Research Article

Karakurt H. U., et al., Evaluation of Differences of Fast and High Accuracy Base Calling Models of Guppy on Variant Calling Using Low Coverage WGS Data. International Journal of Life Sciences and Biotechnology, 2023. 6(3): p. 276-287. DOI: 10.38001/ijlsb.1308355

Evaluation of Differences of Fast and High Accuracy Base Calling Models of Guppy on Variant Calling Using Low Coverage WGS Data

Hamza Umut Karakurt[1,2*], Hasan Ali Pekcan[1], Ayşe Kahraman[1], Muntadher Jihad[1], Bilçağ Akgün[3], Cüneyt Öksüz[4], Bahadır Onay[5]

ABSTRACT

Long-read sequencing technologies such as Oxford Nanopore Technologies (ONT) enabled researchers to sequence long reads fast and cost-effectively. ONT sequencing uses nanopores integrated into semiconductor surfaces and sequences the genomic materials using changes in current across the surface as each nucleotide passes through the nanopore. The default output of ONT sequencers is in FAST5 format. The first and one of the most important steps of ONT data analysis is the conversion of FAST5 files to FASTQ files using "base caller" tools. Generally, base caller tools pre-trained deep learning models to transform electrical signals into reads. Guppy, the most commonly used base caller, uses 2 main model types, fast and high accuracy. Since the computation duration is significantly different between these two models, the effect of models on the variant calling process has not been fully understood. This study aims to evaluate the effect of different models on performance on variant calling. In this study, 15 low-coverage long-read sequencing results coming from different flow cells of NA12878 (gold standard data) were used to compare the variant calling results of Guppy. Obtained results indicated that the number of output FASTQ files, read counts and average read lengths between fast and high accuracy models are not statistically significant while pass/fail ratios of the base called datasets are significantly higher in high accuracy models. Results also indicated that the difference in pass/fail ratios arises in a significant difference in the number of called Single Nucleotide Polymorphisms (SNPs), insertions and deletions (InDels). Interestingly the true positive rates of SNPs are not significantly different. These results show that using fast models for SNP calling does not affect the true positive rates statistically. The primary observation in this study, using fast models does not decrease the true positive rate but decreases the called variants that arise due to altered pass/fail ratios. Also, it is not advised to use fast models for InDel calling while both the number of InDels and true positive rates are significantly lower in fast models. This study, to the best of our knowledge, is the first study that evaluates the effect of different base calling models of Guppy, one of the most common and ONT-supported base callers, on variant calling.

ARTICLE HISTORY
Received
01 June 2023
Accepted
09 July 2023

KEYWORDS
Oxford nanopore sequencing,
variant calling,
long-read sequencing,
base-calling

[1] Idea Technology Solutions R&D Center, Maslak, Istanbul, Turkiye
[2] Department of Bioengineering, Faculty of Engineering, Gebze Technical University, Kocaeli, Turkiye
[3] John P. Hussman Institute for Human Genetics, University of Miami Miller School of Medicine, Miami, FL, USA
[4] Gene2Info A.Ş., Istanbul, Turkiye
[5] Idea Technology Solutions LLC, MA, USA
*Correspondence: hamza.karakurt@ideateknoloji.com.tr

Introduction

Since its development, long-read, single-molecule DNA sequencing Technologies emerged as powerful players in genomics and have proven their ability to resolve some of the most challenging regions of the human genome [1]. Oxford Nanopore Technologies (ONT), especially, provided fast and portable solutions for sequencing. The use of the Oxford Nanopore Sequencing platform has been increasing exponentially for variant calling due to its mobility, easy-to-use structure, accuracy and price. ONT uses electrical signal changes of nucleotides passing through nanopores that are integrated into a semi-conductive surface. Signals are stored as FAST5 files and can be converted to FASTQ files with a procedure called base calling [2]. Guppy, the most common and ONT-approved variant caller which is also used by the MinKNOW operating software of ONT, uses a Hidden Markov Model to generate FASTQ files from FAST5 files [3]. Guppy involves two different built-in models, High Accuracy and Fast models. The fast models are optimized for speed and are designed for applications where quick turn-around times are important, such as in real-time sequencing analysis or rapid diagnostic testing. The high-accuracy models use a more advanced algorithm that provides higher accuracy base-calling but at the expense of longer processing times [4]. These models differ in computation time and computing power requirements. Even though fast models provide significantly faster results, especially when the need for fast result generation, there is not any study that shows the direct effect of different models on variant calling. These kinds of critical cases require clinicians and experimental biologists to know which information on sequencing material they sacrifice to obtain faster results. As a consequence of that, researchers need a guide to have information about the differences between these models. Here, a benchmark study is provided that uses 15 low-coverage human sequencing data sets to provide insight into the model effect on variant calling. In this study, we aimed to investigate the effect of different base-calling models of Guppy on Single Nucleotide Polymorphism and Insertion/Deletion calling by comparing different parameters statistically.

Material and Methods

The study focuses on the “High Accuracy” and “Fast” models of Guppy base caller. Using 15 low-coverage long-read sequencing files (Table 1) from NA12878 Gold Standard Data

[5]; pass/fail ratio, FASTQ quality, true variant discovery and variant quality metrics are compared. FAST5 files were downloaded from Nanopore WGS Consortium [5] using Amazon Web Services (AWS) CLI terminal software [6]. Data sets from different sizes and different laboratories were selected to have a uniform distribution. Data sets are downloaded using AWS S3 Client in Ubuntu 20.04.

Bioinformatic and computational analyses

Base-calling is applied to FAST5 files using Guppy with Fast and High Accuracy built-in models (dna_r9.4.1_450bps_fast and dna_r9.4.1_450bps_hac are used as config files). The base calling process produces 2 different outputs, Pass and Fail. We used FASTQ files in the Pass folder for further steps and then calculated the Pass/Fail Ratios (Supplementary File 2) for each run using R. 16 CPUs are used for base calling processes. In the second step, the number of reads in merged FASTQ files is calculated (Supplementary File 2 / Supplementary Fig 3).

Table 1 Flow Cell Data Used in Guppy Model Analysis

Flowcell ID	Reads	Bases
FAB39075	477495	3014355946
FAB42395	38335	200553219
FAB42260	269507	1583530766
FAB41174	11714	739850920
FAB42476	435934	2655496773
FAB42706	431694	2434471643
FAB43577	427215	2776702333
FAB46664	491945	2335386447
FAB39088	668016	3929822468
FAB39043	442132	2574202451
FAB42316	573736	4047383848
FAB42473	646945	3794243146
FAB42810	322286	2433213020
FAB44989	558539	3962530064
FAB45332	531764	3267600.534

FASTQ files merged using cat command and aligned to the human genome (hg19) using minimap2 [7]. Output SAM files are sorted and indexed using Samtools [8]. Variants are called using Clair3 [9] with default parameters. Called SNPs and InDels are split to separate VCF files using VCFTools [10] (Supplementary File 1). VCF files obtained from Clair3 and filtered using VCFTools are processed using an in-house R function. NA12878 (HG001) truth VCF file [11] is used to compare true and false variants using

Chromosome, Position, Reference Base and Alternative Base (Table 2). For the analysis of InDels, the same procedure as the analysis of SNPs was applied to VCF files (Table 3). For the analysis of false negative rate differences (Supplementary Excel File) between models, the HG001 Truth VCF file is filtered using a BED file, constructed using the regions of VCF files for each dataset. Common variants of each dataset (Table 4 / Supplementary Fig 4-5) are identified and true positive rates for common, "only in fast model output" and "only in high accuracy model output" are calculated.

Results and Discussion

Pass and Fail Ratios

Results indicate that the number of FASTQ files is not significantly changed while Pass/Fail ratios are significantly changed between models (Fig 1 / Supplementary Fig 1). The average of Fold Changes of Pass/Fail Ratios is 0.918 while the P-Value is 0.011 (Effect size is 0.38). The number of generated FASTQ files is not significantly different with a p-value of 0.445 (Effect size is 0.0035). The number of Pass and Fail FASTQ files are also not significantly changed with p-values of 0.078 and 0.077, respectively (With effect sizes of 0.021 and 0.033 respectively).

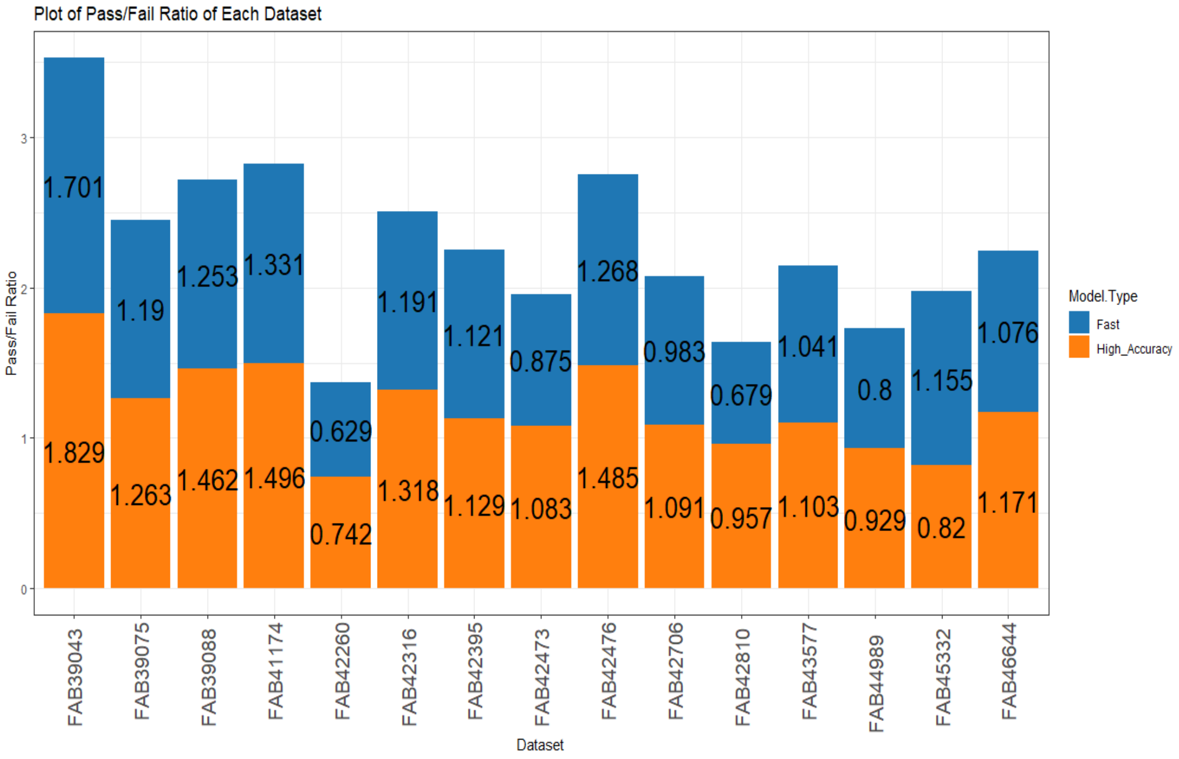

Fig 1 Pass/Fail Ratios of Each Dataset

Comparison of Average Read Lengths of Base Called FASTQ Files

Boxplot of read lengths indicates the average read length between fast and high accuracy models have similar distributions (Supplementary File 2 /Supplementary Fig 2). Paired t-test was applied to the average read lengths and the difference is not significant between models with a p-value of 0.68 (with an effect size of 0.012).

Read Counts in FASTQ Files

The average Fold Change (FC) of read counts is 0.916 while the P-value is 0.24 (effect size is 0.143). Here, it is observed that different models do not have different read counts in FASTQ Files.

Comparison of Single Nucleotide Polymorphisms

The number of called SNPs (Fig 2) is significantly different with 0.475 as Fold Change and 0 as P-value (effect size is 0.44). This result indicated that the number of called SNPs is significantly different between models.

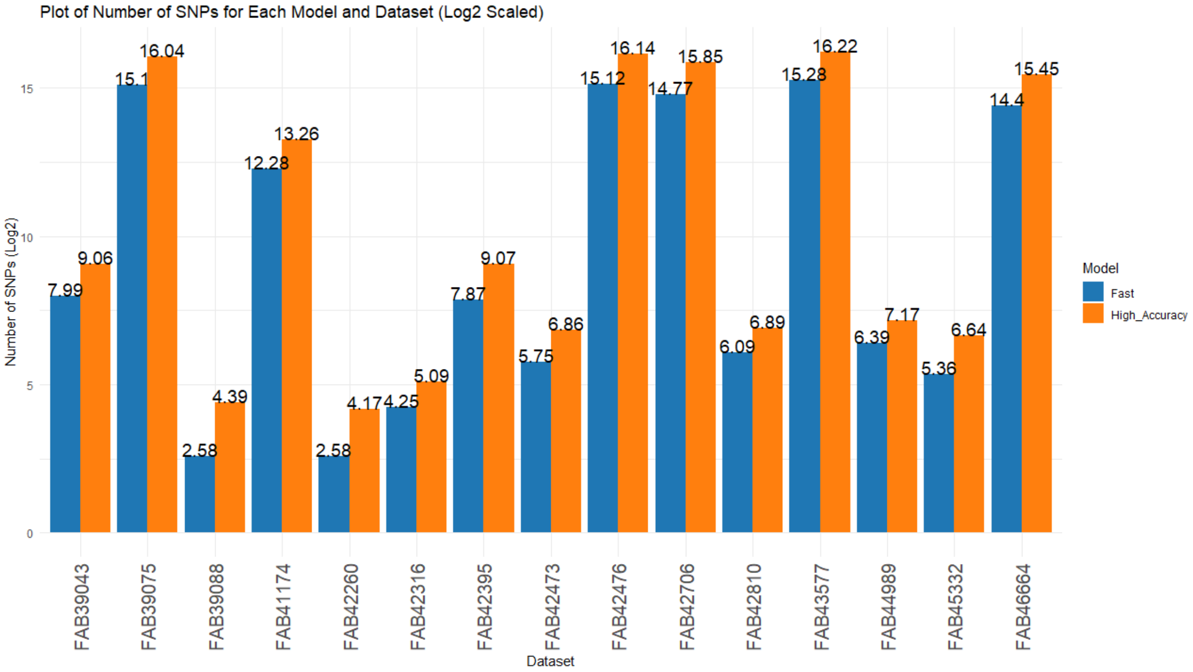

Fig 2 Number of Single Nucleotide Polymorphisms

The same test was applied to true positive rates (Fig 3 / Supplementary Excel File) to test the significance. Even though the number of variants is different, true positive SNP rates are not statistically different between models with 0.97 as Fold Change and 0.22 as P-value (effect size is 0.26).

Due to the high number of variants in the truth VCF file and the dataset's low coverage, the number of false negatives is very high. Even though, different models can be compared since the same methods are applied. Test results indicated that false negative rates are significantly changed with a P-value of 0.013 (effect size = 0.59).

Table 2 Comparison of Variants Obtained with Fast and High Accuracy Models with Truth VCF File (TP: True Positive, FP: False Positive)

Dataset	Model	Number of Variants	Number of TP Variants	Number of FP Variants	TP Ratios
FAB39043	Fast	254	226	28	88.97638
FAB39043	High Accuracy	533	470	63	88.18011
FAB39075	Fast	35192	31144	4048	88.49739
FAB39075	High Accuracy	67430	60538	6892	89.77903
FAB39088	Fast	6	6	0	100
FAB39088	High Accuracy	21	20	1	95.2381
FAB41174	Fast	4989	4210	779	84.38565
FAB41174	High Accuracy	9813	8360	1453	85.19311
FAB42260	Fast	6	3	3	50
FAB42260	High Accuracy	18	12	6	66.66667
FAB42316	Fast	19	17	2	89.47368
FAB42316	High Accuracy	34	34	0	100
FAB42395	Fast	234	205	29	87.60684
FAB42395	High Accuracy	536	466	70	86.9403
FAB42473	Fast	54	43	11	79.62963
FAB42473	High Accuracy	116	103	13	88.7931
FAB42476	Fast	35568	30017	5551	84.39327
FAB42476	High Accuracy	71996	62081	9915	86.2284
FAB42706	Fast	28010	23616	4394	84.31275
FAB42706	High Accuracy	59103	50908	8195	86.13438
FAB42810	Fast	68	62	6	91.17647
FAB42810	High_Accuracy	119	100	19	84.03361
FAB43577	Fast	39673	33783	5890	85.15363
FAB43577	High_Accuracy	76573	65821	10752	85.9585
FAB44989	Fast	84	75	9	89.28571
FAB44989	High_Accuracy	144	122	22	84.72222
FAB45332	Fast	41	33	8	80.4878
FAB45332	High_Accuracy	100	89	11	89
FAB46664	Fast	21621	17779	3842	82.23024
FAB46664	High_Accuracy	44794	37614	7180	83.97107

Comparison of Insertion and Deletions

The number of InDels and true positive rates are significantly different (Figure 4-5) in the context of insertion and deletion calling with p-values as 0, 0.046, respectively (effect sizes are 0.48, 1.16). This result indicates that different models highly affect the results in the context of the number of insertions and deletions more than SNPs and using fast models for InDel calling has more risk to lose variant information.

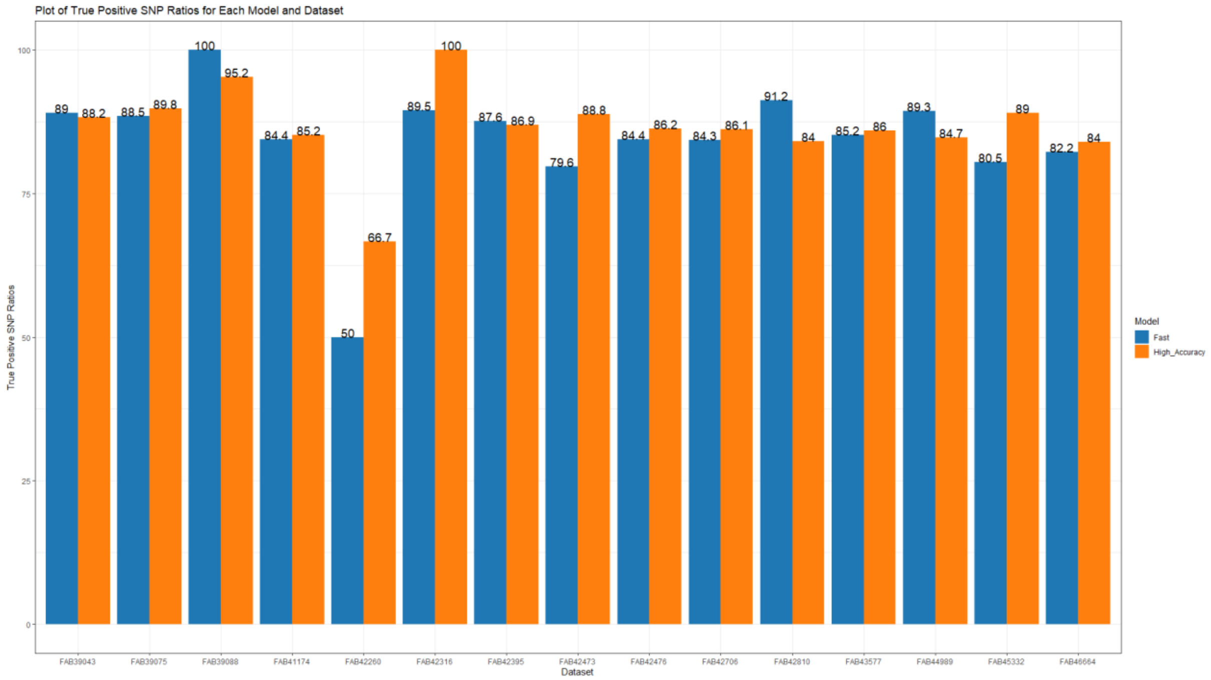

Fig 3 True Positive Ratios of Single Nucleotide Polymorphisms

Table 3 Comparison Results of Called InDels with Truth VCF File (TP: True Positive, FP: False Positive)

Dataset	Model	Number of Variants	Number of TP Variants	Number of FP Variants	TP Ratios
FAB39075	Fast	269	187	82	69.51673
FAB39075	High_Accuracy	493	408	85	82.75862
FAB41174	Fast	17	13	4	76.47059
FAB41174	High_Accuracy	60	46	14	76.66667
FAB42476	Fast	204	154	50	75.4902
FAB42476	High_Accuracy	489	402	87	82.20859
FAB42706	Fast	133	111	22	83.45865
FAB42706	High_Accuracy	361	297	64	82.27147
FAB43577	Fast	248	160	88	64.51613
FAB43577	High_Accuracy	566	406	160	71.73145
FAB46664	Fast	96	69	27	71.875
FAB46664	High_Accuracy	220	180	40	81.81818

Analysis of Common Variants Between Models

Among 21 comparisons (15 SNP and 6 InDel comparisons), it is observed that variants only called in high-accuracy model results have higher true positive rates. True positive rates of variants are only called with fast models and only called with high accuracy models tested using paired t-test and the difference is significant with a p-value of 0.0002 (effect size is 0.98).

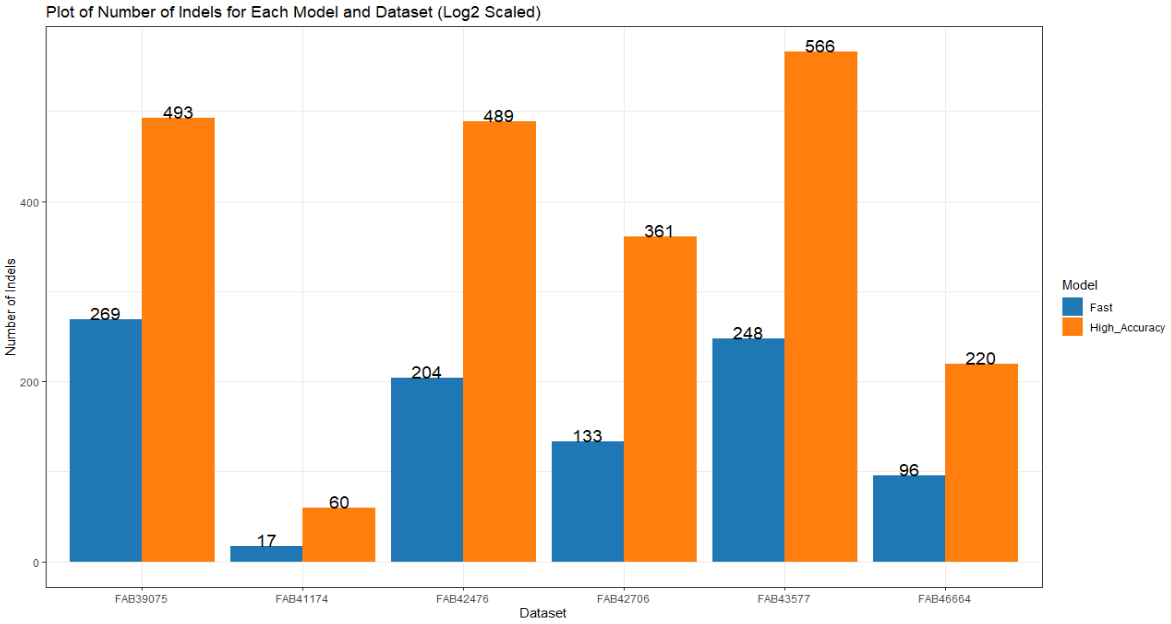

Fig 4 Number of Insertions and Deletions

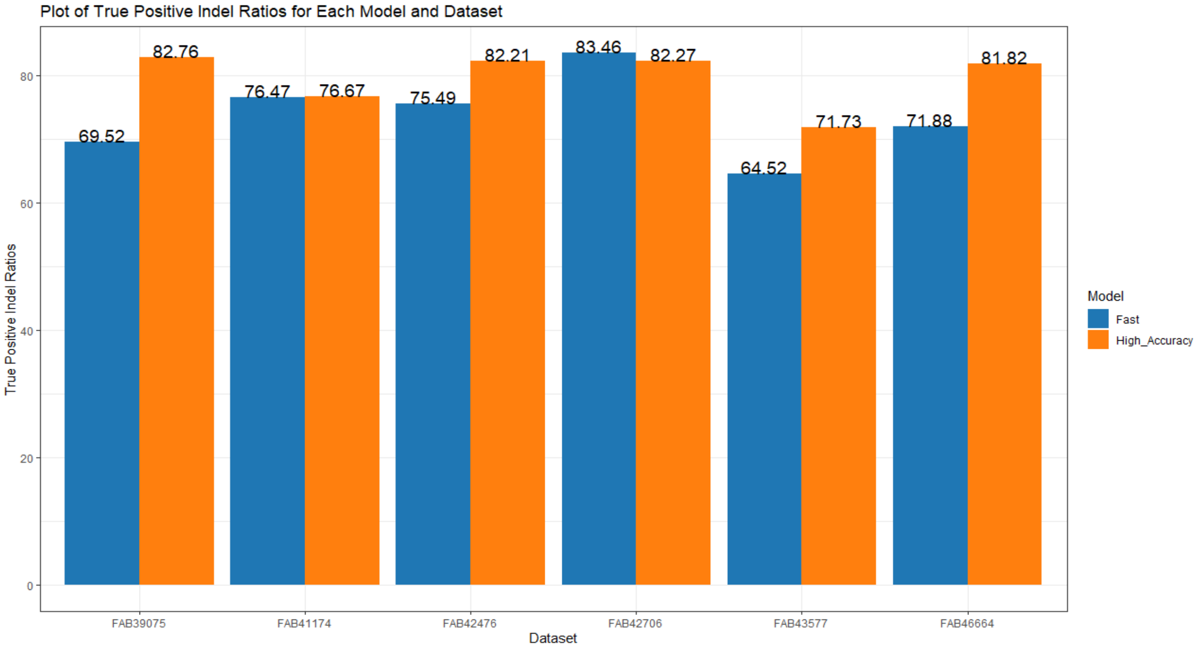

Fig 5 True Positive Ratios of Deletions and Insertions

Table 4 Common Variants Between Models (TP: True Positive, FP: False Positive)

Dataset	Number of Common Variants	TP Rate of Common Variants	Only in Fast Model	Only in High Accuracy Model	TP Rate of Only in Fast Model	TP Rate of Only in High Accuracy Model	Variant Type
FAB39043	150	93.33333	104	383	82.69	86.16	SNP
FAB39075	23725	92.29083	11467	43705	80.65	88.41	SNP
FAB39088	4	100	2	17	100	94.12	SNP
FAB41174	3343	89.41071	1646	6470	74.18	83.01	SNP
FAB42260	3	66.66667	3	15	33.33	66.66	SNP
FAB42316	14	100	5	20	60	100	SNP
FAB42395	155	90.96774	79	381	81.01	85.30	SNP
FAB42473	38	89.47368	16	78	56.25	88.46	SNP
FAB42476	23837	89.7764	11731	48159	73.45	84.47	SNP
FAB42706	18702	90.15079	9308	40401	72.58	84.27	SNP
FAB42810	37	94.59459	31	82	87.1	79.27	SNP
FAB43577	25535	90.75387	14138	51038	75.03	83.56	SNP
FAB44989	41	92.68293	43	103	86.06	81.55	SNP
FAB45332	26	92.30769	15	74	60	87.84	SNP
FAB46664	13816	88.60741	7805	30978	70.94	81.9	SNP
FAB39075	107	89.71963	162	386	56.17	80.83	InDel
FAB41174	14	78.57143	3	46	66.66	76.09	InDel
FAB42476	93	83.87097	111	396	68.47	81.82	InDel
FAB42706	57	96.49123	76	304	73.68	79.60	InDel
FAB43577	98	74.4898	150	468	58	71.15	InDel
FAB46664	40	85	56	180	62.5	81.11	InDel

Analysis of Qualities Common Variants Between Models

The analysis of the qualities of common variants in fast and high-accuracy models indicated that the qualities are not significantly different. (Table 5) and results indicated that the qualities of variants are not significantly different between models.

Conclusion

In this study, 15 different low-coverage data sets from different sequencing experiments (each of them coming from a single flow cell) are used to compare the effects of different built-in base calling models on variant calling. Guppy, the tool that has the best overall performance in benchmark tests [16] and is supported by Oxford Nanopore, is a widely used base caller and to the best of our knowledge, there are not any comparison studies on different base calling models of Guppy. To the best of our knowledge, this study is the first one that analyses the effects of models on variant calling.

This study indicated that the chosen model does not affect true positive and false negative SNP rates significantly while the number of Single Nucleotide Polymorphisms (SNPs), number of Insertions and Deletions (InDels), and true positive and false negative InDel rates are significantly lower in fast models. Also, results indicated that these alterations occur due to significant pass/fail ratio differences, Total read counts and average read lengths of the base called FASTQ files do not significantly change between models.

Table 5 Statistical Analysis Results of Qualities of Common Variants Between Models

Dataset	P-Value	Variant Class
FAB39043	0.845	SNP
FAB39075	0.962	SNP
FAB39088	0.474	SNP
FAB41174	0.95	SNP
FAB42260	0.752	SNP
FAB42316	0.748	SNP
FAB42395	0.72	SNP
FAB42473	0.559	SNP
FAB42476	0.999	SNP
FAB42706	0.169	SNP
FAB42810	0.662	SNP
FAB43577	0.215	SNP
FAB44989	0.818	SNP
FAB45332	0.27	SNP
FAB46664	0.599	SNP
FAB39043	0.389	INDEL
FAB39075	0.555	INDEL
FAB41174	0.065	INDEL
FAB42395	0.232	INDEL
FAB42476	0.288	INDEL
FAB42706	0.832	INDEL
FAB42810	0.935	INDEL
FAB43577	0.3	INDEL
FAB46664	0.512	INDEL

Analyses indicated that High Accuracy and Fast models cause the calling of different numbers of variants but in the context of true positive variants, the difference is not significant for SNPs while it is significant for insertions and deletions. Since there is not a significant difference between read counts and average read lengths, Pass/Fail ratios may be the main reason for this difference. For both models, the differences between false negative SNPs, true positive SNPs and qualities of common variants between models are not significant. It can be concluded, for SNP calling, the usage of fast models in case of lack of computational power and time limitation, does not create a statistical disadvantage. This study can guide researchers about the applications and differences of built-in models of Guppy. As mentioned, Guppy comes with MinKNOW pre-installed and is the most common choice for clinical and scientific research centres without bioinformatics expertise.

This study has limitations on the number of tested samples. Due to the low number of samples, statistical test results may not be generalized but the properties of tested data are held as uniform. Even though the statistical analyses of the study lack generalizability, the differences are clear for the datasets. For further research, the same analyses can be planned and applied to multiple ONT-based Whole Genome Sequencing and Whole Exome Sequencing experiment results.

It should be also noted that the quality of variant calling is directly associated with experimental procedures and properties of genomic locations (high GC content, CpG Islands etc.). Due to this, it is possible to investigate the effects of models based on these parameters and an application procedure based on experimental steps or genomic locations can be developed.

Abbreviations

ONT: Oxford Nanopore Technologies, NGS: Next Generation Sequencing, SNP: Single Nucleotide Polymorphism, InDel: Insertions and Deletions, TP: True Positive, FP: False Positive, VCF: Variant Call Format, FC: Fold Change

Acknowledgements

The Guppy software downloaded by Bilçağ Akgün from Oxford Nanopore Sequencing Official Website with registered user account. We also thank Dr. Pınar Pir for advises on statistical testing.

Funding

This work is partially funded by TUBITAK as a part of 20AG005 Project.

Data Availability statement

The authors declare that all reported results are obtained computationally. No novel experimental data is reported in this publication. Bash and R commands are provided in https://github.com/ideateknoloji/guppy_models_comparison

Compliance with ethical standards

This study does not involve any experiment, all analyses were applied computationally. According to “TRDizin Etik İlkeleri” guide, ethics committee report is not required" olarak değişecek.

Conflict of interest

The authors declare no conflict of interest.

Ethical standards

The study is proper with ethical standards. This study does not involve any experiment. Ethical Standards does not apply.

Authors' contributions

In this work, whole study is managed by Hamza Umut Karakurt, analyses coded and applied by Hamza Umut Karakurt, Hasan Ali Pekcan, Ayşe Kahraman and Muntadher Jihad. All analyses applied under the medical supervision of Bilçağ Akgün and technical supervision of Cüneyt Öksür and Bahadır Onay. All authors contributed to the writing section of this article.

References

1. Logsdon, Glennis A., Mitchell R. Vollger, and Evan E. Eichler. Long-Read Human Genome Sequencing and Its Applications. Nature Reviews Genetics 21, no. 10 (June 5, 2020): 597–614. https://doi.org/10.1038/s41576-020-0236-x
2. Wang, Y., et al., Nanopore Sequencing Technology, Bioinformatics and Applications. Nature Biotechnology 39, no. 11 (November 1, 2021): 1348–65. https://doi.org/10.1038/s41587-021-01108-x.
3. Loman, N. J., and R. A. Quinlan. Poretools: A Toolkit for Analyzing Nanopore Sequence Data. Bioinformatics 30, no. 23 (August 20, 2014): 3399–3401. https://doi.org/10.1093/bioinformatics/btu555.
4. Peresini, P., et al., Nanopore Base Calling on the Edge. Bioinformatics 37, no. 24 (July 27, 2021): 4661–67. https://doi.org/10.1093/bioinformatics/btab528.
5. Jain, M, et al. Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads. Nature Biotechnology 36, no. 4 (January 29, 2018): 338–45. https://doi.org/10.1038/nbt.4060
6. aws/aws-cli: Universal Command Line Interface for Amazon Web Services. https://github.com/aws/aws-cli
7. Li, H., Minimap2: Pairwise Alignment for Nucleotide Sequences. Bioinformatics 34, no. 18 (May 10, 2018): 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
8. Heng, L., et al., The Sequence Alignment/Map Format and SAMtools. Bioinformatics 25, no. 16 (June 8, 2009): 2078–79. https://doi.org/10.1093/bioinformatics/btp352
9. Zheng, Z., et al., Symphonizing Pileup and Full-Alignment for Deep Learning-Based Long-Read Variant Calling. Nature Computational Science 2, no. 12 (December 19, 2022): 797–803. https://doi.org/10.1038/s43588-022-00387-x.
10. Danecek, P., et al., The Variant Call Format and VCFtools. Bioinformatics 27, no. 15 (June 7, 2011): 2156–58. https://doi.org/10.1093/bioinformatics/btr330.
11. Zook, J., et al., Extensive Sequencing of Seven Human Genomes to Characterize Benchmark Reference Materials. Scientific Data 3, no. 1 (June 7, 2016). https://doi.org/10.1038/sdata.2016.25.
12. Ginestet, C. E., ggplot2: Elegant Graphics for Data Analysis. Journal of the Royal Statistical Society 174, no. 1 (January 1, 2011): 245–46. https://doi.org/10.1111/j.1467-985x.2010.00676_9.x.
13. Nan, X., ggsci: Scientific Journal and Sci-Fi Themed Color Palettes for 'ggplot2.' 2023, https://github.com/nanxstats/ggsci.
14. Student. The Probable Error of a Mean. Biometrika 6, no. 1 (March 1, 1908): 1. https://doi.org/10.2307/2331554.
15. Cohen, J., Statistical Power Analysis for the Behavioral Sciences. Routledge EBooks, 2013. https://doi.org/10.4324/9780203771587.
16. Wick, R., et al., Holt. Performance of Neural Network Basecalling Tools for Oxford Nanopore Sequencing. Genome Biology 20, no. 1 (June 24, 2019). https://doi.org/10.1186/s13059-019-1727-y.