Improving variant calling using population data and deep learning

Nae-Chyun Chen[1,‡,*], Alexey Kolesnikov[2], Sidharth Goel[2], Taedong Yun[2], Pi-Chuan Chang[2,†], and Andrew Carroll[2,†,*]

[1]Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA
[2]Google Health, Palo Alto, CA 94304 and Cambridge, MA 02142, USA
corresponding author: cnaechy1@jhu.edu; awcarroll@google.com
[†]These authors contributed equally to this work.
[‡]Work performed while an intern at Google Health.

January 6, 2021

Abstract

Large-scale population variant data is often used to filter and aid interpretation of variant calls in a single sample. These approaches do not incorporate population information directly into the process of variant calling, and are often limited to filtering which trades recall for precision. In this study, we modify DeepVariant to add a new channel encoding population allele frequencies from the 1000 Genomes Project. We show that this model reduces variant calling errors, improving both precision and recall. We assess the impact of using population-specific or diverse reference panels. We achieve the greatest accuracy with diverse panels, suggesting that large, diverse panels are preferable to individual populations, even when the population matches sample ancestry. Finally, we show that this benefit generalizes to samples with different ancestry from the training data even when the ancestry is also excluded from the reference panel.

1 Background

Variant calling [1–3] identifies the positions in an individual genome which differ from a reference or population, and is used to characterize a single sample or build large research cohorts [4, 5]. Variant calling is non-trivial, because of sequencing errors, systematic errors in mapping to repetitive and variable regions [6], and imbalanced sampling of alleles needed to identify a heterozygous variant from a homozygous one.

Variant calling can be improved by jointly genotyping multiple samples together [7–9], but the raw sequence data for a cohort is not always available, and this process is computationally expensive. Instead, large-scale reference panels from a wide range of populations can provide similar information [4, 5]. Recent studies use such information to improve alignment accuracy and reduce biases in alignment [10–12], but there has been little work to incorporate population data with variant calling.

Because far more variants are transmitted than arise de novo, real variants in a population tend to recur at various frequencies [13], while false positives are often either not seen elsewhere in a population, or are seen with a consistent signature [14]. Researchers use this knowledge to filter variant calls, often with rules which lose recall for a gain in precision [15]. More sophisticated machine-learning methods to filter are used in larger cohorts, such as gnomAD, but these also trade recall for precision and also only operate on variant calls and summary information [4].

We reason that including population-level information at an earlier stage in variant calling, when the full read-level data is available, might allow for more effective use of population data. To do this, we adapted DeepVariant [2], which represents BAM information as a multi-dimensional pileup and uses a Convolutional Neural Network (CNN) to call variants. Because DeepVariant learns the features important for variant classification directly from the data, it allows us to feed in the population allele information as an additional channel.

We trained population-aware models and compared them with the default DeepVariant v1.1 models which are agnostic of population information. The population-aware approach reduces the number of errors for all tested datasets, including WGS and WES reads, when using the allele frequencies from 1000Genomes. It also shows stronger error reduction efficacy for lower-coverage read sets. While traditional filtering approaches will increase precision at the expense of recall, we observe improvements to both precision and recall with this method.

When incorporating population data, it is also important for fairness and equity to understand how it changes the accuracy of methods for individuals with ancestries outside of those used in the development of the population resources. It is known that many genomic databases have collected more data for the European population than others [16–18]. We demonstrate that even using frequencies from a genetically distinct population, the population-aware model still performs similarly as the baseline. We find that a reference panel consisting of all ancestries in the 1000 Genomes Project (1000Genomes) outperforms a reference panel with only one of the 1000Genomes population groups, even when that population matches the sample being called. This implies that maximizing the diversity of ancestries in population resources has the potential to improve variant calling for all populations.

The Genome in a Bottle (GIAB) truth sets used to train DeepVariant are from European, Ashkenazi, and Asian ancestry. To assess whether the addition of the reference panel information improves variant calling for populations outside of the popula-

tions represented in training, we use high quality PacBio HiFi [19] data from the Human Genome Structural Variation Consortium for an individual of Puerto Rican ancestry as an evaluation set. We show that an Illumina model using the reference panel has superior concordance with the highly accurate PacBio HiFi variant calls compared to an Illumina model without the reference panel.

2 Results

2.1 Population information improves DeepVariant performance

DeepVariant converts input from a BAM file into a pileup image with 6 channels, representing 1) bases, 2) base qualities, 3) mapping quality, 4) strand, 5) supports variant, and 6) base differs from reference. We modified DeepVariant v1.1 to take an additional input channel, the allele-frequency (AF) of the variant [20]. We trained DeepVariant models with and without the AF channel with the testing samples held out.

We first compared the whole-genome sequencing (WGS) variant calling accuracy for sample HG003, sequenced with 35x coverage from the PrecisionFDA v2 Truth Challenge [21], using the latest GIAB v4.2.1 truth set [22] (Figure 1). HG003 is not used in the training of these DeepVariant models, and so acts as an independent holdout to evaluate their quality.

The population-aware model has superior accuracy than default DeepVariant v1.1 in both precision and recall for both types of variants. It has an overall error reduction of 1514 (4.8%). For SNPs, the error rate (defined as 1-F1 score) decreases from 0.0041 to 0.0038; for indels, the error rate decreases from 0.0044 to 0.0043. Notably, the population-aware model improves SNP false discovery rate (FDR, defined as 1-precision) from 0.0019 to 0.0015, equivalent to an error reduction of 1,096 (17.7%) variants.

We then down-sampled the HG003 reads from 35x to 21x to evaluate the performance of the models with lower-coverage datasets. The population-aware method demonstrates a larger improvement in accuracy over default DeepVariant v1.1 by reducing 5,119 (9.5%) overall errors. The error rate decreases from 0.0062 to 0.0056 for SNPS, and 0.0124 to 0.0113 for indels. Similar to using the 35x read set, the population-aware model shows the strongest improvement to reduce false-positive SNPs, reducing FDR from 0.0040 to 0.0031, equivalent to 3,015 (22.5%) errors.

We further evaluated the performance of the models using two whole-exome sequencing (WES) datasets from a recently released set of genome and exome data [23] (Figure 2). For both WES datasets, the population-aware model outperforms DeepVariant v1.1 in overall number of errors. It has an overall error reduction of 53 (9.9%) for the IDT dataset, and 13 (6.5%) for the Oslo dataset. It has a slightly higher rate for SNPs for the Oslo dataset, from 0.00087 to 0.00092, but the difference is smaller than the gain for indels for that dataset. The population-aware model tends to have a larger lead on precision for both types of variants compared to the baseline, but still has similar or better recall.

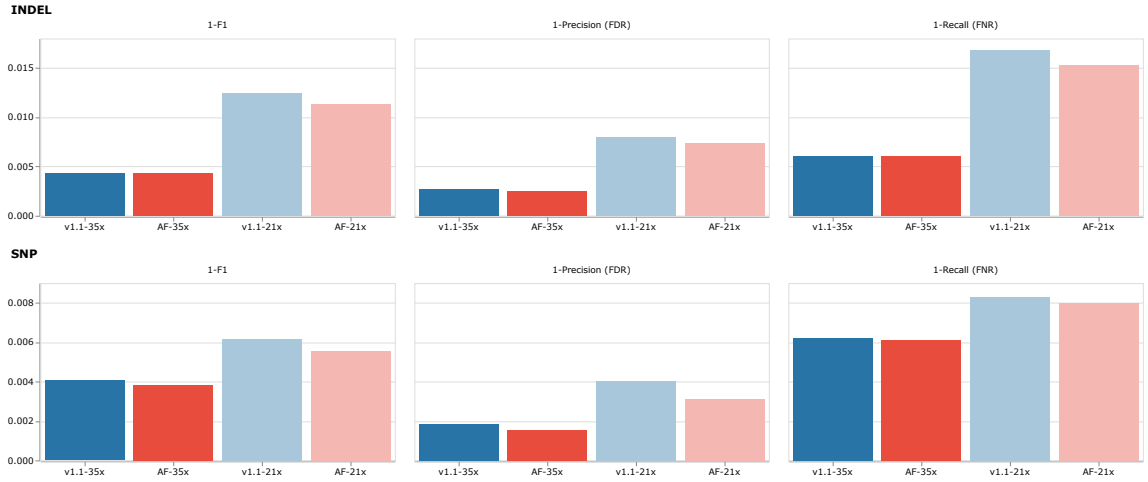

Figure 1: WGS variant calling error rates for HG003. All results are evaluated using the GIAB v4.2.1 truth set in the high-confidence regions. *v1.1*: DeepVariant v1.1; *AF*: the population-aware model that uses the allele-frequency channel. The column label suffixes show the average coverage of the read sets. Lower values correspond to better accuracy.

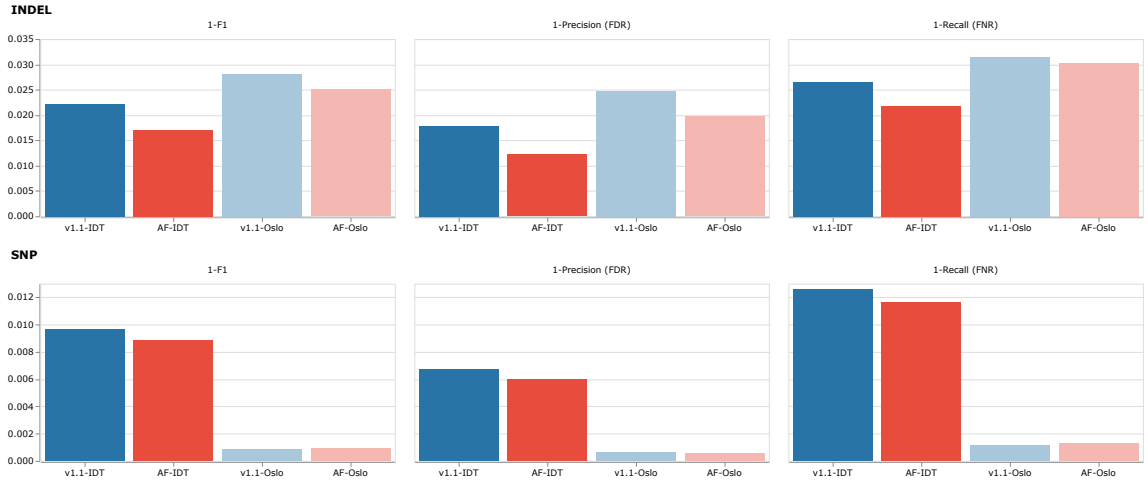

Figure 2: WES variant calling error rate for HG003. The IDT results (*"*-IDT"*) are GRCh38-based and evaluated using the GIAB v4.2.1 truth set; the Oslo datasets (*"*-Oslo"*) are GRCh37-based and evaluated using the GIAB v3.3.2 truth set. *v1.1*: DeepVariant v1.1; *AF*: the population-aware model that uses the allele-frequency channel. Lower values correspond to better accuracy.

2.2 Model-specific errors for population-aware models

Intuitively, population information helps DeepVariant decide whether to make a call based on the commonness of a variant, especially for cases where the variant calling confidence levels are low. With a population-aware model, a variant caller should be more likely to make a positive variant call for a candidate with high allele frequency, and is less likely to make a call when seeing a rare candidate variant.

To understand the influence of allele frequencies in the model, we design an analysis framework to compare a population-agnostic model with a population-aware model. We call this a model-specific error analysis. We stratify the errors into three groups: population-resolved, population-induced and common. The population-resolved variants are called correctly with the allele frequency model, but called incorrectly when using the baseline model. We say such errors are "rescued" by population information. The population-induced errors are specific to the population-aware model, i.e. they are induced by the extra features. The common group contains errors called by both models. The common errors are viewed as ones more difficult to solve without major changes in the data processing pipeline, such as variant caller, upstream computational methods, or sequencing technology. Thus, in this analysis we focus on the first two groups. For simplicity, we only considered bi-allelic calls in this analysis, which are the majority of overall errors.

We used the 35x HG003 WGS dataset to perform the model-specific error analysis. After extracting model-specific erroneous calls, we matched the calls with the 1000Genomes variants to obtain associated allele frequencies. We first examined the relationship between allele frequency (AF) and variant allele fraction (VAF), which is the fraction of reads supporting an alternate allele in a given sample, of each false-positive call. There is an observable distinction between the population-induced group and the population-resolved group in the VAF-AF plots (Figure 3, left and middle panels). Among the population-resolved false-positive errors, more than two third (71.0%) are uncommon (allele frequency $\leq$ 5%) among the 1000Genomes samples, whereas there are only 11.4% uncommon variants for population-induced false positives. This observation supports the hypothesis that the population-aware model uses allele frequency to adjust its variant calls.

We then investigated bi-allelic false-negative errors, as shown in the right panel in Figure 3. Variant allele fraction for false negatives are not always available because many false negatives are not identified as a variant candidate due to reasons including low read coverage, incorrect mapping or insufficient sensitivity in variant candidate discovery. Thus, we only evaluated the allele frequency distribution for false negatives. We noticed a significant difference in the number of common variants (with greater than 5% allele frequency). Among all population-resolved false negatives, 94.6% (1,683 out of 1,780) are common variants. For population-induced false negatives, 59.2% (607 out of 883) are uncommon. The model-specific analysis highlights the difference of the DeepVariant models with or without the AF channel. With the additional population information, DeepVari-

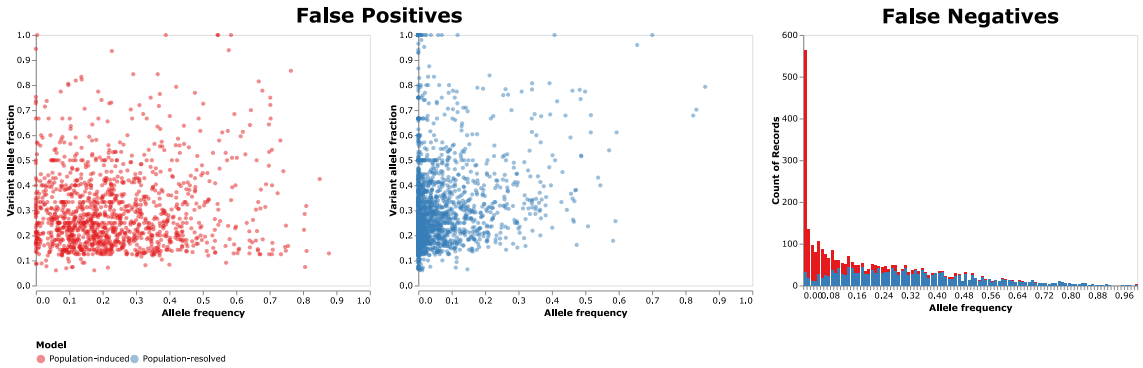

Figure 3: Errors specific to a population-agnostic model (in blue) and a population-aware model (in red) using 35x HG003 WGS data.

ant is capable of adjusting the calls according to the commonness of a variant and shows improvements in both precision and recall.

2.3 Performance on zero-frequency variants

A potential concern for population-aware variant calling models is increasing false negative rate for novel alleles. Since it is not trivial to define a set of truly novel variants in the 1000 Genomes Project, we extracted variants with zero allele frequency to investigate the impact when population information is included in a variant calling model. Using the GIAB v4.2.1 truth set, there are 32,256 (1.0%) SNPs and 3,193 (0.6%) indels that have zero allele frequency for sample HG003. We then use the zero-frequency variant set to evaluate recall of actual variant calls using hap.py [3].

We observed that the recall on zero-frequency variants underperforms the rest using all DeepVariant models, regardless of variant types and whether to utilize population information. With 35x reads, the false-negative rate (FNR, or 1-recall) of the population-agnostic model is 0.1855 for SNPs and 0.2474 for indels (Figure 4). The FNRs further increase to 0.1945 for SNPs and 0.2643 for indels when using the population-aware model. When using 21x reads, the drop in accuracy gets larger for both types of variants. This is consistent with our analysis that the population-aware DeepVariant model requires stronger evidence (higher-quality pileup images) to call zero-frequency variants, thus reducing recall. Further, the population information has a stronger influence in variant calling for low-coverage datasets. Despite the disadvantages, the negative impact on zero-frequency variants is small compared to overall error reduction.

To better understand the zero-frequency variants, we called variants using the DeepVariant PacBio model with the PrecisionFDA v2 35x HG003 reads set sequenced with the PacBio HiFi technology [21]. The FNRs for the zero-frequency variants improve to 0.0481

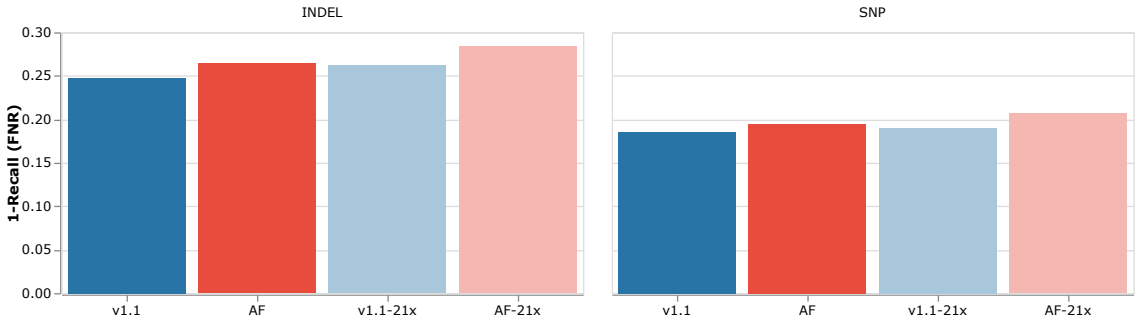

Figure 4: The false negative rate (FNR) of zero-frequency variants for HG003 with different models. Lower values correspond to better accuracy.

for SNPs and 0.0868 for indels. The large difference in recall/FNR indicates that many of the zero-frequency variants are hard to genotype using Illumina reads, and may not be novel mutations relative to samples in reference panels. In the future, reference panels utilizing high-quality long reads will likely provide better allele frequency estimates and improve the population-aware model performance.

2.4 Assessing biases using different 1000Genomes populations

It is important to understand if the inclusion of population information reduces DeepVariant's performance for populations that are not well represented, especially when they have a large genomic difference with the reference panel. We first note that Ashkenazi Jewish, the ethnicity of the HG003, is not among the 26 ethnicities collected by 1000Genomes. Using a testing sample not in the reference panel reduces the risk of bias. Second, we ran inference on the population-aware model using reference panels of alleles frequencies. We split the 1000Genomes sample into five groups based on the superpopulation labels (African, AFR; Admixed American, AMR; East Asian, EAS; European, EUR; South Asian, SAS) and calculated allele frequencies for each super-population. We show that all population-aware approaches outperform for SNPs but underperform for indels when evaluated using HG003 (Figure 5). When considering the overall number of errors, only the model inferred with EAS frequencies calls more errors than the baseline, but the deficit (494, or 1.6%) is small.

We also compared the performance of using different superpopulation frequencies and observed a correlation between variant calling accuracy and the distance between the tested sample and ethnicity groups. According to the principal component (PC) analysis performed by gnomAD v3 [4], Ashkenazi Jewish is closer to the European populations and is farther from East Asian and African in the PC1-PC2 space. We observed that using frequencies from a genetically closer population usually resulted in higher variant calling

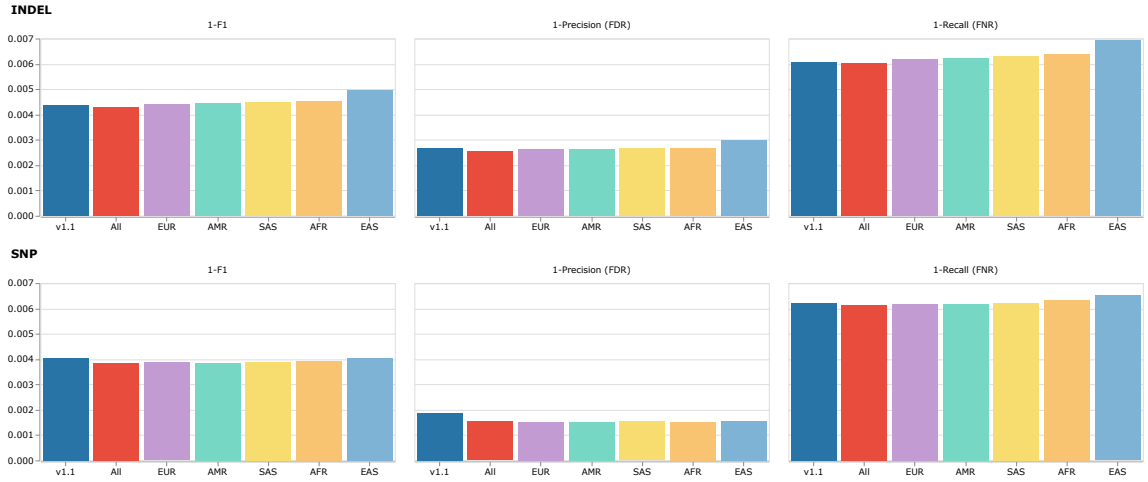

Figure 5: Variant calling accuracy when inferring 35x Illumina reads from HG003 using default DeepVariant v1.1 (*v1.1*), allele frequencies in the entire 1000Genomes (*All*) and five 1000Genomes superpopulations (*EUR*, *AMR*, *SAS*, *AFR* and *EAS*). Lower values correspond to better accuracy.

accuracy. Using EUR frequencies outperforms using other population frequencies, only falling behind using the entire 1000Genomes. On the other hand, using EAS frequencies results in the highest numbers of errors among all population-aware methods.

We point out that using 1000Genomes frequencies from all populations results in the lowest number of errors among all population-aware results, suggesting an advantage to using a diverse population than finding a genetically similar group. This finding echoes our previous statement that we anticipate the population-aware variant calling model to improve further with larger-scaled and more diverse population callsets.

2.5 Silver-standard truth set for HG00733

Genome-in-a-bottle (GIAB) truth variant sets provide gold standards to benchmark variant callers, but until now there are only three samples (HG002-HG003-HG004, the Ashkenazi trio) with curated calls in difficult-to-map regions added in the v4.2.1 release [22]. Further, the samples are from the same ancestry, making it challenging to perform a generalized benchmarking considering the genetic diversity of the human population. To deal with this difficulty, it is desirable to have other high-quality variant sets from non-GIAB samples, preferably from ancestries not covered by GIAB. Thus, we called variants using the DeepVariant PacBio model with 32x high-coverage PacBio HiFi reads [24] for HG00733, a Puerto Rican (labelled as PUR under the AMR superpopulation in 1000Genomes) sample. The DeepVariant PacBio model has a SNP F1 score higher than

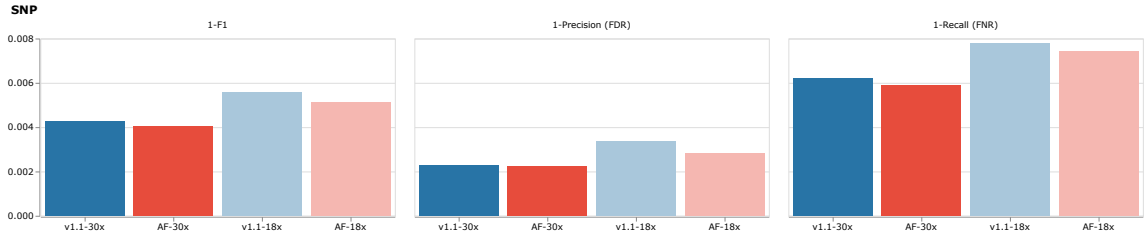

Figure 6: Variant calling results when evaluated using HG00733 data, compared to the PacBio-DeepVariant silver-standard truth set. Lower values correspond to better accuracy.

99.9% and is one of the most accurate models using PacBio HiFi data [22]. We used the DeepVariant HG00733 PacBio SNP calls as a "silver-standard" truth set and benchmarked the performance for models using Illumina reads. We excluded the Puerto Rican population when calculating allele frequencies to avoid biases in favor of the population-aware models. We used 30x Illumina WGS reads sequenced by the New York Genome Center to test all HG00733 models. Because the 1000Genomes has a collection of PUR samples, we excluded all PUR samples and re-calculated allele frequencies for both 1000Genomes and the AMR superpopulation.

The population-aware model has a lower SNP error rate (0.0041 vs. 0.0043), FDR (0.0022 vs. 0.0023) and FNR (0.0059 vs. 0.0062) than the baseline for HG00733 (Figure 6). The number of SNP errors is reduced by 1,353 (4.82%). Similar to the finding using HG003, the population-aware model performs strongly with a down-sampled (18x) read set. The error rate for the 18x read set is reduced from 0.0056 to 0.0051, and the SNP error reduction is 3,145 (8.5%). We also tested the model using different superpopulation frequencies (Figure 7). All but the EAS population-aware model has lower SNP error rates than the baseline. When inferred using the EAS allele frequencies, the SNP error rate increased from 0.0043 to 0.0044, equivalent to 878 (3.1%) more errors. All population-aware models, including EAS, outperform the baseline on FDR and only EAS has a higher FNR than the baseline (0.0066 vs. 0.0062).

3 Discussion

We designed a new population-aware DeepVariant model which can incorporate both base- and read-level information with the population information. We find that population-aware models reduce error rates by 4.9% for WGS and 6.5-9.9% for WES compared to population-agnostic baselines (default DeepVariant v1.1) The relative advantage of the population-aware model increases at lower coverage (4.9% reduction at 35x and 9.5% at 21x). The increased accuracy at lower coverage suggests that population information is

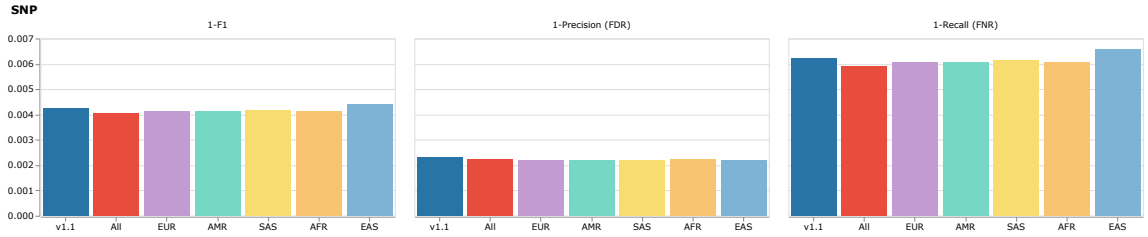

Figure 7: Number of SNP errors when evaluated using 30x WGS reads from a Puerto Rican sample HG00733. All models other than *v1.1* are population-aware, inferred using alleles frequencies from different populations. Lower values correspond to better accuracy.

most valuable in difficult examples, where read-level information alone may not be sufficient for confident calling. In population sequencing projects, this finding could be relevant to the question of whether to sequence more individuals at lower coverage, or fewer at a high coverage. When sequencing for a species without a reference panel, it is possible that sequencing more, diverse individuals at lower coverage could still retain comparable accuracy to traditional methods which do not incorporate population information in calling.

We evaluate potential biases introduced by population information in variant calling by comparing population-aware models that use allele frequencies from different 1000Genomes superpopulation. This experiment simulates a scenario where the tested sample is genetically distinct from the reference panel. Only one population-aware method (inferred with EAS frequencies) underperforms the baseline in total number of errors, but with a small deficit. Furthermore, using allele frequencies calculated from the entire 1000Genomes outperforms population-specific methods. This finding implies that a diverse population can provide more benefits than using a homogeneous one, even when the homogeneous population is more genetically similar with the tested sample. This finding may inform efforts to build population or country-specific resources. Increasing the number of samples for a given population will improve accuracy for that population, but the inclusion of samples from diverse populations will also improve the resource. We believe that the accuracy of the population-aware model can further improve with a larger and more diverse population callset in the future, reinforcing the benefit of collaboration between nation-scale efforts.

We provide an additional "silver-standard" SNP set for a Purto Rican sample, HG00733, a population not present in the labeled training data. We used high-coverage PacBio HiFi reads and an accurate DeepVariant PacBio model to generate this high-quality call set. This method can provide high-confidence SNP calls for non-GIAB samples and increase population diversity when assessing variant calling results. Similar to the results using HG003 data, we show that the proposed model has strong performance compared to the

baseline, and only suffers slight loss of accuracy when inferred using a distinct population. When more high-coverage PacBio HiFi data become available in the future, the high-quality calls generated by DeepVariant can provide a more diversified dataset for variant calling benchmarking and downstream analysis.

Despite greater overall accuracy, we note that the population-aware model underperforms on variants with zero allele frequencies in 1000Genomes. Although the disadvantage is small compared to the overall gain, this results suggests that the decision of whether to use population-aware models should consider the end goal. If reducing potential false positives is a larger concern, the use of a population-aware method could be recommended, but if the goal is to maximize recall of rare or novel variants, traditional methods could be preferred. We also notice that all tested Illumina models performed poorly on the zero-frequency variants, regardless of using population information or not. By analyzing the variants with PacBio reads, we point out many zero-frequency variants in 1000Genomes located in difficult-to-map regions, but likely not genetically novel in the population. This suggests that the power of population-aware methods should increase as large panels of long-read population data become available.

4 Methods

4.1 Training the model

We trained the model following the procedure described in [2], with additional Illumina WGS datasets included [23]. Variants in chromosomes 1 to 19 are used as the training examples, and those in chromosome 21 and 22 are used for tuning. Variants in chromosome 20 are never used in the training process.

4.2 Datasets

The model is evaluated using the GIAB v4.2.1 truth set for HG003 across whole genomes [22]. We also generated another high-quality SNP set using DeepVariant v0.10 and HG00733 PacBio HiFi data [24] across the whole genome. We used the intersection of high-confidence regions of HG002, HG003, and HG004 (GIAB v4.2.1) as the high-confidence regions for the HG00733 SNP set. The read sets used for experiments are listed in Table 1 and the read sets for supporting experiments are provided in Table 2.

4.3 Allele matching algorithm

When incorporating population information in DeepVariant, we need to match a variant candidate with a cohort variant. However, this is not a straightforward task since a variant can be represented in multiple formats [3, 26]. A common approach is to normalize variants, such as using `bcftools norm` [27], but that's not sufficient for complicated

Table 1: Testing datasets.

Sample	Ethnicity	Truth variant	Dataset
HG003	Ashkenazi Jewish	v4.2.1 (GRCh38)	35x Illumina WGS [22] 100x Illumina WES [23]
HG003	Ashkenazi Jewish	v3.3.2 (GRCh37)	300x Illumina WES [25]
HG00733	Puerto Rican	DeepVariant v0.10 PacBio SNP calls (GRCh38)	30x Illumina WGS (NYGC)

Table 2: Other datasets used in this study.

Sample	Ethnicity	Dataset
HG003	Ashkenazi Jewish	35x PacBio HiFi [22]
HG00733	Puerto Rican	32x PacBio HiFi [24]

cases. We designed an algorithm that constructed local haplotypes and performed precise allele matching (Figure 8). The algorithm starts with querying all cohort variants V_C overlapped with a window $[start_v, end_v)$, where startv and endv are the starting and ending positions of a variant candidate v respectively. The queried cohort variants and the candidate variant form set $V \equiv v \cup VC$. Then the window is extended to the smallest starting position and the largest ending position within V, as $[start_V, end_V)$, where $start_V \equiv min(start_u) \forall u \in V$ and $end_V \equiv max(end_w) \forall w \in V$. Local reference haplotype is queried from the reference genome in window $[start_V, end_V]$. For each variant allele in V, its allele haplotype is updated in this window. If there's a perfect match between a cohort allele haplotype and a candidate allele haplotype, the allele frequency of the cohort allele is added to an allele frequency dictionary, using the alternate allele of the candidate variant as key. Afterwards, DeepVariant looks up the dictionary when processing reads overlapped with the candidate variant.

4.4 Allele frequency channel for DeepVariant

To make full advantages of the CNN-based classifier of DeepVariant, allele frequencies need to be encoded in pileup images. We apply a logarithmic transformation to gain resolution for low-frequency signals. For each variant candidate, an additional *allele frequency channel* is added to the pileup image. In this channel, a read is colored by the transformed frequency of its allele at the variant candidate position. A read can carry multiple alternate alleles with different frequencies, so its color intensity may vary across pileup images, where the variant candidates differ. An alternative method to encode allele frequencies is to include the information as features in the fully-connected layers [28], but this approach sacrifices the capability to incorporate allele frequencies with base- and read-level information and thus is not adopted.

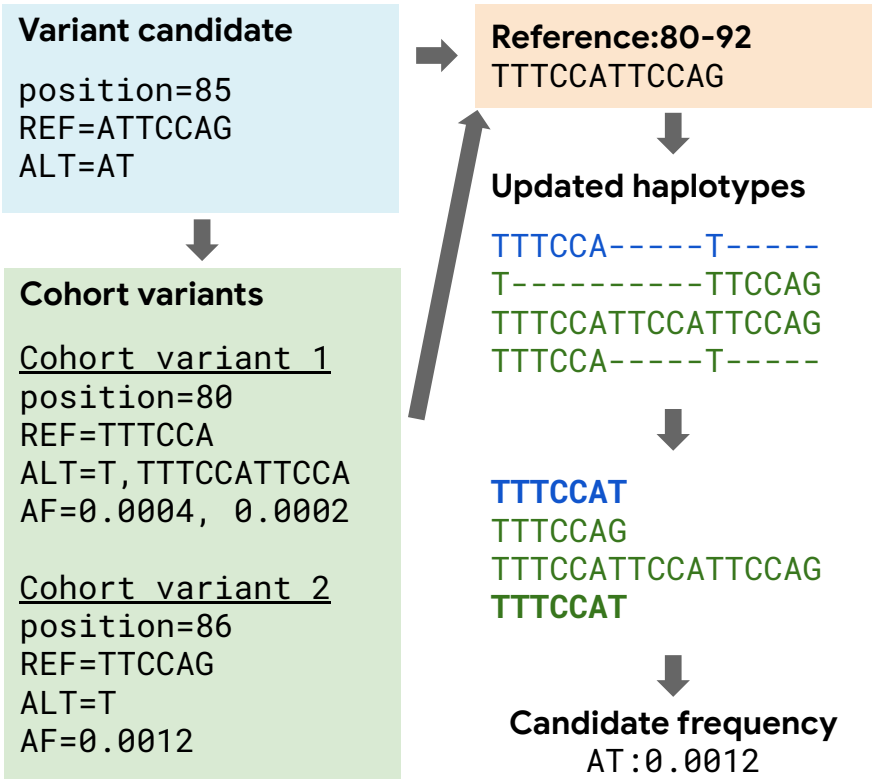

Figure 8: An example for the allele matching algorithm. This algorithm first queries cohort variants overlapped with the variant candidate. These cohort variants and the candidate determine the window where haplotypes are updated. The frequencies of matched allele haplotypes are then updated for the variant candidate as a dictionary. In this diagram, haplotypes are updated with dashes to keep sequenced aligned for better visualization. In practice, dash-free haplotypes are generated by the allele matching algorithm.

To enable the allele frequency channel, users need to enable flag `--use_allele_frequency` and provide DeepVariant cohort variants in VCF format with flag `--population_vcfs`.

4.5 Model-specific error analysis

We compared actual variant calls with GIAB v4.2.1 truth variants using `bcftools isec`. Variants specific to actual calls are regarded as false positives, and those specific to the truth set are regarded as false negatives. We generated the false-positive and false-negative sets for two models, and then applied `bcftools isec` again to obtain model-specific false positives and false negatives. For both sets, we applied the allele matching algo-

rithm to obtain allele frequencies for the variants. For the false-positive sets, we extracted variant allele fractions from the VCF files generated by DeepVariant.

4.6 1000Genomes frequencies from the DeepVariant-GLnexus pipeline

We used the 1000Genomes reference panel generated with the DeepVariant-GLnexus pipeline (v3) [8] for all population-aware experiments, including training and inferring the models. We fill the missing genotypes with the reference genotypes with `bcftools +missing2ref` to make sure all variants have the same denominator.

5 Availability of data and materials

The DeepVariant source code is available at `https://github.com/google/deepvariant` under the BSD-3-Clause License. The PacBio-based HG00733 SNP set is available at `https://console.cloud.google.com/storage/browser/brain-genomics-public/research/allele_frequency/HG00733_SNP_set`. The pre-trained population-aware DeepVariant models are available at `https://console.cloud.google.com/storage/browser/brain-genomics-public/research/allele_frequency/pretrained_model_WGS` (WGS) and `https://console.cloud.google.com/storage/browser/brain-genomics-public/research/allele_frequency/pretrained_model_WES` (WES). The VCF files used in this study are available at `https://console.cloud.google.com/storage/browser/brain-genomics-public/research/cohort/1KGP/cohort_dv_glnexus_opt/v3_missing2ref` (GRCh38) and `https://console.cloud.google.com/storage/browser/brain-genomics-public/research/cohort/1KGP/cohort_dv_glnexus_opt/v3_GRCh37_missing2ref` (GRCh37).

6 Ethics approval and consent to participate

Not applicable.

7 Consent for publication

Not applicable.

8 Competing interests

AK, SG, TY, PC and AC are employees of Google LLC and own Alphabet stock as part of the standard compensation package. This study was funded by Google LLC.

9 Funding

All compute resources used in this work were provided by Google, LLC.

AK, SG, TY, PC and AC are full-time, salaried employees of Google, LLC. NC contributed to this work as a salaried intern of Google, LLC.

10 Acknowledgments

We thank Babak Alipanahi, Gunjan Baid, Daniel Cook, Alexander D'Amour, Hojae Lee, Cory McLean, Maria Nattestad and other colleagues at Google for their feedback on this manuscript and the project in general.

The HG00733 Illumina data were generated at the New York Genome Center with funds provided by NHGRI Grant 3UM1HG008901-03S1.

11 Authors' contributions

NC, AK, PC and AC designed the method. NC, AK and PC implemented the software. NC and PC performed the experiment. NC, AK, SG, TY, PC and AC analyzed the results. NC, PC and AC wrote the manuscript. All authors read and approved the final manuscript.

References

1. DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43,** 491 (2011).
2. Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nature biotechnology* **36,** 983–987 (2018).
3. Krusche, P., Trigg, L., Boutros, P. C., Mason, C. E., Francisco, M., Moore, B. L., Gonzalez-Porta, M., Eberle, M. A., Tezak, Z., Lababidi, S., *et al.* Best practices for benchmarking germline small-variant calls in human genomes. *Nature biotechnology* **37,** 555–560 (2019).
4. Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581,** 434–443 (2020).

5. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

6. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30,** 2843–2851 (2014).

7. Lin, M. F., Rodeh, O., Penn, J., Bai, X., Reid, J. G., Krasheninina, O. & Salerno, W. J. GLnexus: joint variant calling for large cohort sequencing. *BioRxiv,* 343970 (2018).

8. Yun, T., Li, H., Chang, P.-C., Lin, M. F., Carroll, A. & McLean, C. Y. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *bioRxiv* (2020).

9. Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv,* 201178 (2017).

10. Chen, N.-C., Solomon, B., Mun, T., Iyer, S. & Langmead, B. Reducing reference bias using multiple population reference genomes. *BioRxiv* (2020).

11. Rautiainen, M. & Marschall, T. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome biology* **21,** 1–28 (2020).

12. Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature biotechnology* **36,** 875–879 (2018).

13. Witherspoon, D. J., Wooding, S., Rogers, A. R., Marchani, E. E., Watkins, W. S., Batzer, M. A. & Jorde, L. B. Genetic similarities within and between human populations. *Genetics* **176,** 351–359 (2007).

14. Abramovs, N., Brass, A. & Tassabehji, M. Hardy-Weinberg Equilibrium in the Large Scale Genomic Sequencing Era. *Frontiers in Genetics* **11,** 210 (2020).

15. Pedersen, B. S., Brown, J. M., Dashnow, H., Wallace, A. D., Velinder, M., Tvrdik, T., Mao, R., Best, H. D., Bayrak-Toydemir, P. & Quinlan, A. R. Effective variant filtering and expected candidate variant yield in studies of rare human disease. *BioRxiv* (2020).

16. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell* **177,** 26–31 (2019).

17. Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M. & Daly, M. J. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics* **51,** 584–591 (2019).

18. McGuire, A. L., Gabriel, S., Tishkoff, S. A., Wonkam, A., Chakravarti, A., Furlong, E. E., Treutlein, B., Meissner, A., Chang, H. Y., López-Bigas, N., *et al.* The road ahead in genetics and genomics. *Nature Reviews Genetics* **21,** 581–596 (2020).

19. Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature biotechnology* **37,** 1155–1162 (2019).

20. Carroll, A. & Chang, P.-C. *Improving the Accuracy of Genomic Analysis with DeepVariant 1.0* `https://ai.googleblog.com/2020/09/improving-accuracy-of-genomic-analysis.html`. 2020. (accessed: 2020-12-11).

21. Olson, N. D., Wagner, J., McDaniel, J., Stephens, S. H., Westreich, S. T., Prasanna, A. G., Johanson, E., Boja, E., Maier, E. J., Serang, O., *et al.* precisionFDA Truth Challenge V2: Calling variants from short-and long-reads in difficult-to-map regions. *bioRxiv* (2020).

22. Wagner, J., Olson, N. D., Harris, L., Khan, Z., Farek, J., Mahmoud, M., Stankovic, A., Kovacevic, V., Wenger, A. M., Rowell, W. J., *et al.* Benchmarking challenging small variants with linked and long reads. *BioRxiv* (2020).

23. Baid, G., Nattestad, M., Kolesnikov, A., Goel, S., Yang, H., Chang, P.-C. & Carroll, A. An Extensive Sequence Dataset of Gold-Standard Samples for Benchmarking and Development. *bioRxiv.* eprint: `https://www.biorxiv.org/content/early/2020/12/11/2020.12.11.422022.full.pdf`. `https://www.biorxiv.org/content/early/2020/12/11/2020.12.11.422022` (2020).

24. Porubsky, D., Ebert, P., Audano, P. A., Vollger, M. R., Harvey, W. T., Marijon, P., Ebler, J., Munson, K. M., Sorensen, M., Sulovari, A., *et al.* Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nature Biotechnology.* ISSN: 1546-1696. `https://doi.org/10.1038/s41587-020-0719-5` (Dec. 2020).

25. Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C. E., Alexander, N., *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific data* **3,** 1–26 (2016).

26. Sun, C. & Medvedev, P. VarMatch: robust matching of small variant datasets using flexible scoring schemes. *Bioinformatics* **33,** 1301–1308 (2017).

27. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987–2993 (2011).

28. Yi, R., Chang, P.-C., Baid, G. & Carroll, A. Learning from Data-Rich Problems: A Case Study on Genetic Variant Calling. *arXiv preprint arXiv:1911.05151* (2019).