Microbiology and Infectious Disease
Computational and Systems Biology

Benchmarking reveals superiority of deep learning variant callers on bacterial nanopore sequence data

Reviewed Preprint

Published from the original preprint after peer review and assessment by eLife.

About eLife's process

Reviewed preprint version 1
May 20, 2024 (this version)

Sent for peer review
March 29, 2024

Posted to preprint server
March 28, 2024

Michael B. Hall, **Ryan R. Wick, Louise M. Judd, An N. T. Nguyen, Eike J. Steinig, Ouli Xie, Mark R. Davies, Torsten Seemann, Timothy P. Stinear, Lachlan J. M. Coin**

Department of Microbiology and Immunology, The University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Melbourne, Australia • Centre for Pathogen Genomics, The University of Melbourne, Melbourne, Victoria, Australia • Department of Infectious Diseases, The University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Melbourne, Australia • Monash Infectious Diseases, Monash Health, Melbourne, Australia

Abstract

Variant calling is fundamental in bacterial genomics, underpinning the identification of disease transmission clusters, the construction of phylogenetic trees, and antimicrobial resistance prediction. This study presents a comprehensive benchmarking of SNP and indel variant calling accuracy across 14 diverse bacterial species using Oxford Nanopore Technologies (ONT) and Illumina sequencing. We generate gold standard reference genomes and project variations from closelyrelated strains onto them, creating biologically realistic distributions of SNPs and indels.

Our results demonstrate that ONT variant calls from deep learning-based tools delivered higher SNP and indel accuracy than traditional methods and Illumina, with Clair3 providing the most accurate results overall. We investigate the causes of missed and false calls, highlighting the limitations inherent in short reads and discover that ONT's traditional limitations with homopolymerinduced indel errors are absent with high-accuracy basecalling models and deep learning-based variant calls. Furthermore, our findings on the impact of read depth on variant calling offer valuable insights for sequencing projects with limited resources, showing that 10x depth is sufficient to achieve variant calls that match or exceed Illumina.

In conclusion, our research highlights the superior accuracy of deep learning tools in SNP and indel detection with ONT sequencing, challenging the primacy of short-read sequencing. The reduction of systematic errors and the ability to attain high accuracy at lower read depths enhance the viability of ONT for widespread use in clinical and public health bacterial genomics.

eLife assessment

This **important** study provides evidence for a combination of the latest generation of Oxford Nanopore Technology long reads with state-of-the art variant callers enabling bacterial variant discovery at accuracy that matches or exceeds the current "gold standard" with short reads. The evidence supporting the claims of the authors is **convincing**, although the inclusion of a larger number of reference genomes would further strengthen the study. The work will be of interest to anyone performing sequencing for outbreak investigations, bacterial epidemiology, or similar studies.

https://doi.org/10.7554/eLife.98300.1.sa3

Introduction

Variant calling is a cornerstone of bacterial genomics as well as one of the major applications of next generation sequencing. Its downstream applications include identification of disease transmission clusters, prediction of antimicrobial resistance, and phylogenetic tree construction and subsequent evolutionary analyses, to name a few [1–4]. Variant calling is used extensively in public health laboratories to inform decisions on managing bacterial outbreaks [5] and in molecular diagnostic laboratories as the basis for clinical decisions on how to best treat patients with disease [6].

Over the last 15 years, short-read sequencing technologies, such as Illumina, have been the mainstay of variant calling in bacterial genomes, largely due to its relatively high level of basecalling accuracy. However, nanopore sequencing on devices from Oxford Nanopore Technologies (ONT) have emerged as an alternative technology. One of the major advantages of ONT sequencing from an infectious diseases public health perspective is the ability to generate sequencing data in near real-time, as well as the portability of the devices, which has enabled researchers to sequence in remote regions, closer to the source of the disease outbreak [7, 8]. Limitations in ONT basecalling accuracy have historically limited its widespread adoption for bacterial genome variant calling [9]. ONT have recently released a new R10.4 pore, along with a new basecaller[1] with 3 different accuracy modes (fast, high-accuracy [hac] and super high-accuracy [sup]). These new software and hardware bring the ability to identify a subset of paired reads for which both strands have been sequenced (duplex), leading to impressive gains in basecalling accuracy [10–12].

A number of variant callers have been developed for ONT sequencing [13–15]. However, to date, benchmarking studies have focused on human genome variant calling, and have mostly used the older pores, which do not have the ability to identify duplex reads [16–18]. In addition, modern deep learning-based variant callers use models trained on human DNA sequence only, leaving an open question of their generalisability to bacteria [14, 15, 19]. Human genomes have a very different distribution of *k*-mers and patterns of DNA modification, and as such, results from human studies may not directly carry over into bacterial genomics. Moreover there is substantial *k*-mer and DNA modification variation within bacteria, mandating a broad multi-species approach for evaluation [20]. Existing benchmarks for bacterial genomes, while immensely beneficial and thorough, only assess short-read Illumina data [21, 22].

In this study, we conduct a benchmark of SNP and indel variant calling using ONT and Illumina sequencing across a comprehensive spectrum of 14 Gram-positive and Gram-negative bacterial species. We used the same DNA extractions for both Illumina and ONT sequencing to ensure our

results are not biased by acquisition of new mutations during culture. We develop a novel strategy for generating benchmark variant truthsets in which we project variations from different strains onto our gold standard reference genomes in order to create biologically realistic distribution of SNPs and indels. We assess both deep learning-based and traditional variant calling methods and investigate the sources of errors and the impact of read depth on variant accuracy.

Results

We analysed ONT data basecalled with three different accuracy models - fast, high accuracy (hac), and super-high accuracy (sup) - along with different read types - simplex and duplex (Basecalling and quality control). Duplex reads are those in which both DNA strands from a single molecule are sequenced back-to-back, whereas simplex is the standard reads where only a single DNA strand is sequenced. The median (unfiltered) read identities for each basecalling model and read type are shown in ***Figure 1***. Duplex reads basecalled with the sup model had the highest median read identity of 99.93% (Q32), followed by duplex hac (99.79% [Q27]), simplex sup (99.26% [Q21]), simplex hac (98.31% [Q18]), and simplex fast (94.09% [Q12]). Full summary statistics of the reads can be found in Supplementary Table S1.

Genome and variant truthset

Ground truth reference assemblies were generated for each sample using ONT and Illumina reads (Genome assembly).

Creating a truthset of variants for such a benchmark is nontrivial [23, 24]. Theoretically, calling variants with respect to a sample's own reference would yield no variant calls. Therefore, in order to have sufficient variants to assess calling methods, we generated a mutated version of a sample's reference. Rather than simulate mutations randomly, which does not imitate natural mutational patterns, we took a pseudo-real approach [24, 25]. We identified variants between a sample's reference and a donor genome and applied these variants to our reference, thus creating a mutated reference. This approach has the advantage of a simulation, in that we can be certain of the truthset of variants, but with the added benefit of the variants being real differences between two genomes.

For each sample, we chose a (donor) genome with ANI closest to 99.5% (see Truthset and reference generation). We then identified all variants between the sample and this donor using both minimap2 [26] and mummer [27]. We took the intersection of those variant sets and removed overlaps and duplicates, along with indels longer than 50bp. This variant truthset was then applied to the sample's reference to create a mutated reference. This way we know exactly what variants are expected between a sample's reads and the mutated reference, with no complications from large structural differences between the sample and donor. While incorporating structural variation would be an interesting and useful addition to the current work, we chose to focus here on small (< 50bp) variants.

Table 1 shows a summary of the samples used in this work, as well as the number of variants between each sample and its donor genome, along with the ANI between the two. In total, we have 14 samples from 14 different species, which span a wide range of GC content (30-66%). Despite the wide range in the number of SNPs (2102-57887), the number of indels is reasonably consistent (in the same order of magnitude) across the samples (see Suppl. Table S2 for full details).

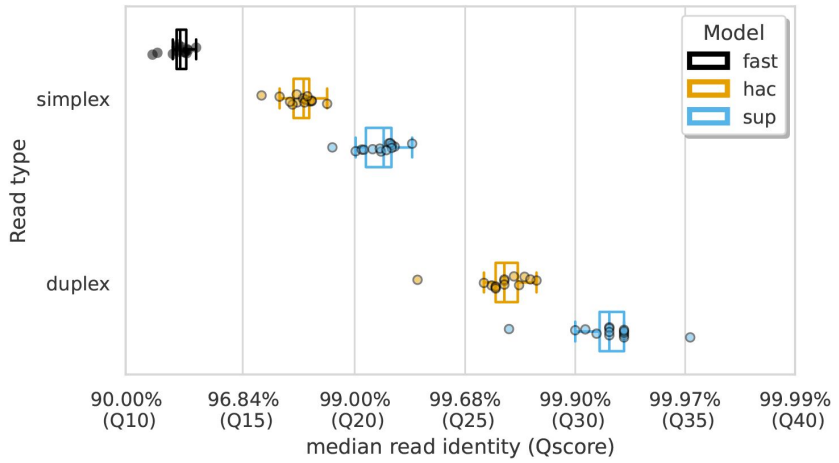

Figure 1.

Median read identity (x-axis) for each sample (points) stratified by basecalling model (colours) and read type (y-axis).

Sample	Species	ANI (%)	GC (%)	SNPs	Insertions	Deletions	Total variants
ATCC_33560__202309	*Campylobacter jejuni*	99.50	30.22	6369	117	106	6592
ATCC_35221__202309	*Campylobacter lari*	98.64	29.81	16541	57	67	16665
ATCC_25922__202309	*Escherichia coli*	99.50	50.42	4531	119	242	4892
KPC2__202310	*Klebsiella pneumoniae*	99.50	57.15	15877	90	78	16045
AJ292__202310	*Klebsiella variicola*	99.50	57.62	22850	95	98	23043
ATCC_19119__202309	*Listeria ivanovii*	99.46	37.13	8451	187	259	8897
ATCC_BAA-679__202309	*Listeria monocytogenes*	99.50	37.98	9090	66	78	9234
ATCC_35897__202309	*Listeria welshimeri*	99.03	36.35	16953	130	133	17216
AMtb_1__202402	*Mycobacterium tuberculosis*	99.73	65.62	2102	95	84	2281
ATCC_10708__202309	*Salmonella enterica*	99.36	52.20	18784	210	189	19183
BPH2947__202310	*Staphylococcus aureus*	99.48	32.80	7894	95	63	8052
MMC234__202311	*Streptococcus dysgalactiae*	99.16	39.49	10474	82	100	10656
RDH275__202311	*Streptococcus pyogenes*	99.50	38.32	5361	60	68	5489
ATCC_17802__202309	*Vibrio parahaemolyticus*	98.75	45.32	57887	280	304	58471

Table 1.

Summary of the ANI and number of variants found between each sample and its donor genome.

Which method is the best?

For this study, we benchmarked the performance of six variant callers on ONT sequencing data: BCFtools (v1.19; [28]), Clair3 (v1.0.5; [14]), DeepVariant (v1.6.0; [19]), FreeBayes (v1.3.7; [29]), Longshot (v0.4.5; [13]), and NanoCaller (v3.4.1; [15]). In addition, we called variants from each sample's Illumina data using Snippy[2] to act as a performance comparison.

Alignments of ONT reads to each sample's mutated reference (Genome and variant truthset) were produced with minimap2 and provided to each variant caller. Variant calls were assessed against the truthset using vcfdist (v2.3.3; [30]), which classifies each called variant as either a true positive (TP) or false positive (FP; i.e., the variant is not in the truthset) and each truth variant as either TP (i.e., the variant was correctly identified by the caller) or false negative (FN; i.e., the caller missed the variant). With these classifications, precision, recall, and the F1 score were calculated at every increment of the VCF quality score for both SNPs and indels. ***Figure 2*** shows the highest

F1 score achieved by each variant caller for every sample, basecalling model, read type and variant type.

The F1 score is the harmonic mean of precision and recall and acts as a good metric for overall evaluation. From ***Figure 2*** we see that Clair3 and DeepVariant produce the highest F1 scores for both SNPs and indels with both read types. Unsurprisingly, the sup basecalling model leads to the highest F1 scores across all variant callers; though hac is not much lower. SNP F1 scores of 99.99% are obtained from Clair3 and DeepVariant on sup-basecalled data. For indel calls, Clair3 achieves F1 scores of 99.53% and 99.20% for sup simplex and duplex, respectively, while DeepVariant scores 99.61% and 99.22%. The higher depth of the simplex reads likely explains why the best duplex indel F1 scores are slightly lower than simplex (see How much read depth is enough?). The precision and recall values at the highest F1 score can be seen in Supplementary Figures S1 and S2 (see Suppl. Table S3 for a summary and S4 for full details). What is clear from ***Figure 2*** is that reads basecalled with the fast model are an order of magnitude worse than the hac and sup models.

Figure 3 shows the precision-recall curves for the sup basecalling model (see Suppl. Figures S3 and S4 for the hac and fast model curves, respectively) for each variant and read type - aggregated across samples to produce a single curve for each variant caller. Due to the right-angle-like shape of the Clair3 and DeepVariant curves, filtering based on low-value variant quality improves precision considerably for variant calls, without losing much recall. A similar pattern holds true for BCFtools SNP calls. The best Clair3 and DeepVariant F1 scores are obtained with no quality filtering on sup data, except for indels from duplex data were a value of 4 provides the best. See Suppl. Table S5 for the full details.

A striking feature of ***Figure 2*** and ***Figure 3*** is the comparison of the deep learning-based variant callers (Clair3, DeepVariant, and NanoCaller) to Illumina. For all variant and read types with hac or sup data - across F1 score, precision, and recall - these deep learning methods are as good as, or better than, Illumina: the median best SNP and indel F1 scores for Illumina are 99.45% and 95.76%, respectively. In the case of Clair3 and DeepVariant, they are an order of magnitude better. Traditional variant callers Longshot, BCFtools and FreeBayes match, or slightly exceed, Illumina for SNP calls with hac and sup model data. FreeBayes also matches Illumina for indel calls, however, BCFtools has reduced indel accuracy compared to Illumina for all models and read types (Longshot does not make indel calls). In all cases, fast model ONT data has a lower F1 score than Illumina, and only achieves parity in the best case for SNPs.

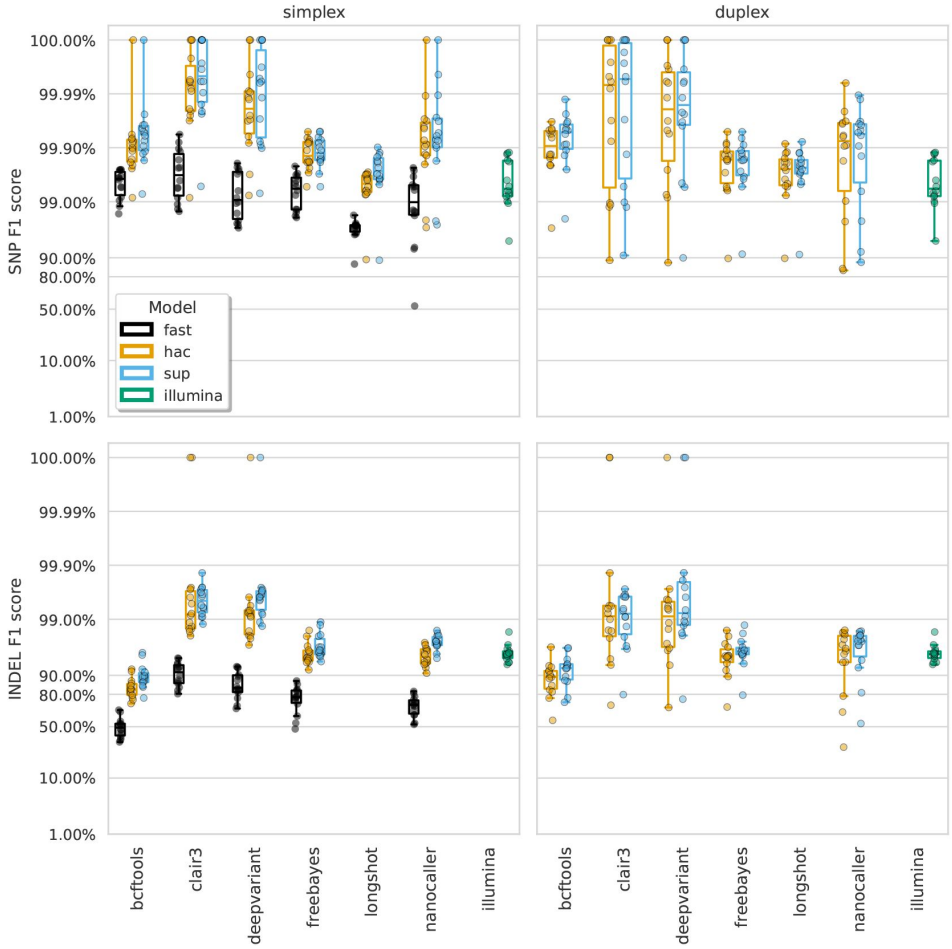

Figure 2.

The highest F1 score for each sample (points), stratified by basecalling model (colours), variant type (rows), and read type (columns). Illumina results (green) are included as a reference and do no have different basecalling models or read types. Note, longshot does not provide indel calls.

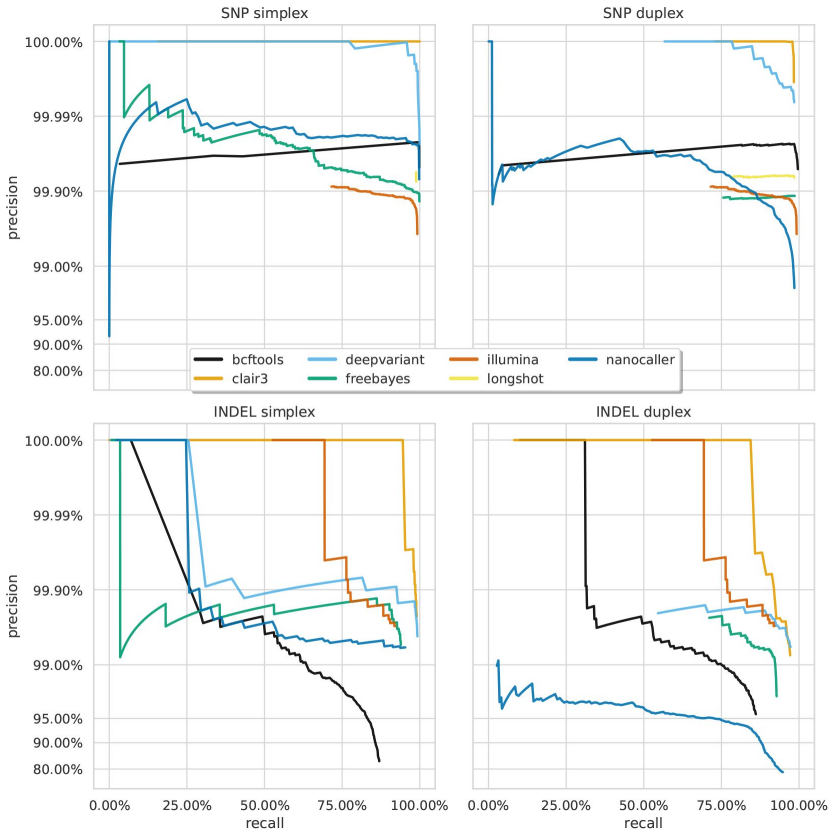

Figure 3.

Precision and recall curves for each variant caller (colours) on sequencing data basecalled with the sup model, stratified by variant type (rows) and read type (columns) and aggregated across samples. The curves are generated by using increasing variant quality score thresholds to filter variants and calculating precision and recall at each threshold. The lowest threshold is the lower right part of the curve, moving to the highest at the top left. Note, Longshot does not provide indel calls.

Understanding missed and false calls

Conventional wisdom will likely leave most readers surprised at finding ONT data can provide better variant calls than Illumina. In order to convince ourselves (and others) of these results, we investigate the main causes for this difference.

In light of the ONT read-level accuracy now exceeding Q20 (***Figure 1***; simplex sup), read length remains the primary difference between the two technologies. Suppl. Figure S2 highlights that the main contributor to the lower F1 score of Illumina is recall, rather than precision (Suppl. Figure S1). Given this, along with the shorter reads, we hypothesized that the Illumina errors could be related to alignment difficulties in repetitive or variant-dense regions.

Figure 4 demonstrates that variant density and repetitive regions account for many of the false negatives (and thus lower recall). We define variant density as the number of variants (missed or called) in a 100bp window centred on each call. ***Figure 4***a, in particular, reveals a bimodal distribution of variant density for Illumina FNs, with the second peak having very minimal overlap with the distribution of TP and FP calls. This secondary peak centres around 20 variants per 100bp. This is in contrast to one of the best-performing ONT variant callers, Clair3, showing no bimodal distribution for any call type (***Figure 4***b), with hardly any missed or false calls at a density of 20. Indeed, the read pileups show that Illumina reads cannot align across these variant-dense regions, while ONT can (Suppl. Figure S5); a density of 20 is a much larger proportion of an Illumina read when compared to an ONT read.

In addition to looking at the impact of variant density, we also assessed the change in F1 score when masking repetitive regions of the genome (Identifying repetitive regions). Again, due the shorter Illumina read length, we expect short reads to have a harder time aligning in repetitive regions than ONT [31]. The read pileups again illustrate the troubles Illumina reads have in repetitive regions, with Suppl. Figure S6 showing a number of missing variants and alignment gaps exclusive to Illumina data. These alignment difficulties are further quantified by the increase in F1 score for Illumina data when repetitive regions are masked (***Figure 4c***). Without masking, the median F1 score across all samples for Illumina is 99.3% and this increases to 99.7% when repetitive regions are masked - Clair3 100x simplex sup data only has an increase of 0.003%.

Indels have traditionally been a systematic weakness for ONT sequencing data; primarily driven by variability in the length of homopolymeric regions as determined by basecallers [9]. Having seen the drastic improvements in read accuracy in ***Figure 1***, we sought to determine whether false positive indel calls are still a byproduct of homopolymer-driven errors.

When analysing the best-performing ONT caller, Clair3, we see that for reads basecalled with the fast model, there is a tendency to get homopolymer lengths wrong by 1 or 2bp (***Figure 5***), though there is an equal amount of non-homopolymeric false indel calls. In contrast, the sup model produces a drastically reduced number of false indel calls of both kinds; with counts and a profile similar to Illumina. The hac model shows a marked improvement over the fast model, though it still produces a noticeable number of false indel calls, predominantly getting homopolymers wrong by 1bp. DeepVariant showed a similar error profile to Clair3 (Suppl. Figure S8), while FreeBayes (Suppl. Figure S9) and NanoCaller (Suppl. Figure S10) were not much worse. However, BCFtools (Suppl. Figure S7) suffers from this homopolymeric indel error bias, even with sup model reads, indicating that while the sup basecaller clearly has a reduced bias, it still exists to a certain extent, and the deep learning methods circumvent this by training models that take these systematic issues into account. Therefore, an honourable mention must go to FreeBayes, which is a traditional variant caller, and therefore handles errors with no *a priori* bias information.

Lastly, we did not see any systematic indel bias in the context of missed calls (Suppl. Figures S11-S15), especially when compared to Illumina indel error profiles.

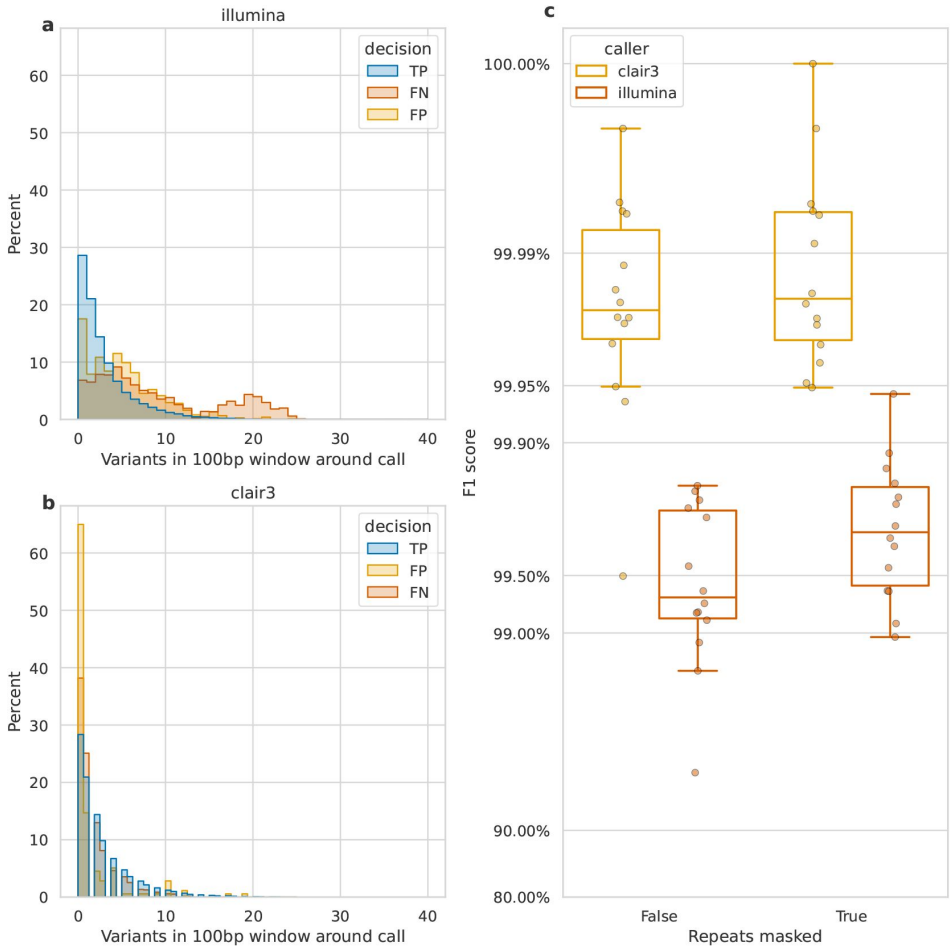

Figure 4.

Impact of variant density and repetitive regions on Illumina variant calling. Variant density is the number of (true or false) variants in a 100bp window centred on a call. a and b) the distribution of variant densities for true positive (TP), false positive (FP) and false negative (FN) calls. The y-axis, percent, indicates the percent of all calls of that decision that fall within the density bin on the x-axis. Illumina calls, aggregated across all samples are shown in a, while b shows Clair3 calls from simplex sup-basecalled reads at 100x depth. c) impact of repetitive regions on the F1 score (y-axis) for Clair3 (100x simplex sup) and Illumina. The x-axis indicates whether variants that fall within repetitive regions are excluded from the calculation of the F1 score. Points indicate the F1 score for a single sample.

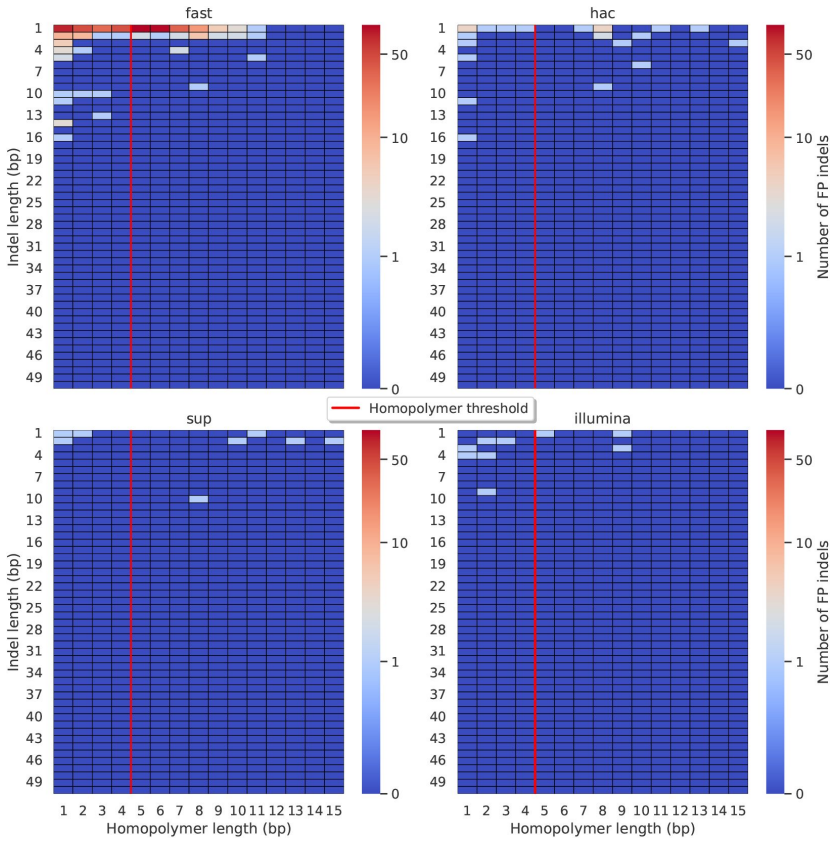

Figure 5.

Relationship between indel length (x-axis) and homopolymer length (y-axis) for false positive (FP) indel calls for Clair3 100x simplex fast (top left), hac (top right), and sup (lower left) calls. Illumina is shown in the lower right for reference. The vertical red line indicates the threshold above which we deem a run of the same nucleotide to be a 'true' homopolymer. Indel length is the number of bases inserted/deleted for an indel, whereas the homopolymer length indicates how long the tract of the same nucleotide is after the indel. The colour of a cell indicates how many FP indels of that indel-homopolymer length combination.

How much read depth is enough?

Having established the accuracy of variant calls from 'full depth' ONT datasets (100x), we turn to answering the question of how much ONT read depth is required to reach a precision or recall value of interest; this value will likely be different for different use cases and different levels of resource access. This question is a somewhat unique aspect of ONT, where sequencing can be analysed, and stopped, in real-time, when a 'sufficient' amount of data has been generated. An important consideration when time and/or resources are constrained.

To answer this question, we randomly subsampled each ONT readset with rasusa (v0.8.0; [32]) to average depths of 5, 10, 25, 50, and 100x and called variants with these reduced sets. Note, due to only one sample having a full readset duplex depth of 100x or more, 50x was the maximum read depth used for that read type, while 100x was used for simplex - these maximum read depths are the read sets that have been used in the preceding sections.

The F1 score, precision, and recall, as a function of read depth, is shown in ***Figure 6*** and ***Figure 7*** for SNPs and indels, respectively. As expected, precision and recall decrease as read depth is reduced. For SNPs, this drop is especially noticeable when going below 25x, with the same being true for indels with sup data. However, perhaps the most compelling result from this analysis is the finding that using Clair3 or DeepVariant on 10x of ONT sup simplex data provides F1 scores consistent with, or better than, full-depth Illumina for both SNPs and indels (see Table S1 for Illumina read depths); the same is also true of duplex hac or sup reads.

With 5x of ONT read depth the F1 score is lower than Illumina for almost all variant caller and basecalling models. However, BCFtools surprisingly produces SNP F1 scores on par with Illumina on duplex sup reads. Despite the inferior F1 scores across the board at 5x, SNP precision remains above Illumina with duplex reads for all methods except NanoCaller, and calls from Clair3 and DeepVariant simplex sup data.

What computational resources do I need?

The final consideration when calling variants is what computational resources are required. For those with the luxury of high-performance computing (HPC) access, this may be a trivial concern. However, due to the small size of bacterial genomes in comparison to eukaryotes, personal computers, such as laptops, are a common platform on which to analyse sequencing data. The two main resource types that can be limiting are memory and runtime. For variant calling, the heaviest steps in analysis are aligning the sequencing reads to a reference, and then taking that alignment and calling variants from it. Though we do note that if the person performing the variant calling has received the raw (pod5) ONT data, basecalling also needs to be accounted for, as depending on how much sequencing was done, this step can also be resource-intensive.

In ***Figure 8*** we present the runtime (expressed as seconds per megabase of sequencing data) and maximum memory usage for read alignment and execution of each variant caller (see Figure S16 and Table S7 for basecalling GPU runtimes). Calling variants with DeepVariant was both the slowest (median 5.7s/Mbp) and most memory-consuming (median 8GB) step. For a 4Mbp genome with 100x of read depth, this equates to a runtime of 38 minutes. However, FreeBayes has the largest variation in runtime, with the maximum (597s/Mbp) equating to 2.75 days on the previous example. For the other best-performing variant caller in this study, Clair3, the median memory usage and runtime were 1.6GB and 0.86s/Mbp (< 6 minutes on the example), respectively. See Suppl. Table S6 for full details.

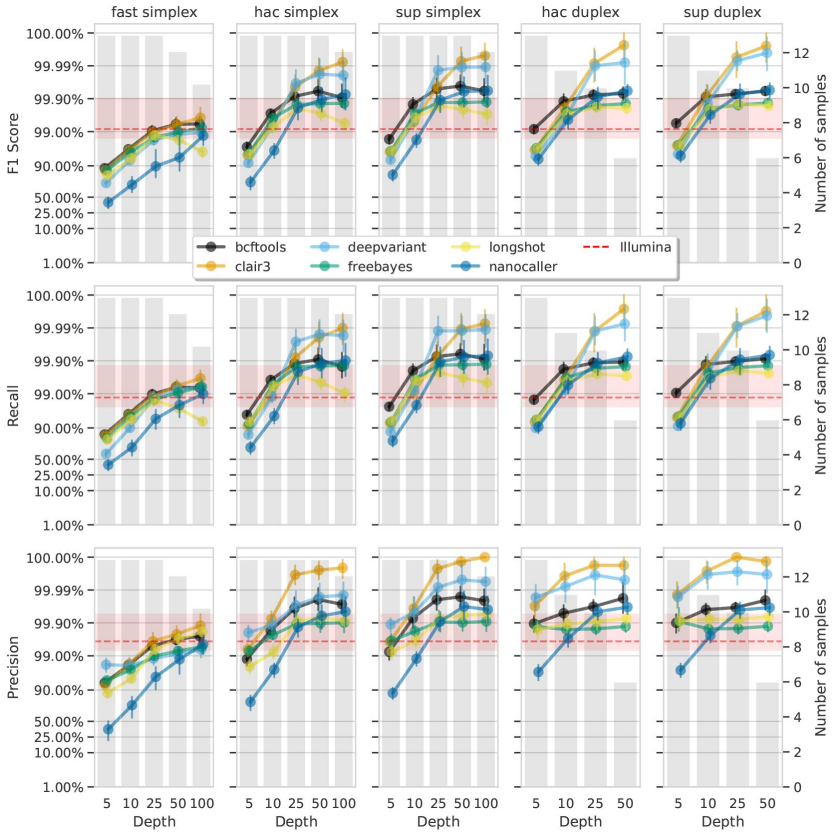

Figure 6.

Effect of read depth (x-axis) on the highest SNP F1 score, and precision and recall at that F1 score (y-axis), for each variant caller (colours). Each column is a basecall model and read type combination. The grey bars indicate the number of samples with at least that much read depth in the full read set. Samples with less than that depth were not used to calculate that depth's metrics. Bars on each point at each depth depict the 95% confidence interval. The horizontal red dashed line is the full-depth Illumina value for that metric, with the red bands indicating the 95% confidence interval.

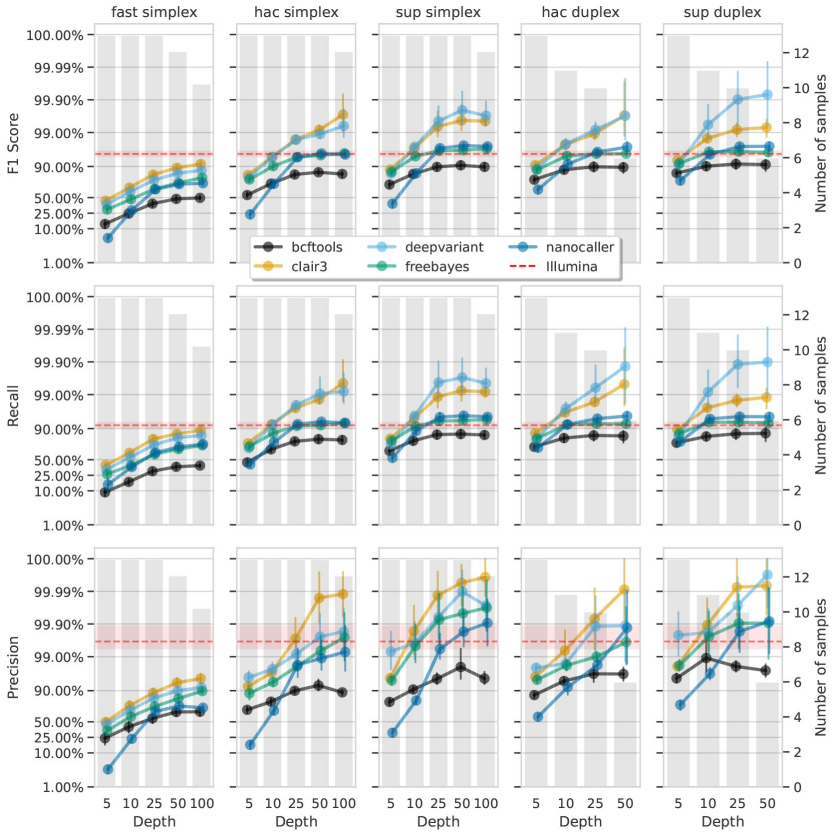

Figure 7.

Effect of read depth (x-axis) on the highest indel F1 score, and precision and recall at that F1 score (y-axis), for each variant caller (colours). Each column is a basecall model and read type combination. The grey bars indicate the number of samples with at least that much read depth in the full read set. Samples with less than that depth were not used to calculate that depth's metrics. Bars on each point at each depth depict the 95% confidence interval. The horizontal red dashed line is the full-depth Illumina value for that metric, with the red bands indicating the 95% confidence interval.

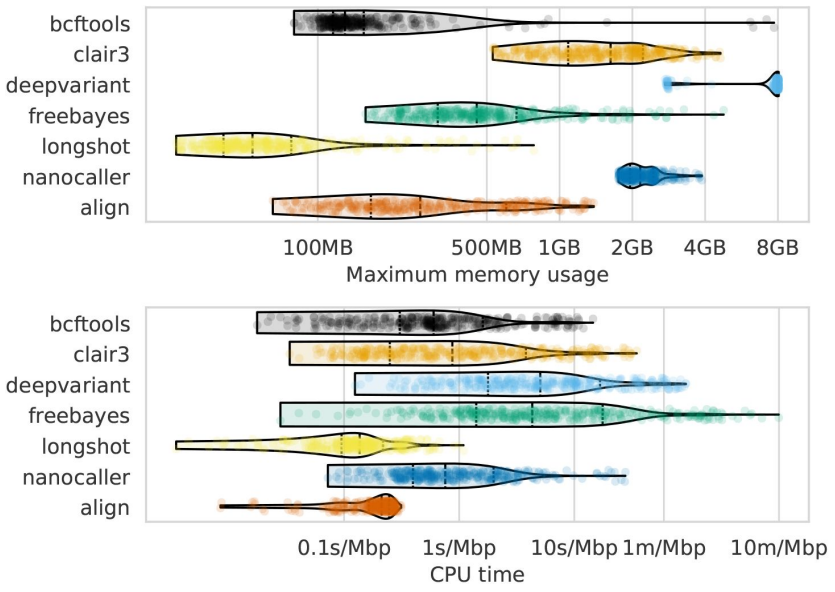

Figure 8.

Computational resource usage of alignment and each variant caller (y-axis and colours). The top panel shows the maximum memory usage (x-axis) and the lower panel shows the runtime as a function of the CPU time (seconds) divided by the number of basepairs in the readset (seconds per megabasepairs; x-axis). Each point represents a single run across read depths, basecalling models, read types, and samples for that variant caller (or alignment). s=seconds; m=minutes; MB=megabytes; GB=gigabytes; Mbp=megabasepairs.

Discussion

In this study, we evaluated the accuracy of bacterial variant calls derived from Oxford Nanopore Technologies (ONT) using both conventional and deep learning-based tools. Our findings show that deep learning approaches, specifically Clair3 and DeepVariant, deliver high accuracy in SNP and indel calls from the latest high-accuracy basecalled ONT data, outperforming Illumina-based methods, with Clair3 achieving median F1 scores of 99.99% for SNPs and 99.53% for indels.

Our dataset comprised deep sequencing of 14 bacterial species using the latest ONT R10.4.1 flowcells, with a 5 kHz sampling rate, and complementary deep Illumina sequencing. Consistent with previous studies [10–12], we observed read accuracies greater than 99.0% (Q20) and 99.9% for simplex and duplex reads, respectively (***Figure 1***).

The high-quality sequencing data enabled the creation of near-perfect reference genomes, crucial for evaluating variant calling accuracy. While not claiming perfection for these genomes, we consider them to be as accurate as current technology allows (or as philosophically possible) [12, 33].

To benchmark variant calling, we utilised a variant truthset generated by applying known differences between closely related genomes to a reference. This pseudo-real method offers a realistic evaluation framework and a reliable truthset for assessing variant calling accuracy [24, 25].

Our comparison of variant calling performances across different methods revealed that deep learning techniques provided the highest F1 scores for SNP and indel detection. This not only highlights the potential of deep learning in genomic analyses but also suggests a paradigm shift towards more sophisticated computational approaches for variant calling. The superior performance of these methods has been established for human variant calls [16, 17]; however, as these models were all trained on human data, we sought to determine if these results generalise to bacteria. The resounding answer from our work is: yes, they certainly do.

The investigation into the causes of missed and false variant calls brings to light the inherent challenges posed by sequencing technology limitations, particularly in relation to read length, alignment issues in complex genomic regions, and variable indel length in homopolymers. Our findings show that variant density and repetitive regions impede the accuracy of Illumina variant calling due to the alignment limitations inherent in short reads. We also found that with the huge improvements in ONT read accuracy, coupled with deep learning-based variant callers, homopolymerinduced false positive indel calls, traditionally a major systematic bias of ONT data [9], are no longer a concern [12].

Having established the accuracy and error sources of modern methods, we examined the impact of read depth on variant calling accuracy, revealing that even at reduced read depths of 10x, high accuracy in variant calling can be achieved, particularly with the use of super high-accuracy basecalling models and deep learning-based variant calling algorithms. This finding has significant implications for sequencing projects with limited resources, suggesting that strategic choices in technology and methodologies can still yield reliable genomic analyses. We found that with duplex super high-accuracy ONT data, even 5x depth was sufficient for SNPs on par with Illumina. Having such confidence in low-depth calls will no doubt be a boon for many clinical and public health applications where sequencing direct-from-sample is desired [2, 34–36].

Lastly, the consideration of computational resource requirements for variant calling highlights a practical aspect of genomic research, particularly for individuals and institutions with limited access to high-performance computing facilities [7, 8, 37]. Our findings reveal a wide range in the computational demands of different variant calling methods, with the worst-case require over two days to generate variant calls (FreeBayes). However, in most cases runtime is less than 40 minutes, with Clair3 having a median runtime in the order of 6 minutes. In terms of memory usage, all methods use less than 8GB, making them compatible with most laptops.

There are two main limitations to this work. The first is that we only assess small variant and ignored structural variants. Zhou *et al.* benchmarked structural variant calling from ONT data [38], though this focused on human sequencing data. Generating a truthset of structural variants between two genomes is, in itself, a difficult task. However, we believe such an undertaking with a thorough investigation of structural variant calling methods for bacterial genomes would be highly beneficial. The second limitation is not using a diverse range ANI values for selecting the variant donor genomes when generating the truthset. Previous work from Bush *et al.* examined different diversity thresholds for selecting reference genomes when calling variants from Illumina data, and found it to be one of the main differentiating factors in accuracy [22]. Our results mirror this to an extent, showing the reduction in Illumina accuracy as the variant density increases, though it would be interesting to determine whether the divergence in reference genomes has an affect on ONT variant calling accuracy. Nevertheless, to maintain our focus on the nuances of variant calling methods, including basecalling models, read types, error types, and the influence of read depth, we decided that introducing another layer of complexity into our benchmark could potentially obscure some of the insights.

In conclusion, this study presents a comprehensive evaluation of the accuracy of bacterial variant calls using Oxford Nanopore Technologies (ONT) and underscores the superior performance of deep learning-based tools, notably Clair3 and DeepVariant, in SNP and indel detection. Our extensive dataset and rigorous benchmarking against nearly perfect reference genomes have demonstrated the significant advancements in sequencing accuracy, especially with the latest ONT technologies. Our analysis revealed the diminishing impact of previously challenging factors like homopolymer-associated errors due to improvements in ONT read accuracy and the application of deep learning variant callers. Additionally, our exploration into the effects of read depth on variant calling accuracy and the computational demands of different variant calling methods has highlighted the feasibility of achieving high accuracy with lower read depths and the practicality of these methods for use in settings with limited computational resources. The capability to achieve reliable variant calling with reduced sequencing depth and computational requirements marks a significant step forward in making advanced genomic analysis more accessible and impactful.

Methods

Sequencing

Bacterial isolates were streaked onto agar plates and grown overnight at 37°C. *Mycobacterium tuberculosis*, *Streptococcus pyogenes*, and *Streptococcus dysgalatiae* subsp. *equisimilis* were grown in liquid media of 7H9 or TSB with shaking until reaching high cell density (OD ~ 1; see Suppl. Section S1 for *Streptococcus* sample selection). The cultures were centrifuged at 13000rpm for 10 minutes and cell pellets were collected. Bacteria was lysed with appropriate enzymatic treatment except for *Mycobacterium* and *Streptococcus*, which were lysed by bead beating (PowerBead, 0.5mm glass beads (13116-50) or Lysing Matrix Y (116960050-CF) and Precellys or Tissue lyser (Qiagen)). DNA extraction was performed by sodium acetate precipitation and further Ampure XP bead purification (Beckman Coulter) or either Beckman Coulter GenFind v2 (A41497) or QIAGEN Blood and Tissue DNEasy kit (69506). Illumina library preparation was performed using Illumina

DNA prep (20060059) using quarter reagents and Illumina DNA/RNA UD Indexes. Short-read whole genome sequencing was performed on an Illumina NextSeq 2000 or MiSeq with a 150bp PE kit. ONT library preparation was performed using either Rapid Barcoding Kit V14 (SQK-RBK114.96) or Native Barcoding Kit V14 (SQK-NBD114.96). Long-read whole genome sequencing was performed on a MinION Mk1b or GridION using R10.4.1 MinION flow cells (FLO-MIN114).

Basecalling and quality control

Raw ONT data were basecalled with Dorado[3] (v0.5.0) using the v4.3.0 models *fast* (dna_r10.4.1_e8.2_400bps_fast@v4.3.0), *hac* (dna_r10.4.1_e8.2_400bps_hac@v4.3.0), and *sup* (dna_r10.4.1_e8.2_400bps_sup@v4.3.0). Duplex reads were additionally generated using the duplex subcommand of Dorado with hac and sup models as fast is not compatible with duplex. Reads shorter than 1000bp or with a quality score below 10 were removed with SeqKit (v2.6.1; [39]) and each readset was randomly subsampled to a maximum mean read depth of 100x with Rasusa (v0.8.0; [32]).

Illumina reads were preprocessed with fastp (v0.23.4; [40]) to remove adapter sequences, trim low-quality bases from the ends of the reads, and remove duplicate reads and reads shorter than 30bp.

Genome assembly

Ground truth assemblies were generated for each sample as per Wick *et al.* [33]. Briefly, the unfiltered ONT simplex sup reads were filtered with Filtlong[4] (v0.2.1) to keep the best 90% (-p 90) and fastp (default settings) was used to process the raw Illumina reads. We performed 24 separate assemblies using the *Extra-thorough assembly* instructions in Trycycler's (v0.5.4; [41]) documentation[5]. Assemblies were combined into a single consensus assembly with Trycycler and Illumina reads were used to polish that assembly using Polypolish (v0.6.0; default settings; [42]) and Py-POLCA (v0.3.1; [43, 44]) with --careful. Manual curation and investigation of all polishing changes was made as per Wick *et al.* [33] (e.g., for very long homopolymers, the correct length was chosen as per Illumina reads support).

Truthset and reference generation

To generate the variant truthset for each sample, we identified all variants between the sample and a *donor* genome. To select the variant-donor genome for a given sample, we downloaded all RefSeq assemblies for that species (up to a maximum of 10000) using genome_updater[6] (v0.6.3; [45]). ANI was calculated between each downloaded genome and the sample reference using skani (v0.2.1; [46]). We only kept genomes with an ANI, a, such that $98.40\% \leq a <= 99.80\%$. In addition, we excluded any genomes with CheckM[47] completeness less than 98% and contamination greater than 5%. We then selected the genome with the ANI closest to 99.50%. Our reasoning for this range exclusion is that genomes with $a < 99.80\%$ are almost always members of the same sequence type (ST)[48, 49], and we found very little variation between them (data not shown).

We then identified variants between the reference and donor genomes using both minimap2 (v2.26; [26]) and mummer (v4.0.0rc1; [27]). We took the intersection of the variants identified by minimap2 and mummer into a single VCF and used BCFtools (v1.19; [28]) to decompose multi nucleotide polymorphisms (MNPs) into SNPs, left-align and normalise indels, remove duplicate and overlapping variants, and exclude any indel longer than 50bp. The resulting VCF file is our truthset.

Next, we generated a mutated reference genome, which we used as the reference against which variants were called by the different methods we assess. BCFtools' consensus subcommand was used to apply the truthset of variants to the sample reference, thus producing a mutated reference.

Alignment and variant calling

ONT reads were aligned to the mutated reference with minimap2 using options --cs, --MD, and -aLx map-ont and output to a BAM alignment file.

Variant calling was performed from the alignment files with BCFtools (v1.19; [28]), Clair3 (v1.0.5; [14]), DeepVariant (v1.6.0; [19]), FreeBayes (v1.3.7; [29]), Longshot (v0.4.5; [13]), and NanoCaller (v3.4.1; [15]). Individual parameters used for each variant caller can be found in the accompanying GitHub repository[7].

Where a variant caller provided an option to set the expected ploidy, haploid was given. In addition, where a minimum read depth or base quality option was available, a value of 2 and 10, respectively, was used in order to try and make downstream assessment and filtering consistent across callers.

For Clair3, the pretrained models for Dorado model v4.3.0 provided by ONT[8] were used. However, as no fast model is available, we used the hac model with the fast-basecalled reads. The pretrained model option --model_type ONT_R104 was used with DeepVariant, and the default model was used for NanoCaller.

For the Illumina variant calls that act as a benchmark to compare ONT against, we chose Snippy[9] due to being tailored for haploid genomes and being one of the best performing variant callers on Illumina data[22]. Snippy performs alignmen of read with BWA-MEM [50] and calls variants with FreeBayes.

Variant call files (VCFs) are then filtered to remove overlapping variants, make heterzygous calls homozygous for the allele with the most depth, normalise and left-align indels, break MNPs into SNPs and remove indels longer than 50bp, all with BCFtools.

Variant call assessment

Filtered VCFs were assessed with vcfdist (v2.3.3; [30]) using the truth VCFs and mutated references from Truthset and reference generation. We disabled partial credit with --credit-threshold 1.0 and set the maximum variant quality threshold (-mx) to the maximum in the VCF being assessed.

Identifying repetitive regions

To identify repetitive regions in the mutated reference, we used the following mummer utilities. nucmer --maxmatch --nosimplify to align the reference against itself and retain non-unique alignments. We then passed the output into show-coords -rTH -I 60 to obtain the coordinates for all alignments with an identity of 60% or greater. Alignments where the start and end coordinates of the alignment do not match are considered as repeats and these are output in the BED format, with intervals being merged with BEDtools [51].

Code availability

All code to perform the analyses in this work are available on GitHub[10] and archived on Zenodo [52].

Data availability

The unfiltered fastq, and assembly files generated in this study have been submitted to the NCBI BioProject database[11] under accession numbers PRJNA1087001 and PRJNA1042815. See Suppl. Table S8 for a list of all BioSample and Run accessions. Variant truthsets and associated data are archived on Zenodo [53].

Acknowledgements

This research was performed in part at Doherty Applied Microbial Genomics, Department of Microbiology and Immunology, The University of Melbourne at the Peter Doherty Institute for Infection and Immunity. This research was supported by the University of Melbourne's Research Computing Services and the Petascale Campus Initiative. We are grateful to Bart Currie and Tony Korman for providing the *Streptococcus dysgalactiae* (MMC234 202311) and *Streptococcus pyogenes* (RDH275 202311) samples. We thank Zamin Iqbal and Martin Hunt for insightful discussions relating to truthset and reference generation. We would also like to thank Romain Guérillot and Miranda Pitt for invaluable feedback throughout the project.

Author contributions

Michael B. Hall: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Software, Writing – original draft, Visualization, Writing – review & editing. Ryan R. Wick: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. Louise M. Judd: Conceptualization, Data curation, Investigation, Methodology, Project administration, Resources, Writing – original draft, Writing – review & editing. An N. T. Nguyen: Investigation, Methodology, Resources, Writing – original draft, Writing – review & editing. Eike J. Steinig: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Writing – review & editing. Ouli Xie: Methodology, Resources, Writing – review & editing. Mark R. Davies: Methodology, Resources, Supervision, Writing – review & editing. Torsten Seemann: Conceptualization, Methodology, Supervision, Writing – review & editing. Lachlan J. M. Coin: Conceptualization, Funding acquisition, Methodology, Resources, Supervision, Writing – original draft, Writing – review & editing. Timothy P. Stinear: Conceptualization, Funding acquisition, Methodology, Resources, Supervision, Writing – review & editing.

References

[1] Stimson James, et al. (2019) **'Beyond the SNP Threshold: Identifying Outbreak Clusters Using Inferred Transmissions'** *Molecular Biology and Evolution* **36**:587–603 https://doi.org/10.1093/molbev/msy242

[2] Sheka Dropen, Alabi Nikolay, Gordon Paul M K (2021) **'Oxford nanopore sequencing in clinical microbiology and infection diagnostics'** *Briefings in Bioinformatics* **22** https://doi.org/10.1093/bib/bbaa403

[3] Walker Timothy M, et al. (2022) **'The 2021 WHO catalogue of Mycobacterium tuberculosis complex mutations associated with drug resistance: a genotypic analysis'** *The Lancet Microbe* https://doi.org/10.1016/s2666-5247(21)00301-3

[4] Bertels Frederic, et al. (2014) **'Automated Reconstruction of Whole-Genome Phylogenies from Short-Sequence Reads'** *Molecular Biology and Evolution* **31**:1077–1088 https://doi.org/10.1093/molbev/msu088

[5] Gorrie Claire L, et al. (2021) **'Key parameters for genomics-based real-time detection and tracking of multidrug-resistant bacteria: a systematic analysis'** *The Lancet Microbe* https://doi.org/10.1016/s2666-5247(21)00149-x

[6] Sherry Norelle L., et al. (2023) **'An ISO-certified genomics workflow for identification and surveillance of antimicrobial resistance'** *Nature Communications* **14** https://doi.org/10.1038/s41467-022-35713-4

[7] Faria Nuno Rodrigues, et al. (2016) **'Mobile real-time surveillance of Zika virus in Brazil'** *Genome Medicine* **8** https://doi.org/10.1186/s13073-016-0356-2

[8] Hoenen Thomas, et al. (2016) **'Nanopore Sequencing as a Rapidly Deployable Ebola Outbreak Tool'** *Emerging Infectious Diseases* **22**:331–334 https://doi.org/10.3201/eid2202.151796

[9] Delahaye Clara, Nicolas Jacques (2021) **'Sequencing DNA with nanopores: Troubles and biases'** *PLOS ONE* **16** https://doi.org/10.1371/journal.pone.0257521

[10] Sanderson Nicholas D., et al. (2023) **'Comparison of R9.4.1/Kit10 and R10/Kit12 Oxford Nanopore flowcells and chemistries in bacterial genome reconstruction'** *Microbial Genomics* **9** https://doi.org/10.1099/mgen.0.000910

[11] Sanderson Nicholas D., et al. (2024) **'Evaluation of the accuracy of bacterial genome reconstruction with Oxford Nanopore R10.4.1 long-read-only sequencing'** *bioRxiv* https://doi.org/10.1101/2024.01.12.575342

[12] Sereika Mantas, et al. (2022) **'Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing'** *Nature Methods* :1–4 https://doi.org/10.1038/s41592-022-01539-7

[13] Edge Peter, Bansal Vikas (2019) **'Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing'** *Nature Communications* **10** https://doi.org/10.1038/s41467-019-12493-y

[14] Zheng Zhenxian, et al. (2022) **'Symphonizing pileup and full-alignment for deep learning-based long-read variant calling'** *Nature Computational Science* **2**:797–803 https://doi.org/10.1038/s43588-022-00387-x

[15] Ahsan Mian Umair, et al. (2021) **'NanoCaller for accurate detection of SNPs and indels in difficult-to-map regions from long-read sequencing by haplotype-aware deep neural networks'** *Genome Biology* **22** https://doi.org/10.1186/s13059-021-02472-2

[16] Olson Nathan D., et al. (2023) **'Variant calling and benchmarking in an era of complete human genome sequences'** *Nature Reviews Genetics* **24**:464–483 https://doi.org/10.1038/s41576-023-00590-0

[17] Olson Nathan D., et al. (2022) **'PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions'** *Cell Genomics* **2** https://doi.org/10.1016/j.xgen.2022.100129

[18] Pei Surui, et al. (2021) **'Benchmarking variant callers in next-generation and third-generation sequencing analysis'** *Briefings in Bioinformatics* **22** https://doi.org/10.1093/bib/bbaa148

[19] Poplin Ryan, et al. (2018) **'A universal SNP and small-indel variant caller using deep neural networks'** *Nature Biotechnology* **36**:983–987 https://doi.org/10.1038/nbt.4235

[20] Tourancheau Alan, et al. (2021) **'Discovering multiple types of DNA methylation from bacteria and microbiome using nanopore sequencing'** *Nature methods* **18**:491–498 https://doi.org/10.1038/s41592-021-01109-3

[21] Bush Stephen J (2021) **'Generalizable characteristics of false-positive bacterial variant calls'** *Microbial Genomics* **7** https://doi.org/10.1099/mgen.0.000615

[22] Bush Stephen J, et al. (2020) **'Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism–calling pipelines'** *GigaScience* **9** https://doi.org/10.1093/gigascience/giaa007

[23] Majidian Sina, et al. (2023) **'Genomic variant benchmark: if you cannot measure it, you cannot improve it'** *Genome Biology* **24** https://doi.org/10.1186/s13059-023-03061-1

[24] Li Heng (2014) **'Toward better understanding of artifacts in variant calling from high-coverage samples'** *Bioinformatics* **30**:2843–2851 https://doi.org/10.1093/bioinformatics/btu356

[25] Li Heng, et al. (2018) **'A synthetic-diploid benchmark for accurate variant-calling evaluation'** *Nature Methods* **15**:595–597 https://doi.org/10.1038/s41592-018-0054-7

[26] Li Heng (2018) **'Minimap2: pairwise alignment for nucleotide sequences'** *Bioinformatics* **34**:3094–3100 https://doi.org/10.1093/bioinformatics/bty191

[27] Marçais Guillaume, et al. (2018) **'MUMmer4: A fast and versatile genome alignment system'** *PLOS Computational Biology* **14** https://doi.org/10.1371/journal.pcbi.1005944

[28] Danecek Petr, et al. (2021) **'Twelve years of SAMtools and BCFtools'** *GigaScience* **10** https://doi.org/10.1093/gigascience/giab008

[29] Garrison Erik, Marth Gabor (2012) **Haplotype-based variant detection from short-read sequencing** *arXiv* https://doi.org/10.48550/arXiv.1207.3907

[30] Dunn Tim, Narayanasamy Satish (2023) **'vcfdist: accurately benchmarking phased small variant calls in human genomes'** *Nature Communications* **14** https://doi.org/10.1038/s41467-023-43876-x

[31] Treangen Todd J., Salzberg Steven L. (2012) **'Repetitive DNA and next-generation sequencing: computational challenges and solutions'** *Nature Reviews Genetics* **13**:36–46 https://doi.org/10.1038/nrg3117

[32] Hall Michael B. (2022) **'Rasusa: Randomly subsample sequencing reads to a specified coverage'** *Journal of Open Source Software* **7** https://doi.org/10.21105/joss.03941

[33] Wick Ryan R., Judd Louise M., Holt Kathryn E. (2023) **'Assembling the perfect bacterial genome using Oxford Nanopore and Illumina sequencing'** *PLOS Computational Biology* **19** https://doi.org/10.1371/journal.pcbi.1010905

[34] Street Teresa L., et al. (2020) **'Optimizing DNA Extraction Methods for Nanopore Sequencing of Neisseria gonorrhoeae Directly from Urine Samples'** *Journal of Clinical Microbiology* **58** https://doi.org/10.1128/jcm.01822-19

[35] Chiu Charles Y., Miller Steven A. (2019) **'Clinical metagenomics'** *Nature Reviews Genetics* **20**:341–355 https://doi.org/10.1038/s41576-019-0113-7

[36] Nilgiriwala Kayzad, et al. (2023) **'Genomic Sequencing from Sputum for Tuberculosis Disease Diagnosis, Lineage Determination, and Drug Susceptibility Prediction'** *Journal of Clinical Microbiology* **61**:e01578–22 https://doi.org/10.1128/jcm.01578-22

[37] Musila Lillian (2022) **'Genomic outbreak surveillance in resource-poor settings'** *Nature Reviews Genetics* :1–1 https://doi.org/10.1038/s41576-022-00500-w

[38] Zhou Anbo, Lin Timothy, Xing Jinchuan (2019) **'Evaluating nanopore sequencing data processing pipelines for structural variation identification'** *Genome Biology* **20** https://doi.org/10.1186/s13059-019-1858-1

[39] Shen Wei, et al. (2016) **'SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation'** *PLOS ONE* **11** https://doi.org/10.1371/journal.pone.0163962

[40] Chen Shifu, et al. (2018) **'fastp: an ultra-fast all-in-one FASTQ preprocessor'** *Bioinformatics* **34**:i884–i890 https://doi.org/10.1093/bioinformatics/bty560

[41] Wick Ryan R., et al. (2021) **'Trycycler: consensus long-read assemblies for bacterial genomes'** *Genome Biology* **22** https://doi.org/10.1186/s13059-021-02483-z

[42] Wick Ryan R., Holt Kathryn E. (2022) **'Polypolish: Short-read polishing of long-read bacterial genome assemblies'** *PLOS Computational Biology* **18** https://doi.org/10.1371/journal.pcbi.1009802

[43] Bouras George, et al. (2024) **'How low can you go? Short-read polishing of Oxford Nanopore bacterial genome assemblies'** *bioRxiv* https://doi.org/10.1101/2024.03.07.584013

[44] Zimin Aleksey V., Salzberg Steven L. (2020) **'The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies'** *PLOS Computational Biology* **16** https://doi.org/10.1371/journal.pcbi.1007981

[45] Piro Vitor C. (2023) **Vitor C. Piro. genome_updater. Version 0.6.3. 2023. DOI: 10.5281/zenodo.8108640.** https://doi.org/10.5281/zenodo.8108640

[46] Shaw Jim, Yu Yun William (2023) **'Fast and robust metagenomic sequence comparison through sparse chaining with skani'** *Nature Methods* **20**:1661–1665 https://doi.org/10.1038/s41592-023-02018-3

[47] Parks Donovan H., et al. (2015) **'CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes'** *Genome Research* **25**:1043–1055 https://doi.org/10.1101/gr.186072.114

[48] Rodriguez-R Luis M., et al. (2023) **'An ANI gap within bacterial species that advances the definitions of intra-species units'** *mBio* **15**:e02696–23 https://doi.org/10.1128/mbio.02696-23

[49] Viver Tomeu, et al. (2024) **'Towards estimating the number of strains that make up a natural bacterial population'** *Nature Communications* **15** https://doi.org/10.1038/s41467-023-44622-z

[50] Li Heng (2013) **'Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM'** *arXiv* https://doi.org/10.48550/arXiv.1303.3997

[51] Quinlan Aaron R., Hall Ira M. (2010) **'BEDTools: a flexible suite of utilities for comparing genomic features'** *Bioinformatics* **26**:841–842 https://doi.org/10.1093/bioinformatics/btq033

[52] Hall Michael B. (2024) **Michael B. Hall. mbhall88/NanoVarBench. 2024. DOI: 10.5281/zenodo.10820970.** https://doi.org/10.5281/zenodo.10820970

[53] Hall Michael B. (2024) **NanoVarBench variant truthset files** *Zenodo* https://doi.org/10.5281/zenodo.10867171

Article and author information

Michael B. Hall
Department of Microbiology and Immunology, The University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Melbourne, Australia
For correspondence: michael.hall2@unimelb.edu.au
ORCID iD: 0000-0003-3683-6208

Ryan R. Wick
Department of Microbiology and Immunology, The University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Melbourne, Australia, Centre for Pathogen Genomics, The University of Melbourne, Melbourne, Victoria, Australia
ORCID iD: 0000-0001-8349-0778

Louise M. Judd
Department of Microbiology and Immunology, The University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Melbourne, Australia, Centre for Pathogen Genomics, The University of Melbourne, Melbourne, Victoria, Australia
ORCID iD: 0000-0003-3613-4839

An N. T. Nguyen
Department of Microbiology and Immunology, The University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Melbourne, Australia
ORCID iD: 0000-0003-4893-6636

Eike J. Steinig
Department of Microbiology and Immunology, The University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Melbourne, Australia
ORCID iD: 0000-0001-5399-5353

Ouli Xie
Department of Infectious Diseases, The University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Melbourne, Australia, Monash Infectious Diseases, Monash Health, Melbourne, Australia
ORCID iD: 0000-0002-5032-1932

Mark R. Davies
Department of Microbiology and Immunology, The University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Melbourne, Australia
ORCID iD: 0000-0001-6141-5179

Torsten Seemann
Department of Microbiology and Immunology, The University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Melbourne, Australia, Centre for Pathogen Genomics, The University of Melbourne, Melbourne, Victoria, Australia
ORCID iD: 0000-0001-6046-610X

Timothy P. Stinear
Department of Microbiology and Immunology, The University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Melbourne, Australia, Centre for Pathogen Genomics, The University of Melbourne, Melbourne, Victoria, Australia
ORCID iD: 0000-0003-0150-123X

Lachlan J. M. Coin
Department of Microbiology and Immunology, The University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Melbourne, Australia
ORCID iD: 0000-0002-4300-455X

Copyright

Editors

Reviewing Editor
Detlef Weigel
Max Planck Institute for Biology Tübingen, Tübingen, Germany

Senior Editor
Detlef Weigel
Max Planck Institute for Biology Tübingen, Tübingen, Germany

Reviewer #1 (Public Review):

Summary:

The authors assess the accuracy of short variant calling (SNPs and indels) in bacterial genomes using Oxford Nanopore reads generated on R10.4 flow cells from a very similar genome (99.5% ANI), examining the impact of variant caller choice (three traditional variant callers: bcftools, freebayes, and longshot, and three deep learning based variant callers: clair3, deep variant, and nano caller), base calling model (fast, hac and sup) and read depth (using both simplex and duplex reads).

Strengths:

Given the stated goal (analysis of variant calling for reads drawn from genomes very similar to the reference), the analysis is largely complete and results are compelling. The authors make the code and data used in their analysis available for re-use using current best practices (a computational workflow and data archived in INSDC databases or Zenodo as appropriate).

Weaknesses:

While the medaka variant caller is now deprecated for diploid calling, it is still widely used for haploid variant calling and should at least be mentioned (even if the mention is only to explain its exclusion from the analysis).

Appraisal:

The experiments the authors engaged in are well structured and the results are convincing. I expect that these results will be incorporated into "best practice" bacterial variant calling workflows in the future.

https://doi.org/10.7554/eLife.98300.1.sa2

Reviewer #2 (Public Review):

Summary:

Hall et al describe the superiority of ONT sequencing and deep learning-based variant callers to deliver higher SNP and Indel accuracy compared to previous gold-standard Illumina short-read sequencing. Furthermore, they provide recommendations for read sequencing depth and computational requirements when performing variant calling.

Strengths:

The study describes compelling data showing ONT superiority when using deep learning-based variant callers, such as Clair3, compared to Illumina sequencing. This challenges the paradigm that Illumina sequencing is the gold standard for variant calling in bacterial

genomes. The authors provide evidence that homopolymeric regions, a systematic and problematic issue with ONT data, are no longer a concern in ONT sequencing.

Weaknesses:

(1) The inclusion of a larger number of reference genomes would have strengthened the study to accommodate larger variability (a limitation mentioned by the authors).

(2) In Figure 2, there are clearly one or two samples that perform worse than others in all combinations (are always below the box plots). No information about species-specific variant calls is provided by the authors but one would like to know if those are recurrently associated with one or two species. Species-specific recommendations could also help the scientific community to choose the best sequencing/variant calling approaches.

(3) The authors support that a read depth of 10x is sufficient to achieve variant calls that match or exceed Illumina sequencing. However, the standard here should be the optimal discriminatory power for clinical and public health utility (namely outbreak analysis). In such scenarios, the highest discriminatory power is always desirable and as such an F1 score, Recall and Precision that is as close to 100% as possible should be maintained (which changes the minimum read sequencing depth to at least 25x, which is the inflection point).

(4) The sequencing of the samples was not performed with the same Illumina and ONT method/equipment, which could have introduced specific equipment/preparation artefacts that were not considered in the study. See for example https://academic.oup.com/nargab/article/3/1/lqab019/6193612.

https://doi.org/10.7554/eLife.98300.1.sa1

Reviewer #3 (Public Review):

Hall et al. benchmarked different variant calling methods on Nanopore reads of bacterial samples and compared the performance of Nanopore to short reads produced with Illumina sequencing. To establish a common ground for comparison, the authors first generated a variant truth set for each sample and then projected this set to the reference sequence of the sample to obtain a mutated reference. Subsequently, Hall et al. called SNPs and small indels using commonly used deep learning and conventional variant callers and compared the precision and accuracy from reads produced with simplex and duplex Nanopore sequencing to Illumina data. The authors did not investigate large structural variation, which is a major limitation of the current manuscript. It will be very interesting to see a follow-up study covering this much more challenging type of variation.

In their comprehensive comparison of SNPs and small indels, the authors observed superior performance of deep learning over conventional variant callers when Nanopore reads were basecalled with the most accurate (but also computationally very expensive) model, even exceeding Illumina in some cases. Not surprisingly, Nanopore underperformed compared to Illumina when basecalled with the fastest (but computationally much less demanding) method with the lowest accuracy. The authors then investigated the surprisingly higher performance of Nanopore data in some cases and identified lower recall with Illumina short read data, particularly from repetitive regions and regions with high variant density, as the driver. Combining the most accurate Nanopore basecalling method with a deep learning variant caller resulted in low error rates in homopolymer regions, similar to Illumina data. This is remarkable, as homopolymer regions are (or, were) traditionally challenging for Nanopore sequencing.

Lastly, Hall et al. provided useful information on the required Nanopore read depth, which is surprisingly low, and the computational resources for variant calling with deep learning

callers. With that, the authors established a new state-of-the-art for Nanopore-only variant, calling on bacterial sequencing data. Most likely these findings will be transferred to other organisms as well or at least provide a proof-of-concept that can be built upon.

As the authors mention multiple times throughout the manuscript, Nanopore can provide sequencing data in nearly real-time and in remote regions, therefore opening up a ton of new possibilities, for example for infectious disease surveillance.

However, the high-performing variant calling method as established in this study requires the computationally very expensive sup and/or duplex Nanopore basecalling, whereas the least computationally demanding method underperforms. Here, the manuscript would greatly benefit from extending the last section on computational requirements, as the authors determine the resources for the variant calling but do not cover the entire picture. This could even be misleading for less experienced researchers who want to perform bacterial sequencing at high performance but with low resources. The authors mention it in the discussion but do not make clear enough that the described computational resources are probably largely insufficient to perform the high-accuracy basecalling required.

https://doi.org/10.7554/eLife.98300.1.sa0

Author response:

> ***eLife assessment***
>
> *This important study provides evidence for a combination of the latest generation of Oxford Nanopore Technology long reads with state-of-the art variant callers enabling bacterial variant discovery at accuracy that matches or exceeds the current "gold standard" with short reads. The evidence supporting the claims of the authors is convincing, although the inclusion of a larger number of reference genomes would further strengthen the study. The work will be of interest to anyone performing sequencing for outbreak investigations, bacterial epidemiology, or similar studies.*

We thank the editor and reviewers for the accurate summary and positive assessment. We address the comment about increasing the number of reference genomes in the response to reviewer 2.

> ***Public Reviews:***
>
> ***Reviewer #1 (Public Review):***
>
> *Summary:*
>
> *The authors assess the accuracy of short variant calling (SNPs and indels) in bacterial genomes using Oxford Nanopore reads generated on R10.4 flow cells from a very similar genome (99.5% ANI), examining the impact of variant caller choice (three traditional variant callers: bcftools, freebayes, and longshot, and three deep learning based variant callers: clair3, deep variant, and nano caller), base calling model (fast, hac and sup) and read depth (using both simplex and duplex reads).*
>
> *Strengths:*
>
> *Given the stated goal (analysis of variant calling for reads drawn from genomes very similar to the reference), the analysis is largely complete and results are compelling. The authors make the code and data used in their analysis available for re-use using current best practices (a computational workflow and data archived in INSDC databases or Zenodo as appropriate).*

> *Weaknesses:*
>
> *While the medaka variant caller is now deprecated for diploid calling, it is still widely used for haploid variant calling and should at least be mentioned (even if the mention is only to explain its exclusion from the analysis).*

We agree that this would be an informative addition to the study and will add it to the benchmarking.

> *Appraisal:*
>
> *The experiments the authors engaged in are well structured and the results are convincing. I expect that these results will be incorporated into "best practice" bacterial variant calling workflows in the future.*

Thank you for the positive appraisal.

> ***Reviewer #2 (Public Review):***
>
> *Summary:*
>
> *Hall et al describe the superiority of ONT sequencing and deep learning-based variant callers to deliver higher SNP and Indel accuracy compared to previous gold-standard Illumina short-read sequencing. Furthermore, they provide recommendations for read sequencing depth and computational requirements when performing variant calling.*
>
> *Strengths:*
>
> *The study describes compelling data showing ONT superiority when using deep learning-based variant callers, such as Clair3, compared to Illumina sequencing. This challenges the paradigm that Illumina sequencing is the gold standard for variant calling in bacterial genomes. The authors provide evidence that homopolymeric regions, a systematic and problematic issue with ONT data, are no longer a concern in ONT sequencing.*
>
> *Weaknesses:*
>
> *(1) The inclusion of a larger number of reference genomes would have strengthened the study to accommodate larger variability (a limitation mentioned by the authors).*

Our strategic selection of 14 genomes—spanning a variety of bacterial genera and species, diverse GC content, and both gram-negative and gram-positive species (including *M. tuberculosis*, which is neither)—was designed to robustly address potential variability in our results. Moreover, all our genome assemblies underwent rigorous manual inspection as the quality of the true genome sequences is the foundation this research is built upon. Given this, the fundamental conclusions regarding the accuracy of variant calls would likely remain unchanged with the addition of more genomes. However, we do acknowledge that a substantially larger sample size, which is beyond the scope of this study, would enable more fine-grained analysis of species differences in error rates.

> *(2) In Figure 2, there are clearly one or two samples that perform worse than others in all combinations (are always below the box plots). No information about species-specific variant calls is provided by the authors but one would like to know if those are recurrently associated with one or two species. Species-specific recommendations could also help the scientific community to choose the best sequencing/variant calling approaches.*

Thank you for highlighting this observation. The precision, recall, and F1 scores for each sample and condition can be found in Supplementary Table S4. We will investigate the samples that consistently perform below expectation to determine if this is associated with specific species, which may necessitate tailored recommendations for those species. Additionally, we will produce a species-segregated version of Figure 2 for a clearer interpretation and will place it in the supplementary materials.

> *(3) The authors support that a read depth of 10x is sufficient to achieve variant calls that match or exceed Illumina sequencing. However, the standard here should be the optimal discriminatory power for clinical and public health utility (namely outbreak analysis). In such scenarios, the highest discriminatory power is always desirable and as such an F1 score, Recall and Precision that is as close to 100% as possible should be maintained (which changes the minimum read sequencing depth to at least 25x, which is the inflection point).*

We agree that the highest discriminatory power is always desirable for clinical or public health applications. In which case, 25x is probably a better minimum recommendation. However, we are also aware that there are resource-limited settings where parity with Illumina is sufficient. In these cases, 10x depth from ONT would provide sufficient data.

The manuscript currently emphasises the latter scenario, but we will revise the text to clearly recommend 25x depth as a conservative aim in settings where resources are not a constraint, ensuring the highest possible discriminatory power for applications like outbreak analysis.

> *(4) The sequencing of the samples was not performed with the same Illumina and ONT method/equipment, which could have introduced specific equipment/preparation artefacts that were not considered in the study. See for example https://academic.oup.com/nargab/article/3/1/lqab019/6193612.*

To our knowledge, there is no evidence that sequencing on different ONT machines or barcoding kits leads to a difference in read characteristics or accuracy. To ensure consistency and minimise potential variability, we used the same ONT flowcells for all samples and performed basecalling on the same Nvidia A100 GPU. We will update the methods to emphasise this.

For Illumina and ONT, the exact machines used for which samples will be added as a supplementary table. We will also add a comment about possible Illumina error rate differences in the 'Limitations' section of the Discussion.

In summary, while there may be specific equipment or preparation artifacts to consider, we took steps to minimise these effects and maintain consistency across our sequencing methods.

> ***Reviewer #3 (Public Review):***
>
> *Hall et al. benchmarked different variant calling methods on Nanopore reads of bacterial samples and compared the performance of Nanopore to short reads produced with Illumina sequencing. To establish a common ground for comparison, the authors first generated a variant truth set for each sample and then projected this set to the reference sequence of the sample to obtain a mutated reference. Subsequently, Hall et al. called SNPs and small indels using commonly used deep learning and conventional variant callers and compared the precision and accuracy from reads produced with simplex and duplex Nanopore sequencing to Illumina data. The authors did not investigate large structural variation, which is a major limitation of the current manuscript. It will be very*

> *interesting to see a follow-up study covering this much more challenging type of variation.*

We fully agree that investigating structural variations (SVs) would be a very interesting and important follow-up. Identifying and generating ground truth SVs is a nontrivial task and we feel it deserves its own space and study. We hope to explore this in the future.

> *In their comprehensive comparison of SNPs and small indels, the authors observed superior performance of deep learning over conventional variant callers when Nanopore reads were basecalled with the most accurate (but also computationally very expensive) model, even exceeding Illumina in some cases. Not surprisingly, Nanopore underperformed compared to Illumina when basecalled with the fastest (but computationally much less demanding) method with the lowest accuracy. The authors then investigated the surprisingly higher performance of Nanopore data in some cases and identified lower recall with Illumina short read data, particularly from repetitive regions and regions with high variant density, as the driver. Combining the most accurate Nanopore basecalling method with a deep learning variant caller resulted in low error rates in homopolymer regions, similar to Illumina data. This is remarkable, as homopolymer regions are (or, were) traditionally challenging for Nanopore sequencing.*
>
> *Lastly, Hall et al. provided useful information on the required Nanopore read depth, which is surprisingly low, and the computational resources for variant calling with deep learning callers. With that, the authors established a new state-of-the-art for Nanopore-only variant, calling on bacterial sequencing data. Most likely these findings will be transferred to other organisms as well or at least provide a proof-of-concept that can be built upon.*
>
> *As the authors mention multiple times throughout the manuscript, Nanopore can provide sequencing data in nearly real-time and in remote regions, therefore opening up a ton of new possibilities, for example for infectious disease surveillance.*
>
> *However, the high-performing variant calling method as established in this study requires the computationally very expensive sup and/or duplex Nanopore basecalling, whereas the least computationally demanding method underperforms. Here, the manuscript would greatly benefit from extending the last section on computational requirements, as the authors determine the resources for the variant calling but do not cover the entire picture. This could even be misleading for less experienced researchers who want to perform bacterial sequencing at high performance but with low resources. The authors mention it in the discussion but do not make clear enough that the described computational resources are probably largely insufficient to perform the high-accuracy basecalling required.*

We have provided runtime benchmarks for basecalling in Supplementary Figure S16 and detailed these times in Supplementary Table S7. In addition, we state in the Results section (P10 L228-230) “Though we do note that if the person performing the variant calling has received the raw (pod5) ONT data, basecalling also needs to be accounted for, as depending on how much sequencing was done, this step can also be resource-intensive.”

Even with super-accuracy basecalling considered, our analysis shows that variant calling remains the most resource-intensive step for Clair3, DeepVariant, FreeBayes, and NanoCaller. Therefore, the statement “the described computational resources are probably largely insufficient to perform the high-accuracy basecalling required”, is incorrect. However, we will endeavour to make the basecalling component and considerations more prominent in the Results and Discussion.

https://doi.org/10.7554/eLife.98300.1.sa4