NanoCaller for accurate detection of SNPs and indels in difficult-to-map regions from long-read sequencing by haplotype-aware deep neural networks

Umair Ahsan[1,#], Qian Liu[1,#], Li Fang[1], Kai Wang[1,2*]

[1] Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

[2] Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

[#] These authors contributed equally to this work.

[*] To whom correspondence should be addressed. Email: wangk@email.chop.edu

Abstract

Variant (SNPs/indels) detection from high-throughput sequencing data remains an important yet unresolved problem. Long-read sequencing enables variant detection in difficult-to-map genomic regions that short-read sequencing cannot reliably examine (for example, only ~80% of genomic regions are marked as "high-confidence region" to have SNP/indel calls in the Genome In A Bottle project); however, the high per-base error rate poses unique challenges in variant detection. Existing methods on long-read data typically rely on analyzing pileup information from neighboring bases surrounding a candidate variant, similar to short-read variant callers, yet the benefits of much longer read length are not fully exploited. Here we present a deep neural network called NanoCaller, which detects SNPs by examining pileup information solely from other nonadjacent candidate SNPs that share the same long reads using long-range haplotype information. With called SNPs by NanoCaller, NanoCaller phases long reads and performs local realignment on two sets of phased reads to call indels by another deep neural network. Extensive evaluation on 5 human genomes (sequenced by Nanopore and PacBio long-read techniques) demonstrated that NanoCaller greatly improved performance in difficult-to-map regions, compared to other long-read variant callers. We experimentally validated 41 novel variants in difficult-to-map regions in a widely-used benchmarking genome, which cannot be reliably detected previously. We extensively evaluated the run-time characteristics and the sensitivity of parameter settings of NanoCaller to different characteristics of sequencing data. Finally, we achieved the best performance in Nanopore-based variant calling from MHC regions in the PrecisionFDA Variant Calling Challenge on Difficult-to-Map Regions by ensemble calling. In summary, by incorporating haplotype information in deep neural networks, NanoCaller facilitates the discovery of novel variants in complex genomic regions from long-read sequencing data.

Introduction

Single-nucleotide polymorphisms (SNPs) and small insertions/deletions (indels) are two common types of genetic variants in human genomes. They contribute to genetic diversity and critically influence phenotypic differences, including susceptibility to human diseases. The detection (i.e. “calling”) of SNPs and indels is thus a fundamentally important problem in using the new generations of high-throughput sequencing data to study genome variations and genome functions. A number of methods have been designed to call SNPs and small indels on Illumina short-read sequencing data. Short reads are usually 100-150 bp long and have per-base error rate less than 1%. Variant calling methods on short reads, such as GATK [1] and FreeBayes [2], achieved excellent performance to detect SNPs and small indels in genomic regions marked as “high-confidence regions” in various benchmarking tests[3-5]. However, since these methods were developed for short-read sequencing data with low per-base error rates and low insertion/deletion errors, they do not work well on long-read sequencing data with high error rates. Additionally, due to inherent technical limitations of short-read sequencing, the data cannot be used to call SNPs and indels in complex or repetitive genomic regions; for example, only ~80% of genomic regions are marked as “high-confidence region” to have reliable SNP/indel calls in the Genome In A Bottle (GIAB) project, suggesting that ~20% of the human genome is inaccessible to conventional short-read sequencing technologies to find variants reliably.

Oxford Nanopore [6] and Pacific Biosciences (PacBio) [7] technologies are two leading long-read sequencing platforms, which have been rapidly developed in recent years with continuously decreased costs and continuously improved read length, in comparison to Illumina short-read sequencing technologies. Long-read sequencing techniques can overcome several challenging issues that cannot be solved using short-read sequencing, such as calling long-range haplotypes, identifying variants in complex genomic regions, identifying variants in coding regions for genes with many pseudogenes, sequencing across repetitive regions, phasing of distant alleles and distinguishing highly homologous regions [8]. To date, long-read sequencing techniques have been successfully used to sequence genomes for many species to powerfully resolve various challenging biological problems such as de novo genome assembly [9-13] and SV detection [14-19]. However, the per-base accuracy of long reads is much lower with raw base calling errors of 3-15% [20] compared with short-read data. The high error rate challenges widely-used variant calling methods (such as GATK [1] and FreeBayes [2]), which were previously designed for Illumina short reads and cannot handle reads with higher error rates. It is also worth noting that (1) HiFi reads after circular consensus sequencing on PacBio long-read sequencing [21] or similar methods on the Nanopore platform can potentially improve the detection of SNPs/indels by adapting existing short-read variant callers, due to its much lower per-base error rates. However, HiFi reads would substantially increase sequencing cost given the same base output, so it may be more suitable for specific application scenarios such as capture-based sequencing or amplicon sequencing. As

more and more long-read sequencing data becomes available, there is an urgent need to detect SNPs and small indels to take the most advantage of long-read data.

Several recent works aimed to design accurate SNP/indel callers on long-read sequencing data using machine learning methods, especially deep learning-based algorithms. DeepVariant [22] is among the first successful endeavor to develop a deep learning variant caller for SNPs and indels across different sequencing platforms (i.e. Illumina, PacBio and Nanopore sequencing platforms). In DeepVariant, local regions of reads aligned against a variant candidate site were transformed into an image representation, and then a deep learning framework was trained to distinguish true variants from false variants that were generated due to noisy base calls. DeepVariant achieved excellent performance on short reads as previous variant calling methods did. Later on, Clairvoyante [23] and its successor Clair [24] implemented variant calling methods using deep learning, where the summary of adjacently aligned local genomic positions of putative candidate sites were used as input of deep learning framework. The three deep learning-based methods can work well on both short-read and long-read data, but they do not incorporate haplotype structure in variant calling; these methods consider each SNP separately, while a recent testing [21] with DeepVariant has shown that a phased BAM with haplotype-sorted readscan improve variant calling accuracy because long reads likely from same haplotype benefits neural network learning from an image of read pileup. However, this testing underutilizes the rich haplotype information from long reads, especially when it is explicitly provided in a phased BAM file as input. Moreover, enough variants need to be known beforehand to phase a BAM file. Two recent works have endeavored to improve variant calling by using phasing information from long-readsequencing data. Longshot [25] uses a pair-Hidden Markov Model (pair-HMM) for a small local window around candidate sites to call SNPs on long-read data, and then improves genotyping of called SNPs using HapCUT2 [26] based on the mostly pair of haplotypes given the current variant genotypes. However, Longshot cannot identify indels. The Oxford Nanopore Technologies company also recently released a SNP/indel caller, i.e. Medaka [27], using deep learning on long-read data. Although not published, based on its GitHub repository, Medaka first predicts SNPs from unphased long reads, and then uses WhatsHap to phase reads. Medaka finally makes SNP and indel calling for each group of phased reads. In both methods, mutual information from long-range haplotype SNPs is ignored. In summary, although several methods for variant detection on long-read sequencing data have become available, there may be room in further improving these approaches especially for difficult-to-map regions. We believe that improved SNP/indel detection on long read data will enable widespread research and clinical applications of long-read sequencing techniques.

In this study, we propose a deep learning framework, NanoCaller, which integrates long-range haplotype structure in a deep convolutional neural network to improve variant detection on long-read sequencing data. It only uses haplotype information in SNP calling (without requiring a phased BAM alignment input) and generates input features of a SNP candidate site using long-range heterozygous SNPs sites, that are

fed into a deep convolutional neural network for SNP calling. Please note that those long-range heterozygous SNPs sites can be hundreds or tens of thousands of bases away and NanoCaller does not use local neighboring bases, which is substantially different from DeepVariant, Clairvoyante [23] and its successor Clair [24], as well as Longshot and Medaka where local neighboring bases of SNP sites were used. After that, NanoCaller uses these predicted SNP calls to phase alignment reads with WhatsHap for indel calling. Local multiple sequence alignment of phased reads around indel candidate sites is used to generate consensus sequence and feature inputs for a deep convolutional neural network to predict indel variant zygosity. We assess NanoCaller on several human genomes, HG001 (NA12878), HG002 (NA24385), HG003 (NA24149), HG004 (NA24143) and HX1 with both Nanopore and PacBio long-read data. In particular, for the Ashkenazim trio (HG002, HG003 and HG004), we evaluate NanoCaller in difficult-to-map genomic regions to investigate the unique advantages provided by long reads. Our evaluation demonstrates competitive performance of NanoCaller against existing tools, with particularly improved performance in complex genomic regions which cannot be reliably called on short-read data. NanoCaller is publicly available at https://github.com/WGLab/NanoCaller.

Results

Overview of NanoCaller

NanoCaller (Figure 1) takes the alignment of a long-read sequencing data aligned against a reference genome as input, and generates a VCF file for called SNPs and indels. For SNP calling in NanoCaller, candidate SNP sites are selected according to the specified thresholds for minimum coverage and minimum frequency of alternative alleles (a fraction of them are likely to be false positives given the relaxed thresholds for candidate identification). Long-range haplotype features for the candidate sites (Figure 2) are generated and fed to a deep convolutional network to distinguish true variants from false candidate sites. The predicted SNPs are phased and then used in the identification of indels together with the long-read alignments. Indel candidate sites are selected according to specified minimum coverage and insertion/deletion frequency thresholds applied to each phased read set. Input features for indel candidate sites are generated using multiple sequencing alignment on the set of entire reads and on each set of phased reads (Figure 3). After that, another deep convolutional neural network is designed to determine indel calls, and allele sequence for the indels is predicted by comparing consensus sequences against reference sequence.

The performance of NanoCaller is evaluated on both Oxford Nanopore and PacBio reads, and compared with performances of Medaka (v0.10.0), Clair (v2.0.1), Longshot (v0.4.1), DeepVariant (v.1.0.0) and

WhatsHap (v1.0) with their default parameters for each type of sequencing technology. The evaluation is on the GRCh38 reference genome by default for benchmark variants in high-confidence regions unless stated otherwise. RTG tools (the *vcfeval* submodule) [28] is used to calculate various evaluation measurements, such as precision, recall and F1. For whole genome analysis, we show each variant caller's performance using its recommended quality threshold if available, e.g. NanoCaller, Clair, Longshot and DeepVariant; for Medaka, we calculate the quality score thresholds that give highest F1 score for each genome using *vcfeval,* and use their average as the final quality score cut-off to report results. More discussion of various quality thresholds for NanoCaller and other variant callers is provided in Supplementary Tables S13-S16. In particular, variant calling performance analysis in difficult-to-map genomic regions requires different quality score cut-offs from whole genome analysis due to highly specific error profiles of these difficult-to-map regions. Therefore, we use the average best quality score cut-off (in the same manner as the quality score cut-off is determined for Medaka) for each variant caller in each type of difficult genomic regions. Below, we present performances of three NanoCaller SNP models: NanoCaller1 (trained on HG001 ONT reads), NanoCaller2 (trained on HG002 ONT reads), and NanoCaller3 (trained on HG003 PacBio CLR dataset), and two NanoCaller indel models: ONT indel model (trained on HG001 ONT reads) and PacBio indel model (trained on HG001 PacBio CCS reads). All NanoCaller models were trained using v3.3.2 of GIAB benchmark variant calls using reads aligned to the GRCh38 reference genome.

Evaluation of NanoCaller on Oxford Nanopore Sequencing

Performance on SNP calling

We evaluated NanoCaller on Oxford Nanopore sequencing reads against several existing tools. The settings of compared SNP callers are given below: NanoCaller was trained on v3.3.2 GIAB benchmark variants, and for performance evaluation we usedalternative allele frequency threshold of 0.15 for SNP candidates, the Clair 1_124x ONT model was trained on 124x coverage HG001 ONT reads using v3.3.2 GIAB benchmark variants, Longshot pair-HMM model is not trained on any genome as it estimates parameters during each run, while Medaka '*r941_min_diploid_snp_model*' model is used for testing, which is trained using several bacteria and eukaryotic read datasets and variant call sets. We compared the performance of those method on five genomes: HG001, HG002, HG003, HG004 and HX1 under two testing strategies: cross-genome testing and cross-reference testing.

Cross-genome testing is critical to demonstrate the performance of a variant caller when used in a real-world scenario: the machine learning model of a variant caller is trained on a set of genomes and tested on other genomes. Under this testing strategy, the performance of SNP calling for NanoCaller, together with four other variant callers, Medaka, Clair, Longshot and WhatsHap, is shown in Table 2 and Figure 4 (a), (b) and (c) on Nanopore reads of five genomes. We have evaluated these performances using the

latest available GIAB benchmark variants for each genome, i.e. v3.3.2 for HG001 and v4.2 for the Ashkenazim trio).

On the Ashkenazim trio (HG002, HG003, HG004), NanoCaller has better performance than Clair and Longshot in terms of both recall and F1-score (F1-scores are NanoCaller1: 97.99, 98.88, 98.77%; NanoCaller2: 97.97, 98.88, 98.82% vs Clair: 97.76%, 98.49, 98.55% and Longshot: 98%, 97.87%, 97.84%, on HG002, HG003 and HG004 respectively), with even higher margins for recall (recalls are NanoCaller1: 97.99%, 98.73%, 98.69%; NanoCaller2: 98.03%, 98.61%, 98.61% vs Clair: 97.28%, 97.94%, 98.11% and Longshot: 97.09%, 96.83%, 96.75% on HG002, HG003 and HG004 respectively), whereas Medaka performs better than NanoCaller in this regard. All methods show similar precision on the trio, except that Longshot has much higher precision on HG002 than other methods. We also evaluated the SNP callings for the four methods on older benchmark v3.3.2 and reads basecalled with older Nanopore basecallers for the Ashkenazim trio HG002-4, as shown in Supplementary Tables S2 and S3. We find that all methods have improved performance; this suggests that improved Nanopore basecalling and increasing number of benchmark variants enhance the evaluation of variant calling: as shown in Table 1, the Ashkenazim trio HG002-4 has a larger variant call set than HG001 (370-400k more SNPs per genome than HG001), and a larger high-confidence region which includes more difficult genomic regions (at least 200mbp larger than HG001 and covering an extra 7% of the reference genome).

We also show that the performance of NanoCaller SNP models is independent of the reference genome used. Under this cross-reference testing, we evaluate NanoCaller1 model (trained on reads aligned against GRCh38) on HG002 ONT reads aligned to both GRCh38 and GRCh37 reference genomes, and calculate the performance using v3.3.2 benchmark variants for HG002, the latest version for which we have benchmark variant calls for GRCh37. For reads aligned to GRCh38, we obtained a precision of 97.49%, a recall of 97.97% and an F1 score of 97.73%, whereas for reads aligned to GRCh37, we obtain a precision of 97.88%, a recall of 98.02% and an F1 score of 97.95%. The competitive performance on GRCh38 and GRCh37 indicates that NanoCaller could be used on alignment generated by mapping to different reference genomes.

Performance on SNP calling in difficult-to-map genomic regions

We further demonstrate that NanoCaller has a unique advantage in calling SNPs in difficult-to-map genomic regions. To have an unbiased testing, we evaluated NanoCaller1 SNP model (trained on HG001 ONT reads with v3.3.2 benchmark variants) on ONT reads of the three genomes of the Ashkenazim trio together with other variant callers, while v4.2 of GIAB benchmark variants for the trio were used for testing with more exhaustive list of true variants and high confidence intervals in difficult-to-map genomic regions. Difficult-to-map genomic regions here are defined by GA4GH Benchmarking Team and the

Genome in a Bottle Consortium, and downloaded as BED files from GIAB v2.0 genome stratification. These regions contain all tandem repeats, all homopolymers >6bp, all imperfect homopolymers >10bp, all low mappability regions, all segmental duplications, GC <25% or >65%, Bad Promoters, and other difficult regions such Major Histocompatibility Complex. We intersected the BED files with high-confidence intervals for each genome in the trio and evaluated SNP performance in the intersected regions. As shown in Table 2, each genome has at least 600k SNPs in the intersection of difficult-to-map regions and high-confidence intervals, which is a significant fraction (18-19%) of all SNPs in the high-confidence regions. The evaluation on these SNPs is shown in Figure 4 (d) and Table 2: NanoCaller1 (F1 score 95.60%, 96.78% and 96.66% for HG002, HG003 and HG004 respectively) performs better than all other variant callers for each genome. F1 scores of NanoCaller exceed Medaka's, Clair's and Longshot's F1 scores by 0.45%, 0.51% and 1.98% on average. In Supplementary Tables S6-S11, we further show a detailed breakdown of performances in the difficult-to-map regions and demonstrate that NanoCaller performs better than other variant callers for SNPs in each of the following difficult-to-map regions: segmental duplications (Supplementary Tables S9), tandem and homopolymer repeats (Supplementary Tables S10), low mappability regions (Supplementary Tables S8), and Major Histocompatibility Complex (Supplementary Tables S11).

NanoCaller team participated in PrecisionFDA truth challenge v2 for difficult-to-map genomic regions (held in July 2020, see https://precision.fda.gov/challenges/10), and submitted variant calls for the Ashkenazim trio made by an ensemble of NanoCaller, Medaka and Clair models described above. The challenge consisted of variant calling evaluation on GIAB v4.1 benchmark variants of HG003 and HG004 that were made public after the challenge ended, and at the conclusion of the challenge GIAB released v4.2 benchmark variants. Our ensemble submission won the award for best performance in Major Histocompatibility Complex (MHC) using Nanopore reads [29]. Figure 4 (e) and (f) show the F1 score of SNPs and overall variants performance of the ensemble, NanoCaller, Medaka and Clair. While the ensemble performs better than all other variant callers in general, NanoCaller's performance on HG002 and HG004 is very close to the ensemble and significantly better than the performances of Medaka and Clair (F1-scores NanoCaller: 98.53%, 99.07% vs ensemble: 98.97%, 99.17%; Medaka: 97.15%, 94.29% and Clair: 97.55%, 98.59% for HG002 and HG004 respectively). NanoCaller always outperforms Longshot for SNP calling in MHC regions. Therefore, this is an independent assessment of the real-world performance of NanoCaller in detecting variants in complex genomic regions.

Performance on Indel Calling

We train NanoCaller indel ONT model on chromosomes 1-22 of HG001 (ONT) with v3.3.2 benchmark variants, and test it on ONT reads of four genomes (HG001, HG002, HG003, HG004) against v3.3.2

benchmark variants for HG001 and v4.2 benchmark variants for the Ashkenazim trio. The settings of this evaluation is given below: haplotype insertion allele frequency at 0.4 and deletion frequency at 0.6 are used to determine indel candidates due to abundance of deletion errors in Nanopore reads. For each genome, we create evaluation regions by removing low complexity homopolymer repeat regions from GIAB high confidence regions, and evaluate the indel performance by RTG *vcfeval*. These low complexity regions consist of perfect homopolymer regions of length greater than or equal to 4bp as well as imperfect homopolymer regions of length greater than 10bp regions provided in GIAB's v2.0 genome stratification. The results are shown in Table 2 and Figure 4 (g), (h) and (1) for NanoCaller together with Medaka and Clair. According to the F1 scores in Table 2, NanoCaller generally performs better than Clair (~10% on HG002, ~7% on HG003 and ~6% on HG004) and competitively against Medaka for HG002 to HG004. It is also worth noting that NanoCaller has a higher recall than Medaka and especially Clair: for example, NanoCaller has ~18% and ~1.8% higher recall than Clair and Medaka respectively on each of the three genomes of the trio. This suggests that NanoCaller complements other existing methods well to identify more indels. Supplementary Figure S1 further shows concordance of ground truth variants in high-confidence regions (including homopolymer repeat regions) of the Ashkenazim trio correctly predicted by NanoCaller, Medaka and Clair. Figure S1 (b) shows each tool has a significant number (ranging from 19k-60k) of correctly predicted indel calls, that are not correctly predicted by other variant callers.

Performance on PacBio Sequencing data

On PacBio datasets of four genomes: HG001, HG002, HG003 and HG004, we evaluated NanoCaller SNP models NanoCaller1 and NanoCaller3 against v3.3.2 benchmark variants for HG001 and v4.2 benchmark variants for the trio. The settings of compared tools are given below: in NanoCaller, minimum alternative allele frequency threshold for SNP calling at 0.15 was used to identify SNP candidates for both PacBio CCS and CLR reads, and NanoCaller models are trained with v3.3.2 benchmark variants. For Clair, the PacBio model trained on HG001 and HG005 was used for testing CCS reads, whereas the model trained on seven genomes HG001-HG007 was used for testing CLR reads; both Clair models used v3.3.2 benchmark variants for training; The provided PacBio model in the new DeepVariant release v1.0.0 is used for testing, and this model was trained on CCS reads of HG001-HG006 with v3.3.2 benchmark variants for HG001, HG005-6 and v4.2 for HG002-HG004.

The results for SNP performance on CCS reads are shown in Table 2 and Figure 5 (a), (b) and (c) along with Clair, Longshot, DeepVariant and WhatsHap. It can be seen from Table 2 and Figure 5 (a), (b) and (c) that 1) both NanoCaller1 and NanoCaller3 have similar performance according to F1 scores, although NanoCaller1 model is trained on Nanopore dataset while NanoCaller3 model is trained on PacBio CLR dataset; thus, NanoCaller is able to perform well cross different sequencing platform; 2) NanoCaller

outperforms Longshot, and 3) NanoCaller shows competitive performance against Clair (99.77%, 99.76%, 99.74% on the trio) and WhatsHap (99.12%, 99.66%, 99.66% on the trio) with difference < 0.3%.

We also evaluated NanoCaller Pacbio indel model on CCS datasets of the Ashkenazim trio: HG002, HG003 and HG004, and the results are shown in Table 2 and Figure 5 (d), (e) and (f) along with Clair and DeepVariant indel performance. The F1-scores on the trio suggest that NanoCaller performs competitively against Clair. As expected, DeepVariant performs very well on CCS reads because CCS reads have much lower error rates.

With Continuous Long Read Sequencing (CLR) datasets, we evaluated NanoCaller SNP models NanoCaller1 and NanoCaller3 on HG001 (reads aligned to GRCh37) and the Ashkenazim trio: HG002, HG003 and HG004 (as shown in Table 2 and Figure 5 (g), (h) and (i)). Due to drastic differences in coverage of CLR datasets, we used a higher NanoCaller quality score cut-off for HG003 and HG004, compared to HG001 and HG002. NanoCaller shows similar performance for the Ashkenazim trio in terms of F1-score (NanoCaller1: 98.26%, 94.33%, 93.47% vs Clair: 98.38%, 94.89%, 94.15% and Longshot: 98.41%, 94.35%, 93.27%), but has worse on HG001.

Novel Variants called by NanoCaller

We also analyzed SNP calls made by NanoCaller on HG002 (ONT reads) that are absent in the GIAB ground truth calls (version 3.3.2) [30], and validated 17 regions of those SNP calls by Sanger sequencing before v4 benchmark for HG002 was made available (Sanger sequencing signals along with inferred sequences of both chromosomes for each region are shown in the supplementary folder of “Sanger Sequencing Files” and *Sanger_sequences.xlsx)*. By deciphering Sanger sequencing results, we identified 41 novel variants (25 SNPs, 10 insertions and 6 deletions), as shown in Table 4. Based on the 41 novel variants, we conducted the variant calling evaluation by different methods on both older ONT HG002 reads and newly released ONT HG002 reads (as described in the Methods section) to see how more accurate long reads improve variant calling. We find that (1) on the newly released ONT HG002 reads, Medaka correctly identified 15 SNPs, 6 insertions and 2 deletions, Clair identified 14 SNPs, 6 insertions and 2 deletions, and Longshot correctly identified 18 SNPs, while NanoCaller was able to correctly identify 20 SNPs,6 insertions and 2 deletions, as shown in the Supplementary Table S12, whereas one of these 2 deletions was not called correctly by other variant callers; and (2) on the older ONT HG002 reads, as shown in Table 4, Medaka correctly identified 8 SNPs, 3 insertions and 1 deletion, and Clair identified 8 SNPs, 2 insertions and 1 deletion, whereas Longshot correctly identified 8 SNPs; In contrast, NanoCaller was able to correctly identify 18 SNPs and 2 insertions, whereas 10 of these 18 SNPs and 1 of these 2 insertions were not called correctly by other variant callers on the older HG002 (ONT reads). This indicates that the improvements in per-base accuracy during base-calling significantly enhances the variant calling performance. Also in Table 4, there are 2 multiallelic SNPs which can be identified by

NanoCaller but cannot be correctly called by all other 3 methods. One of the multiallelic SNPs at chr3:5336450 (A>T,C) is shown in Figure 6, where both the IGV plots and Sanger results clearly show a multiallelic SNP that was correctly identified by NanoCaller but was missed by other variant callers, likely due to the unique haplotype-aware feature of NanoCaller. In summary, the prediction on these novel variants clearly demonstrate the usefulness of NanoCaller for SNP calling.

To demonstrate the performance of NanoCaller for indel calling, we use Figure 7 to illustrate those variants that can be detected by long-read variant callers but cannot be detected by short-read data. In Figure 7, the validated deletion is at chr9:135663795 or chr9:135663804 (there are two correct alignments at the deletion and thus both genomic coordinates are correct.). NanoCaller detects the deletion at chr9:135663805, while Medaka and Clair detect the deletion at chr9:135663799. Although they are several bps away from the expected genomic coordinates (which is normal in long-read based variant calling), the prediction provides accurate information of the deletion compared with short-read data where little evidence supports the deletion as shown in Figure 7 (a). Sanger sequencing signal data, shown in Figure 7 (b), confirms the presence of a heterozygous deletion at the same location which is causing frameshift between the signals from maternal and paternal chromosomes. This is an example to demonstrate how long-read variant callers on long-read data can detect variants that fail to be reliably called on short-read sequencing data.

NanoCaller Runtime Comparison

We assessed NanoCaller's running time in four modes: 'snps_unphased', 'snps', 'indels', and 'both'. In 'snps_unphased' mode, NanoCaller uses deep neural network model to predict SNP calls only, whereas in the 'snps' mode, NanoCaller SNP calling is followed by an additional step of phasing SNP calls by external haplotyping tools such as WhatsHap.. In the 'indels' mode, NanoCaller uses phased reads in a BAM input to predict indels only. The entire NanoCaller workflow is the 'both' mode, where NanoCaller first runs in 'snps' mode to predict phased SNP calls, then uses WhatsHap to phase reads with the SNP calls from the 'snps' mode, followed by running 'indels' mode on phased reads from the previous step. Table 3 shows the wall-clock runtime of each mode of NanoCaller using 16 CPUs (IntelXeon CPU E5-2683 v4 @ 2.10GHz) on 49x HG002 ONT, 35x PacBio CCS and 58x PacBio CLR reads. NanoCaller takes ~18.4 hours and ~2.8 hours to run 'both' and 'snps_unphased' modes on 49x HG002 ONT reads, compared to ~181.6 hours for Medaka and ~5.6 hours for Clair, on the same 16 CPUs. On 35x CCS reads, NanoCaller takes ~11.2 hours and ~2.7 hours to run to run 'both' and 'snps_unphased' modes, compared to ~1.8 hours by Clair and ~11.8 by DeepVariant, on 16CPUs. NanoCaller usually runs faster than other tools. We summarize the runtime of all variant callers in Supplementary Table S17.

Please note that Medaka's first step also produces unphased SNP calls using a recurrent neural network on mixed haplotypes (Medaka later uses WhatsHap to phase SNP calls and reads for haplotype

separated variant calling). Compared with Medaka's first step, NanoCaller's unphased SNP calling not only takes a fraction of the time required for Medaka's first step (~2.8 hours vs ~70.7 hours), but also gives much better performance (precision, recall and F1-score on HG002 are NanoCaller: 98%, 97.99%, 97.99% vs Medaka 98.01%, 92.16%, 94.99%). Similarly, Longshot's first step uses a pair-HMM model to produce SNP calls from mixed haplotypes (Longshot later uses HapCUT2 to update the genotypes of these SNP calls in an iterative manner); on HG002 ONT, Longshot's first step takes 15.2 hours with 93.01%, 95.69%, and 94.33% precision, recall and F1-score respectively. Longshot and WhatsHap cannot use multiple CPUs to produce on SNP calls. With single CPU on 49x HG002 ONT reads, Longshot needs ~49.7 hours and WhatsHap needs ~84.3 hours for SNP calling.

Effects of Various Parameters on NanoCaller's Performance

Strategies of choosing heterozygous SNPs for SNP features generation

In NanoCaller, we generate input features for a SNP candidate site by choosing potentially heterozygous SNP sites that share a read with the candidate site. In the implementation of NanoCaller, at most 20 heterozygous SNP candidates are chosen for downstream and upstream of a candidate site of interest. With an expectation that 1 SNP occurring per 1000bp, a simple way is to include 20kb downstream and upstream sequence centered at the candidate site of interest, and then select 20 nearest heterozygous SNP sites. But in some smaller genomic regions, a cluster of heterozygous SNP candidates may be found but due to the noise with many false positives, and these false SNPs in a very smaller regions would provide a strong co-occurrence evidence for each other but little information for the candidate site of interest. Longshot notices this issue and overcomes high false positive rates due to dense clusters of false positive SNP by simply removing these dense clusters if the number of SNP calls exceeds a certain threshold in a specified range; however applying hard limits like that can lead to missing out on true SNPs that do occur in dense clusters in certain genomic regions.

We decide to use a difference method for selecting nearby potential heterozygous sites by forcing NanoCaller to pick a certain number of site that were some distance away from the candidate site. More precisely, we force NanoCaller to pick 2, 3, 4, 5 and 6 heterozygous SNP sites between the following distances from the candidate site: 0, 2kbp, 5kbp, 10,kbp, 20kbp and 50kbp. This is illustrated in Figure S3 and Table S18 of Supplementary Materials. We found that using this method, we achieve better SNP calling performance for ONT reads; Supplementary Materials Table S21 shows that under this strategy, for each genome in the Ashkenazim trio, we achieve higher precision, recall and F1 score for whole genome analysis as well as in each difficult-to-map genomic regions. On the other hand, SNP calling performance of PacBio CCS and CLR reads is not effected by using this method of selecting

heterozygous SNP sites, as shown in Supplementary Materials Table S22. This might be due to the fact that ONT reads have significantly higher N50 and mean read lengths compared to PacBio CCS and CLR reads, as shown in Supplementary Materials Table S1. Supplementary Materials Figure S2 shows the read lengths distribution of HG004 ONT, CCS and CLR datasets of 88X, 35X and 27X coverages respectively. In these datasets, 99.4% of CCS reads and 97% of CLR reads are shorter than 20,000bp; on the other hand, only 69.3% and 87.4% of the ONT reads are shorter than 20,000bp and 50,000bp. This simple comparison clearly demonstrates that NanoCaller is able to utilize the longer reads to improve SNP calling, and comparatively shorter PacBio reads might in part contribute to the less improvement of NanoCaller. Thus, as read length increases, we expect NanoCaller can have better performance.

Different thresholds for heterozygous SNP sites for SNP features generation

In order to generate haplotype structure features from long reads for a SNP candidate site, we need to select potentially heterozygous sites. Ideally heterozygous sites should have approximately 0.5 alternative allele frequency, which is rarely the case due to alignment and sequencing errors. Therefore, a SNP candidate site is determined to potentially heterozygous if its alternative allele frequency is in a small range centered at 0.5: typically this range is 0.4-0.6 or 0.3-0.7 depending upon the sequencing technology and is called neighbor threshold. Table S19 of Supplementary Materials shows how the choice of this threshold effects SNP calling performance for ONT, PacBio CCS and CLR reads, which have different characteristics of error rates and read lengths (which in turn determines the number of candidate sites). We can observe that, generally, using a narrower range around 0.5% allows higher precision, but recall decreases because not enough heterozygous sites are chosen to give informative features. In particular, the performance is very sensitive to increases in the upper limit of threshold, and decreases drastically when the threshold is increased. We determined that 0.4-0.6 threshold works best for ONT reads, with 0.3-0.7 and 0.3-0.6 being the best thresholds for CCS and CLR reads. Using a narrower threshold for ONT reads makes sense since longer ONT reads give us plenty of heterozygous sites to choose from, compared to CLR or CCS reads. It should be noted that this threshold is used for testing a sequencing data only, and during training of NanoCaller SNP models, we simply use benchmark heterozygous SNPs.

We check how minimum numbers of heterozygous SNP candidates for NanoCaller affect the performance and show the result in Table S20 of Supplementary materials. In NanoCaller, a SNP candidates with less than a minimum number of heterozygous SNP candidates will be considered as false negatives without prediction. By default, a minimum number of heterozygous SNP candidates is 1. In Table S20 of Supplementary materials, different minimum numbers of heterozygous SNP candidates are checked on Nanopore reads and Pacbio reads. On both data, as this minimum threshold increase,

precision increases and recall decreases. The increasing precision suggests that more heterozygous SNP candidates can benefit SNP prediction.

Using WhatsHap 'distrust genotype' option for phasing

WhatsHap is able to call SNPs on ONT and PacBio reads, as shown in Table 2. WhatsHap shows similar performance to NanoCaller, albeit much slowly, while for ONT reads, WhatsHap shows poor performance with F1-scores around 88-93%. In NanoCaller, WhatsHap is used for phasing SNPs and reads but not for variant calling. Further, WhatsHap allows 'distrust genotype' setting for phasing which allows WhatsHap to change genotypes of any SNP, from hetero- to homozygous and vice versa, in an optimal likelihood solution based upon the haplotypes created. In NanoCaller, this setting is disabled by default.

If users of NanoCaller wants to use 'distrust genotype' setting in WhatsHap, negligible effect of SNP calling performance is expected on ONT reads, while an increase in F1-score of 0.15-0.5% and 0.17-1.4% are expected on PacBio CCS and CLR reads, as shown in Table S23 of Supplementary Materials. But please note that using this setting significantly increased the runtime.

Discussion

In this study, we present NanoCaller, a deep learning framework to detect SNPs and small indels from long-read sequencing data. Depending on library preparation and sequencing techniques, long-read data usually have much higher error rates than short-read sequencing data, which poses a significant challenge to variant calling and thus stimulates the development of error-tolerant deep learning methods for accurate variant calling. However, the benefits of much longer read length of long-read sequencing are not fully exploited for variant calling in previous studies. The NanoCaller tool that we present here solely integrates haplotype structure in deep convolutional neural network for the detection of SNPs from long-read sequencing data, and uses multiple sequence alignment to re-align indel candidate sites to generate indel calling. Our evaluations under the cross-genome testing, cross-reference genome testing, and cross-platform testing demonstrate that NanoCaller performs competitively against other long-read variant callers, and outperforms other methods in difficult-to-map genomic regions.

NanoCaller has several advantages to call variants from long-read sequencing data. (1) NanoCaller uses pileup of candidate SNPs from haplotyped set of long-range heterozygous SNPs (with hundreds or thousands bp away rather than adjacent neighborhood local region of a candidate SNP of interest), each of which is shared by a long read with the candidate site. Given a long read with >20kb, there are on averagely >20 heterozygous sites, and evidence of SNPs from the same long reads can thus improve SNP calling by deep learning. Evaluated on several human genomes with benchmarking variant sets, NanoCaller demonstrates competitive performance against existing variant calling methods on long reads

and with phased SNPs. (2) NanoCaller is able to make accurate predictions cross sequencing platforms and cross reference genomes. In this study, we have tested NanoCaller models trained on Nanopore data for performance. We also test NanoCaller models calling variants on PacBio long-read data and achieved similar prediction trained on GRCh38 for GRCh37 and achieve the same level SNP calling performance. (3) With the advantage of long-read data on repetitive regions, NanoCaller is able to detect SNPs/indels outside high-confidence regions which cannot be reliably detected by short-read sequencing techniques, and thus NanoCaller provides more candidate SNPs/indels sites for investigating causal variants on undiagnosed diseases where no disease-causal candidate variants were found by short-read sequencing. (4) NanoCaller has flexible design to call multi-allelic variants, which Clairvoyante and Longshot cannot handle. In NanoCaller, the probability of each nucleotide type is assessed separately, and it is allowed that the probability of 2 or 3 or 4 nucleotide type is larger than 0.5 or even close to 1.0, and thus suggests strong evidence for a specific position with multiple bases in a test genome. Therefore, NanoCaller can easily generate multi-allelic variant calls, where all alternative alleles differ from the reference allele. Furthermore, NanoCaller can be easily configured to call variants for species with polyploidy or somatic mosaic variants when data are available to train an accurate model. (5) NanoCaller uses rescaled statistics to generate pileup for a candidate site, and rescaled statistics is independent on the coverage of a test genome, and thus, NanoCaller is able to handle a test data set with different coverage from the training data set, which might be a challenge for other long-read callers. That is, NanoCaller trained on a whole-genome data has less biases on other data sets with much lower or higher coverage, such as target-sequencing data with thousands folds of coverage. (6) With very accurate HiFi reads (<1% error rate) generated by PacBio, NanoCaller is able to yield competitive variant calling performance.

However, there are several limitations of NanoCaller that we wish to discuss here. One is that NanoCaller relies on the accurate alignment and pileup of long-read sequencing data, and incorrect alignments in low-complexity regions might still occur, complicating the variant calling process. For instance, most variants missed by NanoCaller in MHC region cannot be observed through IGV either due to alignment errors. Both continuingly improved sequencing techniques and improved alignment tools can benefit NanoCaller with better performance. But if the data is targeted at very complicated regions or aligned with very poor mapping quality, the performance of NanoCaller would be affected. Another limitation of NanoCaller is that the indel detection from mononucleotide repeats might not be accurate, especially on Nanopore long-read data which has difficulty in the basecalling of homopolymers [31, 32]. In Nanopore long-read basecalling process, it is challenging to determine how many repeated nucleotides for a long consecutive array of similar Nanopore signals, potentially resulting in false indel calls at these regions, which can be post-processed from the call set.

In summary, we propose a deep-learning tool solely using long-range haplotype information for SNP calling and local multiple sequence alignments for accurate indel calling. Our evaluation on several human genomes suggests that NanoCaller performs competitively against other long-read variant callers, and can generate SNPs/indels calls in complex genomic regions. NanoCaller enables the detection of genetic variants from genomic regions that are previously inaccessible to genome sequencing, and may facilitate the use of long-read sequencing in finding disease variants in human genetic studies.

Methods

Datasets

Long-read data

Five long-read data sets for human genomes are used for the evaluation of NanoCaller: HG001, the Ashkenazim trio (consisting of son HG002, father HG003 and mother HG004), and HX1. For HG001, Oxford Nanopore Technology (ONT) rel6 FASTQ files, basecalled with Guppy 2.3.8, are downloaded from WGS consortium database[9], and aligned to the GRCh38 reference genome using minimap2 [33]; PacBio CCS alignment files for HG001 are downloaded from the GIAB database [30, 34]. For HG002, HG003 and HG004 in the Ashkenazim trio, newly released sets of ONT reads (basecalled by Guppy 3.6) and PacBio CCS reads are obtained from NIST through precisionFDA challenge website, and aligned to the GRCh38 reference genome using minimap2 [33]. Meanwhile, alignment files for an older dataset of ONT reads for HG002, basecalled by Guppy 2.3.5, are downloaded from the GIAB database [30, 34] for analysis. PacBio CLR alignment reads for the Ashkenazim trio files are downloaded from the GIAB database [30, 34]. The fifth genome HX1 was sequenced by us using PacBio [10] and Nanopore sequencing [35]. The long-read data is aligned to the GRCh38 reference genome using minimap2 [33]. Supplementary Table S1 shows the statistics of mapped reads in the five genomes where the coverage of ONT data ranges from 43 to 91 and the coverage of PacBio data is between 27 and 58.

Benchmark variant calls

The benchmark set of SNPs and indels for HG001 (version 3.3.2), and the Ashkenazim trio (v4.2 and v3.3.2) are download from the Genome in a Bottle (GIAB) Consortium [30] together with high-confidence regions for each genome. There are 3,004,071; 3,459,843; 3,430,611; and 3,454,689 SNPs for HG001, HG002, HG003 and HG004 respectively, and 516,524; 587,978; 569,180 and 576,301 indels for them, as shown in Table 1. Benchmark variant calls for HX1 were generated by using GATK on Illumina ~300X reads sequenced by us [10] (Table 1).

NanoCaller framework for variant calling

In the framework of NanoCaller for variant calling, candidate sites of SNPs and indels are defined according to an input alignment and a reference sequence. NanoCaller has two convolutional neural networks, one for SNP calling and the other for indel prediction, each requiring a different type of input. Input pileup images generated for SNP candidate sites only use long-range haplotype information. For indel candidate sites, alignment reads are phased with WhatsHap using SNP calls from NanoCaller, and then NanoCaller uses phased alignment reads to generate input pileup images by carrying out local multiple sequence alignment around each site. Afterwards, NanoCaller combines SNP and indel calls to give a final output. The details are described below.

SNP Calling in NanoCaller

There are four steps in NanoCaller to generate SNP calling result for an input genome: candidate site selection, pileup image generation of haplotype SNPs, deep learning prediction, and phasing of SNP calls.

Candidate site selection: Candidate sites of SNPs are defined according to the depth and alternative allele frequency for a specific genomic position. In NanoCaller, "SAMtools mpileup" [36] is used to generate aligned bases against each genomic position. In NanoCaller, SNP candidate sites are determined using the criteria below. For a genomic site *b* with reference base *R*,

1. Calculate alternative allele frequency $= \frac{\max\{\text{number of reads supporting base B at site } b\}_{\text{B} \in \{\text{A,G,T,C}\} \setminus \text{R}}}{\text{total read depth at site } b}$
2. *b* is considered a SNP candidate site if the total read depth and the alternative allele frequency are both greater than specified thresholds. We set the alternative allele frequency threshold to be 15%.

Pileup image generation: After selecting all SNP candidate sites above, we determine a subset of SNP candidate sites as the set of highly likely heterozygous SNP sites (denoted by V). We extract long-range haplotype information from this subset of likely heterozygous SNP sites to create input images for all SNP candidate sites to be used in a convolutional neural network. This subset consists of SNP candidate sites with alternative allele frequencies in a specified range around 50%, and the default range is 40% to 60% for heterozygous site filtering. This range can be specified by the user depending upon the sequencing technology and read lengths. In detail, the procedure of pileup image generation is described below (as shown in Figure 2). For a SNP candidate site b:

1. We select sites from the set V that share at least one read with b and are at most 50,000bp away from b. For SNP calling on PacBio datasets, we set this limit at 20,000bp.

2. In each direction, upstream or downstream, of the site b, we choose 20 sites from V. If there are less than 20 such sites, we just append the final image with zeros. We denote the set of these potential heterozygous SNP sites nearby b (including b) by Z. An example is shown in Figure 2 (a). More details for how to choose these 40 nearby heterozygous sites from the set V can be found at Supplementary Materials Tables S18-S22 and Figures S2-S3.
3. The set of reads covering b is divided into four groups, $\mathrm{R}_B = \{\text{reads that support base B at b}\}$, $\mathrm{B} \in \{\mathrm{A}, \mathrm{G}, \mathrm{T}, \mathrm{C}\}$. Reads that do not support any base at b are not used.
4. For each read group in R_B with supporting base B, we count the number ($\mathrm{C}^{\mathrm{t}}_{\mathrm{BD}}$) of supporting reads for site $t \in$ Z with base $\mathrm{D} \in \{\mathrm{A}, \mathrm{G}, \mathrm{T}, \mathrm{C}\}$ (As shown in Figure 2(b)).
5. Let $\mathrm{F}^{\mathrm{t}}_{\mathrm{BD}} = \mathrm{C}^{\mathrm{t}}_{\mathrm{BD}} * g(D)$, where $\mathrm{g}(D)$ is a function that returns -1 if D is the reference base at site *t* and 1 otherwise. An example is shown in Figure 2(c).
6. We obtain a 4x41x4 matrix M with entries $[\mathrm{F}^{\mathrm{t}}_{\mathrm{BD}}]_{\mathrm{B,t,D}}$ (as shown Figure 2(d)) where the first dimension corresponds to nucleotide type B at site *b*, second dimension corresponds to the number of sites *t*, and the third dimension corresponds to nucleotide type D at site *t*. Our image has read groups as rows, various base positions as columns, and has 4 channels, each recording frequencies of different bases in the given read group at the given site.
7. We add another channel to our image which is a 4x41 matrix $[\mathrm{Q}^{\mathrm{t}}_{\mathrm{B}}]_{\mathrm{B,t}}$ where $\mathrm{Q}^{\mathrm{t}}_{\mathrm{B}} = 1$ if B is the reference base at site *b* and 0 otherwise (as shown in Figure 2(d)). In this channel, we have a row of ones for reference base at *b* and rows of zeroes for other bases.
8. We add another row to the image which encodes reference bases of site in Z, and the final image is illustrated in Figure 2(e).

Deep learning prediction: In NanoCaller, we present a convolutional neural networks [37] for SNP prediction, as shown in Figure 1. The neural network has three convolutional layers: the first layer uses kernels of three different dimensions and combines the convolved features into a single output: one capture local information from a row, another from a column and the other from a 2D local region; the second and third layers use kernels of size 2x3. The output from third convolutional layer is flattened and used as input for a fully connected network with dropout (using 0.5 drop date). The first fully connected layer is followed by two different neural networks of fully connected layers to calculate two types of probabilities. In the first network, we calculate the probability of each nucleotide type B to indicate that B is present at the genomic candidate site; thus for each nucleotide type B, we have a binary label prediction. The second network combines logit output of first network with output of first fully connected hidden layer to estimate probability for zygosity (homozygous or heterozygous) at the candidate site. The second network is used only in the training to propagate errors backwards for incorrect zygosity predictions. During testing, we infer zygosity from the output of first network only.

In order to call SNPs for a test genome, NanoCaller calculates probabilities of presence of each nucleotide type at the candidate site. If a candidate site has at least two nucleotide types with probabilities exceeding 0.5, it is considered to be heterozygous, otherwise it is regarded as homozygous. For heterozygous sites, two nucleotide types with highest probabilities are chosen with a heterozygous variant call. For homozygous sites, only the nucleotide type with the highest probability is chosen: if that nucleotide type is not the reference allele, a homozygous variant call is made, otherwise, it is homozygous reference. Each of called variants is also assigned with a quality score which is calculated as $-100 \log_{10} Probability\ (1 - P(B))$, where $P(B)$ is the probability of the alternative B (in case of multiallelic prediction we choose B to be the alternative allele with smaller probability) and recorded as a float in QUAL field of the VCF file to indicate the chance of false positive prediction: the larger the score is, the less likely that the prediction is wrong**.**

Phasing of SNP calls: After SNP calling, NanoCaller phases predicted SNP calls using WhatsHap [38]. By default, NanoCaller disables 'distrust-genotypes' and 'include-homozygous' settings of WhatsHap for phasing SNP calls, which would otherwise allow WhatsHap to switch variants from hetero- to homozygous and vice versa in an optimal phasing solution. Enabling these WhatsHap settings has minimal impact on NanoCaller's SNP calling performance (as shown in the Supplementary Tables S2, S3 and S4), but increases the time required for phasing by 50-80%. NanoCaller outputs both unphased VCF file generated by NanoCaller and phased VCF file generated by WhatsHap.

Indel Calling in NanoCaller

Indel calling in NanoCaller takes a genome with phased reads as input and uses several steps below to generate indel predictions: candidate site selection, pileup image generation, deep learning prediction, and then indel sequence determination. In NanoCaller, long reads are phased with SNPs calls that are predicted by NanoCaller and phased by WhatsHap [38] (as described above).

Candidate site selection: Indel candidate sites are determined using the criteria below. For a genomic site *b*,

1. Calculate:
    a. For $i \in \{0, 1\}$, $\text{depth}_i = \text{total number of reads in phase } i \text{ at site } b$
    b. Insertion frequency $= \min\{\frac{\text{number of reads in phase } i \text{ supporting insertion at site } b}{\text{depth}_i}\}_{i \in \{0,1\}}$
    c. Deletion frequency $= \min\{\frac{\text{number of reads in phase } i \text{ supporting deletion at site } b}{\text{depth}_i}\}_{i \in \{0,1\}}$
2. *b* is considered an indel candidate site if:
    a. Both depth_0 and depth_1 are greater than a specified depth threshold
    b. Either insertion frequency is greater than a specified insertion frequency threshold or the deletion frequency is greater than a specified deletion frequency threshold.

Thresholds for alternative allele frequency, insertion frequency, deletion frequency and read depths can be specified by the user depending on coverage and base calling error rate of the genome sequencing data.

Pileup image generation: Input image of indel candidate site is generated using the procedure below as shown in Figure 3. For an indel candidate site b:

1. Denote by S_{all}, S_{phase1} and S_{phase2} the set of all reads, reads in a phase and reads in the other phase at site *b*, respectively.
2. Let seq_{ref} be the reference sequence of length 160bp starting at site *b*
3. For each set $S \in \{S_{all},\ S_{phase1}, S_{phase2}\}$, do the following:
    a. For each read $r \in S$, let seq_r be the 160bp long subsequence of the read starting at the site *b* (for PacBio datasets, we use reference sequence and alignment sequences of length 260bp).
    b. Use MUSCLE to carry out multiple sequence alignment of the following set of sequences $\{seq_{ref}\} \cup \{seq_r\}_{r \in S}$ as shown in Figure 3 a).
    c. Let $\{seq'_{ref}\} \cup \{seq'_r\}_{r \in S}$ be the realigned sequences, where seq'_{ref} denotes the realigned reference sequence, and seq'_r denotes realignment of sequence seq_r. We truncate all sequences at the length 128 from the end.
    d. For $B \in \{A, G, T, C, -\}$ and $1 \leq p \leq 128$, calculate

$$C_{B,p} = \sum_{r \in S} g(\, seq'_r, B, p)$$

$$M_{B,p} = \frac{C_{B,p}}{\#\ reads\ in\ set\ S}$$

where $g(seq'_r, B, p)$ returns 1 if the base at index p of seq'_r is B and 0 otherwise. Figure 3 c) shows raw counts $C_{B,p}$ for each symbol.

    e. Let M be the 5×128 matrix with entries $M_{B,p}$ as shown in Figure 3 d).
    f. Construct a 5×128 matrix Q, with entries $[\mathrm{Q}^{\mathrm{p}}_{\mathrm{B}}]_{\mathrm{B,p}}$, where $\mathrm{Q}^{\mathrm{p}}_{\mathrm{B}} = 1$ if seq'_{ref} has symbol B at index p and 0 otherwise as shown in Figure 3 f). Both matrices M and Q have first dimension corresponding to the symbols $\{A, G, T, C, -\}$, and second dimension corresponding to pileup columns of realigned sequences.
    g. Construct a $5 \times 128 \times 2$ matrix Mat_S whose first channel is the matrix $M - Q$ as shown in Figure 3 e) and the second channel is the matrix Q.
4. Concatenate the three matrices $Mat_{S_{all}}, Mat_{S_{phase1}}$ and $Mat_{S_{phase2}}$ together to get a $15 \times 128 \times 2$ matrix as input to convolutional neural network.

Deep learning prediction: In NanoCaller, we present another convolutional neural networks [37] for indel calling, as shown in Figure 1. This neural network has a similar structure as SNP calling, and the difference is the fully connected network: for indel model, the first fully connected layer is followed by two fully connected hidden layers to produce probability estimates for each of the four zygosity cases at the candidate site: homozygous-reference, homozygous-alternative, heterozygous-reference and heterozygous-alternative (i.e. heterozygous with no reference allele).

Indel sequence determination: After that, NanoCaller calculates the probabilities for four cases of zygosities: homozygous-reference, homozygous-alternative, heterozygous-reference and heterozygous-alternative. No variant call is made if the homozygous-reference label has the highest probability. If homozygous-alternative label has the highest probability, we determine consensus sequence from the multiple sequence alignment of S_{all}, and align it against reference sequence at the candidate site using BioPython's pairwise2 local alignment algorithm with affine gap penalty. Alternative allele is inferred from the indel of the pairwise alignment of the two sequences. In case either of the heterozygous predictions has the highest probability, we use S_{phase1} and S_{phase2} to determine consensus sequences for each phase separately and align them against reference sequence. Indel calls from both phases are combined to make a final phased indel call at the candidate site.

Training and testing

For SNP calling, we have trained three convolutional neural network models on three different genomes that users can choose from. NanoCaller1 and NanoCaller2 are models trained on ONT reads of chr1-22 of HG001 (basecalled by Guppy v2.3.8) and HG002 (basecalled by Guppy v2.3.5.) respectively. NanoCaller3 is a model trained on chr1-22 of HG003 PacBio continuous long reads. For indel calling, we have trained two models on chr1-22 of HG001 using ONT (basecalled by Guppy v2.3.8) and PacBio CCS reads. NanoCaller uses the former model when calling indels on ONT datasets and the latter for PacBio datasets. All training sequencing datasets were aligned to GRCh38, and v3.3.2 GIAB's benchmark variants were used for training all SNP and indel models.

In NanoCaller, the SNP and indel models have 137,678 parameters in total, a significantly lower number than Clair [24](2,377,818) Clairvoyante[23] (1,631,496). All parameters in NanoCaller are initiated by Xavier's method [39]. Each model was trained for 100 epochs, using a learning rate of 1e-3 and 1e-4 for SNP and indel models respectively. We also applied L2-norm regularization, with coefficients 1e-3 and 1e-4 for SNP and indel models respectively, to prevent overfitting of the model.

To use NanoCaller on a test genome, it is reasonable that the test genome has different coverage as the genome used for training NanoCaller. To reduce the bias caused by different coverages, after generating

pileup images for SNP calling, NanoCaller by default scales the raw counts of bases in pileup images to adjust for the difference between coverages of the testing genome and the genome used for training of the model selected by user, i.e we replace the counts $C_{B,p}$ shown in Figure 2 (b) by

$$C_{B,p} \times \frac{coverage\ of\ training\ genome}{coverage\ of\ testing\ genome}$$

Performance measurement

The performance of SNP/indel calling by a variant caller is evaluted against the benchmark variant tests. Several measurements of performance evaluation are used, such as precision (p), recall (r) and F1 score as defined below.

$$p = \frac{TP}{TP + FP}$$
$$r = \frac{TP}{TP + FN}$$
$$F1 = \frac{2 * p * r}{p + r}$$

where TP is the number of benchmark variants correctly predicted by a variant caller, and FP is the number of miscalled variants which are not in benchmark variant sets, FN is the number of benchmark variants which cannot be called by a variant caller. $F1$ is the weighted average of p and r, a harmonic measurement of precision and recall. The range of the three measurements is $[0, \quad 1]$: the larger, the better.

Sanger validation of selected sites on HG002

To further demonstrate the performance of NanoCaller and other variant callers, we select 17 genomic regions whose SNPs/indels are not in the GIAB ground truth calls (version 3.3.2), and conduct Sanger sequencing for them on HG002. Firstly, we design PCR primers within ~400 bp of a select site of interest and then use a high-fidelity PCR enzyme (PrimeSTAR GXL DNA Polymerase, TaKaRa) to amplify each of the target selected repeat regions. The PCR products are purified using AMPure XP beads and sequenced by Sanger sequencing. We then decipher two sequences from Sanger results for variant analysis. The data and deciphered sequences are in the supplementary files. Please note that more than 17 variant sites are detected in the Sanger results, because each PCR region can contain 1+ variants.

Acknowledgements

The authors would like to thank members of the Wang lab for valuable comments and feedback. We would like to thank GIAB and nanopore-wgs-consortium for providing the sequencing data sets and the gold standard variant call data for use in our evaluation. We would like to thank PrecisionFDA team for organizing the variant calling Truth Challenge on difficulty-to-map regions and for scoring the submissions. This study is in part supported by NIH/NIGMS grant GM132713 to KW.

Competing Interests

The authors declare no competing interests.

Author contributions

UA and QL developed the computational method and drafted the manuscript. UA implemented the software tool and evaluated its performance. LF conducted wet-lab experiments of Sanger sequencing of candidate variants. KW conceived the study, advised on model design and guided implementation/evaluation. All authors read, revised, and approved the manuscript.

References

1. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res* 2010, **20:**1297-1303.
2. Garrison E, G. M: **Haplotype-based variant detection from short-read sequencing.** *arXiv* 2012, **1207.3907**.
3. Krusche P, Trigg L, Boutros PC, Mason CE, De La Vega FM, Moore BL, Gonzalez-Porta M, Eberle MA, Tezak Z, Lababidi S, et al: **Best practices for benchmarking germline small-variant calls in human genomes.** *Nat Biotechnol* 2019, **37:**555-560.
4. Zook JM, McDaniel J, Olson ND, Wagner J, Parikh H, Heaton H, Irvine SA, Trigg L, Truty R, McLean CY, et al: **An open resource for accurately benchmarking small variant and reference calls.** *Nat Biotechnol* 2019, **37:**561-566.
5. Cameron DL, Di Stefano L, Papenfuss AT: **Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software.** *Nature Communications* 2019, **10:**3240.
6. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X, et al: **The potential and challenges of nanopore sequencing.** *Nat Biotechnol* 2008, **26:**1146-1153.

7. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al: **Real-time DNA sequencing from single polymerase molecules.** *Science* 2009, **323:**133-138.
8. Mantere T, Kersten S, Hoischen A: **Long-Read Sequencing Emerging in Medical Genetics.** *Front Genet* 2019, **10:**426.
9. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, et al: **Nanopore sequencing and assembly of a human genome with ultra-long reads.** *Nat Biotechnol* 2018, **36:**338-345.
10. Shi L, Guo Y, Dong C, Huddleston J, Yang H, Han X, Fu A, Li Q, Li N, Gong S, et al: **Long-read sequencing and de novo assembly of a Chinese genome.** *Nat Commun* 2016, **7:**12065.
11. Pendleton M, Sebra R, Pang AW, Ummat A, Franzen O, Rausch T, Stutz AM, Stedman W, Anantharaman T, Hastie A, et al: **Assembly and diploid architecture of an individual human genome via single-molecule technologies.** *Nat Methods* 2015, **12:**780-786.
12. Seo JS, Rhie A, Kim J, Lee S, Sohn MH, Kim CU, Hastie A, Cao H, Yun JY, Kim J, et al: **De novo assembly and phasing of a Korean human genome.** *Nature* 2016, **538:**243-247.
13. Cho YS, Kim H, Kim HM, Jho S, Jun J, Lee YJ, Chae KS, Kim CG, Kim S, Eriksson A, et al: **An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes.** *Nat Commun* 2016, **7:**13637.
14. Stephens Z, Wang C, Iyer RK, Kocher JP: **Detection and visualization of complex structural variants from long reads.** *BMC Bioinformatics* 2018, **19:**508.
15. Heller D, Vingron M: **SVIM: structural variant identification using mapped long reads.** *Bioinformatics* 2019, **35:**2907-2915.
16. Jiang T, Liu B, Jiang Y, Li J, Gao Y, Cui Z, Liu Y, Wang Y: **Long-read-based Human Genomic Structural Variation Detection with cuteSV.** *bioRxiv* 2019**:**780700.
17. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC: **Accurate detection of complex structural variations using single-molecule sequencing.** *Nat Methods* 2018, **15:**461-468.
18. Fang L, Hu J, Wang D, Wang K: **NextSV: a meta-caller for structural variants from low-coverage long-read sequencing data.** *BMC Bioinformatics* 2018, **19:**180.
19. Gong L, Wong CH, Cheng WC, Tjong H, Menghi F, Ngan CY, Liu ET, Wei CL: **Picky comprehensively detects high-resolution structural variants in nanopore long reads.** *Nat Methods* 2018, **15:**455-460.
20. Ameur A, Kloosterman WP, Hestand MS: **Single-Molecule Sequencing: Towards Clinical Applications.** *Trends Biotechnol* 2019, **37:**72-85.
21. Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al: **Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome.** *Nat Biotechnol* 2019, **37:**1155-1162.
22. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, et al: **A universal SNP and small-indel variant caller using deep neural networks.** *Nat Biotechnol* 2018, **36:**983-987.
23. Luo R, Sedlazeck FJ, Lam TW, Schatz MC: **A multi-task convolutional deep neural network for variant calling in single molecule sequencing.** *Nat Commun* 2019, **10:**998.
24. Luo R, Wong C-L, Wong Y-S, Tang C-I, Liu C-M, Leung C-M, Lam T-W: **Exploring the limit of using a deep neural network on pileup data for germline variant calling.** *Nature Machine Intelligence* 2020, **2:**220-227.
25. Edge P, Bansal V: **Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing.** *Nat Commun* 2019, **10:**4660.

26. Edge P, Bafna V, Bansal V: **HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies.** *Genome research* 2017, **27:**801-812.
27. **medaka: Sequence correction provided by ONT Research** [https://github.com/nanoporetech/medaka]
28. Cleary JG, Braithwaite R, Gaastra K, Hilbush BS, Inglis S, Irvine SA, Jackson A, Littin R, Rathod M, Ware D, et al: **Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines.** *bioRxiv* 2015**:**023754.
29. Olson ND, Wagner J, McDaniel J, Stephens SH, Westreich ST, Prasanna AG, Johanson E, Boja E, Maier EJ, Serang O, et al: **precisionFDA Truth Challenge V2: Calling variants from short- and long-reads in difficult-to-map regions.** *bioRxiv* 2020**:**2020.2011.2013.380741.
30. Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, et al: **Extensive sequencing of seven human genomes to characterize benchmark reference materials.** *Scientific Data* 2016, **3:**160025.
31. Rang FJ, Kloosterman WP, de Ridder J: **From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy.** *Genome Biol* 2018, **19:**90.
32. Zascavage RR, Thorson K, Planz JV: **Nanopore sequencing: An enrichment-free alternative to mitochondrial DNA sequencing.** *Electrophoresis* 2019, **40:**272-280.
33. Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics* 2018, **34:**3094-3100.
34. **GENOME IN A BOTTLE** [https://jimb.stanford.edu/giab]
35. Liu Q, Fang L, Yu G, Wang D, Xiao CL, Wang K: **Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data.** *Nat Commun* 2019, **10:**2449.
36. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.
37. Krizhevsky A, Sutskever I, Hinton GE: **ImageNet classification with deep convolutional neural networks.** *Commun ACM* 2017, **60:**84-90.
38. Patterson M, Marschall T, Pisanti N, van Iersel L, Stougie L, Klau GW, Schonhuth A: **WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads.** *J Comput Biol* 2015, **22:**498-509.
39. Glorot X, Bengio Y: **Understanding the difficulty of training deep feedforward neural networks.** In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (Yee Whye T, Mike T eds.), vol. 9. pp. 249--256. Proceedings of Machine Learning Research: PMLR; 2010:249--256.

Tables

Table 1. Statistics of benchmark variants in chromosomes 1-22 of each genome aligned to the GRCh38 reference genome. Four genomes with GIAB benchmark variant calls, with v3.3.2 for HG001, and v4.2 for HG002, HG003 and HG004, statistics within the high confidence regions are also given. For HX1, high confidence regions are created by removing GIAB low complexity regions from the GRCh38 reference genome.

	Whole genome		High confidence region				Non homo -polymer region
Genome	**SNPs**	**Indels**	**SNPs**	**Indels**	**Total Length**	**% of genome**	**Indels**
HG001	3,004,071	516,524	2,961,527	483,630	2,329,784,734	81.03	180,844
HG002	3,459,843	587,978	3,365,334	525,466	2,542,282,520	88.43	210,354
HG003	3,430,611	569,180	3,331,481	504,910	2,530,414,070	88.01	199,530
HG004	3,454,689	576,301	3,355,595	511,524	2,530,026,864	88.00	201,110
HX1	3,282,242	687,501	2,788,450	176,587	2,356,619,870	75.77	--

Table 2. Performances (Precision/Recall/F1 percentages) of SNP and indel predictions by NanoCaller1, NanoCaller2 and NanoCaller3 on ONT and PacBio (CCS and CLR) data and on difficult-to-map genomic regions along with the performance of existing variant callers. These evaluation is bsed on v3.3.2 benchmark variants for HG001 and v4.2 benchmark variants for the Ashkenazim trio (HG002, HG003, HG004).

Pred		HG001	HG002	HG003	HG004	HX1
For SNPs on ONT data	NanoCaller1	97.29/93.59/95.40	98.00/97.99/97.99	99.02/98.73/98.88	98.85/98.69/98.77	90.60/89.96/90.28
	NanoCaller2	96.64/93.21/94.89	97.90/98.03/97.97	99.16/98.61/98.88	99.03/98.61/98.82	92.10/89.80/90.94
	Medaka	98.29/96.13/97.20	98.38/98.66/98.52	99.01/99.05/99.03	99.07/99.00/99.03	97.24/95.16/96.19
	Clair	98.79/94.39/96.54	98.24/97.28/97.76	99.05/97.94/98.49	99.00/98.11/98.55	96.38/93.66/95.00
	Longshot	98.62/95.66/97.12	98.93/97.09/98.00	98.93/96.83/97.87	98.96/96.75/97.84	95.40/93.81/94.60
	WhatsHap		96.33/84.08/89.79	97.96/88.34/92.90	97.65/80.56/88.28	
For SNPs on ONT data in difficult regions	Total SNPs		628,157	611,270	619,719	
	NanoCaller1		96.16/95.04/95.60	97.30/96.27/96.78	97.23/96.14/96.68	
	Medaka		94.74/95.48/95.11	96.14/96.54/96.34	96.41/96.12/96.27	
	Clair		94.68/95.33/95.01	96.47/96.11/96.29	96.23/96.22/96.22	
	Longshot		97.48/91.12/94.19	96.84/92.34/94.53	96.87/92.06/94.40	
For indels on ONT data	NanoCaller1	49.23/58.09/53.30	82.22/77.40/79.74	79.22/83.34/81.23	77.12/82.49/79.71	
	NanoCaller2	48.69/56.86/52.46	82.45/77.05/79.66	79.27/83.06/81.12	77.22/82.35/79.70	
	Medaka	75.45/57.09/65.00	82.99/75.10/78.85	86.47/82.37/84.37	85.00/80.38/82.63	
	Clair	81.02/49.75/61.65	85.88/58.24/69.41	86.00/65.30/74.23	85.33/64.70/73.60	
For all variants on ONT data in MHC	Ensemble		98.30/95.34/96.79	98.30/96.56/97.42	97.94/96.12/97.02	
	NanoCaller1		97.20/95.22/96.20	97.78/96.21/96.99	97.48/96.28/96.87	
	Medaka		97.19/92.99/95.05	98.06/97.62/97.84	96.16/89.20/92.55	
	Clair		96.92/94.18/95.53	97.51/95.73/96.61	97.70/95.46/96.57	
For SNPs on PacBio CCS data	NanoCaller1	96.97/99.37/98.15	99.31/99.60/99.45	99.36/99.63/99.49	99.37/99.63/99.50	
	NanoCaller3	96.69/99.36/98.01	99.16/99.56/99.36	99.22/99.58/99.40	99.23/99.59/99.41	
	Clair	99.59/99.77/99.68	99.83/99.71/99.77	99.82/99.70/99.76	99.80/99.68/99.74	
	Longshot	99.73/99.04/99.39	99.81/98.26/99.03	99.82/98.25/99.03	99.82/98.22/99.02	
	WhatsHap	98.86/99.63/99.24	99.73/98.52/99.12	99.74/99.59/99.66	99.75/99.57/99.66	
	DeepVariant	99.75/99.86/99.81	99.9/99.81/99.85	99.87/99.8/99.83	99.89/99.8/99.84	
For indels on PacBio CCS data	NanoCaller1	94.39/93.78/94.09	94.74/94.07/94.40	92.42/94.67/93.53	92.02/94.76/93.37	
	NanoCaller3	94.33/93.71/94.02	94.65/93.96/94.30	92.32/94.55/93.42	91.90/94.66/93.26	
	Clair	96.14/94.60/95.36	94.98/95.04/95.01	96.64/96.17/96.40	96.57/96.06/96.31	
	DeepVariant	97.68/97.52/97.60	98.14/98.39/98.26	98.31/98.71/98.51	98.28/98.66/98.47	
For SNPs on PacBio CLR data	NanoCaller1	97.85/87.08/92.15	98.64/96.58/97.60	94.48/91.83/93.14	93.26/91.03/92.13	
	NanoCaller3	97.54/85.66/91.21	98.43/94.78/96.57	93.35/88.77/91.00	92.10/88.01/90.01	
	Clair	99.10/92.77/95.83	98.22/98.54/98.38	95.69/94.11/94.89	95.03/93.28/94.15	
	Longshot	99.50/94.26/96.81	99.02/97.80/98.41	99.4/89.79/94.35	99.38/87.87/93.27	

Table 3. Wall-clock runtime in hours of different modes of NanoCaller using 16 CPUs on 49x ONT, 35x CCS and 58x CLR reads of HG002.

NanoCaller Mode	snps_unphased	snps	indels	both
Description	Only SNP calling by NanoCaller deep convolutional network model	'snps_unphased' mode followed by WhatsHap SNP call phasing	Only indel calling by NanoCaller using phased BAM input	'snps' mode, followed by read phasing by WhatsHap, and 'indels' mode
ONT	3.8	5.0	12.6	18.4
CCS	2.7	3.2	7.6	11.2
CLR	3.6	4.8	--	--

Table 4. Novel variants in HG002 genome, missing in v3.3.2 benchmark variant, discovered by Sanger sequencing together with the prediction information by NanoCaller and other variant callers using ONT reads basecall with Guppy 2.3.4.

Chrom	Position	REF	ALT	Genotype	Nanocaller	Medaka	Clair	Longshot
chr1	78883942	C	T	1|0	Correct	--	--	--
chr1	78883952	ATATATATTTATCCTTTATATATATTCTT	A	1|0	--	--	--	--
chr2	227913871	A	ATATCTATCTATC	1|0	Correct	Correct	Correct	--
chr2	227913885	G	A	1|0	Correct	Correct	Correct	Correct
chr2	227913889	G	A	1|0	Correct	Correct	Correct	Correct
chr2	227913928	T	TA	1|0	wrong allele	Correct	--	--
chr2	227913931	A	T	1|0	Correct	--	--	--
chr3	5336450	A	T,C	1|2	Correct	wrong allele	wrong allele	wrong allele
chr3	5336452	C	CGCGT	0|1	--	--	--	--
chr3	5336465	ACACACACACACG	A	0|1	wrong variant type	wrong allele	--	--
chr3	5336477	GCA	G	1|0	wrong allele	Correct	wrong allele	--
chr3	5336487	A	G	0|1	wrong variant type	wrong variant type	--	--
chr6	160009985	C	CTTAA	0|1	wrong allele	--	wrong allele	--
chr6	160009986	C	A	0|1	wrong zygosity	wrong allele	wrong variant type	wrong allele and zygosity
chr6	167130970	G	GGGCCCCCCTCCCTCCGGGACTCCTCCCTCT	0|1	--	--	--	--

chr6	167130972	GA	G	1|0	--	wrong zygosity	wrong zygosity	--
chr6	167130973	A	G	0|1	--	--	--	--
chr6	167130976	A	C	1|1	Correct	wrong zygosity	Correct	Correct
chr6	167130986	T	C	1|0	Correct	wrong allele	--	wrong allele
chr6	167130989	A	G	1|1	Correct	Correct	Correct	Correct
chr6	167130990	C	A	1|0	--	--	--	--
chr6	167130992	C	T	1|0	Correct	--	Correct	Correct
chr9	134784949	C	T	0|1	Correct	Correct	Correct	Correct
chr9	134784951	G	T	0|1	--	--	--	--
chr9	134784955	G	GGGGGGCA	0|1	--	--	--	--
chr9	134784956	T	G	0|1	Correct	wrong variant type	wrong variant type	--
chr9	135663795	ACAGAGGGG GACCTGGAG GGGCAGAGG AGAGACCTG TGGGG	A	0|1	--	--	--	--
chr9	135663892	A	G	1|1	wrong zygosity	Correct	Correct	Correct
chr9	135663893	T	A,G	1|2	Correct	wrong allele	wrong allele	wrong allele
chr11	113466435	G	GC	1|1	--	Correct	Correct	--
chr11	113466437	A	T	1|1	Correct	wrong allele	wrong allele	wrong allele
chr12	100940063	A	AT	0|1	Correct	--	--	--
chr12	100940065	G	C	0|1	Correct	wrong allele	--	--
chr14	75318035	C	T	1|0	Correct	Correct	Correct	Correct
chr14	75318038	AT	A	1|0	--	wrong allele	Correct	--
chr14	75318052	T	C	1|0	Correct	--	--	--
chr14	75318054	T	G	1|0	--	--	--	--
chr20	11064571	T	TGA	0|1	--	wrong allele	--	--
chr20	11064574	A	ATTTTCAAG ACTATTGTG ACTATGAC	0|1	--	Correct	--	--
chr20	11064578	A	T	0|1	Correct	--	--	--
chr20	11064579	C	T	0|1	Correct	Correct	wrong variant type	--

Figures

Figure 1. The deep-learning framework for SNP and indel calling. **a) An illustration of the convolutional neural network model for SNP calling; b) An illustration of the convolutional neural network model for indel calling. In both models, first convolutional layer uses 3 kernels of sizes 1x5, 5x1 and 5x5, whereas the second and third convolutional layers use kernels of size 2x3. Output of third convolutional layer is flattened and fed to a fully connected layer followed by a hidden layer with 48 nodes with a 50% dropped rate. In a), output of the hidden layer is split into two independent pathways: one for calculating probabilities of each base and the other for calculating zygosity probabilities. Zygosity probability is only used in the training process. In b), output of first fully connected layer is fed into two fully connected hidden layers to produce probabilities of four possible cases of zygosities.**

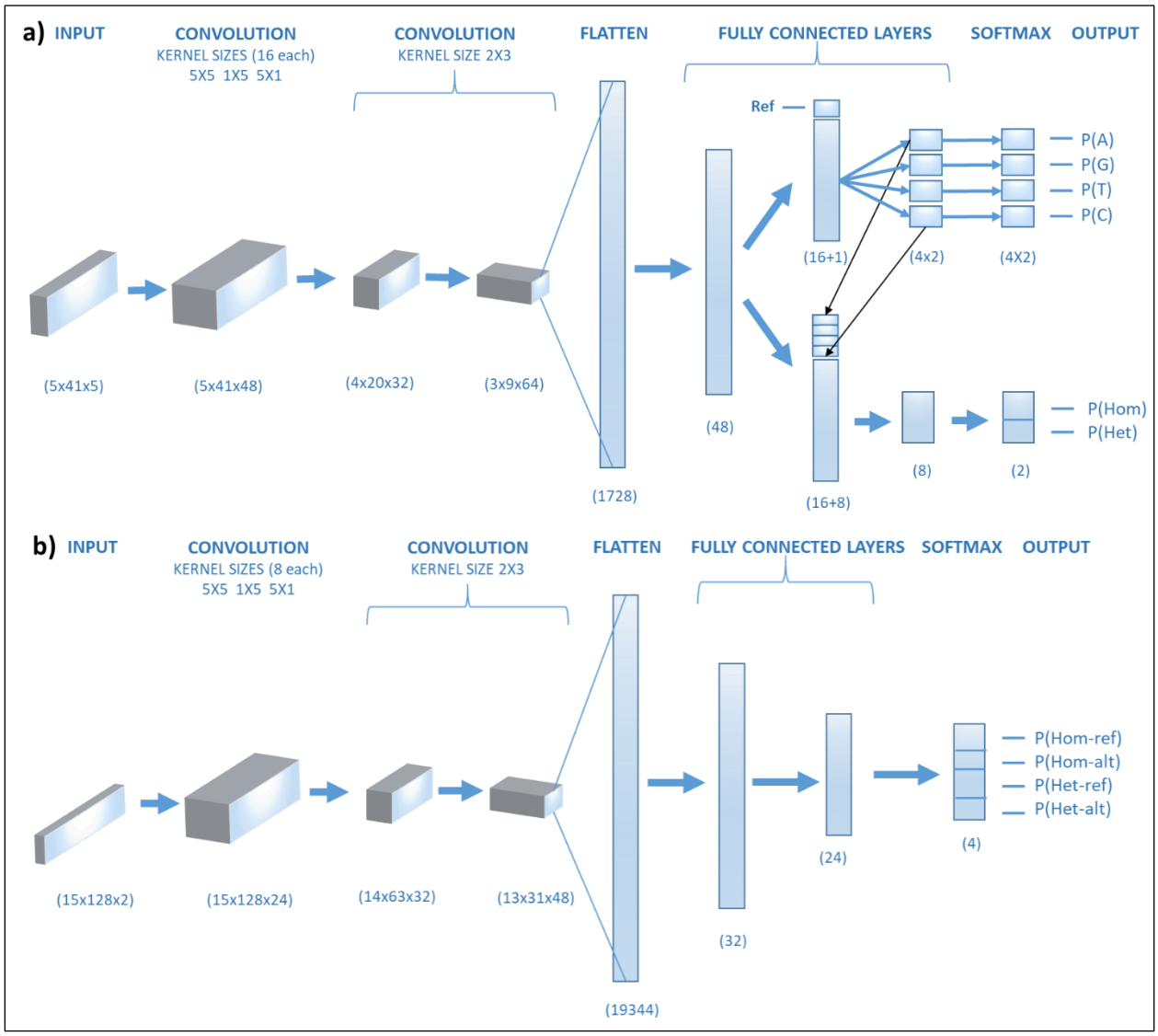


Figure 2. An example on how to construct image pileup for a SNP candidate site. a) reference sequence and reads pileup at site b and sites in set Z; b) raw counts of bases at sites in Z in each read group split by the nucleotide types at site b; c) frequencies of bases at sites in Z with negative signs for reference bases; d) flattened pileup image with fifth channel after reference sequence row is added; e) pileup image used as input for NanoCaller.

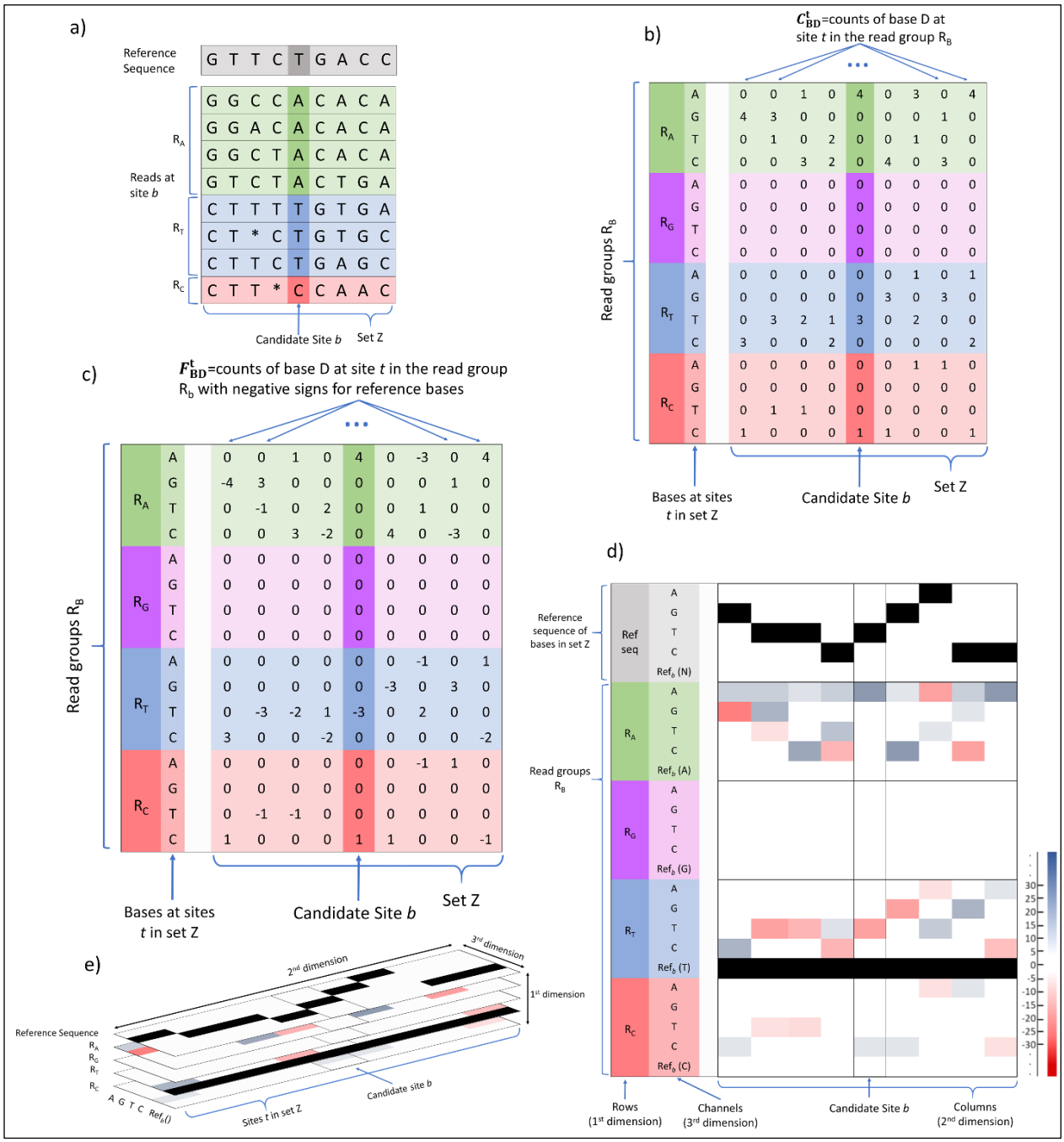
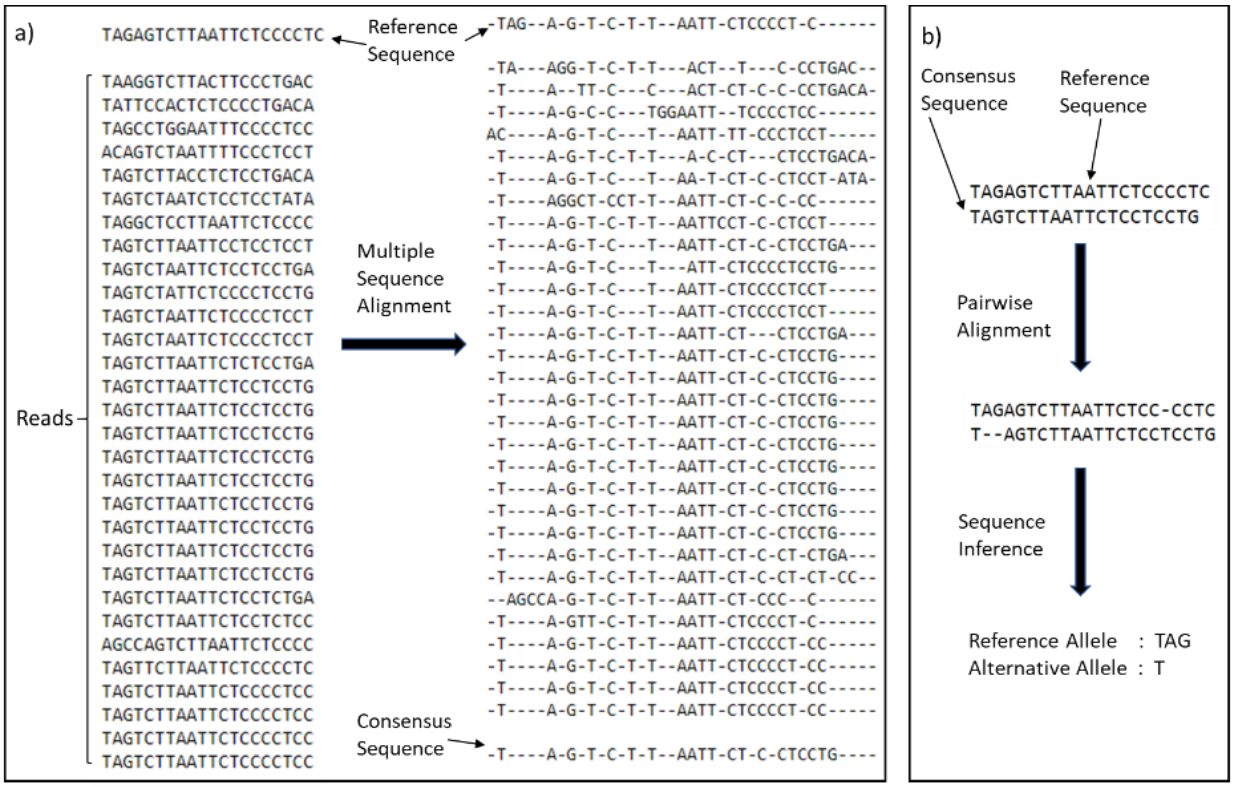


Figure 3. An example on how to construct image pileup for an indel site. a) reference sequence and reads pileup at the candidate site before and after multiple sequence alignment, and the consensus sequence after realignment; b) reference sequence and consensus sequence at the candidate site before and after pairwise alignment, and the inferred sequence; c) shows raw count of each symbol at each column of multiple sequence alignment pileup; d) matrix M, showing frequency of each symbol at each column of multiple sequence alignment pileup; e) first channel of input image, matrix M minus Q (one-hot encoding of realigned reference sequence); f) matrix Q, one-hot encoding of realigned reference sequence which forms the second channel of input image.



c)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
A	1	0	2	0	0	0	30	0	0	0	0	0	0	0	0	0	0	0	0	26	30	0	0
G	0	0	0	1	0	0	0	2	29	0	0	0	0	0	0	0	0	1	1	0	0	0	0
T	0	28	0	0	0	0	0	0	0	2	29	0	0	0	22	0	29	0	0	0	0	26	29
C	0	1	0	0	1	1	0	0	0	1	1	0	30	1	0	0	1	0	0	0	0	2	1
-	29	1	28	29	29	29	0	28	1	27	0	30	0	29	8	30	0	29	29	4	0	2	0
ref	-	T	A	G	-	-	A	-	G	-	T	-	C	-	T	-	T	-	-	A	A	T	T

d)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
A	0.033	0	0.067	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0.867	1	0	0
G	0	0	0	0.033	0	0	0	0.067	0.967	0	0	0	0	0	0	0	0	0.033	0.033	0	0	0	0
T	0	0.933	0	0	0	0	0	0	0	0.067	0.967	0	0	0	0.733	0	0.967	0	0	0	0	0.867	0.967
C	0	0.033	0	0	0.033	0.033	0	0	0	0.033	0.033	0	1	0.033	0	0	0.033	0	0	0	0	0.067	0.033
-	0.967	0.033	0.933	0.967	0.967	0.967	0	0.933	0.033	0.9	0	1	0	0.967	0.267	1	0	0.967	0.967	0.133	0	0.067	0
ref	-	T	A	G	-	-	A	-	G	-	T	-	C	-	T	-	T	-	-	A	A	T	T

e)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
A	0.033	0	-0.933	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.133	0	0	0
G	0	0	0	-0.967	0	0	0	0.067	-0.033	0	0	0	0	0	0	0	0	0.033	0.033	0	0	0	0
T	0	-0.067	0	0	0	0	0	0	0	0.067	-0.033	0	0	0	-0.267	0	-0.033	0	0	0	0	-0.133	-0.033
C	0	0.033	0	0	0.033	0.033	0	0	0	0.033	0.033	0	0	0.033	0	0	0.033	0	0	0	0	0.067	0.033
-	-0.033	0.033	0.933	0.967	-0.033	-0.033	0	-0.067	0.033	-0.1	0	0	0	-0.033	0.267	0	0	-0.033	-0.033	0.133	0	0.067	0
ref	-	T	A	G	-	-	A	-	G	-	T	-	C	-	T	-	T	-	-	A	A	T	T

f)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
A	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
G	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	1	1
C	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
-	1	0	0	0	1	1	0	1	0	1	0	1	0	1	0	1	0	1	1	0	0	0	0
ref	-	T	A	G	-	-	A	-	G	-	T	-	C	-	T	-	T	-	-	A	A	T	T

Figure 4. Performance of NanoCaller and state-of-the-art variant callers on five whole-genome Oxford Nanopore sequencing data sets. **The performance of SNP predictions on ONT reads: a) precision, b) recall, c) F1 score. d) The performance of SNP predictions on HG002 (ONT), HG003 (ONT) and HG004 (ONT) in difficult-to-map genomic regions. The performance of variant predictions on HG002 (ONT), HG003 (ONT) and HG004 (ONT) in Major Histocompatibility Complex regions: e) SNPs, f) overall variants. The performance of indel predictions on ONT reads in non-homopolymer regions: g) precision, h) recall, i) F1 score. For HX1, the variants, which were called on high-coverage short-read data are used as benchmark with complement of difficult-to-map genomic regions used in d) as high-confidence regions. Benchmark variants v3.3.2 for HG001 and v4.2 for the Ashkenazim trio (HG002, HG003, HG004) are used for evaluation.**

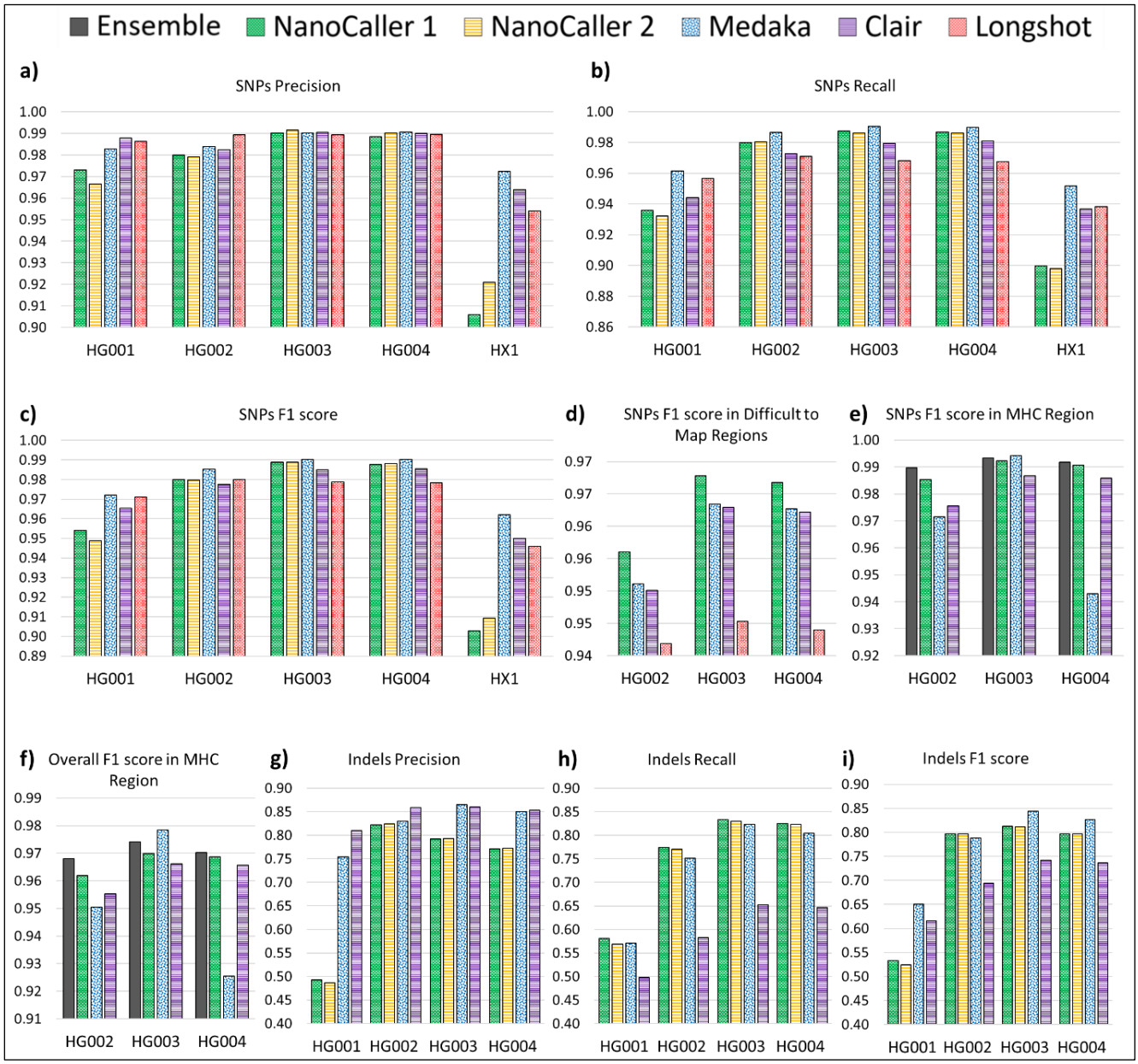


Figure 5. Performance of NanoCaller and state-of-the-art variant callers on PacBio sequencing data sets. The performance of SNP predictions on PacBio CCS reads: a) precision, b) recall, c) F1 score. The performance of indel predictions on PacBio CCS reads: d) precision, e) recall, f) F1 score. The performance of SNP predictions on PacBio CLR reads: g) precision, h) recall, i) F1 score. Benchmark variants v3.3.2 for HG001 and v4.2 for the Ashkenazim trio (HG002, HG003, HG004) are used for evaluation.

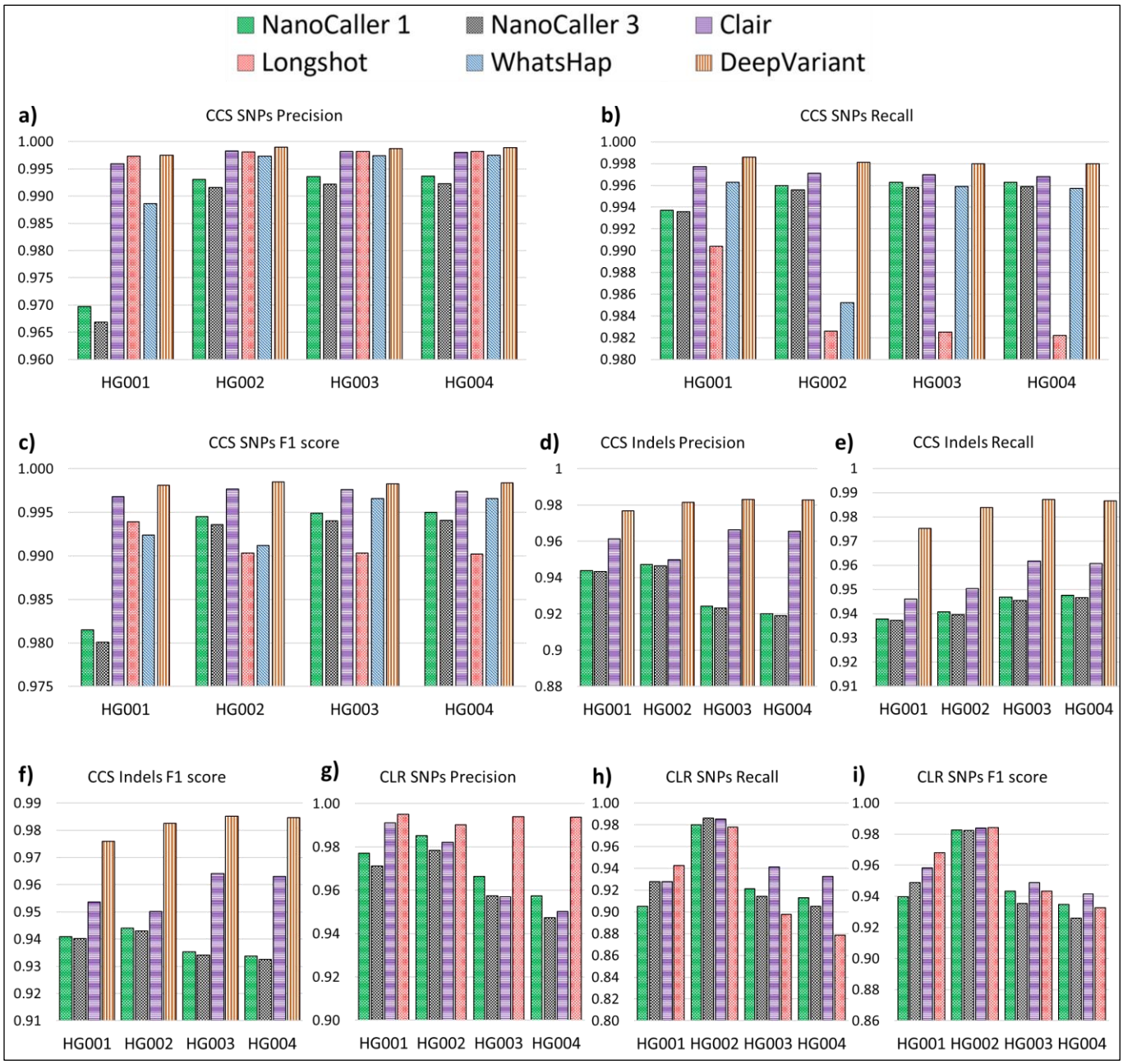


Figure 6. Evidence for novel multiallelic SNP. a) IGV plots of Nanopore, PacBio CCS and Illumina reads of HG002 genome at chr3:5336450-5336480. b) Sanger sequencing signal data for the same region. NanoCaller on older HG002 data correctly identified the multi-allelic SNP at chr3:5336450 (A>T,C) shown in black box.

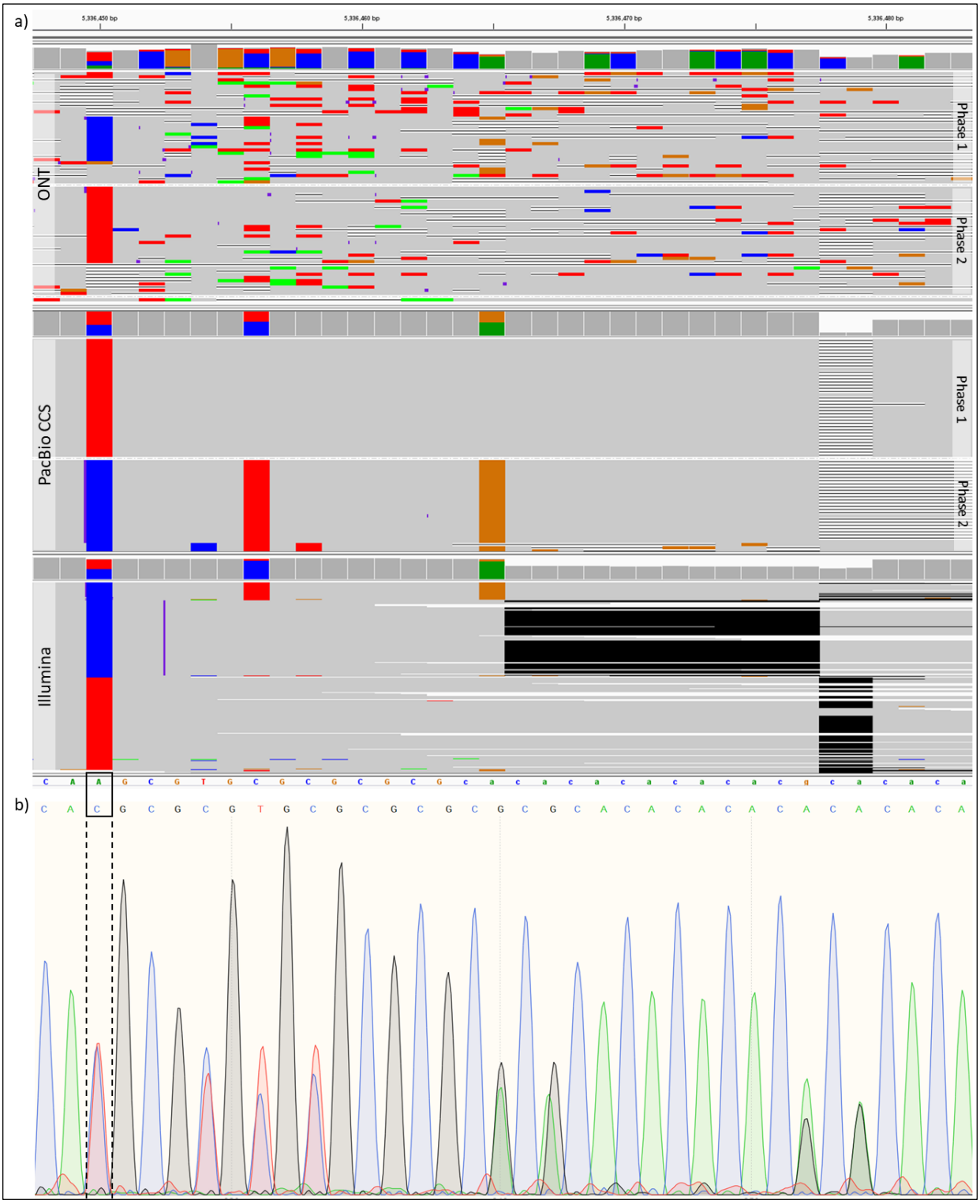
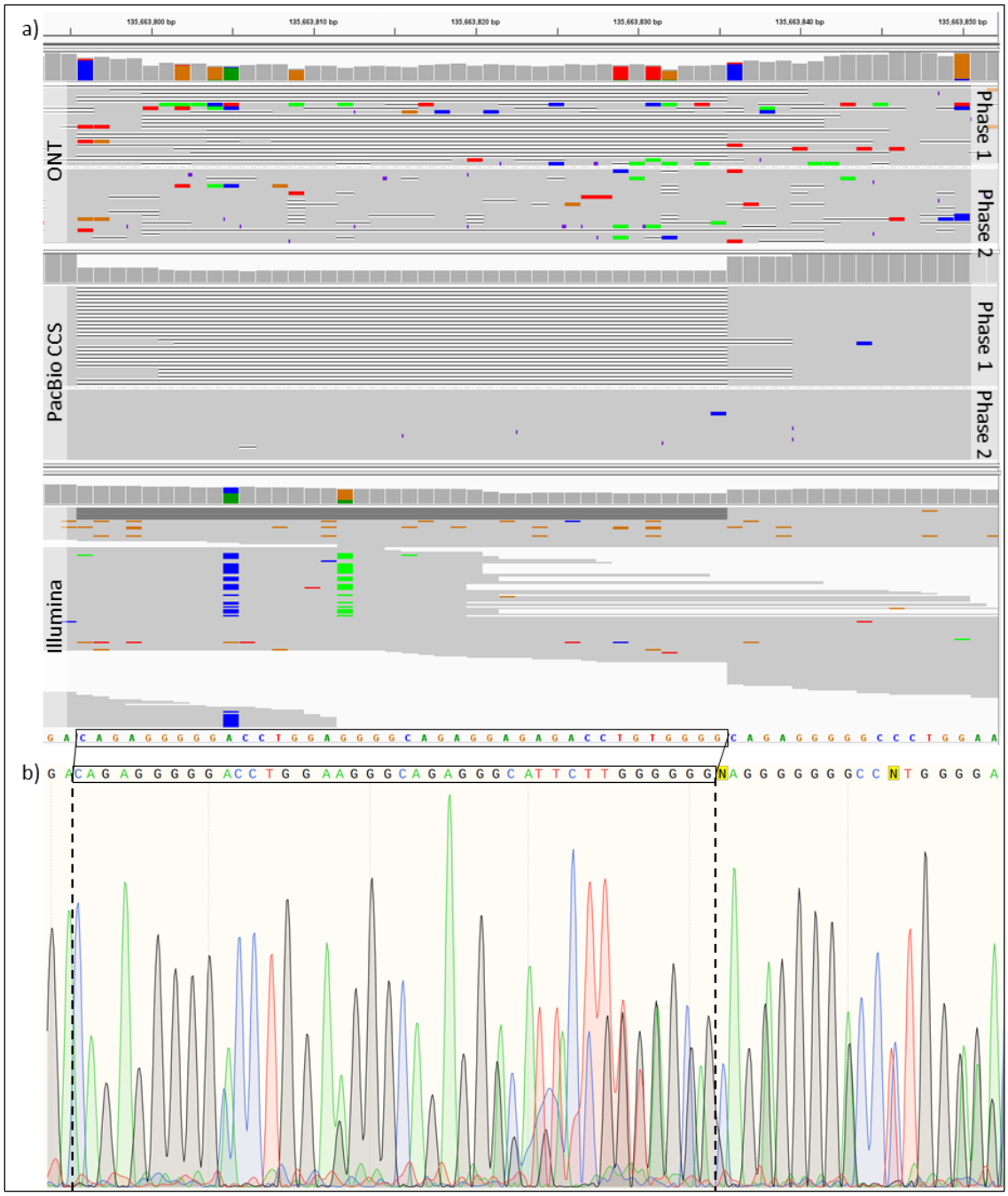


Figure 7. Evidence for novel deletions. a) IGV plots of Nanopore, PacBio CCS and Illumina reads of HG002 genome at chr9:135663780-chr9:135663850. The 40bp long deletion shown below in black box was identified using Sanger sequencing at chr9:135663795 or chr9:135663804 (both are correct and the difference is due to two different alignments). b): Sanger sequencing signal data around the deletion.



Supplementary Materials of

NanoCaller for accurate detection of SNPs and indels in difficult-to-map regions from long-read sequencing by haplotype-aware deep neural networks

Umair Ahsan[1,#], Qian Liu[1,#], Li Fang, Kai Wang[1,2*]

[1] Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

[2] Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

[#] These authors contributed equally to this work.

[*] To whom correspondence should be addressed. Email: wangk@email.chop.edu

Statistics of long-reads datasets

Table S5. Whole genome statistics of five data sets on human genomes by Nanopore and PacBio sequencing. Each genome is aligned to the GRCh38 reference genome, and only the mapped reads were used to calculate the statistics. Total number of bases is calculated as the sum of length of all mapped reads, and coverage is defined as number of mapped bases divided by the reference genome length.'Mean' and 'Median' represent the mean and median of read length in long reads for a genome.

Platform	Genome	# reads	# bases	N50	Mean	Median	Coverage
ONT	HG001	15,666,888	132.9 Gb	13,630	8,485	5,387	43X
ONT	HG002	7,788,089	144.2 Gb	52,549	18,514	6,672	49X
ONT	HG003	13,601,716	249.4 Gb	46,163	18,335	8,168	85X
ONT	HG004	12,530,839	251.3 Gb	49,563	20,051	8,371	88X
ONT	HX1	20,497,769	271.8 Gb	22,273	13,261	10,123	88X
PacBio (CLR)	HG001	23,808,790	106.3 Gb	6,309	4,466	3,180	37X
PacBio (CLR)	HG002	21,511,145	179.0 Gb	11,264	8,323	7,500	58X
PacBio (CLR)	HG003	10,564,465	85.3 Gb	10,943	8,078	7,251	28X
PacBio (CLR)	HG004	10,369,228	83.4 Gb	10,869	8,040	7,545	27X
PacBio (CCS)	HG001	8,487,716	84.6 Gb	10,004	9,963	9,973	29X
PacBio (CCS)	HG002	7,979,372	102.6 Gb	12,883	12,855	12,781	36X
PacBio (CCS)	HG003	6,872,414	100.5 Gb	14,761	14,627	14,627	35X
PacBio (CCS)	HG004	6,578,333	98.9 Gb	15,100	15,031	14,949	34X

Performance of NanoCaller and other variant callers on old Nanopore datasets of the Ashkenazim trio

Table S6. Performances of SNP predictions of NanoCaller1 and NanoCaller2 SNP models, along with existing variant callers, on old ONT data of the Ashkenazim trio, evaluated against v3.3.2 of GIAB benchmark variants.

	HG002			HG003			HG004		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
NanoCaller1	97.99	96.55	97.27	97.76	96.51	97.14	97.61	96.33	96.97
NanoCaller2	97.84	96.40	97.12	97.77	96.27	97.02	97.60	96.12	96.86
Medaka	97.75	96.92	97.34	97.86	96.89	97.38	97.73	96.67	97.20
Clair	98.85	96.67	97.75	99.02	97.30	98.16	98.97	97.04	98.00
Longshot	99.05	96.53	97.78	99.13	95.16	97.11	99.05	94.98	96.98

Table S7. Performances of SNP predictions of NanoCaller1 and NanoCaller2 SNP models, along with existing variant callers, on old ONT data of the Ashkenazim trio, evaluated against v4.2 of GIAB benchmark variants.

	HG002			HG003			HG004		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
NanoCaller1	98.90	96.77	97.83	99.03	96.77	97.89	98.96	96.50	97.72
NanoCaller2	98.79	96.63	97.70	98.87	96.75	97.80	98.85	96.47	97.65
Medaka	98.81	96.98	97.89	98.95	97.04	97.99	99.02	96.84	97.92
Clair	99.01	96.61	97.80	99.07	97.51	98.29	99.09	97.24	98.16
Longshot	98.97	95.89	97.41	99.11	94.60	96.81	99.13	94.32	96.67

Table S8. Performances of indel predictions in non-homopolymer regions of NanoCaller, along with existing variant callers, on old ONT data of the Ashkenazim trio, evaluated against v3.3.2 of GIAB benchmark variants.

	HG002			HG003			HG004		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
NanoCaller	77.37	59.65	67.37	75.90	60.09	67.08	73.56	58.30	65.05
Medaka	76.56	63.64	69.51	78.27	65.98	71.61	77.13	63.90	69.90
Clair	84.92	55.26	66.96	85.39	58.53	69.46	85.76	56.48	68.11

Table S9. Performances of indel predictions in non-homopolymer regions of NanoCaller, along with existing variant callers, on old ONT data of the Ashkenazim trio, evaluated against v4.2 of GIAB benchmark variants.

	HG002			HG003			HG004		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
NanoCaller	76.87	59.10	66.83	73.26	59.18	65.47	70.60	57.30	63.26
Medaka	76.34	62.62	68.81	77.78	65.67	71.22	76.74	63.77	69.66
Clair	84.40	54.59	66.30	85.57	58.28	69.34	85.74	56.53	68.14

SNP performance of NanoCaller and other variant callers in difficult-to-map regions on Nanopore reads of the Ashkenazim trio

BED file sources of different difficult to map regions with respect to GRCh38 reference genome:

1. All difficult regions in Table 3 of NanoCaller manuscript.
   Source: ftp://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v2.0/GRCh38/union/GRCh38_alldifficultregions.bed.gz
2. Low mappability regions in Table S12:
   Source: ftp://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v2.0/GRCh38/mappability/GRCh38_lowmappabilityall.bed.gz
3. Segmental duplications in Table S13Table S14.
   Source: ftp://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v2.0/GRCh38/SegmentalDuplications/GRCh38_segdups.bed.gz
4. Tandem and homopolymer repeats in Table S14.
   Source: ftp://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v2.0/GRCh38/LowComplexity/GRCh38_AllTandemRepeatsandHomopolymers_slop5.bed.gz
5. Major Histocompatibility Complex (chr6:28510020-33480577) in Table S15.
   Source: ftp://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v2.0/GRCh38/OtherDifficult/GRCh38_MHC.bed.gz

Table S10. Statistics of number ground truth variants in GIAB v4.2 benchmark of the Ashkenazim trio within various difficult genomic regions identified in GIAB genome stratification v2.0. Numbers shown below are for SNPs, with the exception of MHC region for which both SNP and indel counts are shown.

Region	HG002	HG003	HG004
All Difficult Regions	628,157	611,270	619,719
Tandem and Homopolymer Repeats	182,901	166,398	166,807
Segmental Duplications	121,179	122,393	125,711
Low Mappability regions	192,749	193,153	195,332
MHC (SNPs)	20,139	19,730	20,576
MHC (indels)	1667	1655	1738

Table S11. Performances of SNP predictions in all difficult genomic regions by NanoCaller1 SNP models, along with existing variant callers on ONT data, using 4.2 benchmark variants.

		NanoCaller1	Medaka	Clair	Longshot
	Threshold	144	3	370	118
Precision	HG002	96.16	94.74	94.68	97.48
	HG003	97.30	96.14	96.47	96.84
	HG004	97.23	96.41	96.23	96.87
Recall	HG002	95.04	95.48	95.33	91.12
	HG003	96.27	96.54	96.11	92.34
	HG004	96.14	96.12	96.22	92.06
F1	HG002	**95.60**	95.11	95.01	94.19
	HG003	**96.78**	96.34	96.29	94.53
	HG004	**96.68**	96.27	96.22	94.40

Table S12. Performances of SNP predictions in low mappability regions by NanoCaller1 SNP models, along with existing variant callers on ONT data, using 4.2 benchmark variants.

		NanoCaller1	Medaka	Clair	Longshot
	Threshold	200	8	713	172
Precision	HG002	96.73	95.33	95.9	98.70
	HG003	96.84	95.27	96.69	97.42
	HG004	96.86	96.03	96.52	97.66
Recall	HG002	97.31	97.59	97.43	88.13
	HG003	98.32	98.46	98.09	90.42
	HG004	98.28	98.38	98.26	90.78
F1	HG002	**97.02**	96.45	96.66	93.12
	HG003	**97.57**	96.84	97.39	93.79
	HG004	**97.57**	97.19	97.38	94.09

Table S13. Performances of SNP predictions in segmental duplication regions by NanoCaller1 SNP models, along with existing variant callers on ONT data, using 4.2 benchmark variants.

		NanoCaller1	Medaka	Clair	Longshot
	Threshold	228	11	768	186
Precision	HG002	94.19	90.79	93.07	97.99
	HG003	94.43	92.2	94.55	95.86
	HG004	94.58	93.64	94.3	96.30
Recall	HG002	95.64	96.54	96.23	84.99
	HG003	97.19	97.46	97.16	88.9
	HG004	97.17	97.31	97.27	89.27
F1	HG002	**94.91**	93.58	94.62	91.03
	HG003	95.79	94.75	**95.84**	92.25
	HG004	**95.86**	95.44	95.76	92.66

Table S14. Performances of SNP predictions in tandem and homopolymer repeat regions by NanoCaller1 SNP models, along with existing variant callers on ONT data, using 4.2 benchmark variants.

		NanoCaller1	Medaka	Clair	Longshot
	Threshold	93	2	300	115
Precision	HG002	93.20	90.09	90.48	93.96
	HG003	95.48	93.49	93.38	93.23
	HG004	95.20	92.63	92.85	92.70
Recall	HG002	88.53	88.30	89.05	89.67
	HG003	90.30	90.14	90.29	92.31
	HG004	89.78	89.57	90.01	91.81
F1	HG002	90.81	89.19	89.76	**91.76**
	HG003	**92.82**	91.78	91.81	92.76
	HG004	**92.41**	91.07	91.41	92.25

Table S15. Performances of SNPs, indels and overall variants predictions in Major Histocomplatibility Complex by NanoCaller1 SNP models, along with existing variant callers on ONT data, using 4.2 benchmark variants. Overall variant performance is calculated by combining the SNPs and indels performances shown.

SNPs Performance						
		Ensemble	NanoCaller1	Medaka	Clair	Longshot
	Threshold	--	145	0	272	15
Precision	HG002	99.44	98.88	99.02	97.88	98.77
	HG003	99.58	99.37	99.61	98.76	99.10
	HG004	99.51	99.18	97.92	98.77	99.00
Recall	HG002	98.50	98.18	95.34	97.23	52.09
	HG003	99.11	99.09	99.23	98.60	57.24
	HG004	98.84	98.97	90.93	98.40	51.05
F1	HG002	98.97	**98.53**	97.15	97.55	68.21
	HG003	99.34	99.23	**99.42**	98.68	72.57
	HG004	99.17	**99.07**	94.29	98.59	67.37

Indels Performance						
		Ensemble	NanoCaller1	Medaka	Clair	Longshot
	Threshold	--	205	19	509	--
Precision	HG002	79.40	72.77	74.02	81.00	--
	HG003	80.23	75.02	80.20	78.99	--
	HG004	76.39	74.59	75.80	81.32	--
Recall	HG002	57.21	59.47	64.78	57.31	--
	HG003	66.18	61.89	78.47	61.59	--
	HG004	64.00	64.55	68.84	60.86	--
F1	HG002	66.51	65.45	**69.09**	67.13	--
	HG003	72.53	67.82	**79.33**	69.21	--
	HG004	69.64	69.21	**72.15**	69.62	--

Overall Variants Performance						
		Ensemble	NanoCaller1	Medaka	Clair	Longshot
Precision	HG002	98.30	97.20	97.19	96.92	--
	HG003	98.30	97.78	98.06	97.51	--
	HG004	97.94	97.48	96.16	97.70	--
Recall	HG002	95.34	95.22	92.99	94.18	--
	HG003	96.56	96.21	97.62	95.73	--
	HG004	96.12	96.28	89.20	95.46	--
F1	HG002	96.79	**96.20**	95.05	95.53	--
	HG003	97.42	96.99	**97.84**	96.61	--
	HG004	97.02	**96.87**	92.55	96.57	--

Novel variants validated by Sanger sequencing

Table S16. Sanger validated variants, and performance of predictions in HG002 genome by various variant callers using new ONT reads. These variants are missing in GIAB v3.3.2 of HG002.

Chrom	Position	REF	ALT	Genotype	Nanocaller	Medaka	Clair	Longshot
chr1	78883942	C	T	1|0	Correct	Incorrect variant type	Correct	--
chr1	78883952	ATATATATTTATCCTTTATATATATTCTT	A	1|0	Incorrect allele	Incorrect allele	--	--
chr2	227913871	A	ATATCTATCTATC	1|0	Correct	Correct	Correct	--
chr2	227913885	G	A	1|0	Correct	Correct	Correct	Correct
chr2	227913889	G	A	1|0	Correct	Correct	Correct	Correct
chr2	227913928	T	TA	1|0	Correct	Correct	Correct	--
chr2	227913931	A	T	1|0	Correct	Correct	Correct	Correct
chr3	5336450	A	T,C	1|2	Incorrect allele and zygosity	Incorrect allele and zygosity	Incorrect allele and zygosity	Incorrect allele and zygosity
chr3	5336452	C	CGCGT	0|1	--	--	--	--
chr3	5336465	ACACACACACACG	A	0|1	--	--	--	--
chr3	5336477	GCA	G	1|0	Correct	Incorrect allele	Incorrect allele	--
chr3	5336487	A	G	0|1	Incorrect variant type	--	Correct	Correct
chr6	160009985	C	CTTAA	0|1	Incorrect allele	Incorrect allele	Correct	--
chr6	160009986	C	A	0|1	Incorrect zygosity	Incorrect allele	Incorrect variant type	Incorrect zygosity
chr6	167130970	G	GGGCCCCCCTCCCTCCGGGACTCCTCCCTCT	0|1	--	--	--	--
chr6	167130972	GA	G	1|0	Correct	Correct	Correct	--
chr6	167130973	A	G	0|1	Incorrect allele and zygosity	Incorrect allele	Correct	Incorrect allele and zygosity
chr6	167130976	A	C	1|1	Correct	Correct	Correct	Correct

chr6	167130986	T	C	1|0	Correct	Correct	Incorrect variant type	Correct
chr6	167130989	A	G	1|1	Correct	Correct	Correct	Correct
chr6	167130990	C	A	1|0	Correct	Correct	--	Correct
chr6	167130992	C	T	1|0	Correct	--	--	Correct
chr9	134784949	C	T	0|1	Correct	Incorrect variant type	--	Correct
chr9	134784951	G	T	0|1	--	--	--	--
chr9	134784955	G	GGGGGGCA	0|1	Incorrect allele	--	Incorrect allele	--
chr9	134784956	T	G	0|1	Correct	Incorrect variant type	Incorrect variant type	Correct
chr9	135663795	ACAGAGGGGGACCTGGAGGGGCAGAGGAGAGACCTGTGGGG	A	0|1	--	Correct	--	--
chr9	135663892	A	G	1|1	Correct	Correct	Correct	Correct
chr9	135663893	T	A,G	1|2	Correct	Correct	Correct	Incorrect allele and zygosity
chr11	113466435	G	GC	1|1	Correct	Correct	Correct	--
chr11	113466437	A	T	1|1	Correct	Correct	Incorrect variant type	Correct
chr12	100940063	A	AT	0|1	Correct	Correct	Correct	--
chr12	100940065	G	C	0|1	Correct	Correct	Correct	Correct
chr14	75318035	C	T	1|0	Correct	Correct	Correct	Correct
chr14	75318038	AT	A	1|0	Incorrect allele	Incorrect allele	Correct	--
chr14	75318052	T	C	1|0	Correct	--	Correct	Correct
chr14	75318054	T	G	1|0	Correct	Correct	Correct	Correct
chr20	11064571	T	TGA	0|1	Correct	Correct	--	--
chr20	11064574	A	ATTTTCAAGACTATTGTGACTATGAC	0|1	Correct	Correct	Correct	--
chr20	11064578	A	T	0|1	Correct	Correct	--	--
chr20	11064579	C	T	0|1	Correct	Correct	Incorrect variant type	Correct

Quality score thresholds for filtering variant calls

We provide precision/recall/F1 statistics for each variant caller with respect to the recommended quality score, or by taking average over quality scores giving highest F1-score when evaluated by RTG's *vcfeval*. Here we show the quality thresholds we used for different Nanopore models on various datasets to show how the quality of reads, depth and quality of ground truth variants effects the optimal quality score. The range of SNP quality scores is 30-999. **Table S17** shows that SNP F1-scores of both NanoCaller1 and NanoCaller2 SNP models are very resilient to small changes on quality score thresholds. For NanoCaller1 and NanoCaller3 SNP models, we get the best performance on PacBio CCS reads when we do not use any threshold, which corresponds to a quality score threhsold of 30. For PacBio CLR reads, genome coverage significantly effects SNP calling performance, thus we choose different thresholds for HG001/HG002 and HG003/HG004, as shown in **Table S18**.

Table S17. F1 scores of NanoCaller1 and NanoCaller2 SNP models on ONT datasets. We recommend 162 and 78 as quality thresholds for NanoCaller1 and NanoCaller2 models, respectively. The column 'Quality Score Range' shows the range of quality scores that give the optimal SNP F1-score shown in the 'Best F1' column. Third colum under each model shows the F1-score when we use the recommended tquality threshold.

	NanoCaller1			NanoCaller2		
	Quality Score Range	Best F1	F1 with Quality Score=162	Quality Score Range	Best F1	F1 with Quality Score=78
HG001	120-120	95.54	95.40	65-70	94.93	94.89
HG002	179-201	98.03	97.99	99-117	98.03	97.97
HG003	157-185	98.88	98.88	62-85	98.88	98.88
HG004	166-191	98.78	98.77	71-87	98.82	98.82
HX1	191-194	90.52	90.28	93-97	91.07	90.94

Table S18. F1 scores of NanoCaller1 and NanoCaller3 SNP models on CLR datasets. We show recommended thresholds for higher coverage genomes HG001/HG002 and lower coverage genomes HG003/HG004. The column 'Quality Score Range' shows the range of quality scores that give the optimal SNP F1-score shown in the 'Best F1' column. Third and fourth columns under each model shows the F1-score when we use the recommended quality threshold.

	NanoCaller1				NanoCaller3			
	Quality Score Range	Best F1	Reported Quality Cut-off	F1 score using reported Cut-off	Quality Score Range	Best F1	Reported Quality Cut-off	F1 score using reported Cut-off
HG001	30	93.97	None	93.97	30-37	94.95	55	94.90
HG002	32-34	98.26	None	98.26	79-111	98.29	55	98.24
HG003	108-114	94.33	110	94.33	202-210	93.53	210	93.53
HG004	118-132	93.49	110	93.47	209-218	92.61	210	92.59

Table S19. F1 scores of NanoCaller ONT and PacBio indel models. We recommend 144 and 25 as quality thresholds for ONT and PacBio models, respectively. The column 'Quality Score Range' shows the range of quality scores that give the optimal indel F1-score shown in the 'Best F1' column. Third colum under each model shows the F1-score when we use the recommended quality threshold.

	ONT model			PacBio model		
	Quality Score Range	Best F1	F1 with Quality Score=144	Quality Score Range	Best F1	F1 with Quality Score=25
HG001	243-247	56.79	53.30	2-8	94.16	94.09
HG002	145-152	79.75	79.74	2-5	94.52	94.40
HG003	101-101	81.31	81.23	52-55	93.68	93.53
HG004	101-104	79.76	79.71	56-64	93.61	93.37

Table S20. Quality thresholds used for different variant callers for different sequencing technologies. For Clair and PacBio we use the developer recommended thresholds, whereas DeepVariant developers do not recommend using any threshold. For Medaka, we calculated the average of best quality scores for SNPs over five ONT genomes, and used the average as final SNP quality cut-off, and repeated the same procedure for indels as well. For WhatsHap, best results were obtained without any quality cut-off as well.

Variant Caller	Sequencing Type	Threshold
Medaka	ONT	2 (SNPs), 23 (indels)
Clair	ONT	748
	PacBio	113
Longshot	ONT	65
	PacBio	50
WhatsHap	ONT, PacBio	None
DeepVariant	PacBio	None

Runtime Comparisons

Table 21. Wall-clock runtimes for various variant callers using Intel Xeon CPU E5-2683 v4 @ 2.10GHz. For variant callers that support parallelization we used 16 CPUs. We used HG002 ONT 49X, PacBio CCS 35X and PacBio CLR 58X datasets for evaluation.

	Variant Caller	Time (hours)	Types of variants	Number of CPUs	Parallelization Supported
ONT 49X	NanoCaller	18.4	SNPs and Indels	16	Yes
	Medaka	181.6	SNPs and Indels	16	Yes
	Clair	5.6	SNPs and Indels	16	Yes
	Longshot	49.7	SNPs	1	No
	WhatsHap	84.3	SNPs	1	No
CCS 35X	NanoCaller	11.2	SNPs and Indels	16	Yes
	Clair	1.8	SNPs and Indels	16	Yes
	Longshot	14.4	SNPs	1	No
	WhatsHap	57.3	SNPs	1	No
	DeepVariant	11.8	SNPs and Indels	16	Yes
CLR 58X	NanoCaller	4.8	SNPs	16	Yes
	Clair	3.2	SNPs and Indels	16	Yes
	Longshot	11.1	SNPs	1	No

Selection of nearby potentially heterozygous sites for SNP calling feature generation

Table S22. Number of potentially heterozygous Sites chosen for each candidate site under two methods. Method 1 is used for SNP calling ONT reads and selects the given number of sites from each range given. Method 2 is used for SNP calling on PacBio reads. In each range for either method, we chose the specified number of sites that are closest to the candidate site because they share more reads with the candidate site. The design of these two methods was motivated by the difference in read length distribution of ONT and PacBio reads, shown in Figure S9. An illustration of the two methods is shown in Figure S10.

Method 1 (used for SNP calling on ONT reads)		Method 2 (used for SNP calling on PacBio reads)	
Range	Number of Potentially Heterozygous Sites Chosen	Range	Number of Potentially Heterozygous Sites Chosen
-50kbp to -20kbp	6		
-20kbp to -10kbp	5	-20kbp to 0	20
-10kbp to -5kbp	4		
-5 kbp to -2kbp	3		
-2kbp to 0	2		
SNP Candidate Site		SNP Candidate Site	
0 to +2kbp	2	0 to +20kbp	20
+2kbp to +5kbp	3		
+5kbp to +10kbp	4		
+10kbp to +20kbp	5		
+20kbp to +50kbp	6		

Table S23. Performance of NanoCaller SNP calling on HG004 ONT, CCS and CLR reads with various thresholds used for define heterozygous SNPs. All the results show the best F1-score achieved using each threshold, on v4.2 benchmark variants.

Oxford Nanopore Reads			
Threshold	Precision	Recall	F1
0.30-0.60	0.9846	0.9843	0.9845
0.30-0.70	0.9840	0.9833	0.9837
0.30-0.80	0.9814	0.9797	0.9806
0.35-0.65	0.9878	0.9857	0.9868
0.40-0.60	0.9888	0.9867	**0.9878**
0.40-0.70	0.9882	0.9847	0.9865
0.40-0.80	0.9845	0.9790	0.9818
0.45-0.55	0.9880	0.9851	0.9866
PacBio CLR Reads			
Threshold	Precision	Recall	F1
0.20-0.70	0.9527	0.8829	0.9165
0.30-0.60	0.9595	0.9114	**0.9349**
0.30-0.70	0.9559	0.8971	0.9256
0.30-0.80	0.9410	0.8730	0.9058
0.35-0.65	0.9534	0.8989	0.9254
0.40-0.60	0.9457	0.8906	0.9174
0.40-0.70	0.9409	0.8689	0.9035
0.40-0.80	0.9215	0.8400	0.8789
0.45-0.55	0.9228	0.8607	0.8907
PacBio CCS Reads			
Threshold	Precision	Recall	F1
0.30-0.60	0.9920	0.9944	0.9932
0.30-0.70	0.9936	0.9963	**0.9950**
0.30-0.80	0.9935	0.9967	0.9951
0.35-0.65	0.9932	0.9950	0.9941
0.40-0.60	0.9900	0.9899	0.9900
0.40-0.70	0.9936	0.9950	0.9943
0.40-0.80	0.9935	0.9957	0.9946
0.45-0.55	0.9769	0.9702	0.9735

Table S24. Performance of NanoCaller SNP calling on HG004 ONT and CCS reads with various thresholds for minimum number (first column) of heterozygous SNPs required for a a candidate site. All the results show the best F1-score achieved using each threshold, on v4.2 benchmark variants.

Oxford Nanopore Reads			
Minimum Number	Precision	Recall	F1
1	98.88	98.67	98.78
2	98.84	98.69	98.77
3	98.84	98.65	98.75
5	98.90	98.35	98.63
7	98.94	97.66	98.30
9	98.96	96.43	97.68
13	98.93	92.21	95.46
PacBio CCS			
Minimum Number	Precision	Recall	F1
1	**99.37**	**99.63**	**99.50**
3	99.41	95.54	97.44
5	99.44	90.53	94.77
7	99.44	87.04	92.83
9	99.44	84.79	91.53

Table S25. Performance of NanoCaller SNP calling, in whole genome and various difficult genomic regions, for HG002, HG003 and HG004 ONT reads using both, Method1 and Method2 described in Table S22 and Error! Reference source not found.**. All the results show the best F1-score achieved using each method for both NanoCaller1 and NanoCaller2 SNP models, with evaluation on v4.2 benchmark variants.**

NanoCaller1							
Genomic Regions	Genomes	Method1 (for ONT reads)			Method2 (for PacBio reads)		
		Precision	Recall	F1	Precision	Recall	F1
Whole genome	HG002	98.23	97.82	98.03	98.19	97.44	97.82
	HG003	99.00	98.76	98.88	98.98	98.51	98.75
	HG004	98.88	98.67	98.78	98.80	98.45	98.63
All difficult regions	HG002	96.59	94.68	95.63	96.19	94.65	95.42
	HG003	97.13	96.44	96.79	97.02	96.34	96.68
	HG004	97.19	96.19	96.69	97.03	96.07	96.55
Low Mappability Regions	HG002	96.63	97.40	97.02	96.61	97.14	96.88
	HG003	96.77	98.37	97.57	96.80	98.18	97.49
	HG004	97.00	98.15	97.58	96.80	98.14	97.47
Tandem and Homopolymer Repeats	HG002	93.87	88.04	90.87	92.98	88.27	90.57
	HG003	95.32	90.47	92.84	95.04	90.33	92.63
	HG004	94.99	89.99	92.43	94.83	89.66	92.18
Segmental Duplications	HG002	94.07	95.81	94.94	93.83	95.72	94.77
	HG003	94.65	96.99	95.81	94.51	96.97	95.73
	HG004	94.51	97.26	95.87	94.41	97.20	95.79
MHC	HG002	98.89	98.18	98.54	98.83	98.12	98.48
	HG003	99.36	99.11	99.24	99.39	98.88	99.14
	HG004	99.21	98.96	99.09	99.21	98.82	99.02
NanoCaller2							
Genomic Regions	Genomes	Method1 (for ONT reads)			Method2 (for PacBio reads)		
		Precision	Recall	F1	Precision	Recall	F1
Whole genome	HG002	98.26	97.79	98.03	98.09	97.37	97.73
	HG003	99.03	98.72	98.88	98.95	98.40	98.68
	HG004	98.97	98.66	98.82	98.90	98.29	98.60
All difficult regions	HG002	96.07	94.65	95.36	95.99	94.26	95.12
	HG003	96.93	96.21	96.57	96.91	95.91	96.41
	HG004	96.91	96.09	96.50	96.87	95.77	96.32
Low Mappability Regions	HG002	96.11	97.17	96.64	96.01	96.86	96.44
	HG003	96.18	98.21	97.19	96.26	97.90	97.08
	HG004	96.24	98.25	97.24	96.27	97.88	97.07
Tandem and Hoomopolymer Repeats	HG002	93.88	87.39	90.52	93.92	86.76	90.20
	HG003	95.56	89.51	92.44	95.35	89.25	92.20
	HG004	95.25	89.01	92.03	95.17	88.64	91.79
Segmental Duplications	HG002	92.73	95.63	94.16	92.24	95.65	93.92
	HG003	93.21	97.14	95.14	93.25	96.85	95.02
	HG004	93.59	97.12	95.33	93.50	96.91	95.18

MHC	HG002	98.90	98.25	98.58	98.88	97.95	98.42
	HG003	99.43	98.98	99.21	99.47	98.69	99.08
	HG004	99.43	98.94	99.19	99.38	98.83	99.11

Table S26. Performance of NanoCaller whole genome SNP calling for HG002, HG003 and HG004 PacBio CCS and CLR reads using both, Method1 and Method2 described in Table S22 and Error! Reference source not found.**. All the results show the best F1-score achieved using each method with NanoCaller1 SNP model, with evaluation on v4.2 benchmark variants.**

PacBio CCS Reads							
	Genomes	Method1 (for ONT reads)			Method2 (for PacBio reads)		
		Precision	Recall	F1	Precision	Recall	F1
Whole genome	HG002	99.27	99.53	99.40	99.31	99.60	99.45
	HG003	99.33	99.56	99.45	99.36	99.63	99.49
	HG004	99.33	99.57	99.45	99.37	99.63	99.50
PacBio CLR Reads							
	Genomes	Method1 (for ONT reads)			Method2 (for PacBio reads)		
		Precision	Recall	F1	Precision	Recall	F1
Whole genome	HG002	98.68	97.73	98.20	98.55	97.98	98.27
	HG003	96.15	92.47	94.28	96.46	92.28	94.33
	HG004	95.80	91.15	93.42	95.95	91.14	93.49

Effects of allowing WhatsHap to change genotype on SNP calling performance

Table S27. Performance of SNP calling by NanoCaller with and without the use of 'distrust genotypes' for phasing by WhatsHap. All the results shown elsewhere besides this table do not use 'distrust genotype' option whiich allows WhatsHap to change genotypes. NanoCaller users can enable 'distrust genotype' option by setting 'enable_whatshap' flag in NanoCaller run. By default this option is turned off in NanoCaller.

Oxford Nanopore Reads												
	NanoCaller1						NanoCaller2					
	Without 'enable_whatshap'			With 'enable_whatshap'			Without 'enable_whatshap'			With 'enable_whatshap'		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
HG001	96.50	94.59	95.54	96.53	94.90	95.71	96.13	93.75	94.93	96.20	94.07	95.13
HG002	98.23	97.82	98.03	98.24	97.87	98.06	98.26	97.79	98.03	98.19	97.92	98.06
HG003	99.00	98.76	98.88	98.93	98.80	98.87	99.03	98.72	98.88	99.05	98.68	98.87
HG004	98.88	98.67	98.78	98.83	98.70	98.77	98.97	98.66	98.82	98.94	98.66	98.80
HX1	92.61	88.51	90.52	92.05	88.41	90.20	93.39	88.85	91.07	92.89	88.89	90.85
PacBio CCS Reads												
	NanoCaller1						NanoCaller3					
	Without 'enable_whatshap'			With 'enable_whatshap'			Without 'enable_whatshap'			With 'enable_whatshap'		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
HG001	96.97	99.37	98.15	97.44	99.72	98.57	96.69	99.36	98.01	97.28	99.81	98.53
HG002	99.31	99.60	99.45	99.56	99.79	99.68	99.16	99.56	99.36	99.52	99.83	99.68
HG003	99.36	99.63	99.49	99.58	99.78	99.68	99.22	99.58	99.40	99.54	99.82	99.68
HG004	99.37	99.63	99.50	99.58	99.77	99.67	99.23	99.59	99.41	99.53	99.81	99.67
PacBio CLR Reads												
	NanoCaller1						NanoCaller3					
	Without 'enable_whatshap'			With 'enable_whatshap'			Without 'enable_whatshap'			With 'enable_whatshap'		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
HG001	97.72	90.51	93.97	98.24	90.98	94.47	96.78	93.20	94.95	97.68	94.01	95.81
HG002	98.55	97.98	98.27	98.78	98.25	98.51	98.11	98.46	98.29	98.17	99.09	98.63
HG003	96.46	92.28	94.33	96.56	93.42	94.97	95.58	91.56	93.53	95.93	92.52	94.20
HG004	95.95	91.14	93.49	96.24	92.22	94.19	94.98	90.34	92.61	95.50	91.23	93.32

Figures

Figure S8. Concordance of ground truth variants correctly predicted by various variant callers on Nanopore reads of the Ashkenazim trio. Venn diagrams show the overlap of v4.2 ground truth variant calls predicted correctly by NanoCaller, Medaka and Clair. a) SNPs, b) indels. All variants are inside high-confidence regions.

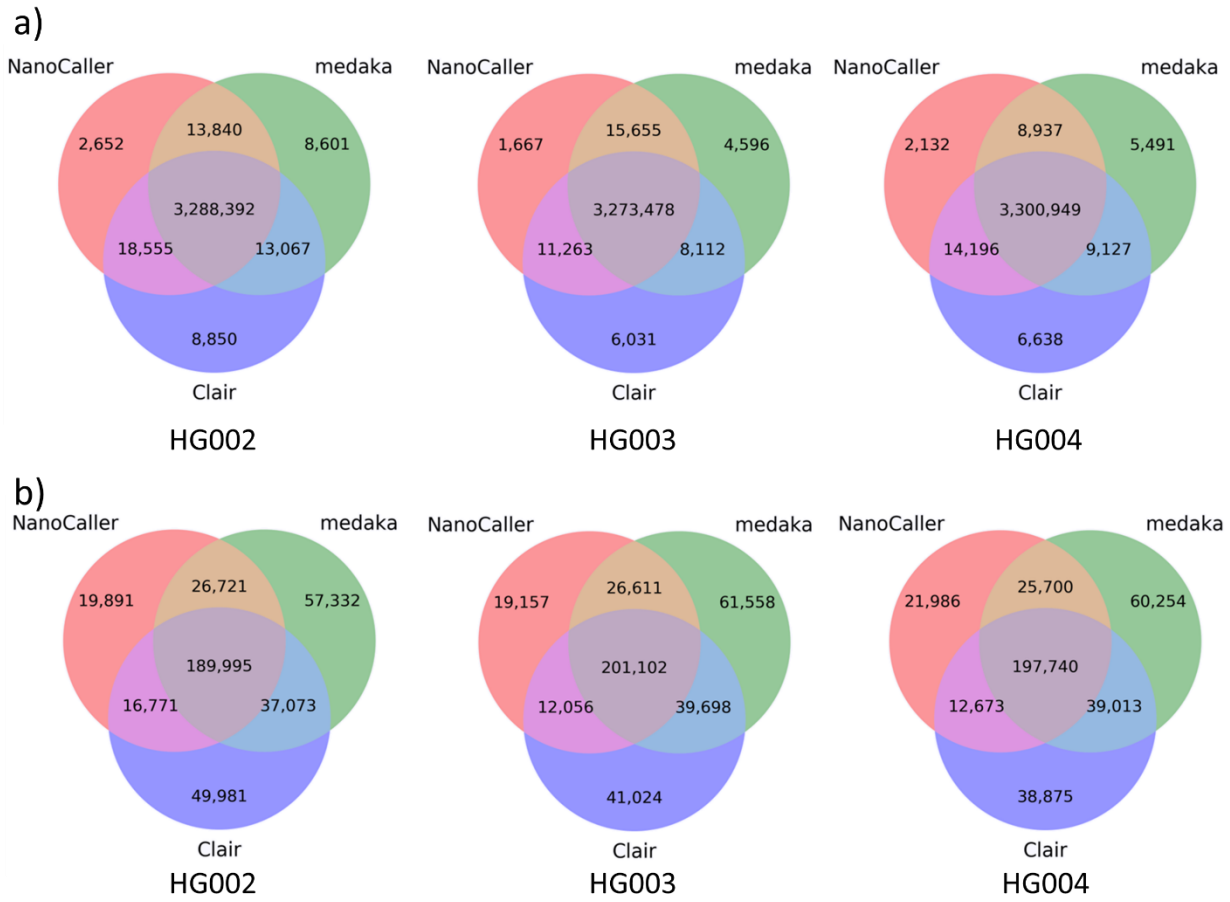


Figure S9. Read length distributions of 88X ONT, 35X PacBio CCS and 27X PacBio CLR reads of HG004.

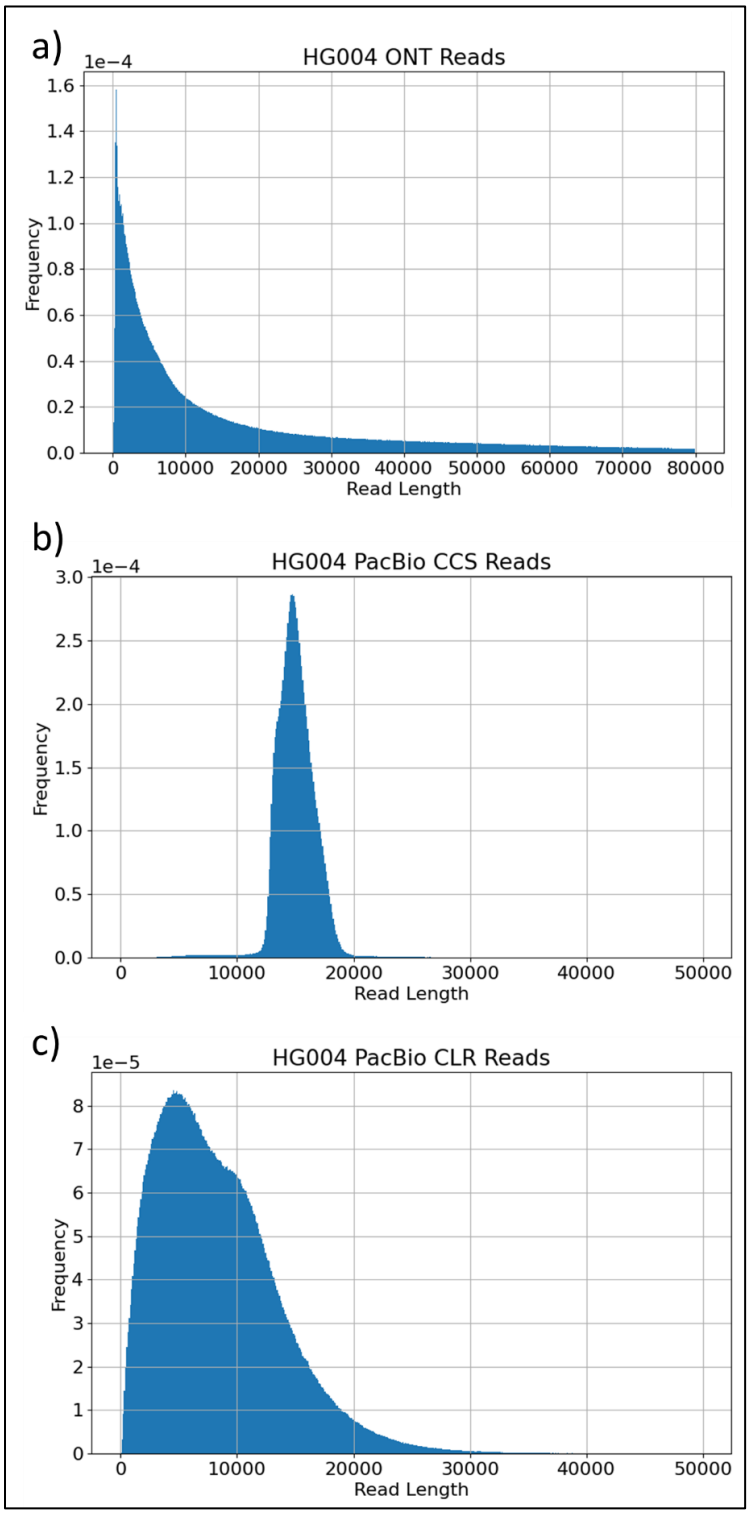
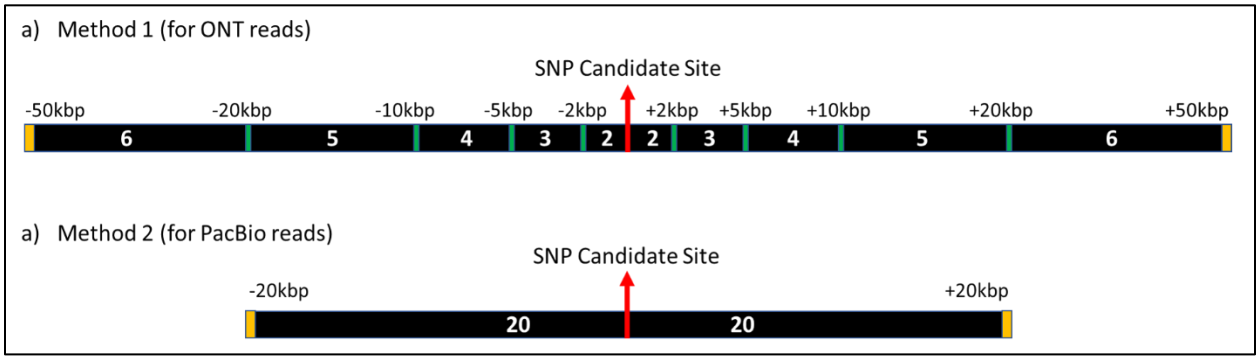



Figure S10. Illustration of number of potentially heterozygous SNP sites chosen in each range for the two methods.