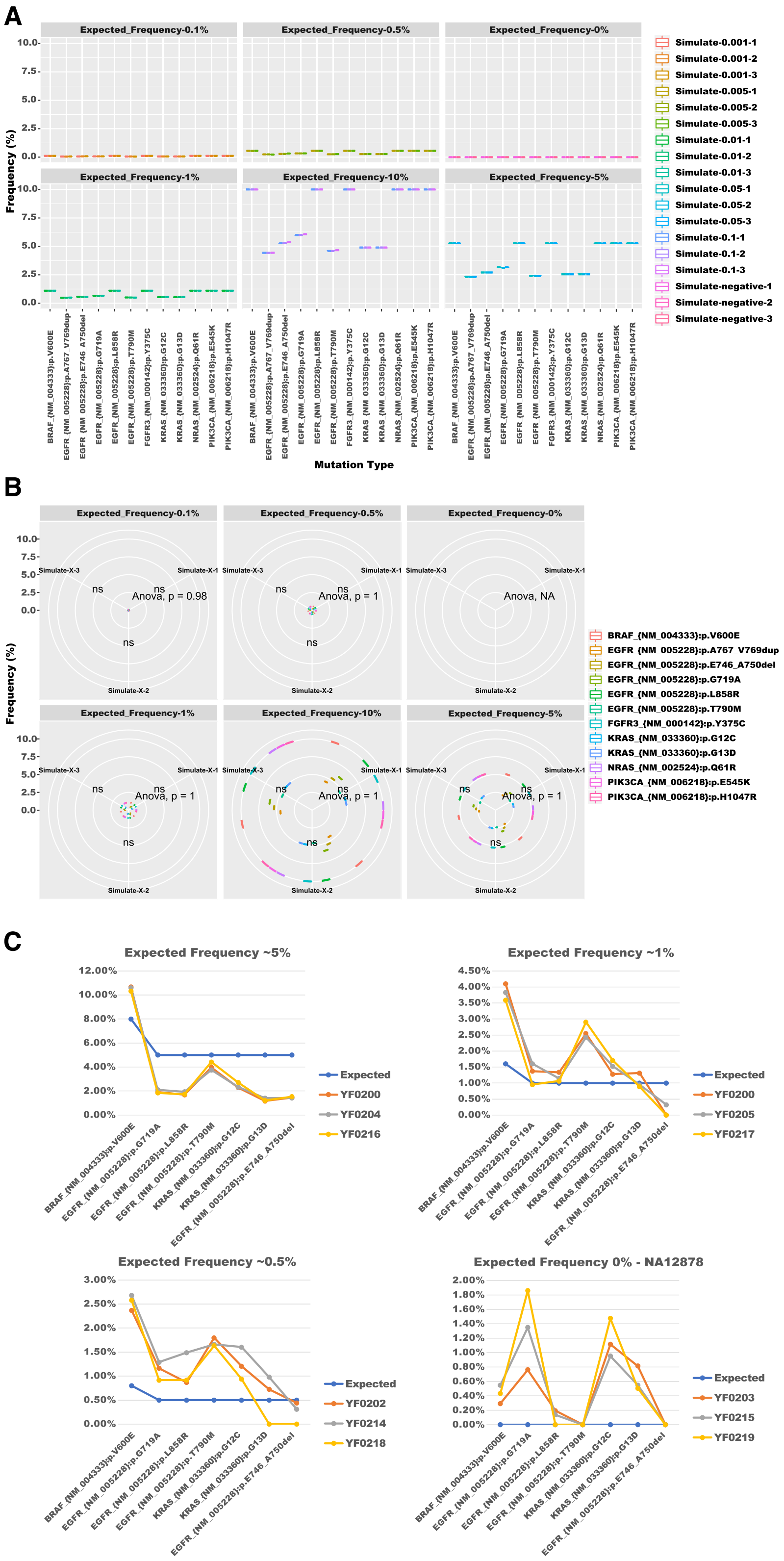

Figure 2. Comparison of Nano2NGS-Muta + Freebayes, Longshot, PEPPER-Margin-DeepVariant, and iGDA on simulated data and standard experimental data. (**A**) Distribution of mutations detected by Nano2NGS-Muta + Freebayes on simulated reads. (**B**) Significance of differences in the mutation frequency between replicates detected by Nano2NGS-Muta + Freebayes. (**C**) Performance of four detection methods on standard experimental data.

the threshold of mutation frequency is detected, the false positive rate and false negative rate could be better controlled, which more samples need to be collected for evaluation. Therefore, for the application scenarios of the detection method combined with the data characteristics should be emphatically considered. For example, the default LoD of Deepvariant-pepper-margin may be 20%, and the LoD of Nano2NGS-Muta may be 2% or even 5%, which means that mutation detection of nanopore long-read sequencing may only be used to detect the high-frequency mutations or low-frequency rare mutations such as genetic diseases. These will also be the focus on further development and optimization of the Nano2NGS-Muta framework. The detection of hotspot mutations by iGDA and PEPPER-Margin-DeepVariant algorithms is dependent on the training dataset with consistent data characteristics and distribution, whereas our own experimental data were insufficient as training dataset. Therefore, the default training datasets published with software were used to do the data analysis, but it may be inappropriate to use such results for comparison. Although we simulated ONT or ONT-like data, they were not generated on the same experimental and sequencing platforms. We will regenerate training datasets for these two algorithms based on lots of experimental data to improve the accuracy of evaluation.

Though the overall analysis performance of Nano2NGS-Meta was close to conventional analytical algorithms for nanopore sequencing data, there were differences in the calculated relative abundance and problems of missed or false-positive species. For the undetected species, to be noted, we checked the analyzed process files and found that some results were not detected in all the extracted data, they were just filtered out when integrating them. For example, *Salmonella enterica* was not detected by the Nano2NGS-Meta + Metaphlan in simulated data-1. In fact, this strain was detected in 4 of 30 extracted NGS-liked short read sets, the relative abundance value was 0.129, 0.355, 0.247 and 0.123, respectively, and the remain results were 0. With a confidence level of 0.05, the 95% confidence interval is [-0.0008, 0.0578], so the four detected results were regarded as outliers. Certainly, the reason may also be insufficient randomly simulated data, loss of long-read information during NGS-liked short-read extraction or inaccurate taxonomical classification due to the similar conserved regions of selected species along with the presence of mismatches and indels. This issue required more detailed analysis and discussion. Moreover, the construction of standard databases for alignment is critical, and the databases used by conventional software have specific characteristics and modifications. Probably, none of these databases include all species, which requires laboratory-built databases for taxonomical classification. Selecting the threshold and filtering the relative abundance of detected species are also important. Based on the analysis of published data, some true-positive species could be detected but showed very low relative abundance ($1e-16$), and such a low threshold would inevitably lead to very high false-positive rate. If a high threshold is used for detection, some key pathogenic species might be missed in scenarios with strong host backgrounds (e.g. alveolar lavage fluid) (41,42). The host genome not only reduces the relative proportion of metagenomic DNA and the volume of metagenomic data in subsequent analyses but also increases sequencing data volume for validation, leading to high sequencing cost and extremely low cost-effectiveness. In data analysis, it is important to identify microbial taxa with high sensitivity and a well-controlled false-positive rate and to address inefficient data analysis caused by the intrinsic characteristics of nanopore sequencing data. Therefore, we suggest that different thresholds of relative abundance should be used in different application scenarios to balance false-positive and false-negative results. Undoubtedly, there are also significant challenges in the analytical methods for metagenomic data from nanopore sequencing platform.

In summary, the Nano2NGS-Muta as an analytical framework requires further development and optimization regarding computing resource consumption, running time, and statistical algorithms, among others, to minimize running time and resource consumption while ensuring the accuracy of analysis. In the future, we will continue to extend this concept to develop algorithms for the detection of copy number variation, structural variation, gene fusion and gene expression, and integrate them into a big framework, named Nano2NGS. While enriching the functions of the Nano2NGS framework, it also expands more applications in scientific research and clinical practice. Deep learning models or algorithms such as Convolutional Neural Networks (CNN), support vector machines (SVM) and bootstrap will also be incorporated into the framework to improve the performance and accuracy of analysis. We are also working on optimizing the Nano2NGS-Muta and Nano2NGS-Meta to improve the performance of nanopore sequencing data analysis. These efforts will accelerate the application and popularization of nanopore long-read sequencing, so that TGS can better serve the development of the sequencing industry and the commercialization of precision medicine.

CONCLUSIONS

The Nano2NGS-Muta framework converts nanopore sequencing data into NGS-liked short-read data and can be compatible with the NGS data processing algorithms and/or software for hotspot mutation detection, and shows higher sensitivity and specificity. Nano2NGS-Muta can improve effective utilization of nanopore sequencing data, and effective in solving the problems of low detection accuracy and limited applications of nanopore sequencing data analysis tools. Nano2NGS-Muta is highly extensible and accelerates the application of nanopore sequencing technology in scientific research and clinical diagnosis.

DATA AVAILABILITY

The download link of the strain reference genome sequences is https://www.ncbi.nlm.nih.gov/genome/; the download link of the NA12878's Nanopore data is https://github.com/nanopore-wgs-consortium/NA12878; the download link of the NA12878′s data analysis result file is ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome/project.NIST.hc.snps.indels.vcf. FASTQ data files for this study can be found in the NCBI Sequence Read Archive (SRA) database

(BioProject ID: PRJNA779570). The codes are available at https://github.com/langjidong/Nano2NGS.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

Author Contributions: Jidong Lang designed the project, analyzed the data and wrote the manuscript. Zhi Yang, Yanmei Chen, Lei Huang and Ping Li collected the data, did the experiments and sequencing. Jiguo Sun, Lei He, Yu He, Jialin Li and Liu Qin analyzed the data.

FUNDING

No external funding.
Conflict of interest statement. None declared.

REFERENCES

1. Eid,J., Fehr,A., Gray,J., Luong,K., Lyle,J., Otto,G., Peluso,P., Rank,D., Baybayan,P., Bettman,B. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.
2. Clarke,J., Wu,H.C., Jayasinghe,L., Patel,A., Reid,S. and Bayley,H. (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.*, **4**, 265–270.
3. Goyal,P., Krasteva,P.V., Van Gerven,N., Gubellini,F., Van den Broeck,I., Troupiotis-Tsaïlaki,A., Jonckheere,W., Péhau-Arnaudet,G., Pinkner,J.S., Chapman,M.R. *et al.* (2014) Structural and mechanistic insights into the bacterial amyloid secretion channel csgG. *Nature*, **516**, 250–253.
4. Ip,C.L.C., Loose,M., Tyson,J.R., de Cesare,M., Brown,B.L., Jain,M., Leggett,R.M., Eccles,D.A., Zalunin,V., Urban,J.M. *et al.* (2015) MinION analysis and reference consortium: phase 1 data release and analysis. *F1000Res*, **4**, 1075.
5. Istace,B., Friedrich,A., d'Agata,L., Faye,S., Payen,E., Beluche,O., Caradec,C., Davidas,S., Cruaud,C., Liti,G. *et al.* (2017) de novo assembly and population genomic survey of natural yeast isolates with the Oxford nanopore MinION sequencer. *Gigascience*, **6**, 1–13.
6. Magi,A., Semeraro,R., Mingrino,A., Giusti,B. and D'Aurizio,R. (2018) Nanopore sequencing data analysis: state of the art, applications and challenges. *Brief Bioinform*, **19**, 1256–1272.
7. Wang,Y., Zhao,Y., Bollas,A., Wang,Y. and Au,K.F. (2021) Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.*, **39**, 1348–1365.
8. Loman,N.J., Quick,J. and Simpson,J.T. (2015) A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods*, **12**, 733–735.
9. Taylor,T.L., Volkening,J.D., DeJesus,E., Simmons,M., Dimitrov,K.M., Tillman,G.E., Suarez,D.L. and Afonso,C.L. (2019) Rapid, multiplexed, whole genome and plasmid sequencing of foodborne pathogens using long-read nanopore technology. *Sci. Rep.*, **9**, 16350.
10. Senol Cali,D., Kim,J.S., Ghose,S., Alkan,C. and Mutlu,O. (2019) Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions. *Brief Bioinform*, **20**, 1542–1559.
11. Miga,K.H., Koren,S., Rhie,A., Vollger,M.R., Gershman,A., Bzikadze,A., Brooks,S., Howe,E., Porubsky,D., Logsdon,G.A. *et al.* (2020) Telomere-to-telomere assembly of a complete human x chromosome. *Nature*, **585**, 79–84.
12. Jung,H., Winefield,C., Bombarely,A., Prentis,P. and Waterhouse,P. (2019) Tools and strategies for long-read sequencing and de novo assembly of plant genomes. *Trends Plant Sci.*, **24**, 700–724.
13. Nurk,S., Koren,S., Rhie,A., Rautiainen,M., Bzikadze,A.V., Mikheenko,A., Vollger,M.R., Altemose,N., Uralsky,L., Gershman,A. *et al.* (2022) The complete sequence of a human genome. *Science*, **376**, 44—53.
14. Lindner,M.S., Strauch,B., Schulze,J.M., Tausch,S.H., Dabrowski,P.W., Nitsche,A. and Renard,B.Y. (2017) HiLive: real-time mapping of illumina reads while sequencing. *Bioinformatics*, **33**, 917–319.
15. Tausch,S.H., Strauch,B., Andrusch,A., Loka,T.P., Lindner,M.S., Nitsche,A. and Renard,B.Y. (2018) LiveKraken–real-time metagenomic classification of illumina data. *Bioinformatics*, **34**, 3750–3752.
16. Loka,T.P., Tausch,S.H. and Renard,B.Y. (2019) Reliable variant calling during runtime of illumina sequencing. *Sci. Rep.*, **9**, 16502.
17. Charalampous,T., Kay,G.L., Richardson,H., Aydin,A., Baldan,R., Jeanes,C., Rae,D., Grundy,S., Turner,D.J., Wain,J. *et al.* (2019) Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nat. Biotechnol.*, **37**, 783–792.
18. Gu,W., Deng,X., Lee,M., Sucu,Y.D., Arevalo,S., Stryke,D., Federman,S., Gopez,A., Reyes,K., Zorn,K. *et al.* (2021) Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids. *Nat. Med.*, **27**, 115–124.
19. Edge,P. and Bansal,V. (2019) Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nat. Commun.*, **10**, 4660.
20. Luo,R., Wong,C.-L., Wong,Y.-S., Tang,C.-I., Liu,C.-M., Leung,C.-M. and Lam,T.-W. (2020) Exploring the limit of using a deep neural network on pileup data for germline variant calling. *Nat. Mach. Intell.*, **2**, 220–227.
21. Shafin,K., Pesout,T., Chang,P.C., Nattestad,M., Kolesnikov,A., Goel,S., Baid,G., Kolmogorov,M., Eizenga,J.M., Miga,K.H. *et al.* (2021) Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat. Methods*, **18**, 1322–1332.
22. Magi,A., Giusti,B. and Tattini,L. (2017) Characterization of MinION nanopore data for resequencing analyses. *Brief Bioinform*, **18**, 940–953.
23. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, **25**, 1754–1760.
24. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and Genome Project Data Processing, S. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
25. Garrison,E. and Marth.,G. (2012) Haplotype-based variant detection from short-read sequencing. bioRxiv doi: https://doi.org/10.48550/arXiv.1207.3907, 20 July 2012, preprint: not peer reviewed.
26. Wang,K., Li,M. and Hakonarson,H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.
27. Feng,Z., Clemente,J.C., Wong,B. and Schadt,E.E. (2021) Detecting and phasing minor single-nucleotide variants from long-read sequencing data. *Nat. Commun.*, **12**, 3032.
28. Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.
29. Tarasov,A., Vilella,A.J., Cuppen,E., Nijman,I.J. and Prins,P. (2015) Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, **31**, 2032–2034.
30. Pollard,M.O., Gurdasani,D., Mentzer,A.J., Porter,T. and Sandhu,M.S. (2018) Long reads: their purpose and place. *Hum. Mol. Genet.*, **27**, R234–R241.
31. Bentley,D.R., Balasubramanian,S., Swerdlow,H.P., Smith,G.P., Milton,J., Brown,C.G., Hall,K.P., Evers,D.J., Barnes,C.L., Bignell,H.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.
32. Goodwin,S., McPherson,J.D. and McCombie,W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, **17**, 333–351.
33. Jeon,S.A., Park,J.L., Kim,J.H., Kim,J.H., Kim,Y.S., Kim,J.C. and Kim,S.Y. (2019) Comparison of the MGISEQ-2000 and illumina hiseq 4000 sequencing platforms for RNA sequencing. *Genomics Inform.*, **17**, e32.
34. Fehlmann,T., Reinheimer,S., Geng,C., Su,X., Drmanac,S., Alexeev,A., Zhang,C., Backes,C., Ludwig,N., Hart,M. *et al.* (2016) cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. *Clin Epigenetics*, **8**, 123.
35. Rothberg,J.M., Hinz,W., Rearick,T.M., Schultz,J., Mileski,W., Davey,M., Leamon,J.H., Johnson,K., Milgrew,M.J., Edwards,M.

et al. (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, **475**, 348–352.

36. Heather,J.M. and Chain,B. (2016) The sequence of sequencers: the history of sequencing DNA. *Genomics*, **107**, 1–8.
37. Amarasinghe,S.L., Su,S., Dong,X., Zappia,L., Ritchie,M.E. and Gouil,Q. (2020) Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.*, **21**, 30.
38. Thirunavukarasu,D., Cheng,L.Y., Song,P., Chen,S.X., Borad,M.J., Kwong,L., James,P., Turner,D.J. and Zhang,D.Y. (2021) Oncogene concatenated enriched amplicon nanopore sequencing for rapid, accurate, and affordable somatic mutation detection. *Genome Biol.*, **22**, 227.
39. Martignano,F., Munagala,U., Crucitta,S., Mingrino,A., Semeraro,R., Del Re,M., Petrini,I., Magi,A. and Conticello,S.G. (2021) Nanopore sequencing from liquid biopsy: analysis of copy number variations from cell-free DNA of lung cancer patients. *Mol. Cancer*, **20**, 32.
40. Gorzynski,J.E., Goenka,S.D., Shafin,K., Jensen,T.D., Fisk,D.G., Grove,M.E., Spiteri,E., Pesout,T., Monlong,J., Baid,G. *et al.* (2022) Ultrarapid nanopore genome sequencing in a critical care setting. *N. Engl. J. Med.*, **386**, 700–702.
41. Couto,N., Schuele,L., Raangs,E.C., Machado,M.P., Mendes,C.I., Jesus,T.F., Chlebowicz,M., Rosema,S., Ramirez,M., Carrico,J.A. *et al.* (2019) Author correction: critical steps in clinical shotgun metagenomics for the concomitant detection and typing of microbial pathogens. *Sci. Rep.*, **9**, 6406.
42. Miller,S., Naccache,S.N., Samayoa,E., Messacar,K., Arevalo,S., Federman,S., Stryke,D., Pham,E., Fung,B., Bolosky,W.J. *et al.* (2019) Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. *Genome Res.*, **29**, 831–842.