BMC Research Notes

RESEARCH NOTE Open Access

A universal indel filtering workflow for both long-read and short-read NGS data

Md. Shariful Islam Bhuyan[1] and M. Sohel Rahman[1*]

Abstract
Accurate detection of insertions and deletions (indels) is critical for applications in disease genomics, population genetics, and personalized healthcare. Despite advancements in sequencing technologies, indel detection remains challenging, particularly in difficult-to-map genomic regions. In this study, we present a universal machine learning-based filtering workflow that significantly improves indel detection accuracy for both long-read and short-read sequencing data, utilizing only publicly available genomic annotation datasets, eliminating the need for sequencing workflow-specific information, such as read depth. Our method (a gradient-boosting classifier powered by XGBoost) enhances precision by ~26% for long-read and ~24% for short-read data while maintaining high recall rates (~90%). We validate our approach using the Genome in a Bottle (GIAB) dataset and 62 indel call sets from the precisionFDA Truth Challenge V2, demonstrating its effectiveness in handling complex genomic regions and diverse sequencing workflows. Our tool is open-access and workflow-agnostic, making it a valuable resource for improving indel calling accuracy across various applications.

Keywords Long-read sequencing, Short-read sequencing, Indel detection, Gradient-boosting, Variant filtering

Introduction

Accurately identifying short genomic insertions or deletions (indels) remains challenging for conventional alignment-based variant calling methods [1]. The focus on indel detection has increased due to advancements in long-read sequencing technologies [2–6], such as Pacific Biosciences [7] and Oxford Nanopore Technologies [8]. Indels are variations involving the insertion or deletion of fewer than 50 base pairs [9], and although less common than SNPs, they account for 16%-25% of genetic variation [10, 11]. Indels are linked to numerous genetic disorders, including Mendelian diseases, fragile X syndrome [12], and cancers such as lung cancer and AML [13, 14]. Additionally, by altering promoter structures, indels can impact gene expression [15]. Indels also contribute to genetic diversity across populations [16], making their accurate detection crucial for clinical genomics and population genetics, especially in rare somatic mutations and de novo germline mutations.

Next-generation sequencing (NGS) technologies, particularly long-read sequencing, are advancing and becoming more accessible in clinical applications. The "Genome in a Bottle" (GIAB) benchmark dataset primarily uses long-read sequencing data [17], achieving high SNP detection accuracy (99th percentile) and improved indel detection in easily mappable regions. However, in difficult-to-map regions, indel detection accuracy remains limited, with common indel callers achieving around 60% accuracy across the entire genome [18, 19]. In the PrecisionFDA truth challenges V1 [20]and V2 [21], tools like DeepVariant exhibit precision rates above 99% for high-confidence regions (HCRs), but this drops

*Correspondence:
M. Sohel Rahman
sohel.kcl@gmail.com
[1]CSE Department, ECE Building, Bangladesh University of Engineering and Technology, West Palashi, Dhaka, Bangladesh

to 63% genome-wide [22]. Indel detection accuracy is influenced by sequencing workflows, aligners, callers, and genomic context, with different workflows producing unique error patterns [23].

Many tools (e.g., GATK [24], Samtools [25], SOAPIndel [26], Pindel [27], SplazerS [28], VarScan [29], FreeBayes [30], and more) have been developed for indel detection, but our approach differs by serving as a post-filter applied to any variant calling method. This filter reduces false positives by incorporating prior knowledge to improve precision without sacrificing sensitivity. Lately, deep learning methods have been integrated into genomic variant calling for both short (Illumina, Complete Genomics) and long (Oxford Nanopore, Pacific Biosciences) sequencing read data. Notable tools for short-read data include GATK4 [31], DeepVariant [23], NeuSomatic [32], and for long-read data, Clair [33], Clairvoyante [34], Sentieon DNAscope LongRead [35], NanoCaller [36] can be cited. Their performances have also been demonstrated in other bioinformatics classification problems, such as predicting piRNAs, sumoylation sites, and RNA 5-methyluridine (m5U) modifications [37–39]. Although deep learning models are capable to learn effective features from the training, all three above-mentioned papers emphasize the importance of workflow-independent feature extraction and selection, which is also a cornerstone of our method.

This paper demonstrates that enhancing indel caller precision for both short-read and long-read platforms is possible using a gradient-boosting-based filtering scheme trained on high-fidelity datasets and genomic features. This method maintains high sensitivity and is particularly useful for analyzing publicly available datasets like Haplotype Reference Consortium [40] and Exome Aggregation Consortium [41]. We validate our method using the GIAB dataset and 62 indel call sets from the precisionFDA Truth Challenge V2, showcasing its ability to refine datasets by eliminating false positives and enriching genuine indels.

In summary, our main contributions in this paper can be summarized as follows:

- Building upon our prior work [22], we demonstrate the efficacy of our machine learning-driven probabilistic intelligent filtering to effectively detect inaccurate short indel calls for both long-read and short-read sequencing platforms. This enhancement substantially boosts performance, especially when dealing with whole-genome call sets, even those encompassing non-high confidence regions (NHCR). Furthermore, we have made this tool open-access to researchers and practitioners for further research and development.
- We experimentally affirm that performance shows improvement as more training samples are added, albeit with a rapidly diminishing improvement rate. This provides us with an estimate of the optimal number of training samples needed to achieve a specific level of performance. Additionally, we demonstrate that this performance enhancement is independent of the particular samples used for training; it is sample-agnostic.
- Furthermore, we verify the effectiveness of our method, particularly on long-read sequencing technologies, by examining submissions in the precisionFDA Truth Challenge V2, which exclusively featured sequencing reads generated from Oxford Nanopore and Pacific Biosciences. We also validate whether our approach can handle complex genomic regions, resulting in an average precision improvement of around 25% when evaluating 62 indel call sets submitted in the precisionFDA Truth Challenge V2.

Results

Improvement in performance following variant filtering

We define pre-filter performance as the precision of the indel callsets prior to applying our filtering model, and post-filter performance as the precision after filtering. The effectiveness of our approach is examined by comparing the pre-filter and post-filter precision metrics. In Fig. 1, we present pre-filter precision (computed before filtering) and post-filter precision (computed after filtering) for eight distinct variant calling workflows. These workflows encompass combinations of sequenced base callers, read aligners, and variant callers, including three long-read sequencing platforms. We evaluate these workflows across three types of indels, as discussed below, and also outlined in Supplementary Table 2.

1. Homopolymers: Homopolymers are the type of indels where both the added or removed nucleotides and the neighboring nucleotide (either prefix or/and suffix) are composed exclusively of a specific nucleotide. For instance: [G → GG], [TT → T].
2. Tandem repeats: These are not homopolymers; instead, they involve multiple inserted or deleted nucleotides preceded or followed by a similar sequence. For example: [CG → CGCG].
3. Aperiodic indels: Every deletion or insertion that is not a tandem repeat or a homopolymer belongs to this group. For instance: [C → CAT].

For each of these eight workflows, we have access to four samples for which gold-standard data is available by GIAB, enabling us to utilize one for testing and three samples for training purposes. Additionally, we assessed

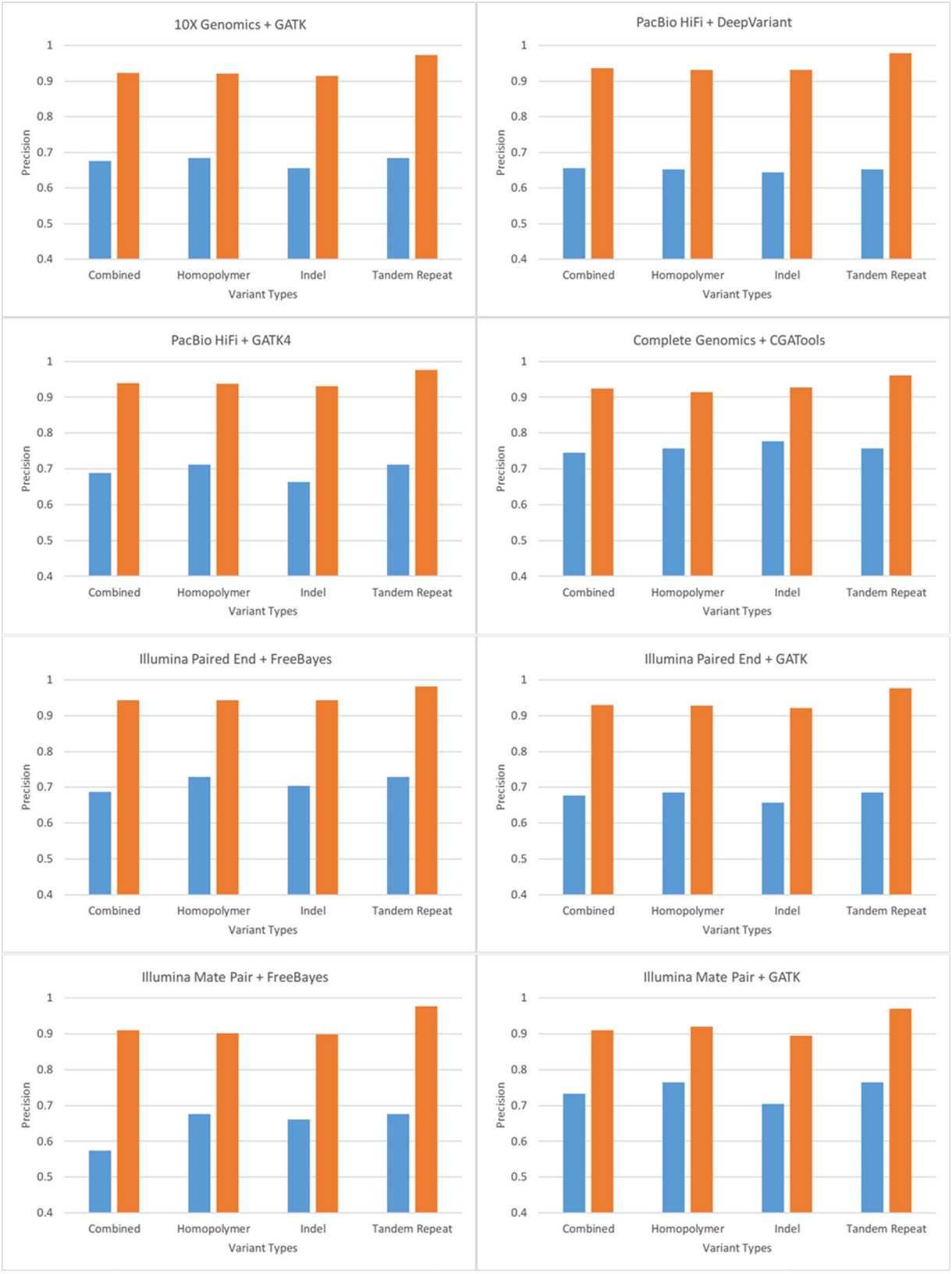

Fig. 1 Improvement of precision (Y-axis) across eight different workflows (names mentioned in the top part of each subfigure) and four variant categories (X-axis). Blue columns denote pre-filter precision, and the orange columns denote post-filter precision improvement

the performance across all indels collectively, denoted as 'Combined' in later discussions, and Fig. 1.

The Complete Genomics pipeline with CGATools achieved the highest pre-filter precision of 77.65% for aperiodic indels, followed by Illumina (mate pair reads) and GATK at around 70.45%, and Illumina (paired-end reads) with BWA-MEM and FreeBayes at 70.40%. PacBio (long-read) with DeepVariant had the lowest pre-filter precision at 64.33%. After applying our filter, all platforms saw performance improvements, with PacBio showing the largest gain of 28.73%, reaching 93.06% post-filter precision and 89.30% recall. Illumina (paired-end reads) using FreeBayes achieved the highest post-filter precision at 94.36% with 91.24% recall. Illumina (mate pair reads) with BWA-MEM and GATK had the lowest post-filter precision of 89.43% and a recall of 90.55%. Further details can be found in Supplementary Table 2. For combined indels, including homopolymers, precision increased, with some workflows, like Illumina and FreeBayes, showing a 33.49% improvement while maintaining 90.86% recall. Tandem repeat workflows consistently achieved over 90% post-filter precision with over 88% recall.

Performance vs. Training data size

We examine how varying the training sample size affects precision in the context of four distinct (comprising three long-read and one short-read) sequencing workflows. This investigation holds significance because, in the future, more samples with a gold standard callset may become accessible. Our study encompassed training sample sizes ranging from one to six.

In Fig. 2, the initial data point, i.e., zero in the X-axis, represents the pre-filter precision. It is evident from the figure that the initial training samples lead to the most significant improvement in precision. Gradually the improvement rate decelerates, and after the inclusion of the sixth sample, the improvement tends to level off. This pattern holds true across all four workflows. For a detailed breakdown of the precise enhancements attributable to each additional sample, please refer to the Supplementary Table 3.

In Fig. 3, we demonstrate that the precision of a particular workflow remains similar across different training samples for all four sequencing workflows (comprising three long-read and one short-read workflows). This clearly indicates that the performance is not affected significantly by any particular training sample. In this context, the training sample (European) and the test samples (Ashkenazi Jews and Han Chinese) refer to distinct individuals from different populations. When the test and training samples differ, there is still a possibility of the existence of a set of insertions and deletions (indels) that are present in both the test and training datasets. For samples from a similar population, this shared set of

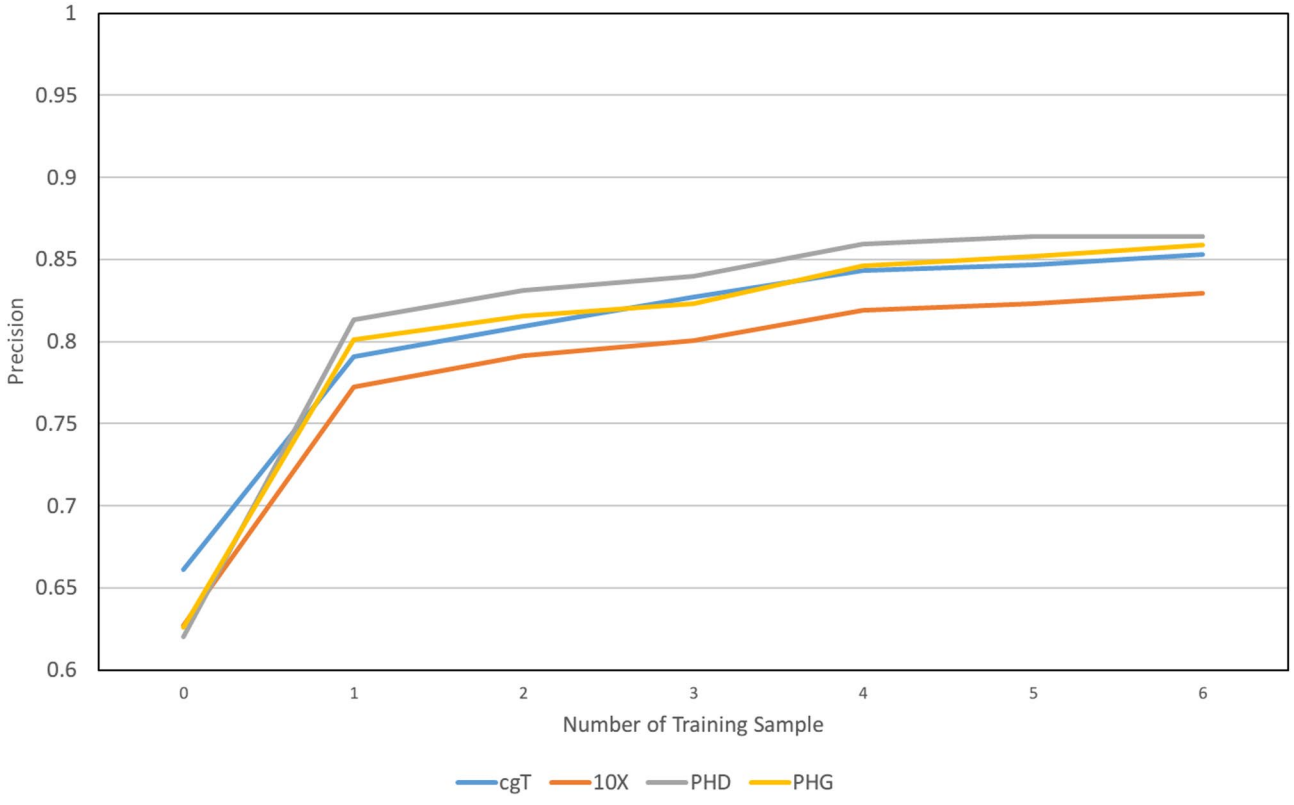

Fig. 2 Precision vs. Size of Training Samples for Different Workflows. True Positives are taken from True Workflow Calls

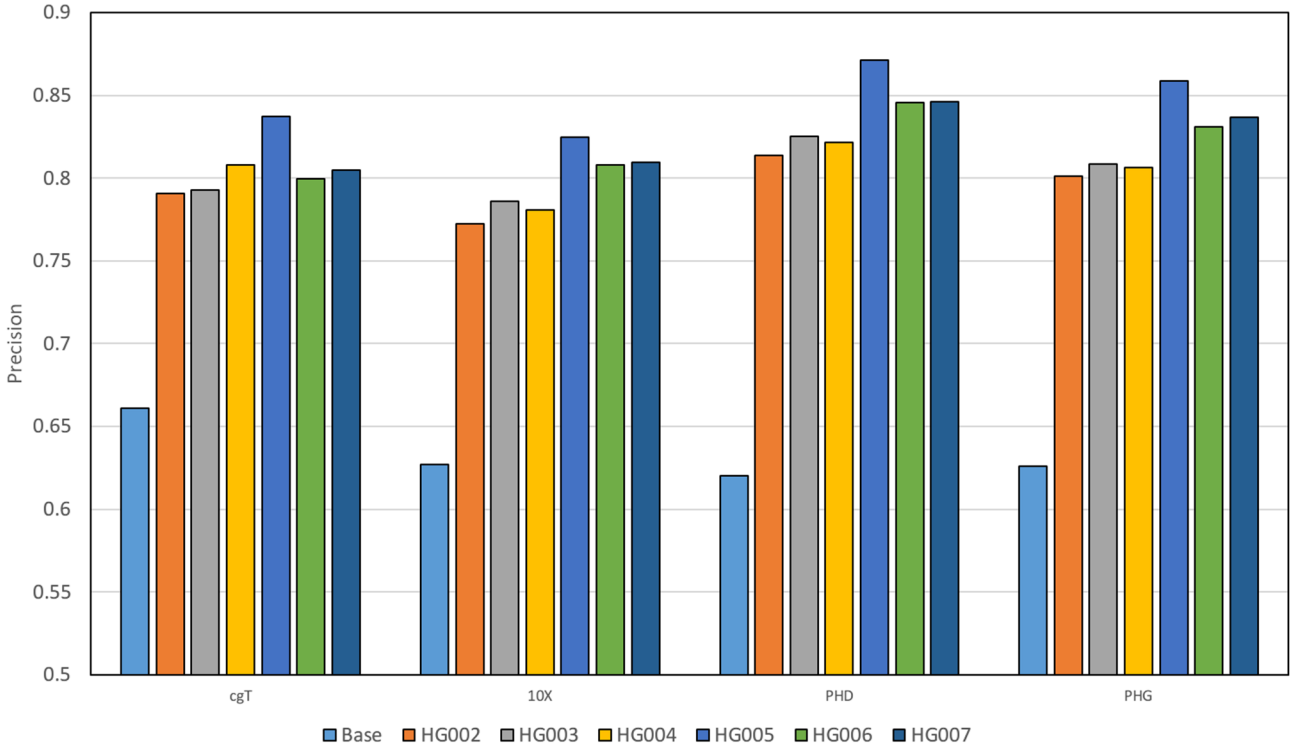

Fig. 3 Pre-filter (Base) Precision of the Training Sample (HG001) and Post-filter Precisions of Test Samples (HG002, HG003, HG004, HG005, HG006, HG007) for Four Different Workflows

indels will be bigger, which can positively influence the performance of the classifier. To mitigate this potential bias, we employ HG001 as the test sample, which has a different ethnicity compared to the other samples, helping to reduce the influence of shared indels on classifier performance.

If we compare with Fig. 1, the exclusion of the four Illumina combinations (iI2F, iI2G, iMF, iMG) from Figs. 2 and 3 is based on two key reasons. First, we wanted to focus on the long-read sequencing workflows, which are more error-prone. Second, we have only included workflows with all seven training samples used by GIAB to ensure a robust and unbiased evaluation of the method's performance.

Discrimination ability

While we opt for precision as our primary performance metric, the precision-recall graph from Fig. 4, clearly illustrates that our workflow has consistently yielded a robust AUC (Area Under the Curve) across all eight distinct sequencing workflows under consideration. These workflows encompass three long-read and five short-read methods, with the highest AUC recorded at 0.9642 for the Complete Genomics workflow. Given that AUC serves as a measure of the capability to distinguish true indels from false ones, the filtering scheme is demonstrated to work effectively as a classifier.

Our approach permits the adjustment of this threshold to align with specific precision and recall preferences, thereby catering to diverse practical scenarios. In our analysis, we set the threshold at 0.5, implying that if the likelihood of an indel being a true positive exceeds 0.5, we accept it. The highest F1 score is achieved at this threshold. For certain applications where we need to prioritize sensitivity over specificity, a lower threshold value can be chosen. In extreme cases, such as a threshold of zero, our filter would allow all variants to pass through. Conversely, clinical applications that demand heightened precision can opt for a higher threshold probability, albeit at the expense of sensitivity.

Performance over PrecisionFDA truth challenge V2 benchmark dataset

We not only extensively analyse the eight GIAB workflows, but also conduct a thorough examination of the combined indel call sets consisting of 62 submissions as part of the precisionFDA Truth Challenge V2. Our findings regarding the precision enhancement for this combined indel callset are detailed in Supplementary Table 4. In our analysis, we trained our models using only a single training sample for each workflow (since the gold standard data was provided for only one sample in the challenge). Among the 62 workflows assessed, an impressive 58 of them displayed a precision improvement of over

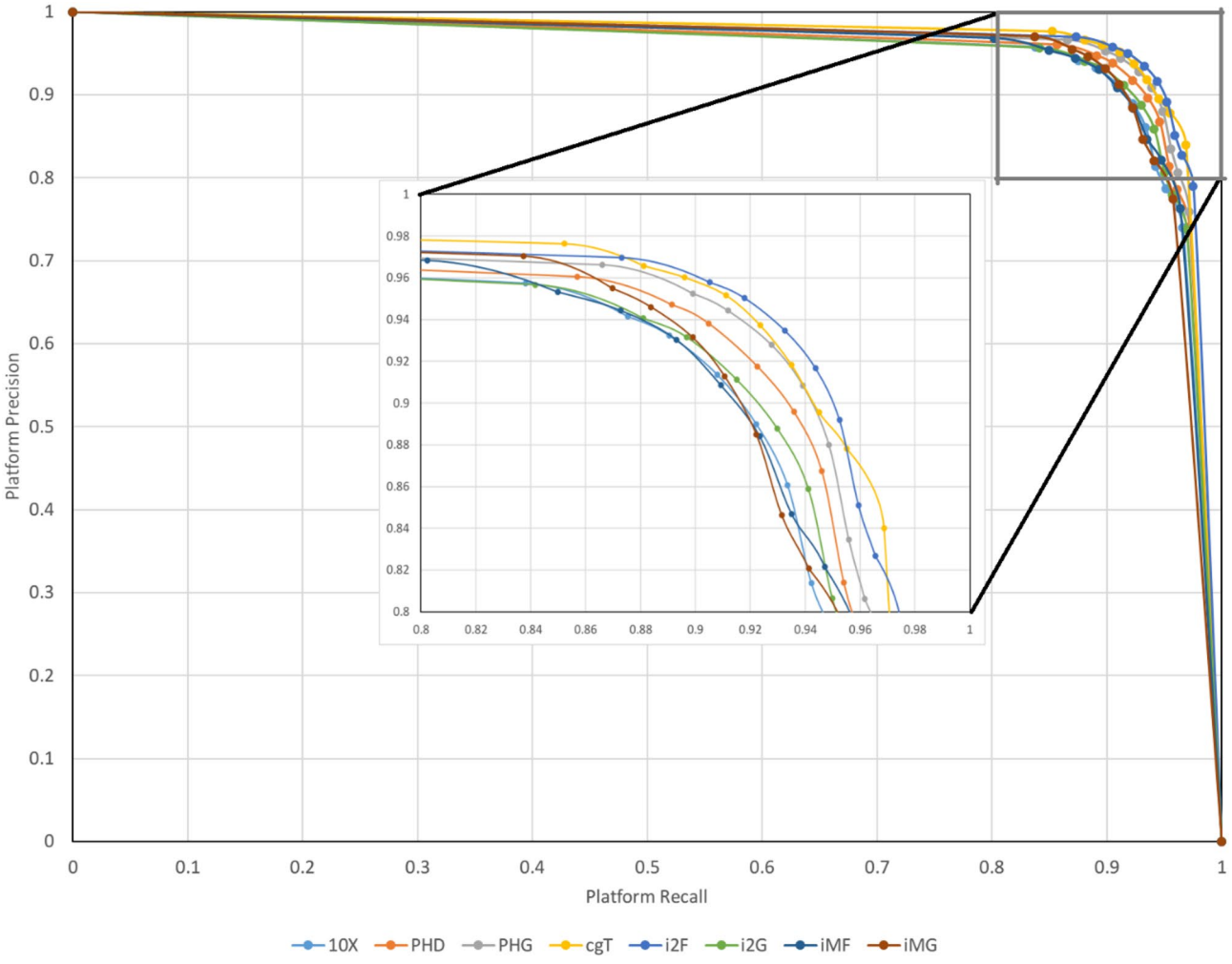

Fig. 4 A Precision-Recall Curve for Different Platforms. Upper-Right Corner of the Curves is Expanded to Show Small Differences among Platforms AUC (Area Under Curve)

20% (and all of them having positive improvement), with an average improvement exceeding 25%. Figure 5 visually represents the histogram of pre-filter and post-filter precisions, clearly illustrating a shift towards higher precision values and a reduction in variance.

Discussions

The accurate identification of indel mutations holds great importance in various applications within personal genomics. Thanks to the widespread adoption of long-read sequencing technologies, we now have extensive repositories of indel and structural variant data at our disposal. However, most of these datasets lack publicly available detailed sequence read information. Our sequencing-read-independent method can detect potentially true indels from these large datasets. We can compute required features from publicly available annotation databases.

Performance comparison between long vs. Short read sequencing technologies

In our study, we compare the performance of long-read sequencing technologies against short-read sequencing technologies. Regardless of the pre-filter precision, the specific sequencing technology used, or the workflow tools applied, our tool demonstrates a significant improvement in indel mutation calling accuracy. Notably, it enhances the performance of less robust workflows to a greater extent. Consequently, employing our tool as a filter ensures consistent and reliable performance for any workflow that leverages the indel callset of one or more GIAB samples. Supplementary Table 2 reveals that long-read technologies tend to benefit slightly more from our filtering workflow. For instance, PacBio, a long-read technology, combined with DeepVariant, exhibits the most substantial improvement, elevating the performance of this particular workflow from the lower ranks to the top. Additionally, as the number of training samples

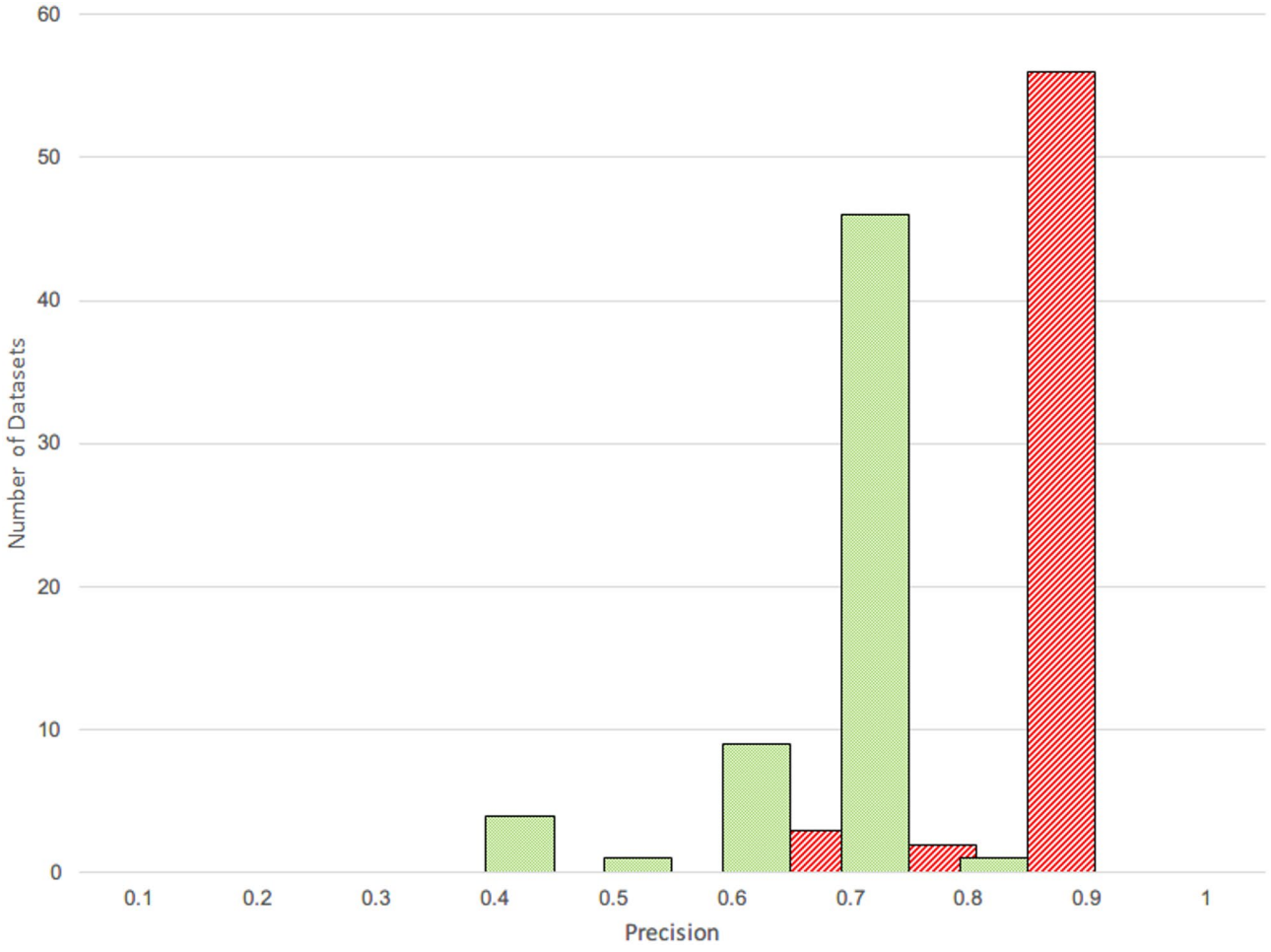

Fig. 5 Histogram of the Pre-filter Precisions (Green) and Post-filter Precisions (Red) of 62 Indel Call Sets Submitted in the precisionFDA Truth Challenge V2

increases, the performance of the same workflow experiences the most significant enhancement, which saturates eventually.

Different performance measure

Similar to our previous work [22] using short-read sequencing technologies, our primary focus is on eliminating false positive results from detected indels rather than attempting to identify every true positive. Consequently, we emphasize on precision as the key performance metric. However, it is worth noting that since our filtering process learns a logistic regression model with binary outcome, it does not maximize precision specifically. We also include other standard performance metrics, such as F1-score, the area under the precision-recall curve (AUPRC) and recall according to their applicability. Our filtering approach demonstrates good performance across all these metrics.

Our reported recall measures explicitly the proportion of true indels supported by our workflow (those identified by the workflow) that our filter correctly identifies as positives. It does not calculate the percentage of true indels backed by the GIAB gold standard callsets, part of which the workflow may not detect. This approach is logical because our filter does not have the opportunity to assess an indel that was not called by the workflow in the first place. Consequently, meaningfully determining the true recall of our filter is challenging, if not impossible, further underscoring our choice to prioritize precision as our primary metric.

Model interpretability

A major trend in genomic data analysis is the extensive use of deep learning architectures. We have already discussed the potential of DeepVariant, which has shown promising outcomes for variant calling. However, a significant drawback of deep neural network models is their inherent black-box nature, making it challenging to interpret their inner workings. Nevertheless, lots of research are currently addressing this interpretability issue [42]. In contrast to other domains, interpretability holds great importance in biology, as it forms the basis for a deeper comprehension of living systems.

Another drawback lies in the presence of adversarial examples—samples with minimal noise that the model misclassifies disproportionately. Additionally, the complexity is compounded by the need for substantial amount of training data (no clear and established method are there for augmenting sequencing data), high-performance GPUs, and intricate hyperparameter tuning. In our context, gradient tree boosting model directly provides the importance of features, facilitating model interpretation. Furthermore, annotation data used for learning does not change with an increase in samples since we do not employ read data or model-specific error patterns. Increasing the amount of training data should contribute to refining our knowledge about the genomic contexts associated with potentially false indel calls generated from a specific sequencing workflow.

Limitations

While our proposed filtering workflow demonstrates significant improvements in indel detection accuracy, several limitations and areas for future work should be acknowledged as discussed below.

1. **Sensitivity to Rare or Novel Variants**:

   Our method may struggle to distinguish rare or novel genetic variants from sequencing errors, as it does not utilize sequencing-read-specific information (e.g., read depth or alignment quality). In this context, our filtering approach shares similarities with a masking method, although there are some significant distinctions as follows. Firstly, our filter is not rigid in its determination of whether an insertion or deletion at a specific genomic location is true or false; this determination can change if additional data suggests otherwise. Secondly, the determination varies from one analysis workflow to another, depending on the error patterns specific to each workflow.

2. **Dependence on High-Quality Training Data**:

   As with any machine learning model, the performance of our model relies on the availability of high-quality and unbiased training data. While our method is sample-agnostic, its effectiveness may vary when applied to datasets with lower-quality or less comprehensive gold-standard callsets. Expanding the training data to include more diverse populations and sequencing platforms could further improve generalizability.

3. **Limited Scope to Indels**:

   Currently, our method focuses exclusively on indel detection and does not address other types of genomic variations, such as structural variants or complex rearrangements. Extending the framework to include these variants would broaden its applicability in genomic research and clinical diagnostics.

4. **Lack of Direct Comparisons**:

   To our knowledge, there are no directly comparable methods that serve as post-filters for indel detection. It is important to note that our tool is not a variant caller but can enhance the performance metrics of any caller when used with the appropriate dataset. While this highlights the novelty of our approach, it also limits our ability to benchmark against similar tools. Future research could focus on developing standardized evaluation frameworks for indel filtering methods.

Uniqueness of our universal method

The proposed universal method distinguishes itself from other methods by its workflow independence, reliance on publicly available genomic features, and ability to act as a post-filter for any variant calling workflow. These characteristics make it broadly applicable to diverse datasets and sequencing platforms, while also improving performance in challenging genomic regions. The method's interpretability and flexibility further enhance its utility for clinical genomics and population genetics.

Methods

To find more details about our methodology, please see Supplementary Methods. Our approach involves formalizing the detection of erroneous indel calls as a supervised classification problem. We conduct our training and testing using the "Genome-in-a-Bottle" (GIAB) dataset, specifically its latest release. For annotation purposes, we employ RepeatMasker [43] to identify different types of repetitive regions, such as tandem and interspersed repeats. To assess the level of conservation, we utilize GERP scores [44]. We.

compute all the DNA sequence contexts based on the hg19 reference genome [45]. However, for the precisionFDA Truth Challenge V2 dataset, we switch to the hg38 reference genome. In our quest to identify a robust set of genomic features independent of specific workflows, we experiment with various genome annotations and combinations thereof. It is worth noting that we can generate all these features using publicly available annotation and sequence databases. We intentionally avoid using population or workflow-specific features like genotype calls, allele frequency, read alignment, read quality

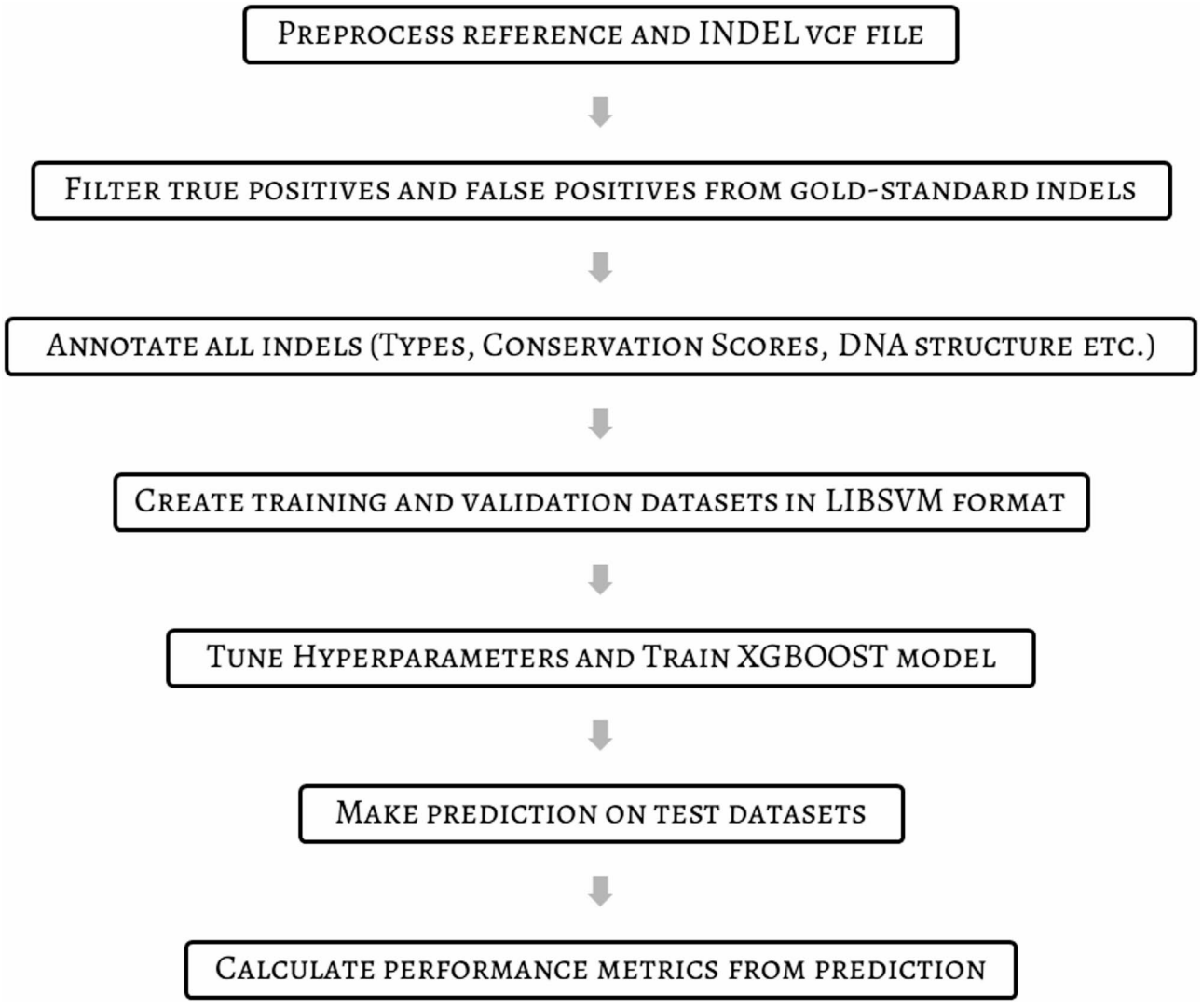

Fig. 6 Flow Chart of the Filtering Steps

and read depth. A Flow-Chart of our filtering steps has been depicted in Fig. 6.

In our experiments, we harness the power of XGBoost [46], an efficient and scalable implementation of gradient tree boosting. XGBoost is designed to handle sparsity and takes advantage of cache optimization, making it well-suited for distributed computing. It also employs a second-order approximation (Taylor series) of a regularized loss function to expedite computation. To evaluate our model's performance, we primarily report precision, also known as positive predictive value (PPV). For our experiments, we tune the XGBoost hyperparameters as follows: tree depth = 16, learning rate = 0.25 and boosting rounds = 250. However, for a single training sample, we set the boosting rounds to 100.

The computational complexity of our proposed method is primarily determined by two main components: feature annotation and model training/prediction. The annotation process involves extracting genomic features whose time complexity is linear with respect to the number of input instances. Time complexity of model training and prediction with XGBoost is $O(T \cdot d \cdot n \log n)$, where T = number of trees, d = maximum depth of each tree, and n = number of training samples.

Conclusions

In our previous work [22], we showed the effectiveness of our intelligent filtering tool using the gold standard GIAB dataset, which focused on short-read technologies like Illumina and Complete Genomics. Now, with updated data from additional short-read and newly added data from long-read technologies, e.g., Oxford Nanopore and Pacific Biosciences, this paper demonstrates that our method works just as well, if not better, for long-read sequencing.

Recently, there has been a rise in machine learning-based variant callers, especially for difficult-to-map regions, as seen in the precisionFDA Truth Challenge V2, which used both long- and short-read data. Our previous work showed that our filter was highly effective for difficult regions in GIAB short-read datasets. This paper validates its ability to enhance variant calling in challenging regions regardless of the sequencing technologies (long-read or short-read) or variant calling workflows (based on deep learning, graph algorithm, etc.).

Our findings confirm the method's utility for both long-read and short-read sequencing, helping researchers identify reliable indels. Future plans include extending the filter to detect more complex variants, such as rare, somatic, and multi-allelic indels, and potentially using sequencing information like read depth. While we've had promising initial results, further exploration is needed. Another goal is establishing reliable indel callsets from public databases, even without sequencing-read-specific data.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13104-025-07445-3.

Supplementary Material 1

Acknowledgements
Not Applicable.

Author contributions
M.S.I.B. initiated, planned, executed, and wrote the paper. M.S.R. guided, supervised, and reviewed the paper.

Funding
Not Applicable.

Data availability
The datasets generated during and/or analysed during the current study are available in the **GIAB** repository, https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/

Declarations

Ethics approval and consent to participate
Not Applicable.

Consent for publication
Not Applicable.

Competing interests
The authors declare no competing interests.

Declaration of generative AI and AI-assisted technologies in the writing process
During the preparation of this work the author(s) used ChatGPT 3.5 from OpenAI to enhance linguistic presentation within some part of this paper. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Received: 28 August 2024 / Accepted: 11 August 2025
Published online: 18 February 2026

References

1. Wala JA, Bandopadhayay P, Greenwald NF, O'Rourke R, Sharpe T, Stewart C, Schumacher S, Li Y, Weischenfeldt J, Yao X, Nusbaum C, Campbell P, Getz G, Meyerson M, Zhang C-Z, Imielinski M, Beroukhim R. SvABA: genome-wide detection of structural variants and indels by local assembly. Genome Res. 2018;28:581–91. https://doi.org/10.1101/gr.221028.117.
2. Marx V. Method of the year: long-read sequencing, Nature Methods 2022 20:1 20 (2023) 6–11. https://doi.org/10.1038/s41592-022-01730-w
3. Warburton PE, Sebra RP, Long-Read DNA, Sequencing. Recent advances and remaining challenges. Annu Rev Genomics Hum Genet. 2023;24:109–32. https://doi.org/10.1146/ANNUREV-GENOM-101722-103045/CITE/REFWORKS.
4. Mahmoud M, Huang Y, Garimella K, Audano PA, Wan W, Prasad N, Handsaker RE, Hall S, Pionzio A, Schatz MC, Talkowski ME, Eichler EE, Levy SE, Sedlazeck FJ. Utility of long-read sequencing for all of Us. Nat Commun 2024. 2024;15:115. https://doi.org/10.1038/s41467-024-44804-3.
5. Hu J, Wang Z, Liang F, Liu SL, Ye K, Wang DP. NextPolish2: A Repeat-aware Polishing tool for genomes assembled using HiFi long reads. Genomics Proteom Bioinf. 2024;22. https://doi.org/10.1093/GPBJNL/QZAD009.
6. Hu J, Wang Z, Sun Z, Hu B, Ayoola AO, Liang F, Li J, Sandoval JR, Cooper DN, Ye K, Ruan J, Le Xiao C, Wang D, Wu DD, Wang S. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. Genome Biol. 2024;25:1–19. https://doi.org/10.1186/S13059-024-03252-4/FIGURES/3.
7. Abde Aliy M, Bayeta S, Takale W. Pacific bioscience sequence technology: review. Int J Veterinary Sci Res. 2022;8. https://doi.org/10.17352/ijvsr.000108.
8. Lu H, Giordano F, Ning Z. Oxford nanopore minion sequencing and genome assembly. Genomics Proteom Bioinf. 2016;14. https://doi.org/10.1016/j.gpb.2016.05.004.
9. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011;12:363–76. https://doi.org/10.1038/nrg2958.
10. Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res. 2006;16:1182–90. https://doi.org/10.1101/gr.4565806.
11. Mullaney JM, Mills RE, Pittard WS, Devine SE. Small insertions and deletions (INDELs) in human genomes. Hum Mol Genet. 2010;19:R131–6. https://doi.org/10.1093/hmg/ddq400.
12. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2015. https://doi.org/10.1093/nar/gkv1070.
13. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. The next-generation sequencing revolution and its impact on genomics. Cell. 2013;155:27–38. https://doi.org/10.1016/j.cell.2013.09.006.
14. Wang Z, Liu X, Yang B-Z, Gelernter J. The role and challenges of exome sequencing in studies of human diseases. Front Genet. 2013;4:160. https://doi.org/10.3389/fgene.2013.00160.
15. Cheung VG, Spielman RS. Genetics of human gene expression: mapping DNA variants that influence gene expression. Nat Rev Genet. 2009;10:595–604. https://doi.org/10.1038/nrg2630.
16. Vali U, Brandstrom M, Johansson M, Ellegren H. Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. BMC Genet. 2008;9:8. https://doi.org/10.1186/1471-2156-9-8.
17. Wagner J, Olson ND, Harris L, Khan Z, Farek J, Mahmoud M, Stankovic A, Kovacevic V, Yoo B, Miller N, Rosenfeld JA, Ni B, Zarate S, Kirsche M, Aganezov S, Schatz MC, Narzisi G, Byrska-Bishop M, Clarke W, Evani US, Markello C, Shafin K, Zhou X, Sidow A, Bansal V, Ebert P, Marschall T, Lansdorp P, Hanlon V, Mattsson CA, Barrio AM, Fiddes IT, Xiao C, Fungtammasan A, Chin CS, Wenger AM, Rowell WJ, Sedlazeck FJ, Carroll A, Salit M, Zook JM. Benchmarking challenging small variants with linked and long reads. Cell Genomics. 2022;2. https://doi.org/10.1016/j.xgen.2022.100128.
18. Hasan MS, Wu X, Zhang L. Performance evaluation of indel calling tools using real short-read data. Hum Genomics. 2015;9:20. https://doi.org/10.1186/s40246-015-0042-2.
19. Cornish A, Guda C. A comparison of variant calling pipelines using genome in a bottle as a reference. Biomed Res Int. 2015;2015:1–11. https://doi.org/10.1155/2015/456479.
20. Krusche P, Trigg L, Boutros PC, Mason CE, De La Vega FM, Moore BL, Gonzalez-Porta M, Eberle MA, Tezak Z, Lababidi S, Truty R, Asimenos G, Funke B, Fleharty M, Chapman BA, Salit M, Zook JM. Best practices for benchmarking germline small-variant calls in human genomes. Nat Biotechnol. 2019;37:555–60. https://doi.org/10.1038/s41587-019-0054-x.
21. Olson ND, Wagner J, McDaniel J, Stephens SH, Westreich ST, Prasanna AG, Johanson E, Boja E, Maier EJ, Serang O, Jáspez D, Lorenzo-Salazar JM, Muñoz-Barrera A, Rubio-Rodríguez LA, Flores C, Kyriakidis K, Malousi A, Shafin K,

Pesout T, Jain M, Paten B, Chang PC, Kolesnikov A, Nattestad M, Baid G, Goel S, Yang H, Carroll A, Eveleigh R, Bourgey M, Bourque G, Li G, Ma CX, Tang LQ, Du YP, Zhang SW, Morata J, Tonda R, Parra G, Trotta JR, Brueffer C, Demirkaya-Budak S, Kabakci-Zorlu D, Turgut D, Kalay Ö, Budak G, Narcı K, Arslan E, Brown R, Johnson IJ, Dolgoborodov A, Semenyuk V, Jain A, Tetikol HS, Jain V, Ruehle M, Lajoie B, Roddey C, Catreux S, Mehio R, Ahsan MU, Liu Q, Wang K, Ebrahim Sahraeian SM, Fang LT, Mohiyuddin M, Hung C, Jain C, Feng H, Li Z, Chen L, Sedlazeck FJ, Zook JM. PrecisionFDA truth challenge V2: calling variants from short and long reads in difficult-to-map regions. Cell Genomics. 2022;2. https://doi.org/10.1016/j.xgen.2022.100129.

22. Bhuyan MSI, Pe'er I, Sohel Rahman M. SICaRiO: short indel call filtering with boosting. Brief Bioinform. 2021;22. https://doi.org/10.1093/bib/bbaa238.
23. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, Gross SS, Dorfman L, McLean CY, DePristo MA. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36:983. https://doi.org/10.1038/nbt.4235.
24. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The genome analysis toolkit: A mapreduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303. https://doi.org/10.1101/gr.107524.110.
25. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter Estimation from sequencing data. Bioinformatics. 2011;27:2987–93. https://doi.org/10.1093/bioinformatics/btr509.
26. Li S, Li R, Li H, Lu J, Li Y, Bolund L, Schierup MH, Wang J. SOAPindel: efficient identification of indels from short paired reads. Genome Res. 2013;23:195–200. https://doi.org/10.1101/gr.132480.111.
27. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009;25:2865–71. https://doi.org/10.1093/bioinformatics/btp394.
28. Emde A-K, Schulz MH, Weese D, Sun R, Vingron M, Kalscheuer VM, Haas SA, Reinert K. Detecting genomic indel variants with exact breakpoints in single- and paired-end sequencing data using splazers. Bioinformatics. 2012;28:619–27. https://doi.org/10.1093/bioinformatics/bts019.
29. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009;25:2283–5. https://doi.org/10.1093/bioinformatics/btp373.
30. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing, (2012). http://arxiv.org/abs/1207.3907 (accessed July 19, 2020).
31. Heldenbrand JR, Baheti S, Bockol MA, Drucker TM, Hart SN, Hudson ME, Iyer RK, Kalmbach MT, Kendig KI, Klee EW, Mattson NR, Wieben ED, Wiepert M, Wildman DE, Mainzer LS. Recommendations for performance optimizations when using GATK3.8 and GATK4. BMC Bioinformatics. 2019;20. https://doi.org/10.1186/s12859-019-3169-7.
32. Sahraeian SME, Liu R, Lau B, Podesta K, Mohiyuddin M, Lam HYK. Deep convolutional neural networks for accurate somatic mutation detection. Nat Commun. 2019;10. https://doi.org/10.1038/s41467-019-09027-x.
33. Luo R, Wong C-L, Wong Y-S, Tang C-I, Liu C-M, Leung C-M, Lam T-W. Exploring the limit of using a deep neural network on pileup data for germline variant calling. Nat Mach Intell. 2020;2:220–7. https://doi.org/10.1038/s42256-020-0167-4.
34. Luo R, Sedlazeck FJ, Lam T-W, Schatz MC. A multi-task convolutional deep neural network for variant calling in single molecule sequencing. Nat Commun. 2019;10:998. https://doi.org/10.1038/s41467-019-09025-z.
35. Freed D, Rowell WJ, Wenger AM, Li Z. Sentieon DNAscope LongRead – A highly Accurate, Fast, and Efficient Pipeline for Germline Variant Calling from PacBio HiFi reads, BioRxiv (2022).
36. Ahsan MU, Liu Q, Fang L, Wang K. NanoCaller for accurate detection of SNPs and indels in difficult-to-map regions from long-read sequencing by haplotype-aware deep neural networks. Genome Biol. 2021;22:1–33. https://doi.org/10.1186/S13059-021-02472-2/TABLES/5.
37. Noor S, Naseem A, Awan HH, Aslam W, Khan S, AlQahtani SA, Ahmad N. Deep-m5U: a deep learning-based approach for RNA 5-methyluridine modification prediction using optimized feature integration. BMC Bioinformatics. 2024;25:1–23. https://doi.org/10.1186/S12859-024-05978-1/TABLES/12.
38. Khan S, AlQahtani SA, Noor S, Ahmad N. PSSM-Sumo: deep learning based intelligent model for prediction of sumoylation sites using discriminative features. BMC Bioinformatics. 2024;25:1–15. https://doi.org/10.1186/S12859-024-05917-0/TABLES/6.
39. Khan S, Khan M, Iqbal N, Abd Rahman MA, Karim MKA. Deep-piRNA: Bi-Layered prediction model for PIWI-Interacting RNA using discriminative features, computers. Mater Continua. 2022;72:2243–58. https://doi.org/10.32604/CMC.2022.022901.
40. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Kang HM, Fuchsberger C, Danecek P, Sharp K, Luo Y, Sidore C, Kwong A, Timpson N, Koskinen S, Vrieze S, Scott LJ, Zhang H, Mahajan A, Veldink J, Peters U, Pato C, van Duijn CM, Gillies CE, Gandin I, Mezzavilla M, Gilly A, Cocca M, Traglia M, Angius A, Barrett JC, Boomsma D, Branham K, Breen G, Brummett CM, Busonero F, Campbell H, Chan A, Chen S, Chew E, Collins FS, Corbin LJ, Smith GD, Dedoussis G, Dorr M, Farmaki A-E, Ferrucci L, Forer L, Fraser RM, Gabriel S, Levy S, Groop L, Harrison T, Hattersley A, Holmen OL, Hveem K, Kretzler M, Lee JC, McGue M, Meitinger T, Melzer D, Min JL, Mohlke KL, Vincent JB, Nauck M, Nickerson D, Palotie A, Pato M, Pirastu N, McInnis M, Richards JB, Sala C, Salomaa V, Schlessinger D, Schoenherr S, Slagboom PE, Small K, Spector T, Stambolian D, Tuke M, Tuomilehto J, Van den Berg LH, Van Rheenen W, Volker U, Wijmenga C, Toniolo D, Zeggini E, Gasparini P, Sampson MG, Wilson JF, Frayling T, de Bakker PIW, Swertz MA, McCarroll S, Kooperberg C, Dekker A, Altshuler D, Willer C, Iacono W, Ripatti S, Soranzo N, Walter K, Swaroop A, Cucca F, Anderson CA, Myers RM, Boehnke M, McCarthy MI, Durbin R. Nat Genet. 2016;48:1279–83. https://doi.org/10.1038/ng.3643. G. Abecasis, J. Marchini, A reference panel of 64,976 haplotypes for genotype imputation.
41. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won H-H, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG. Analysis of protein-coding genetic variation in 60,706 humans, Nature 536 (2016) 285–291. https://doi.org/10.1038/nature19057
42. Shrikumar A, Greenside P, Kundaje A. Learning Important Features Through Propagating Activation Differences, (2017). http://arxiv.org/abs/1704.02685 (accessed April 5, 2019).
43. Tempel S. Using and Understanding repeatmasker. Methods Mol Biol. 2012;859:29–51. https://doi.org/10.1007/978-1-61779-603-6_2.
44. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010;6:e1001025. https://doi.org/10.1371/journal.pcbi.1001025.
45. Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, Harte RA, Heitner S, Hickey G, Hinrichs AS, Hubley R, Karolchik D, Learned K, Lee BT, Li CH, Miga KH, Nguyen N, Paten B, Raney BJ, Smit AFA, Speir ML, Zweig AS, Haussler D, Kuhn RM. Kent, the UCSC genome browser database: 2015 update. Nucleic Acids Res. 2015;43:D670–81. https://doi.org/10.1093/nar/gku1177.
46. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, ACM Press, New York, New York, USA, 2016: pp. 785–794. https://doi.org/10.1145/2939672.2939785

Publisher's note