StratoMod: Predicting sequencing and variant calling errors with interpretable machine learning

Nathan Dwarshuis[1], Peter Tonner[1], Nathan D. Olson[1], Fritz J Sedlazeck[2,3], Justin Wagner[1,*], Justin M. Zook[1,*]

1. Material Measurement Laboratory, National Institute of Standards and Technology, 100 Bureau Dr, MS8312, Gaithersburg, MD 20899
2. Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA
3. Department of Computer Science, Rice University, 6100 Main Street, Houston, TX, 77005, USA

*Authors contributed equally

Abstract

The structure of the genome, specifically repetitive and duplicated regions, leads to difficulty in accurately characterizing variation across a variety of sequencing technologies and variant detection methods. The Genome in a Bottle consortium generates variant benchmarks for DNA Reference Materials from seven human cell lines to reliably identify errors in comparison variant sets, as well as stratification files for error analysis in specific genomic regions. Current genome stratifications include discrete bins of homopolymer size, tandem repeat size, mappability, and segmental duplications focusing on regions subject to bias in sequencing technologies and/or variant calling methods. To move to a data-driven approach for genome stratification development and interrogation of variant calling errors, we developed a stratification model - StratoMod. StratoMod is built with an interpretable machine learning classifier, Explainable Boosting Machines (EBMs), to predict variant calling errors from features derived from genomic and sequencing data. EBMs quantify univariate effects and pairwise interactions, which facilitates understanding how each feature contributes to a prediction. Here, we describe the design, training, testing, and application of StratoMod, including demonstrating its ability to identify characteristics of sequencing reads and repetitive genome regions that are likely to generate variant calling errors. Specifically, StratoMod identified distinct associations with errors for A/T vs. G/C homopolymer lengths, and quantified sources of error for a new sequencing technology. We also demonstrated that the model could predict clinically-relevant variants that may be missed by certain methods, using DeepVariant calls from Illumina as an example. For this example, we produced a resource of difficult-to-map genes with challenging variants and large challenging INDELs. Using this model with GIAB benchmarks facilitates a deeper quantitative understanding of sources of variant calling errors from a wide variety of methods.

Introduction

The current era of sequencing offers an array of short read and long read technologies to identify the bases of a DNA molecule, each with its own strengths and weaknesses.[1] Biases arise from characteristics of these technologies, such as PCR-based library preparation, sequencing chemistry, and read length. In addition, the bioinformatics methods used to map or assemble reads and filter false positive variants can either mitigate or exacerbate these biases when making variant calls. These biases can theoretically be overcome by combining complementary technologies; this recently enabled the first complete human genome assembly[2], as well as ongoing work from the Human Pangenome Reference Consortium to develop 350 reference quality assemblies.[3] Furthermore, many challenging medically relevant genes occur in challenging regions.[4–7] To fully take advantage of rapidly advancing technologies, it is important to identify and understand their biases.

The Genome in a Bottle consortium (GIAB) established and continually updates a set of variant benchmarks for DNA reference materials derived from a large batch of seven human cell lines.[7–9] Each benchmark is a variant call format (VCF) file representing the differences between any of the GIAB samples and a given reference such as GRCh38 within the benchmark regions. GIAB develops these benchmarks by using the strengths of different sequencing technologies, which often have varying error rates and mappability depending on the region being sequenced. These benchmarks then can be used to reliably identify errors in sequencing technologies and variant calling methods. The current small variant benchmark regions include ~94% and ~92% of the reference sequence for the HG002 GIAB sample on GRCh37 and GRCh38, respectively. This includes more difficult regions than previous benchmarks but still excludes the most difficult regions and variants.[8] To compare these benchmarks with a candidate set of variants, best practices typically utilize a sophisticated bioinformatics tool to account for different variant representations.[10]

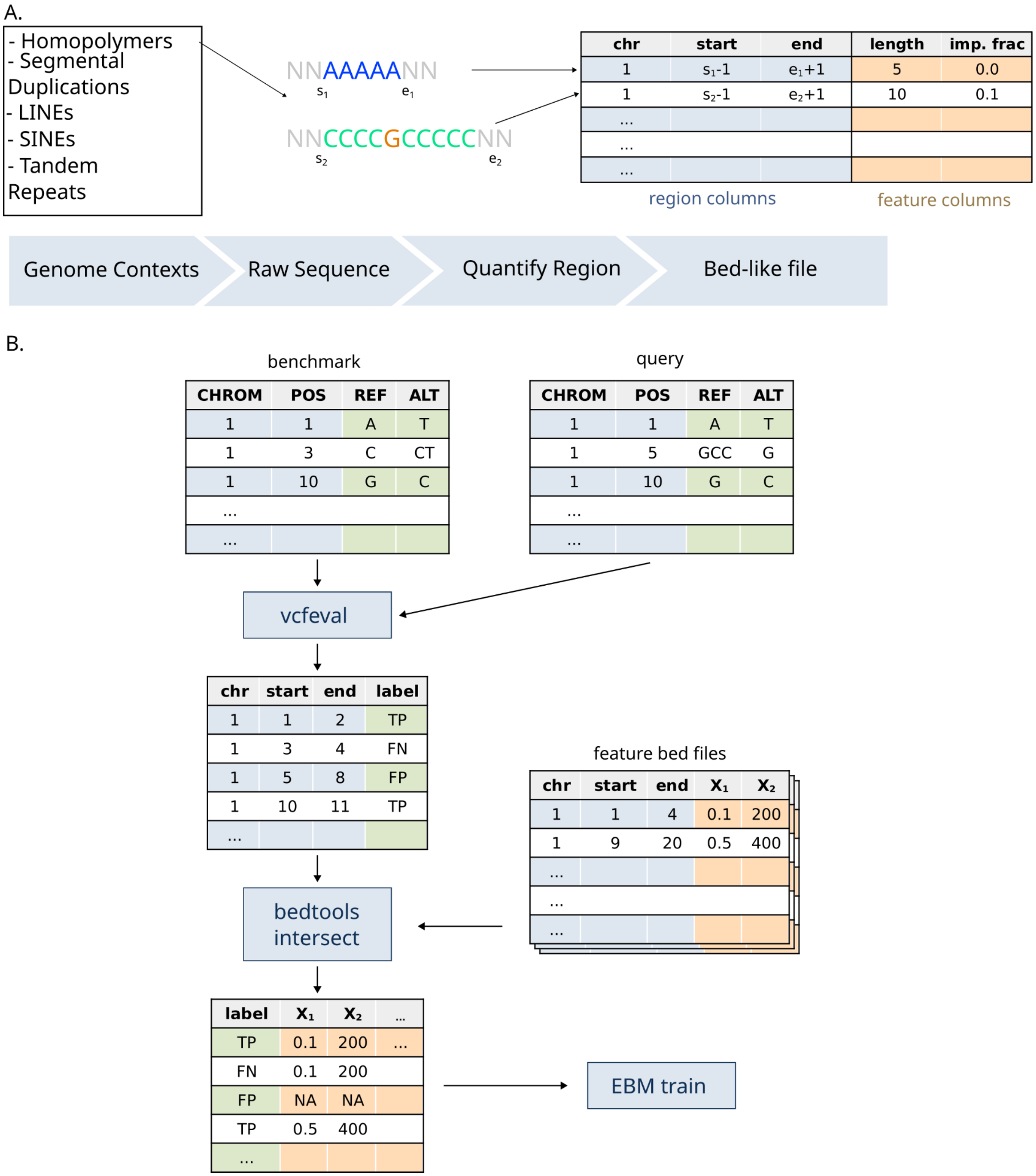
Since these benchmarks include variants from regions of varying difficulty, it is often useful to bin performance assessments by genome context (eg, if the variant is in a homopolymer, segmental duplication, etc) and variant type (SNV or INDEL). GIAB maintains a set of bed files to perform this performance assessment which are called “stratifications.” As an example, we used the GIAB stratifications in the precisionFDA Truth Challenge V2 to compare the strengths and weaknesses in different genomic contexts of sequencing technologies.[1] These stratifications can be binned into the following categories, are expected to have an impact on variant call accuracy: Low Complexity, Functional Technically Difficult, Genome Specific, Functional Regions, GC content, mappability,[11] Other Difficult or erroneous reference regions,[7] Segmental Duplications, Union of multiple categories, Ancestry of the reference,[12] and sex chromosomes. By using these stratifications, we were able to show that Oxford Nanopore Technologies (ONT) reads had higher performance in segmental duplications and hard-to-map-regions whereas Illumina excelled in low complexity regions.

While stratifications can be useful in assessing performance, they themselves do not provide a model for where errors are likely to occur. To this end, a variety of approaches have been used to model sequencing errors, mostly as part of variant calling in order to filter false positives. For example, GATK Variant Quality Score Recalibration uses Gaussian Mixture Models to identify abnormal read characteristics.[13] More recently several deep learning models have been developed to minimize the need for expert-curated features by taking in sequences from the reference and characteristics of aligned reads in a small region around each candidate variant.[14–17] Additional methods have been designed by clinical laboratories to understand which variants are enriched for false positives and should be orthogonally confirmed by another method like Sanger sequencing.[18,19] All of these methods have been very useful, particularly for filtering false positives in variant calls, but they have important limitations. For example, deep learning and many machine learning methods lack interpretability, and all of the above methods focus mostly on sequencing read characteristics at the expense of genome context, and they do not predict false negatives.

Clinical assays test for thousands to millions of different variants, so their validation relies on ‘representative variants’: categories of variants of different types and in different genome contexts. The Association for Molecular Pathology recommends the use of representative variants in their bioinformatics guidelines[20], and these must satisfy FDA’s requirements for regulatory submissions.[21] While GIAB’s stratifications could be used to select categories of representative variants, there is no systematic approach to select which of the numerous stratifications are important for a particular assay beyond general recommendations to select some variants from ‘challenging types’ and ‘challenging genome contexts’.

In this work, our goal was to develop an interpretable model to predict the likelihood of calling a variant correctly given its genome context. Interpretability was desired to allow end users of our model to understand how each feature (which corresponds to an aspect of genome context) contributes to a given prediction.[22] To this end, we chose Explainable Boosting Machines (EBMs)[23], which are a specific implementation of generalized additive models (GAMs) where model predictions are derived from additive effects of univariate and pairwise functions of dataset features (see methods for equation form). Each of these functions can be plotted individually to assess its impact on the response. EBMs have previously been shown to identify patterns in data that were obscured by other models, including confounding effects that can only be explained by domain experts.[24] In our use case, this aspect is especially important for clinicians who generally are required to justify their decisions to patients or other stakeholders (beyond simply saying “the model told me”).

This modeling approach using EBMs with genomic context features, which we call StratoMod, offers several advantages over the current strategy for assessing performance based on GIAB stratifications. First, StratoMod is much more precise. In the case of homopolymers, the genome stratification approach would have required a decision to threshold discrete bins of homopolymer lengths such as 4 to 6 bp, 7 to 10 bp, >10 bp, and >20 bp; in this case errors can only be reported in terms of these discrete bins. In contrast, the EBM model reports errors in terms of a continuous scale (log-odds), in which users can more precisely identify the impact of homopolymer length on likelihood of an error which in turn can highlight biases or strength of different sequencing technologies. Second, this model approach allows multiple genome contexts to be assessed simultaneously. Since EBMs can also include bivariate terms, we can inspect the interaction between homopolymer and INDEL length for instance. Since INDELs themselves can have varying difficulty depending on their length, sign (e.g. insertion or deletion), and method by which they are measured, it would be useful to understand how homopolymers (or other genome context) modulates this difficulty. This would also be important for assessing structural variants (INDELs >50bp), which we did not address in this work but are nonetheless of interest to the field.



chr	start	end	length	imp. frac
1	s_1-1	e_1+1	5	0.0
1	s_2-1	e_2+1	10	0.1
...				
...				
...				

benchmark

CHROM	POS	REF	ALT
1	1	A	T
1	3	C	CT
1	10	G	C
...			
...			

query

CHROM	POS	REF	ALT
1	1	A	T
1	5	GCC	G
1	10	G	C
...			
...			

chr	start	end	label
1	1	2	TP
1	3	4	FN
1	5	8	FP
1	10	11	TP
...			

feature bed files

chr	start	end	X_1	X_2
1	1	4	0.1	200
1	9	20	0.5	400
...				
...				
...				

label	X_1	X_2	...
TP	0.1	200	...
FN	0.1	200	
FP	NA	NA	
TP	0.5	400	
...			

Figure 1: Graphical overview of explainable boosting machines framework for genomic context. a) Conceptual framework for mapping “genomic context” to a machine-understandable value (with homopolymer length as an example) b) Flow chart for the analysis pipeline.

Results

Figure 1 shows the overall approach of StratoMod. For this study, we trained multiple iterations of StratoMod to assess distinct research questions as outlined in the following sections. In each case, we trained two individual models for INDELs and SNVs separately. Furthermore, all false positive (FP) and false negative (FN) errors were determined with respect to the GIAB v4.2.1 benchmark VCFs for the corresponding genome sample using GRCh38 as the reference. Lastly we simplified the analysis by removing SVs, MNVs, multiallelic variants, genotype errors, and any variants that appeared in the MHC regions from both the benchmark and query VCF input files.

Use case 1: Predicting False Positives in PCR-free vs PCR-plus

Model training and feature interpretation

We first asked if StratoMod could be used to predict where Illumina PCR-plus and PCR-free sequencing technologies have higher sequencing or mapping error rates that could produce false positive variant calls (**Fig 2a**). This choice in comparison was motivated by the fact that PCR amplification is known to produce insertions and deletions in homopolymer repeats (stutter[25]) (**Fig 2b**) and thus we hypothesized that the model would be able to precisely show the effect of homopolymer length on error rate, in addition to other repetitive genomic contexts.

We trained two models (for SNVs and INDELs) using PCR-free/plus VCF files with all DeepVariant candidate variants (no filtering) from HG004 compared against the GIAB HG004 v4.2.1 benchmark VCF. We used candidate variants for this use case in order to assess the relationship between the features and sequencing/mapping errors, independent of the variant filtering process. Each model was trained on false positive (FP, defined as extra variants not matching a benchmark variant and genotype but in the benchmark regions)[10] and true positive (TP, defined as variants matching a benchmark variant and genotype) variant call classifications, as reported by vcfeval. The model utilized 24 main effect (univariate) features, including 1 categorical feature denoting the sequencing technology (PCR-free/plus). These main effect features included quantified characteristics of homopolymers, tandem repeats, segmental duplications, and other repetitive elements, as well as depth of coverage (DP) and variant allele fraction (VAF). Each model also included interaction terms between the 24 main effects and the PCR-free/plus categorical feature, allowing the model to show the behavior of each main effect conditional on technology (see Methods and **Supp Table 1** for complete list of features). Each model was trained using an 80/20 test split on HG004 (Ashkenazi Jewish ancestry); we additionally tested the model on HG007 (Han Chinese ancestry) to assess its generalizability to other genomes. We observed that both precision and recall (measured by AUC) were similar between HG004 and HG007 (with HG007 lagging slightly behind as expected given it was the holdout dataset) (**Fig 2c, Supp Fig 1**). The negative class in the training sets for SNV and INDEL were 63% and 85% percent respectively (**Supp Table 2**).

The EBM underlying StratoMod is a GAM, meaning a prediction from the model is determined by additive contributions of different genomic and sequencing features, which we refer to as the score for each feature value. In general, the score indicates the feature's contribution to the log odds of the variant being a TP vs. an FP (e.g. a decrease in score of 1 means the odds of a FP occurring are 2.7 times more likely). Overall, the largest driving features (unsurprisingly) were VAF and DP as read from the input VCF files (**Supp Fig 2**), as many errors had low VAF and abnormally low or high DP (**Supp Fig 3**). However, other features with large effect included homopolymer length and homopolymer imperfect fraction (**Supp Fig 3**). When observing the homopolymer length feature profiles directly from the model, we found that the likelihood of a FP generally increased with increasing length as expected from PCR stutter and sequencing biases. For INDELs, PCR-plus generally predicted more errors, with relatively small interactions between PCR and homopolymer length (**Fig 2d**). Most homopolymers fell between the lengths of 0 to 50 or 0 to 15 bp for A/T and G/C homopolymers respectively, with the number of TPs and FPs decreasing exponentially with increasing length and ~100x more TPs in A/T than in G/C homopolymers longer than 10 bp (**Supp Fig 4**). Previous work had found that the number of FP INDELs in A/T homopolymers was much larger than in G/C homopolymers,[26] which was reflected in the feature rankings by Stratomod, but Stratomod also showed that G/C homopolymer length similarly predicts higher FP rates. Additionally, these feature plots provide more precise information regarding the length at which a certain relative error threshold will be crossed. In the case of SNVs, A homopolymers became more error-prone compared to the non-homopolymer baseline (the dotted lines in **Fig 2d**) after 10 bp; G homopolymers of any length were more error prone (note that both A and G homopolymer profiles were similar to their complements, see **Supp Fig 4**). Increased SNV error rates in G/C homopolymers have been attributed to inhibition of base elongation in GC-rich regions during sequencing by synthesis[27] or to formation of non-B-DNA stem-loop motifs at G quadruplexes.[28] For INDELs, these thresholds were conditional on the sequencing technology, where PCR-free and PCR-plus were more error-prone than baseline after 13 and 11 bp respectively for A homopolymers. For G homopolymers this drop off occurred around 10 bp for both PCR-free and PCR-plus (note that in the case of C homopolymers these thresholds were 12 and 10 bp for PCR-free and PCR-plus respectively, see **Supp Fig 4**).

For INDELs, we also observed an unexpected increase from baseline (i.e., a higher likelihood of TP vs. FP relative to non-homopolymers) in both A and G homopolymers for short lengths of between 4 and ~10 bp, with a peak around 8 bp. Because the EBM score is a function of both TP and FP rates, we hypothesized that the rate of true variants (TPs) increases faster than the rate of sequencing errors (FPs) in short homopolymer regions. Supporting this hypothesis, the ratio of TPs to FPs was higher for short homopolymers than for non-homopolymers, which may be caused by the higher rate of true INDEL variants in homopolymers (e.g., 39% of benchmark INDELs are in homopolymers 7 to 10 bp, while only 1.7% of the benchmark regions are in these homopolymer regions). Indeed, when plotting the TP and FP rates per base pairs covered by each homopolymer size, we saw that the TP rate increased faster than the FP rate for small homopolymers. As the homopolymer length increased, the FP/bp rate increased more than the TP/bp rate, which was reflected in the decreased EBM score. Interestingly, the TP/bp and FP/bp rates decreased for very large homopolymers, likely because large homopolymers are more likely to be excluded from the v4.2.1 GIAB benchmark, reflecting a limitation of the current training dataset (**Supp Fig 5**). These results explain the increase in EBM score for short homopolymers, followed by a decrease, before flattening out due to the small number of very long homopolymers included in the v4.2.1 benchmark. This deep-dive into a counter-intuitive result highlights both the challenges in interpreting the model's results, particularly that it is modeling the ratio of TPs to FPs rather than the FP rate per genomic bp, as well as its power in identifying unexpected associations of features with error rates.

We also noticed that the EBM INDEL scores had some sharp downward peaks for particular values of segmental duplication length and identity (**Supp Fig 6a**). When examining variants in the 2 largest peaks near 20 kbp, we found that they were caused by segmental duplication between chr7:142,450,000-142,526,000 on GRCh38 inside the T cell receptor beta locus. This region has a known issue in GRCh38, and the patch contains a ~20 kbp insertion, which is an extra tandem copy of the segmental duplication and causes many FP variant calls in Illumina and HiFi (**Supp Fig 6b**). It also intersects with 2 types of problematic reference regions identified in the recent T2T variants work: GRCh38 collapsed duplications and gnomAD inbreeding coefficient FPs, both of which annotate regions with FPs due to reads from extra copies of the region that are in most genomes but missing from the reference.[12] This result highlights a strength of this model to identify unexpected relationships between features and errors, and also suggests the possibility of adding new features associated with reference errors to future versions of the model.

These results indicate that these models can be used to precisely quantify FP error rates with respect to a meaningful, interpretable genomic context, which in turn could be useful in defining more accurate stratifications.

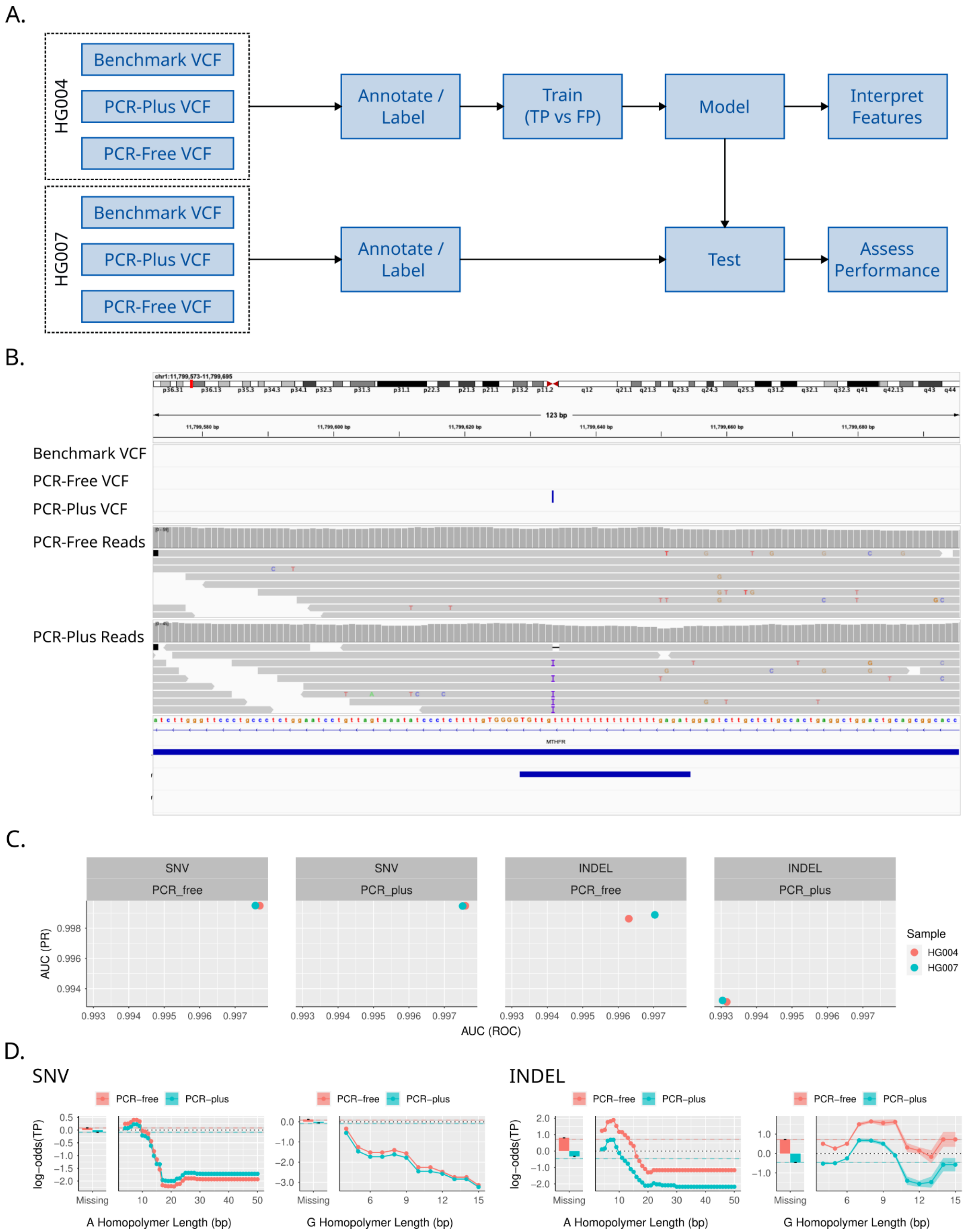


Figure 2: StratoMod revealed context-specific regions where false positives are likely to occur and show relative performance of PCR-free and PCR-plus technologies. a) Overview of

experimental setup. Two VCFs from PCR-free and PCR-plus were compared to the GIAB benchmark, concatenated, and annotated before fitting SNVs and INDELs in the EBM framework. HG005 was annotated and used to test the EBM model. b) IGV session depicting a false positive call identified by this model c) performance characteristics of HG004 (train) and HG007 (test). d) EBM plots showing A and G homopolymer profiles. The x axis in all plots was truncated to only show homopolymers <50bp and <15bp for A and G respectively. The bar plots on the left of each plot show the value for non-homopolymers (“missing”). The y axis in the top row is the log odds of a TP outcome. Each colored dotted line in the top row is the “baseline” error rate for the corresponding sequencing technology. The bottom plots show the distribution of homopolymer lengths (TPs are above the dotted line and FPs are below).

Comparison of EBM performance to DeepVariant performance within candidate regions

We next compared the accuracy of StratoMod’s FP vs TP classifications to DeepVariant’s, for the candidates generated by DeepVariant. We first examined the calibration of DeepVariant’s genotype quality score (GQ) (**Fig 3a**) and StratoMod’s probability score (**Fig 3b**). As expected from DeepVariant’s richer information used for classification, it provided useful phred-based quality scores up to empirical scores of >50 (1 in 100,000 error rate) vs. the v4.2.1 benchmark, though it was somewhat overconfident for INDELs. StratoMod provides well-calibrated scores up to about 35 (1 in 3000 error rate) for SNVs and 25 (1 in 300 error rate) for INDELs. In **Figure 3b**, we assign a label on “1” to any StratoMod probability > .5 and “0” otherwise. We then made a Venn diagram showing intersections between DeepVariant and StratoMod predicted TP labels against the benchmark when combining SNVs and indels for PCR-free and PCR-plus Illumina. 99.4% of the benchmark variants were classified as TPs by both StratoMod and DeepVariant. Of the remaining 0.6%, most (36,675) were classified incorrectly as FPs by StratoMod and as TPs by DeepVariant, whereas 1,765 were correctly classified as TPs by StratoMod and not DeepVariant. In addition, StratoMod incorrectly classified 44,629 variants as TPs that were correctly filtered by DeepVariant, many more than the 3,930 uniquely incorrect TPs in DeepVariant. When intersecting these uniquely mis-classified variants with features, DeepVariant uniformly performed better, but ~25% of false variants uniquely classified as TPs by DeepVariant were in regions difficult to map with 250 bp reads, suggesting DeepVariant may benefit from additional genome mappability features.

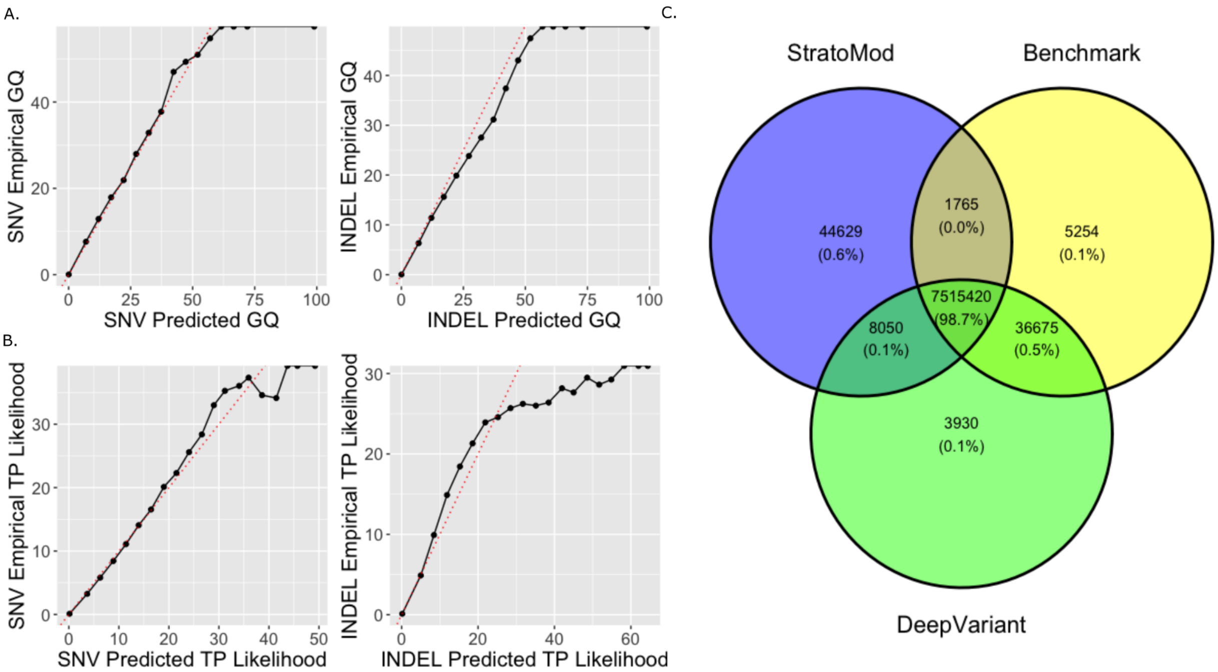


Figure 3: Comparison of Stratomod calibration and accuracy relative to deep learning-based method DeepVariant. a) Predicted genotype quality score (GQ) from DeepVariant plotted against an empirically derived GQ measure, -10*log10(FP/(FP+TP)), using the GIAB v4.2.1 small variant benchmark. b) PHRED-scaled plots depicting calibration for how well StratoMod confidently predicts TPs (or values near 1 for its predictions) similar to a typical genotype quality score c) Comparison of StratoMod and DeepVariant performance against the benchmark values for the candidate sites from DeepVariant used in our model.

Comparison of EBMs to other commonly used models

We compared the performance of an EBM model to XGBoost (XGB), Random Forest (RF), Logistic Regression (LR), and Decision Tree (DT) models in classifying variants as errors using the same training and testing data. In all cases, performance was comparable when examining the area under the curve (AUC) for the receiver operator characteristic (ROC) curve and precision-recall (PR) curves (**Table 1**). Note that the RF SNV model did not complete within the constraints imposed by our compute cluster.

While similar, StratoMod performed slightly better than both DT and LR models and slightly worse than RF and XGB models for both SNVs and INDELs. This is not surprising considering that RF and XGB are much more flexible than EBMs, and EBMs in turn are more flexible than LR models. DT models might be more flexible than EBMs given that they have less restrictions when building trees, but DT models also tend to be brittle as they are not ensemble models (unlike EBMs). Notably, the models that did perform slightly better than EBMs were also blackbox models (e.g. unable to be inspected analogously to EBMs).

These data demonstrated that for this use case, using EBMs for the classification algorithm in StratoMod performed similarly to other commonly used models. Slight performance benefits may be had with blackbox models such as random forest or XGBoost, albeit with a loss of interpretability and potentially much higher compute requirements.

Table 1: Comparison of EBM performance to those of other popular machine-learning methods. OOM: out of memory

Variant	Metric	StratoMod	DT	LR	RF	XGB
INDEL	ROC	99.5%	99.2%	99.2%	99.8%	99.8%
INDEL	PR	99.7%	99.4%	99.5%	99.9%	99.9%
SNV	ROC	99.8%	99.4%	99.5%	OOM	99.9%
SNV	PR	99.9%	99.8%	99.9%	OOM	100.0%

Use case 2: Predicting False negatives in Illumina PCR-free and PacBio HiFi variant calling pipelines

Model training and feature interpretation

We also produced similar models to the previous section, but for FNs vs. TPs and comparing Illumina PCR-free to PacBio HiFi DeepVariant variant calls. In contrast to the previous section, we compared the variant calls from the two technologies after DeepVariant's filtering, so that we could understand FNs caused by filtering. The model was trained analogously using the GIAB v4.2.1 small variant benchmarks (**Fig 4a**), except we used HG002 as the training sample, HG004 and HG007 as the testing samples, and predicted TP and FN (as opposed to FP) variant call labels. We also did not include DP and VAF features so that we could understand how well FNs for a variant calling method could be predicted from sequence context alone. Finally, after training the model, we used it to predict FN rates for variants in ClinVar, to understand variants that a given method might miss; note that the GIAB benchmarks do not have sufficient pathogenic and likely pathogenic variants to validate model performance for these variants, so our validation strategy included manually inspecting a subset of outcomes.

The human genome contains many copies of transposable elements such as LINEs, LTRs, and SINEs, which can cause mapping challenges, particularly for short reads. These elements often overlapped our difficult-to-map regions, but we added them as additional features to understand their effect on FN rate beyond our difficult-to-map features. Any variant in a LINE was more likely to be missed than those not in LINEs, and the difference in predicted score between Illumina and HiFi became larger as length increased (**Fig 4c**). There was a spike in the FN rate around 6000 bp long (corresponding to full-length L1 LINEs), possibly because full-length LINEs are more recent and similar to each other; this may be an indication that adding additional features such as subclasses of LINEs may increase the accuracy and interpretability of the model. Interestingly, SINEs had the opposite effect as LINEs and predicted fewer SNV FNs in Illumina than HiFi, possibly because SINEs typically have a long homopolymer at one end (**Fig 4c**).

Similar to transposable elements but with fewer copies, segmental duplications are large (>1000 bp) genomic regions with at least 85% sequence similarity, causing challenges with short and long read mapping. The SNV model showed that any segmental duplication size predicts more FNs than non-segmental duplications, and increases in segmental duplication size predict more FNs, as expected (**Fig 4d**). Furthermore, the FN rate for HiFi was uniformly lower than that of

Illumina; however, this difference was less for increasing segmental duplication lengths, possibly because the longer HiFi read length has less advantage when the segmental duplication is longer than the read length, particularly if it is not also in one of the difficult to map region features. The FN rate also increases with increasing segmental duplication similarity, particularly above 98%, with a roughly constant difference between HiFi and Illumina (**Fig 4d**).

Both HiFi and Illumina were more likely to miss variants in hard-to-map regions as expected (**Fig 4e**). However, the decrease in TP log-likelihood was much higher in Illumina (about 4 and 5 for 100 and 250 bp respectively) compared to HiFi (about 2 and 4 for 100 and 250 bp respectively). Particularly in the case of 100bp sized regions, our model predicted that the likelihood of correctly calling a SNV in a hard-to-map region with HiFi was about the same as the baseline of Illumina.

For INDELs, we investigated homopolymers and tandem repeats (**Fig 4f**). For T homopolymers specifically (which were similar to A homopolymers in **Supp Fig 7**), we found that the overall error rate of HiFi is much higher than that of Illumina, which was expected given the known error profile of HiFi with respect to INDELs. Furthermore, the rate of FN likelihood in HiFi began accelerating relative to Illumina at around 10 bp for T homopolymers. For tandem repeats, FNs were less likely in Illumina at lengths less than 125bp, but then HiFi had fewer errors for lengths greater than this cutoff. This underscores the importance of long reads for finding variants in longer repeats; while HiFi may have an intrinsically higher error rate for INDELs relative to Illumina, this disadvantage is negated in longer repeats where Illumina reads become shorter than the repeat so that HiFi is less prone to have mapping errors. This effect was likely not seen in homopolymers due to their lengths being much shorter than the Illumina read lengths.

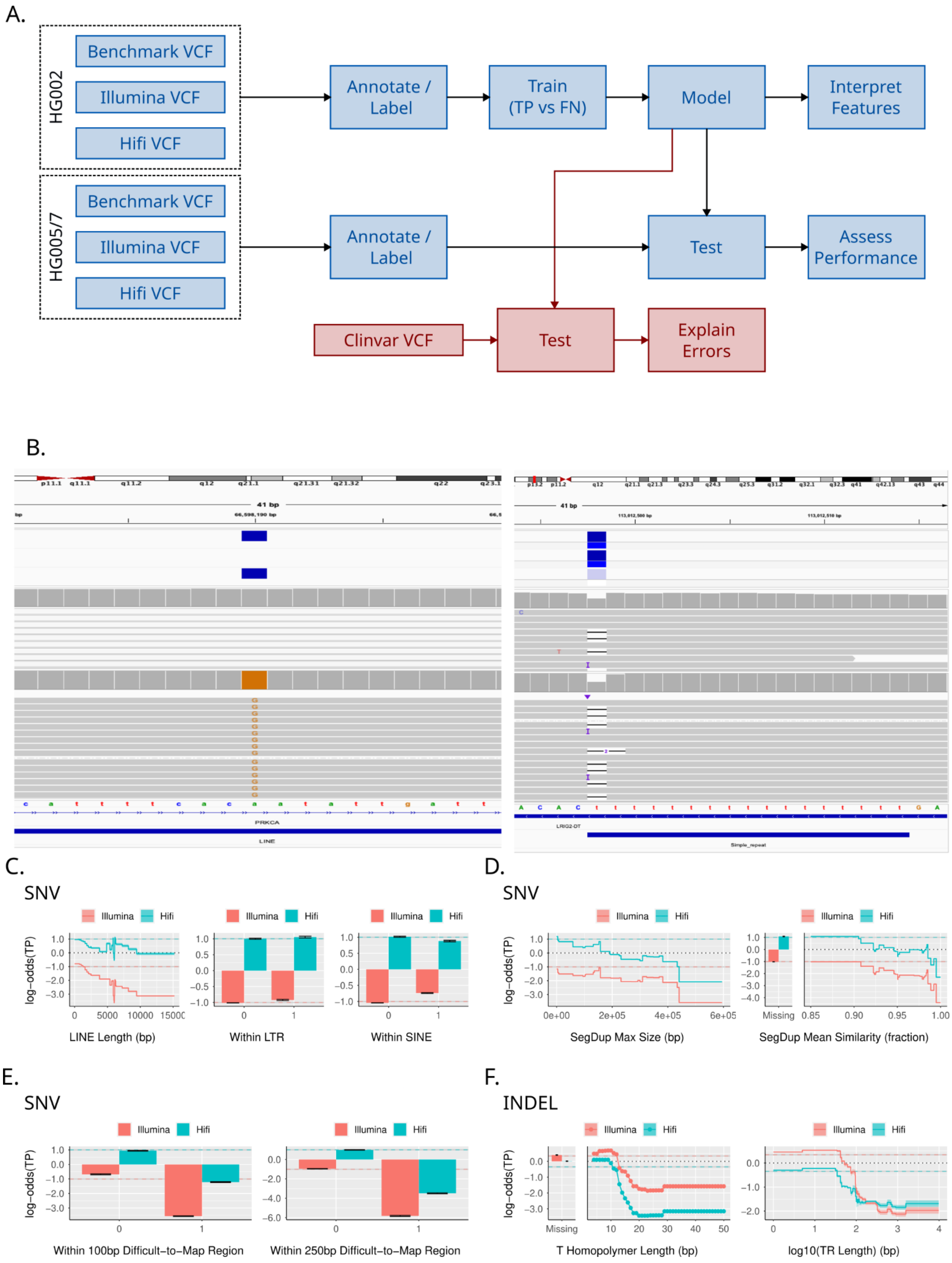


Figure 4. StratoMod found context-specific regions where variants are likely to be missed in

either HiFi or Illumina analysis pipelines. a) Experimental overview b) IGV session depicting false negative calls identified by this model c-f) EBM interaction feature plots for SNV segmental duplications (c), SNV transposable elements (d), SNV hard-to-map regions (e) and INDEL homopolymers (f).

Performance assessment and prediction of missing clinically-relevant variants

We tested these models on both HG004 and HG007 to assess generalizability to different genomes. Overall, the model performance was worse between training and testing on HiFi compared to Illumina (**Supp Fig 1, Supp Fig 8a**), most likely because the VCFs used for training and testing had inconsistent coverage (**Supp Table 5**). For Illumina, we observed that performance (AUC under PR and ROC curves) was comparable between all samples, indicating generalizability of our model (**Fig 5a**) despite the fact that the negative class for these models was <1% (**Supp Table 2**). However, we also observed that in the case of the INDEL model, the two testing samples actually performed better than the training sample. While this is likely due to inherent differences between samples, data, and/or benchmarks that we are not capturing adequately with our model (for example, HG007 and HG002 are of different ancestries which may have variants of varying difficulty and/or heterozygosity in the regions we included in our model), these performance metrics showed that at least the models did not totally fail when tested on an entirely different human genome.

We next assessed the ability of our model to predict the likelihood of missing clinically relevant variants with Illumina-DeepVariant, focusing on this callset because it has a higher FN rate than HiFi-DeepVariant. We fed a ClinVar VCF through our model (without labels since our benchmarks have very few ClinVar variants) and extracted the probabilities of missing a variant. We also examined the contribution of each feature to those probabilities to “explain” why some variants were likely to be missed. To focus on variants of more clinical interest, we only considered variants marked as “pathogenic” or “likely pathogenic” in the ClinVar VCF. In the case of SNVs and INDELs this resulted in 99905 and 64583 variants (**Supp Fig 6b**). We then filtered variants whose likelihood of being missed was greater than 10% and plotted them in a heatmap showing the local contribution of each feature (**Fig 5b, Supp Files 1 and 2**). The vast majority of variants were called with relatively high confidence of being a TP (**Fig 5c**) and the model was well-calibrated at the 10% FN cutoff we used (**Supp Fig 9**).

For SNVs, all predicted FN calls except two were explained both by low mappability and being in a highly similar/duplicated segmental duplication, with more than 20 FN SNVs predicted in genes *CYP21A2*, *CHEK2*, *NEB*, *PMS2*, *TUBB8*, *SMN1*, *STRC*, *PIK3CA*, and *ABCC6*, as well as genes that are falsely duplicated in GRCh38 (CBS, with fewer in *KCNE1*, *U2AF1*, and *CRYAA*), with full list of genes and counts of predicted FNs in **Supplementary File 3**. This is plausible given that these two regions types should overlap significantly and also should lead to mismapped reads and hence missed variant calls. The two predicted FNs not in segmental duplications were a cryptic splice acceptor at the edge of a full-length LINE in an intron of ATM (NM_000051.4(ATM):c.2466+1552G>C), and a variant from one submitter in a VNTR in an intron of PRPF31 (NM_015629.4(PRPF31):c.1374+654C>G).

Predicted INDEL FNs were also primarily in segmental duplications; however other features such as INDEL length, and tandem repeat length explained some of the errors, indicating that multiple error mechanisms may be at play. For example, potential FNs were predicted for INDELs in VNTRs in *ACAN*, *COL6A1*, *F7*, *MYO7A*, *AGRN*, and *CDKN1C* (some of which are inside an exon and some partly exonic but mostly intronic), as well as large insertions in trinucleotide tandem repeats like *DMPK*, *PHOX2B*, and *RUNX2*, where expansions are associated with disorders. The remaining predicted FNs were large INDELs in non-repetitive regions (a few incidentally in homopolymers), except for a cluster FNs in a few hundred bp region of *ANKRD11* that is identical to a region on chrX, but can be mapped accurately with most paired-end reads. The full list of genes and counts of predicted FNs is in **Supplementary File 4.**

Many of the above identified genes, tandem repeat regions, and variant types have previously been known to be challenging. In addition to highlighting general genes and characteristics of variants, this model predicted particular pathogenic variants that may be missed by a particular sequencing and bioinformatics method (in this case, 35x Illumina-DeepVariant PCR-free WGS), which may not be representative of all Illumina-based methods. This model could be used similarly to predict variants that might be missed by any method that has been used to call variants from benchmark samples like those available from GIAB.

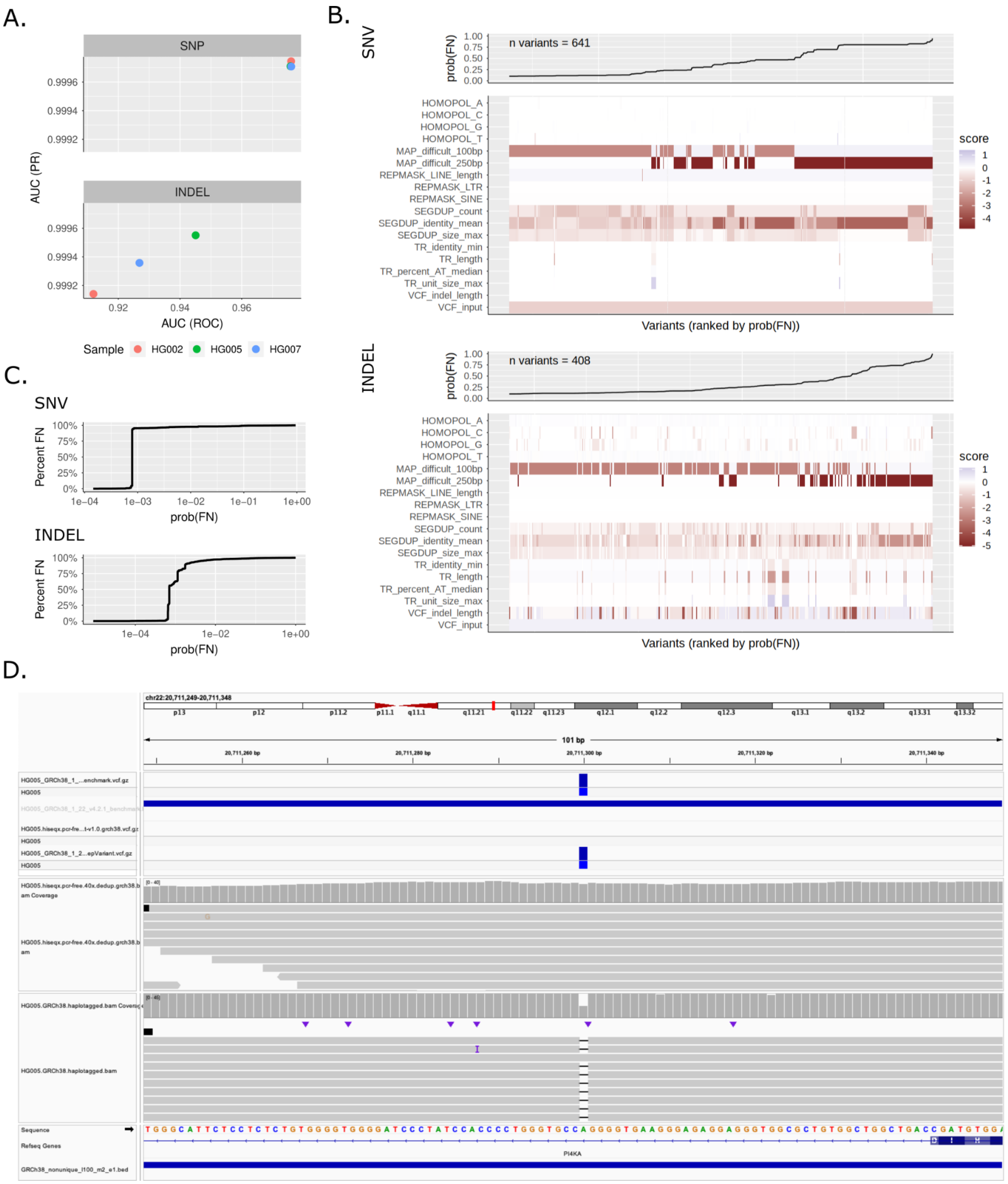


Figure 5: Assessment of FN Clinvar variants using Illumina reads. A) Performance of models on each genome sample (note 80% of HG002 was used for training B) Heatmap of variants with a 10% or higher likelihood of being predicted as FN showing contributions of each feature within the model. Note that all interaction terms were added to their univariate main effect features to simplify visualization, and also note that negative scores contributed positively to a FN. C) Cumulative distribution of FN variants vs probability of the model predicting a FN.

Use case 3: Assessment of new sequencing technologies

New sequencing technologies are under active development, so we next tested if StratoMod could be used to understand strengths and weaknesses of one new sequencing method from Ultima that promises to be less expensive.[29] We compared Illumina PCR-free variant calls to those of Ultima, both filtered by DeepVariant analogously to the FN model outlined in use-case 2, except we trained/tested 80/20 on HG004 and tested entirely on HG007. The output labels to be predicted were TP and FN, and thus this model was intended to show where variants might be missed between the two technologies.

Performance between HG004 and HG007 for both SNV and INDEL models was comparable (**Supp Fig 1, Supp Fig 10a**). However, performance for these models was overall comparably lower than that for the PCR-free/plus model (despite PCR-free being the same VCF in both models), likely due to the larger class imbalance between FN and TP vs FP and TP (**Supp Table 2**).

Our analysis provides additional insight compared to what Ultima released in their white paper.[29] In the case of homopolymers, we found that for SNVs, A homopolymers greater than 8bp have higher FN rates compared to baseline; in G homopolymers this threshold is 6bp (**Supp Fig 10b**), with similar results for T and G homopolymers, respectively (**Supp Fig 10b** top row). For INDELs, A homopolymers >11 bp are more error-prone than baseline, and 7bp in the case of G homopolymers. The Ultima white paper broadly excludes all homopolymers >11bp regardless of whether it is a A/T or G/C homopolymer or SNV or INDEL.

Additionally, in the case of tandem repeats we found that for SNVs and INDELs, the error rate in A/T rich repeats drops precipitously above ~80%, which is much lower than the 95% threshold excluded in the Ultima white paper (**Supp Fig 10b** bottom row). Also regarding tandem repeats, we found that repeats longer than 30bp for both SNVs and INDELs had precipitously higher error rates with increasing length (although the difference between Illumina and Ultima in these cases was less substantial than for the A/T fraction and homopolymer length).

Compared to Illumina, the features we investigated that seemed to negatively diverge were A/T percent for tandem repeats and homopolymer length. For other measures, Ultima and Illumina were comparable, and in the case of longer tandem repeat lengths actually converged to Illumina for higher lengths.

For all of these results, it is important to note that particular conclusions about technologies are likely to change over time as methods improve. However, these data showed how this modeling approach can be used to precisely identify the strengths and weaknesses of novel sequencing techniques relative to existing tools, and models can be updated as new sequencing technologies and variant calling models likely improve over time.

Discussion

In this work, we demonstrated that an inherently interpretable model with expert-designed features for genome context can be used to gain a deeper understanding of sequencing and variant calling errors.

Predicting false negatives from short and long reads enabled us to gain a quantitative understanding of strengths and weaknesses of each technology with DeepVariant for different types of repeats. For example, difficult-to-map regions and repetitive elements that cause challenges in mapping predict lower SNV FN rates for HiFi-DeepVariant, whereas longer homopolymers and shorter tandem repeats predict lower INDEL FN rates for Illumina-DeepVariant.

StratoMod newly enables prediction of variants that might be missed by a particular method (sequencing and variant caller) if that method was used to sequence one or more benchmark samples like those from GIAB. While previous studies have outlined challenging genes and challenging types of variants in general, StratoMod can predict variants that might be missed by a particular method along with the features that cause each variant to be challenging.

Finally, we showed that StratoMod can be used to understand the strengths and weaknesses of new technologies, which is particularly important given recent innovations in short- and long-read platforms. These technologies and associated variant callers are under active development; StratoMod could be used by developers to understand how new innovations improve results and where there is room for ongoing improvement.

We selected a relatively small set of features to maintain interpretability and reduce correlations between related features as this reduces interpretability. We also evaluated a limited number of interactions between features, because large numbers of feature interactions are challenging to interpret. We expect adding and fine-tuning features and adding more interactions could improve model performance, but would also create more challenges in interpretation and visualization of results. The tree-based modeling approach enabled us to identify some discontinuities in scores that point to possible improvements in feature design, such as identifying peaks in segmental duplication length related to particular segmental duplications with errors in GRCh38, and a peak in LINE length related to full-length LINEs. However, the tree-based nature of the model also makes robust uncertainty estimates challenging relative to other GAM-based approaches. These results also provide suggestions for features that could be added to other models like those used by variant callers to filter false positives.

While we have demonstrated that StratoMod can be useful for diagnosing and understanding variant calling errors, it was not intended to be a replacement for any existing variant caller. In particular, up to 20% of FP variants are not assigned any genome context features, so StratoMod uses only depth of coverage and variant allele fraction for classification. This limits StratoMod's ability to be as comprehensive as expected from a variant caller (**Supp Fig 10**). Adding slop (i.e., extra bases around the repeat) can be important for some features, and we experimented with adding 1 to 5 bp slop around homopolymers. We decided to use 1 bp to simplify interpretation around nearby homopolymers, but this increased the number of variants without a genome context feature, particularly for GC-rich regions. Despite this lack of comprehensiveness, the model's interpretability enables a more detailed understanding of how types of repeats and their characteristics predict sequencing and mapping error rates for PCR-free vs. PCR-plus Illumina sequencing. Understanding differences in sequencing and mapping errors between technologies can be challenging due to differences in how candidate variants are generated, so we did not use the model to predict false positives across technologies, but instead predicted differences in false negatives after filtering for DeepVariant-based callsets from Illumina and HiFi in the second use case.

The work presented here supports future stratification development in a more data-driven way. More comprehensive benchmarks, such as those based on de novo assembly, will also provide more accurate models of variant call errors, particularly in more difficult regions and for larger variants. We also expect these models along with manual curation to help systematize the creation of new benchmarks by helping to understand the strengths and weaknesses of each technology and variant caller so that we can know which method to trust when they differ. Models like StratoMod will provide a new approach both for developing better benchmarks and for using these benchmarks to understand strengths and weaknesses of a method and predict which clinically-relevant variants may be missed.

References

1. Olson, N. D. *et al.* PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions. *Cell Genom* **2**, (2022).
2. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
3. Liao, W.-W. *et al.* A Draft Human Pangenome Reference. *bioRxiv* 2022.07.09.499321 (2022) doi:10.1101/2022.07.09.499321.
4. Goldfeder, R. L. *et al.* Medical implications of technical accuracy in genome sequencing. *Genome Med.* **8**, (2016).
5. Mandelker, D. *et al.* Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. *Genet. Med.* **18**, 1282–1289 (2016).
6. Lincoln, S. E. *et al.* One in seven pathogenic variants can be challenging to detect by NGS: an analysis of 450,000 patients with implications for clinical sensitivity and genetic test implementation. *Genet. Med.* **23**, 1673–1680 (2021).
7. Wagner, J. *et al.* Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat. Biotechnol.* 1–9 (2022).
8. Wagner, J. *et al.* Benchmarking challenging small variants with linked and long reads. *Cell Genomics* **2**, (2022).
9. Zook, J. M. *et al.* A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-0538-8.
10. Krusche, P. *et al.* Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* **37**, 555–560 (2019).
11. Derrien, T. *et al.* Fast computation and applications of genome mappability. *PLoS One* **7**, e30377 (2012).

12. Aganezov, S. *et al.* A complete reference genome improves analysis of human genetic variation. *Science* **376**, eabl3533 (2022).

13. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

14. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983 (2018).

15. Shafin, K. *et al.* Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat. Methods* **18**, 1322–1332 (2021).

16. Sahraeian, S. M. E. *et al.* Deep convolutional neural networks for accurate somatic mutation detection. *Nat. Commun.* **10**, 1041 (2019).

17. Luo, R., Sedlazeck, F. J., Lam, T.-W. & Schatz, M. C. A multi-task convolutional deep neural network for variant calling in single molecule sequencing. *Nat. Commun.* **10**, 998 (2019).

18. Holt, J. M. *et al.* Reducing Sanger confirmation testing through false positive prediction algorithms. *Genet. Med.* **23**, 1255–1262 (2021).

19. Lincoln, S. E. *et al.* A Rigorous Interlaboratory Examination of the Need to Confirm Next-Generation Sequencing–Detected Variants with an Orthogonal Method in Clinical Genetic Testing. *J. Mol. Diagn.* **21**, (2019).

20. Roy, S. *et al.* Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. *J. Mol. Diagn.* **20**, 4–27 (2018).

21. Balasubramaniam, S. *et al.* FDA Approval Summary: Rucaparib for the Treatment of Patients with Deleterious BRCA Mutation-Associated Advanced Ovarian Cancer. *Clin. Cancer Res.* **23**, 7165–7170 (2017).

22. Lipton, Z. C. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queueing Syst.* **16**, 31–57 (2018).

23. Lou, Y., Caruana, R., Gehrke, J. & Hooker, G. Accurate intelligible models with pairwise

interactions. in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* 623–631 (Association for Computing Machinery, 2013).

24. Caruana, R. *et al.* Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1721–1730 (Association for Computing Machinery, 2015).
25. Gymrek, M. PCR-free library preparation greatly reduces stutter noise at short tandem repeats. *bioRxiv* 043448 (2016) doi:10.1101/043448.
26. Fang, H. *et al.* Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Med.* **6**, 89 (2014).
27. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36**, e105 (2008).
28. Weissensteiner, M. H. *et al.* Distinct sequencing success at non-B-DNA motifs. *bioRxiv* 2022.06.13.495922 (2022) doi:10.1101/2022.06.13.495922.
29. Almogy, G. *et al.* Cost-efficient whole genome-sequencing using novel mostly natural sequencing-by-synthesis chemistry and open fluidics platform. *bioRxiv* 2022.05.29.493900 (2022) doi:10.1101/2022.05.29.493900.
30. Nori, H., Jenkins, S., Koch, P. & Caruana, R. InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv [cs.LG]* (2019).

Methods

Variant labeling

The overall process to label a query VCF file such that it can be understood by the EBM model is given in **Fig 1b**. The following is a more detailed description of this process:

Preprocessing

To treat DeepVariant's filtered variants as variant when doing the comparison with vcfeval, we converted genotypes from ./. to 0/1 and 0/0 to 0/1. Furthermore, we removed all chromosomes except 1-22 (as the GIAB benchmarks we used as the truth sets did not have X/Y or any alternate chromosomes). Finally, we removed the MHC region from the benchmark and query VCF files before comparison.

Comparison

We generated TP, FP, and FN labels using vcfeval with `--refoverlap –all-records` to preserve all filtered variants (which were either kept or removed depending on the desired analysis). The output vcf files corresponding to TP, FP, and FN labels were then concatenated and converted to a bed file with an additional column holding the corresponding label for each variant. During this step we also computed all VCF_* features (see below) from the VCF file itself. We also removed multi-allelic variants (to simplify the analysis), multinucleotide variants (variants whose REF and ALT were both >1bp and equal to each other) and structural variants (those whose REF or ALT columns were longer than 50 bp). Importantly, we shifted the start/end columns resulting from the VCF file leftwise by 1 to make the final result 0-based instead of 1-based for proper intersection with BED files.

After this step we split the resulting bed files into SNV and INDEL streams.

Annotation

After generating all feature bed files (see below section) we merged the SNV and INDEL label bed files from the previous step using multiple rounds of bedtools intersect with the -loj flag.

Pre-train processing

Prior to use in the EBM for training, we converted the labels (TP, FP, FN) to a 0/1 feature as required for binary classification. For the FP model, we simply removed the FN label and mapped FP -> 0 and TP -> 1. For the FN model(s), we wanted to remove the effect of including filtered variants in the training set (as if we had never used `--all-records` with vcfeval) and thus for all non-PASS variants, we mapped FP -> TN and TP -> FN, filtered only FN and TP labels, and then mapped TP -> 1 and FN -> 0.

Since we used the -loj flag in the previous annotation step, many variants did not have any values for a given feature (since on average a given variant will only intersect with a few regions corresponding to our feature categories). For HOMOPOL_imperfect_frac and TR_percent_AT_median we filled these missing values with -1 since 0 had real meaning for these features. For all other features we filled in missing values with 0; for these features this corresponded to 0-length or 0-count, and thus made numerical sense as a missing substitute.

Feature Engineering

Model features were grouped into 5 main categories. See **Supplemental Table 1** for a list of all features. The prefix of each feature corresponds to its category. An overview of each feature category and their method of generation follows:

VCF features (prefix = VCF)

VCF_DP and VCF_VAF were taken from the query VCF file without modification. VCF_input was used as an index to track the query VCF file where multiple inputs were used in the model. VCF_indel length was taken as the difference between the ALT and REF columns (thus positive values represented insertions).

Homopolymers (prefix = HOMOPOL)

Perfect homopolymers (eg homopolymers with no other interrupting bases) with lengths >= 4bp were generated directly from the reference using an in-house python script and saved to a bed file. This bed file was then split into each of the four bases. Each individual homopolymer-per-base file was then merged using bedtools with -d 1 (to get “imperfect” homopolymers which have at least two stretches of the same base >= 4pb with one different base in between). We then added 1bp slop to each end of the merged regions in order to detect errors immediately adjacent to the homopolymer itself. For each homopolymer region, we used the length (without slop) as well as the fraction of imperfectness (the number of non-homopolymer bases over the length). Note that in our formulation, imperfect fraction approaches a theoretical maximum of 20% in the limit as length increases.

Tandem Repeats (prefix = TR)

Tandem repeat features were based off of the simple repeat finder UCSC database. We used bedtools merge to summarize a subset of the columns in this database (period, copyNum, perMatch, perIndel, and score; note that we renamed these in our feature set to better distinguish them). For each of these columns we computed the min, max, and median value when merging, and also stored the count of the number of merged repeats. Furthermore, we computed the percent of GC and AT content in each region using the individual base percent columns present in the database, and merging analogously to the previous columns. Finally, computed the length of the tandem repeat region directly using the coordinates present in the database file. We added 5bp slop to each region, and removed all regions with period/unitsize == 1 as these corresponded to homopolymers which were represented in a different feature category.

Segmental Duplications (prefix = SEGDUP)

Segmental duplication features were merged in a similar manner to the tandem repeat columns, except we used the genomicSuperDups database from UCSC. We only used the alignL and fracMatchIndel columns, and computed the min, max, and mean of these as well as the count of regions that were merged. We did not add slop to these features.

Repeat Masker (prefix = REPMASK)

Repeat masker features were based on the rmsk database from UCSC. For entries whose class was SINE, LINE, or LTR, we filtered by class and merged using bedtools. We then calculated the length of each merged region (conditioned on class). We did not add slop to these regions. In the case of SINE and LTR, we converted each feature to binary by setting each length to 1 (and then any non-intersecting variants would get a 0 representing that they did not intersect this region).

Hard-to-map (prefix = MAP)

We used the GEM-based[11] lowmappabilityall.bed.gz (100bp) and nonunique_l250_m0_e0.bed.gz (250bp) from the GIAB v3.0 stratifications[1] as the basis for this feature. For both of these bed files, we simply appended a column filled entirely with 1's representing a binary feature where 1 means "in a hard to map region" and 0 otherwise.

Model Training

We used the interpretml package[30] from Microsoft Research to train the EBM models. We specifically used the ExplainableBoostingClassifier which has the following form:

$$g(x) = f_1(x_1) + f_1(x_1) + \dots + f_{1,2}(x_1, x_2) + \dots$$

Here g is the logit link function and each f_i is either a univariate or bivariate decision tree in terms of its input(s).

These were trained using a random 80/20 train/test split. For interactions, we specifically included all interactions that included VCF_input as well as indel_length vs all four of the homopolymer length features. All other settings were left at default.

Model Comparisons

To compare EBMs with other commonly used models in the machine learning space, we used the FP/TP EBM model with Illumina PCR-Free/Plus and ran the associated data through the following algorithms with the following hyperparameter tuning schemes:

Model	Implementation	Hyperparameter levels
Decision Tree	rpart (R)	Cost_complexity: 0.00001, 0.0001, 0.001, 0.01, 0.1
Logistic Regression	glmnet (R)	Penalty: 0.000001, 0.000001, 0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10 Mixture: 0, 0.5, 1
Random Forest	ranger (R)	mtry : 1, 4, 7 trees: 500, 1000, 2000
XGBoost	xbgoost (python/gpu accel)	max_depth : 3, 6, 9 n_estimators: 100, 500, 1000 gamma: 1, 10, 100

All models were trained on a compute cluster with 512GB memory, 2 20 core Intel Xeon E52698 v4 CPUs and 8 Nvidia Tesla V100 (per node). Each job was allowed 3 days of compute time. Of all the algorithms used (including EBMs) only xgboost was able to take advantage of GPU acceleration.

ClinVar Analysis

We used the November 5 2022 release of the ClinVar VCF (clinvar_20221105.vcf.gz). We added a FORMAT (GT) and SAMPLE (0/1) column to the VCF and used hap.py (https://github.com/Illumina/hap.py) to compare the v4.2.1 benchmark VCF and BED (https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/ChineseTrio/HG005_NA24631_son/NISTv4.2.1/GRCh38/). Restricting to "likely pathogenic" or "pathogenic" variants in the ClinVar VCF, this resulted in 133 matching variants from HG005.

Software Packages

See **Supplemental Table 3** for a list of software packages and their versions that were used in this work.

The source for StratoMod can be found on github: https://github.com/ndwarshuis/stratomod

Acknowledgements

We thank Jennifer McDaniel for developing many of the stratifications used as a basis for this work. Certain commercial equipment, instruments, or materials are identified to specify adequately experimental conditions or reported results. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the equipment, instruments, or materials identified are necessarily the best available for the purpose.