Journal Title Here, 2022, 1–8
doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

SNooPy: a statistical framework for long-read metagenomic variant calling

Roland Faure [1,2,3,*] Ulysse Faure,[4] Tam Truong,[2] Alessandro Derzelle,[3] Dominique Lavenier,[2] Jean-François Flot[3] and Christopher Quince[5,6,7]

[1]Sequence Bioinformatics, Department of Computational Biology, Institut Pasteur, Paris, France, [2]University Rennes, Inria, CNRS, IRISA - UMR 6074, Rennes, France, [3]Service Evolution Biologique et Ecologie, Université libre de Bruxelles (ULB), Brussels, Belgium, [4]Dep. of Mathematics, ETH Zürich, Switzerland, [5]Organisms and Ecosystems, Earlham Institute, Norwich, UK, [6]Gut Microbes and Health, Quadram Institute, Norwich, UK, and [7]School of Biological Sciences, University of East Anglia, Norwich, UK

*Corresponding author. roland.faure@pasteur.fr

Received YYYY-MM-DD; Revised YYYY-MM-DD; Accepted YYYY-MM-DD

Abstract

Current long-read single-nucleotide variant callers were designed primarily for genomic data—particularly human genomes. While some have been used on metagenomic data, their underlying assumptions and training procedures fail to account for the inherent complexity of metagenomic samples. To date, no long-read variant caller has been purpose-built for metagenomic applications. To address this gap, we present SNooPy, a SNP-calling tool that implements a new statistical framework tailored to long-read metagenomic data. Unlike previous genomic methods, our approach makes no assumptions about the number of haplotypes present, their evolutionary relationships, or their sequence divergence. We demonstrate that SNooPy outperforms both traditional statistical and deep learning–based SNP callers. Our results suggest that future integration of this framework with deep learning approaches could further enhance variant calling performance.

Introduction

A fundamental problem in the analysis of genomics data is the detection of variants: given a consensus genome from one individual and sequence reads from a related but not identical organism, how to compare the two. Differences can comprise single nucleotide variants (SNVs) or more complex structural variations such as rearrangements, insertions and deletions. Variant calling is the first step in numerous genomics applications, from the construction of genealogies to genome wide association studies (GWAS) [22, 25]. It is also a key step in metagenome analysis, where reads derive from multiple different organisms, typically microbes, and encompass not only inter-species diversity but also intra-species strain variation [23, 18]. Increasingly, metagenome analysis focuses on this strain-level, requiring variant identification within species [19, 16].

Not surprisingly given the importance of this problem, many programs for genomic variant calling have been developed. At present, two main families of variant callers exist. Historically, the first approach, statistical variant callers,employed probabilistic models to differentiate true genetic variants from sequencing errors [10, 11, 6, 7, 8]. These models were built upon assumptions regarding the data, typically that sequencing errors were independent or that the sequenced samples were diploid. A second paradigm emerged following the introduction of DeepVariant in 2016 [17], leveraging deep neural networks [14, 28, 1]. These replaced the statistical tests with "black-box" machine-learning, which requires training using ground-truth datasets of known variants. While these callers have achieved state-of-the-art performance for human genomes, their accuracy remains heavily dependent on the training data, sometimes underperforming when calling variants outside of the set of species on which they were trained [27].

In metagenomes, multiple strains of the same species may exist with very different abundances. The species genomes may also represent novel, previously unseen diversity, generated by binning de novo assemblies into metagenome-assembled genomes (MAGs) [19]. These specificities of metagenomes compared to single genomes can break assumptions behind some statistical models (typically the assumption that the sample is diploid or polyploid). Neural network callers will have been typically trained on known genomes (e.g. human) and are also implicitly learning the biases of their training data, which might be specific to the genome and not apply to metagenome use-case. Metagenomics hence calls for a specific approach to variant calling. Several short-read variant callers have been developed to address this challenge [2, 16, 6]. However, to the best of our knowledge, no long-read

variant callers have been specifically developed for metagenomic applications.

To fill this methodological gap, we present SNooPy, a novel statistical SNP caller tailored for long-read metagenomic datasets. Similar to existing tools such as Longshot [8] and NanoCaller [1], SNooPy exploits the statistical dependencies among reads that arise from the inherent population structure of the sequenced sample. However, unlike Longshot and NanoCaller, SNooPy's statistical framework is designed for metagenomic samples and does not make any assumption about ploidy. To do so, we implement a new statistical test inspired from previous work on haplotype assembly [9], which makes no assumptions on the number of haplotypes, their sequencing depth, or the sequencing error profiles: its only assumption is that sequencing errors occur independently across distinct reads.

Because of the lack of existing specialized metagenomic long-read SNP callers, we chose to compare SNooPy (0.3.13) with the widely used genomic SNP callers bcftools (1.22), Longshot (1.0.0), Nanocaller (3.6.2), Clair3 (1.0.10), and Deepvariant (1.9.0). These were therefore run slightly outside of their intended application area but it is common in metagenomics applications to use genomic variant callers. As detailed below, this demonstrated that SNooPy significantly outperformed not only statistical methods but also deep-learning once (when applied without retraining, as is usually the case), and, hence, provides a fast and effective means to detect variants even in noisy long-read metagenome data.

MATERIALS AND METHODS

Multi-loci variant calling

SNooPy employs a multi-locus analysis strategy, i.e., it does not attempt to detect individual variants but rather identify and validates variant groups. This approach leverages the statistical principle that sequencing and alignment errors occur (nearly) independently; therefore, the probability of observing correlated sequencing errors across multiple reads at numerous loci is vanishingly small. When correlated patterns are detected, they more likely indicate reads originating from a distinct strain carrying true variants, rather than sequencing artifacts.

This work builds upon ideas developed in the field of metagenome assembly, more specifically upon the software HairSplitter [9] and is based on the code of Strainminer [24]. The complete variant-calling algorithm includes additional validation steps and recovery mechanisms, detailed in subsection *Implementation details*. This section focuses on the core variant-calling procedure, which comprises two primary steps: (1) identification of correlated loci representing candidate groups of single nucleotide polymorphisms (SNPs), and (2) statistical validation of these candidate groups.

SNooPy processes BAM files through a sliding window approach. Within each window, the pileup data is transformed into a binary matrix M, where rows represent individual reads, columns correspond to genomic loci, and each entry M_{ij} equals 1 if read i contains the reference allele at locus j, or 0 if it contains an alternative base. To identify candidate variants, the algorithm starts by computing pairwise correlations between all columns using chi-square tests, yielding a p-value for each pair of columns. This p-value is used to perform complete-linkage hierarchical clustering with a p-value cutoff of 0.05 to group highly correlated columns. The choice of complete-linkage clustering ensures that all pairs of columns correlate well. From these groups, we identify variant patterns as subsets of loci and reads where all bases show a non-reference allele.

Building upon the statistical framework established in HairSplitter [9], we developed a statistical test which evaluates whether observed variant patterns represent authentic variants (alternative hypothesis) or result purely from coincidental sequencing errors (null hypothesis). When a pattern passes the test, the corresponding variants are outputted in a VCF file.

Let a be the number of reads and b the number of loci in the pattern, among a matrix totalling n reads and m loci. Let s denote an upper bound on the per-base sequencing error probability, estimated from the alignment data as described below.

In a matrix of size $n \times m$, there are $\binom{n}{a}\binom{m}{b}$ submatrices of size $a \times b$. In any of these submatrix, under our independence assumption, the probability that *all* the bases of the submatrix are sequencing errors is $\leq s^{ab}$. The union bound gives us a bound on the probability p of observing such a pattern *at least once* among the $\binom{n}{a}\binom{m}{b}$ submatrices under our null hypothesis: $p \leq s^{ab}\binom{n}{a}\binom{m}{b}$. We reject the null hypothesis when $p \leq 0.001$.

To enhance computational efficiency and statistical power, we only include loci where alternative alleles appear in more than 5% of reads. This filtering substantially reduces the matrix dimensions, accelerating computations while strengthening the statistical test through a reduced value of m. The error rate is estimated as the divergence between the reads and the reference. To account for error-prone regions such as homopolymers and ensure that our error rate parameter s always remains higher than the actual local error rate, we empirically set s as three time the measured error rate.

For illustrating the statistical strength of the procedure, consider a pileup of 100 Oxford Nanopore reads with error rate 0.05 (hence $s = 3 \times 0.05 = 0.15$) spanning 5,000 base pairs. Let us imagine that among these 5,000 loci, 500 exhibit an alternative allele in more than 5% of reads and that we observe 5 reads sharing alternative bases at 10 loci. The probability that this pattern arises from sequencing errors alone is $\leq 3 \times 10^{-17}$, providing strong evidence for genuine variants. A graphical illustration of the test is shown on Figure 1.

Rescuing SNPs

Our multi-locus variant calling approach has three potential limitations that could result in missed SNPs. We have developed specific rescue strategies to address each limitation.

Detection of isolated SNPs

Multi-locus variant calling relies on correlations between SNPs to achieve statistical power. However, isolated SNPs lacking correlation with other variants cannot benefit from this approach. While our statistical test remains valid for single SNPs, its power is substantially reduced.

To rescue isolated SNPs, we implement a position-specific analysis using a simple binomial model. For each genomic position, we model the expected number of sequencing errors as following a binomial distribution with parameters c (the coverage at that

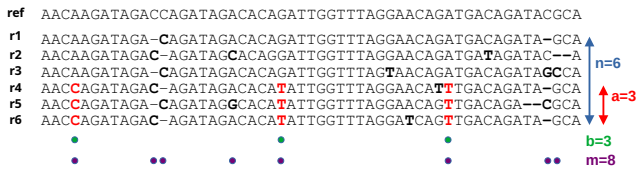

Fig. 1: Statistical foundation of the SNooPy algorithm. SNooPy starts by identifying groups of correlating columns, highlighted by green circles (b), among all columns potentially containing variants, highlighted by purple circles (m). n is the total number of reads and a the number of reads bearing the tested variants. The error rate s is over-estimated as three times the divergence of the reads to the reference. Here there are 31 non-reference bases out of 336, hence $s = 3 \times 31/336 = 0.28$. Our statistical test states that the probability of observing this pattern due to independent sequencing errors is $\leq s^{ab}\binom{n}{a}\binom{m}{b} = 3 \times 10^{-9}$.

position) and s (the maximum error rate). When an alternative allele appears at a frequency that yields a p-value smaller than 0.001 under this null model, we classify the position as an "obvious" SNP, bypassing the need for multi-locus validation.

Recovery of high-noise correlated variants

Some true SNPs may be difficult to detect due to elevated local error rates. During initial hierarchical clustering, such positions are excluded from variant groups to preserve the statistical power of our multi-locus test, as including high-noise positions can weaken the statistical test by excluding rows containing errors from the tested submatrix.

After establishing high-confidence variant calls, we perform a recovery phase where all loci are tested for correlation with confirmed SNPs using chi-square tests. Positions demonstrating significant correlation (p-value $< 10^{-6}$) with established variants are rescued and called as SNPs.

Detection of multi-allelic sites

At positions harboring multiple alternative alleles, our primary algorithm typically identifies only the most frequent variant, potentially masking additional polymorphisms. To address this limitation, we perform iterative variant calling. After the initial round, all called variants are masked (i.e., converted to reference-like status in the binary matrix), and the algorithm is re-executed on the modified data. This iterative process reveals previously hidden alternative alleles by removing the dominant variant signal, allowing detection of secondary polymorphisms at multi-allelic sites.

Implementation details

Transforming a read alignment into matrices

The algorithm begins by transforming read alignments from BAM format into binary matrices suitable for variant calling. To ensure that multi-locus analysis operates on consistent read sets, we partition the reference genome into fixed-length windows and only consider reads that span the full window. This partitioning is essential because our method requires that all loci within a group be covered by the same set of reads. We set the window length to half the median read length, which balances the need for sufficiently long windows covering multiple loci with the requirement to maintain adequate coverage across loci.

Read mapping patterns provide additional information for strain identification. Reads frequently map to only partial segments of the reference genome rather than aligning end-to-end, especially when they come from strains with structural variations regarding the reference. To exploit this signal, we cluster reads within each window based on their mapping coordinates. Groups of at least 5 reads sharing exactly the same mapping coordinates are grouped together, as they potentially originate from the same strain. Conversely, reads mapping to different coordinates within the window are analyzed separately, as they probably originate from distinct strains with different genomic architectures or insertion/deletion patterns. This coordinate-based clustering serves as a pre-filtering step that segregates reads before variant calling. By analyzing each read cluster independently, we prevent the conflation of signals from different strains and enhance the algorithm's ability to detect strain-specific variants. This approach is particularly effective when strains exhibit structural variations that cause their reads to map to different reference coordinates, even within the same genomic window.

RESULTS

Benchmark description

We compared SNooPy (0.3.13) with the widely used SNP callers bcftools (1.22), longshot (1.0.0), Nanocaller (3.6.2), Deepvariant (1.9.0) and Clair3 (1.0.10), all run with recommended or default options.

We benchmarked SNooPy on three sequencing datasets. The first one is a commercially available mock community, named Zymobiomics Gut Microbiome Standard, sequenced using ONT R10.4.1 (SRR17913199). This community has the particularity of containing five strains of *Escherichia coli* which are mostly collapsed in the metaFlye assembly, in which we expect to observe variants. The second and third ones are a human stool and a soil sample sequenced in [4], also sequenced using the latest ONT R10.4.1 flow cells. Both of these datasets were chosen because PacBio HiFi sequencing of the same samples were conducted [3] and were thus available to evaluate the quality of the calls .

All datasets were assembled using metaFlye [12]. We then called the variants using the assembly as a reference. For the Zymobiomics Gut Microbiome Standard, we report recall and precision only for the *E. coli* strains, to measure the ability of the SNP callers to call variants in a muli-strains context. The soil dataset presented computational challenges due to its very large size (6.8G), which exceeded the processing capacity of all tools within our one-week runtime constraint. To address this limitation, we randomly selected 921 contigs (with an N50 of 56kb) and conducted our analysis exclusively on this subset.

We encountered a technical issue with DeepVariant, which crashes when processing loci containing multiple alternative alleles. We have reported this bug to the DeepVariant GitHub repository. As an interim solution, we excluded the problematic contigs from our DeepVariant analysis. Since this exclusion did not significantly alter the performance metrics of other variant callers, we report statistics across the complete dataset for all tools except DeepVariant, for which we report performance on the reduced contig set.

To confirm our analyses, we created simulated datasets, following the same protocol as in a previous article [9], and benchmarked the variant callers on them. More precisely, we selected 10 *E. coli* genomes spread across the phylogenetic tree of *E. coli*. We then simulated sequencing using Badreads [26] with the error model "nanopore2023", varying the number of strains, coverage, error rate of the reads, as detailed in Supplementary Table 1.

Evaluation metrics

We assessed the recall and precision of the SNP callers. Variant call comparison is challenging because SNP callers implicitly assume that reads align on the reference end-to-end, an assumption that fails with highly divergent sequences or structural variants. We therefore excluded variants longer than 5bp from our analysis (less than 0.2% of the variants), as different SNP callers may legitimately disagree on these calls. For transparency and reproducibility, all comparison scripts used in this analysis are available in our GitHub repository (github.com/RolandFaure/SNooPy).

For the Zymobiomics dataset, we aligned the reference genomes against the assembly and used this alignment to build a set of ground truth SNPs.

For the other two datasets, we employed PacBio HiFi sequencing reads, which we mapped on the assemblies using minimap2 [13], and validated or invalidated variants based on their presence in the alignment of the set of HiFi reads on the same assemblies. This approach has inherent limitations: namely, both sequencing experiments might have only sequenced a (different) sample of the overall diversity. Nevertheless, the detection of a variant using both technologies strongly supports its validity. For these two datasets, we defined recall as the proportion of HiFi-confirmed SNPs that each software successfully identified, relative to the total pool of HiFi-confirmed SNPs called by any software. This definition intentionally excludes variants observed exclusively in HiFi data. The precision metric requires careful interpretation in this context. When SNPs called from ONT data lack confirmation in HiFi data, we classify them as false positives. However, this classification may be overly stringent: some of these variants may represent true polymorphisms that simply were not captured in the HiFi sequencing. Indeed, manual investigation using Logan-Search [5] against SRA revealed that several supposed "false positives" had been previously observed in other datasets, suggesting they are likely genuine variants. The limited throughput of Logan-Search did not allow us to conduct a systematic analysis of all these putative false positives.

The scripts to normalize, merge and compare the obtained VCFs to either the ground truth or the HiFi mapping results are available at `https://github.com/rolandfaure/snoopy`.

SNooPy excels on deeply sequenced complex communities

Figure 2 present the recall and precision metrics for all evaluated variant-calling tools (the full results are presented in the supplementary data). SNooPy consistently outperformed other tools in terms of recall across all datasets, demonstrating its effectiveness for metagenomic variant calling.

The performance evaluation identified two distinct groups of variant callers. The first group comprising Clair3, Nanocaller, and Longshot showed limited recall in metagenomic contexts, in line with their documentation that specifically targets diploid applications. The second group—including SNooPy and DeepVariant—demonstrated superior performance on metagenomic datasets, achieving over 80% recall and precision on the mock community. Bcftools occupied an intermediate position between these two groups.

Although all tools demonstrated comparable precision with minimal differences (which may not be significant given our inexact methodology), their recall rates exhibited substantial variation across datasets. DeepVariant matched SNooPy's performance on the mock community and soil sample but showed significantly lower recall on the stool sample. Further analysis of the soil data revealed a coverage-dependent performance pattern: SNooPy recalled 10% more variants than DeepVariant on contigs with >40x coverage, while DeepVariant recalled 7% more variants than SNooPy on contigs with <40x coverage. This explains the performance discrepancy of the two tools between soil and stool samples—the soil sample dataset had an average coverage of 11x, while the stool sample was sequenced at a much higher average depth of 82x.

Our simulated datasets (Figure 3) confirmed this coverage-dependent behavior:

- in experiment (i), where multiple *E. coli* strains were each sequenced at 20x, DeepVariant's recall began declining when the combined coverage exceeded 120x (more than 6 strains);
- in experiment (iv), which involved a mixture of three strains at 20x each plus strain EC590 at varying coverage, SNooPy demonstrated superior performance in identifying EC590-specific SNPs;
- by contrast, experiment (iii), which simulated a single strain at progressively lower coverage levels, showed comparable performance between DeepVariant and SNooPy.

These results collectively indicate that SNooPy excels at identifying even rare variants in high-coverage scenarios, while DeepVariant has a slight advantage in low-coverage conditions. We hypothesize that the good performance of DeepVariant stems from its deep neural network architecture, which can effectively "correct" and compensate for alignment difficulties in complex genomic regions. In contrast, SNooPy's approach remains more directly dependent on the input alignment quality, requiring a sufficient quantity of cleanly aligned reads to confidently call variants.

Another interesting result of our simulated experiments was that DeepVariant precision plummeted above 2% error rate (Figure 3 (ii)). However, as the error rates of ONT reads have now dropped below this threshold, we did not observe this effect in our real sequencing datasets.

DISCUSSION

In this study, we present a new approach for long-read metagenomic variant calling based on a simple, non-parametric test of correlation among reads. To our knowledge, this represents the first statistical variant-calling framework for long reads built on assumptions sufficiently general to hold across virtually all types of sequencing experiments, including metagenomic data. We implemented this test in SNooPy, a metagenomic variant caller that is on par with the deep-learning state-of-the-art, and performed best among the methods tested, except when coverage depth was low. This benchmark shed light on the limitations of

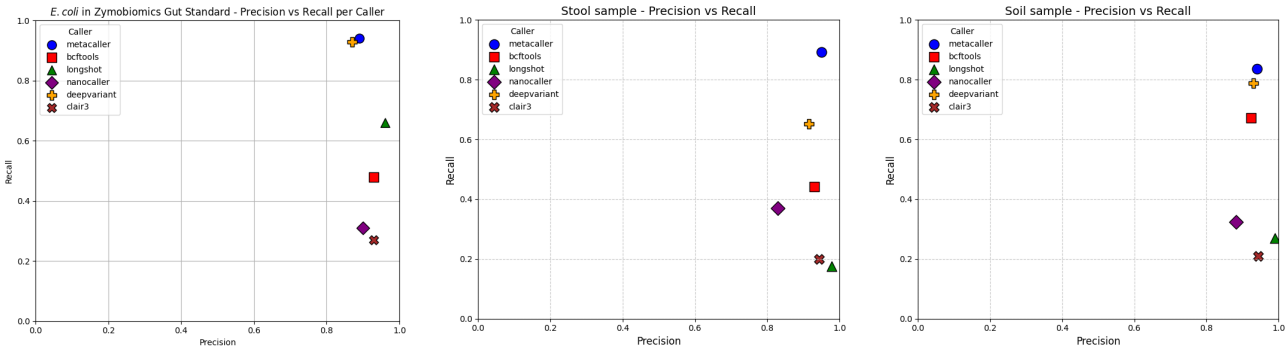

Fig. 2: Precision and recall of benchmarked metagenomic tools on three datasets. **Left:** ONT sequencing of a synthetic gut microbiome mock community. Metrics are calculated only against known *E. coli* genomes. **Middle and right:** ONT R10.4.1 sequencing of a human stool sample and a soil sample. Metrics are evaluated using a PacBio HiFi sequencing of the same sample. Recall is computed w.r.t. the union of all called variants confirmed by HiFi, as some variants may be present exlusively in the HiFi sequencing.

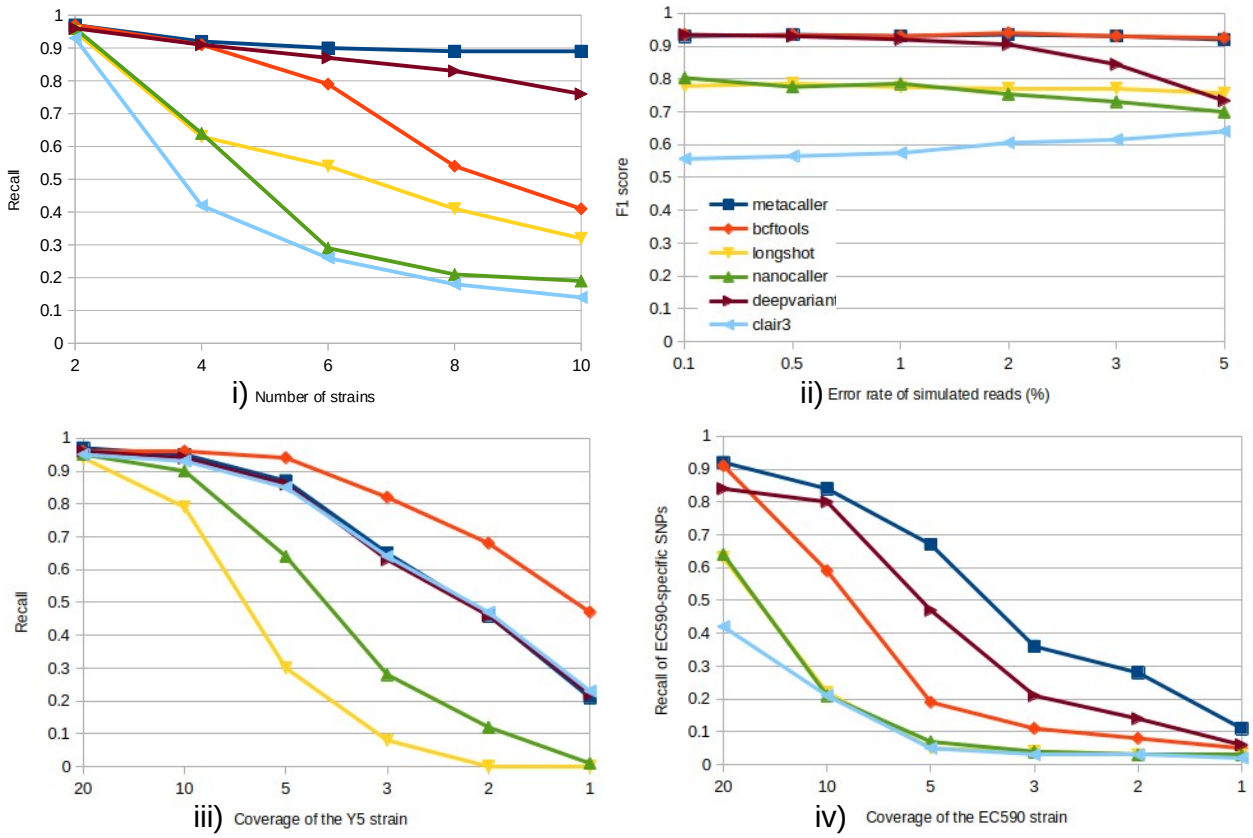

Fig. 3: F1 metric (i.e. harmonic mean of precision and recall) of variant callers evaluated on simulated read datasets with varied parameters: (i) Mixtures of two to ten strains, with reads simulated at 2% error rate and 20× coverage for all strains. Variants called against metaFlye assembly. (ii) Four-strain mixtures, with error rates varied and 20× coverage for all strains; variants called against metaFlye assembly; Note that the F1-score is reported, as DeepVariant's precision plummets above 2% error rate. (iii) Y5 strain sequenced with 2% error at different coverages; variant called against H5 genome; (iv) Four-strain mixture with 2% error, three strains at 20× coverage, and EC590 at varying coverage. Variants called against metaFlye assembly. For (iv), only recall of EC590-specific SNPs is reported.

the majority of existing long-read variant callers when applied to metagenomics data. This was unsurprising given the data they were trained on and their underlying assumptions, but represent an important, under-appreciated caveat for practitioners developing their own long-read metagenomics analysis pipelines. For instance, Longshot has been frequently used in metagenomics contexts [21, 20] but performed poorly in our tests. DeepVariant stands out as an exception, providing good precision and recall in the metagenomic context. However, the current release (v1.9.0) suffers from a bug that hampers its practical application in metagenomic analyses.

The strength of our pileup-based statistical variant calling is that it is grounded in a solid statistical framework explicitly designed for metagenomics. By contrast, the strength of deep neural networks, which take whole alignments as input rather than considering only pileups at few distinct loci, lies in its ability to draw information even out of noisy, low-coverage alignments [14]. We believe that both approaches are in essence orthogonal and could be combined to exploit the strengths of both strategies. For instance, the information of co-occurring variants used in our statistical test could be incorporated as an input feature of a deep learning variant caller. The result could be a method explicitly developed for long-read metagenomics that combines the effectiveness of SNooPy at high-coverage depths with that of DeepVariant for low-coverage genomes, achieving better performance than both overall.

Our benchmark on the soil sequencing dataset shows that even the best-performing tool, SNooPy, missed close to 20% of the SNPs. Furthermore, this is a lower estimate as our metric did not account for SNPs missed by all callers. Given the increasing importance of both long-read sequencing and strain analysis in metagenomics, and the potential for improvement that this indicates, the development of dedicated long-read metagenomic variant callers is likely to remain an active research field in the coming years.

Data availability statement

SNooPy is freely available at `https://github.com/rolandfaure/SNooPy`.

Supplementary Data Statement

Supplementary Data are available at NAR Online.

Author Contributions Statement

Roland Faure: Investigation, Conceptualization, Software, Writing. **Ulysse Faure:** Conceptualization. **Tam Truong:** Software. **Alessandro Derzelle:** Investigation. **Dominique Lavenier:** Conceptualization, Supervision. **Jean-François Flot:** Conceptualization, Supervision. **Christopher Quince:** Conceptualization, Supervision, Writing.

Funding

R.F. is supported by the Horizon Europe ERC grant number 101088572 “IndexThePlanet”. C.Q. acknowledges the support of the Biotechnology and Biological Sciences Research Council (BBSRC), part of UK Research and Innovation; Earlham Institute Strategic Programme Grant (Decoding Biodiversity) BBX011089/1 and its constituent work package BBS/E/ER/230002C; the Core Strategic Programme Grant (Genomes to Food Security) BB/CSP1720/1 and its constituent work packages BBS/E/T/000PR9818 and BBS/E/T/000PR9817; and the Core Capability Grant BB/CCG2220/1.

ACKNOWLEDGEMENTS

We acknowledge the GenOuest bioinformatics core facility (`https://www.genouest.org`) for providing the computing infrastructure. Many thanks to Rumen Andonov for his feedback, brainstorming and careful proofreading. The program Tablet [15] was used to visualize data while developing SNooPy. For the purpose of Open Access, a CC-BY public copyright licence has been applied by the authors to the present document and will be applied to all subsequent versions up to the Author Accepted Manuscript arising from this submission.

Conflict of interest statement.
None declared.

References

1. Mian Ahsan, Qian Liu, Li Fang, and Kai Wang. 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, 09 2021.
2. Sergio Andreu-Sánchez, Lianmin Chen, Wang Daoming, Hannah Augustijn, Alexandra Zhernakova, and Jingyuan Fu. A benchmark of genetic variant calling pipelines using metagenomic short-read sequencing. *Frontiers in Genetics*, 12:648229, 05 2021.
3. Gaëtan Benoit, Sébastien Raguideau, Robert James, Adam M Phillippy, Rayan Chikhi, and Christopher Quince. High-quality metagenome assembly from long accurate reads with metaMDBG. *Nature Biotechnology*, pages 1–6, 2024.
4. Gaëtan Benoit, Robert James, Sébastien Raguideau, Georgina Alabone, Tim Goodall, Rayan Chikhi, and Christopher Quince. High-quality metagenome assembly from nanopore reads with nanoMDBG. *biorXiv*, 04 2025.
5. Rayan Chikhi, Téo Lemane, Raphaël Loll-Krippleber, Mercè Montoliu-Nerin, Brice Raffestin, Antonio Pedro Camargo, Carson J. Miller, Mateus Bernabe Fiamenghi, Daniel Paiva Agustinho, Sina Majidian, Greg Autric, Maxime Hugues, Junkyoung Lee, Roland Faure, Kristen D. Curry, Jorge A. Moura de Sousa, Eduardo P. C. Rocha, David Koslicki, Paul Medvedev, Purav Gupta, Jessica Shen, Alejandro Morales-Tapia, Kate Sihuta, Peter J. Roy, Grant W. Brown, Robert C. Edgar, Anton Korobeynikov, Martin Steinegger, Caleb A. Lareau, Pierre Peterlongo, and Artem Babaian. Logan: Planetary-scale genome assembly surveys life's diversity. *bioRxiv*, 2025.
6. Paul Costea, Robin Munch, Luis Pedro Coelho, Lucas Paoli, Shinichi Sunagawa, and Peer Bork. metaSNV: A tool for metagenomic strain level analysis. *PLOS ONE*, 12:e0182392, 07 2017.
7. Mark DePristo, Eric Banks, Ryan Poplin, Kiran Garimella, Jared Maguire, Christopher Hartl, Anthony Philippakis, Guillermo del Angel, Manuel Rivas, Matt Hanna, Aaron McKenna, Tim Fennell, Andrew Kernytsky, Andrey Sivachenko, Kristian Cibulskis, Stacey Gabriel, David Altshuler, and Mark Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43:491–8, 05 2011.
8. Peter Edge and Vikas Bansal. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nat. Commun.*, 10(1):4660, October 2019.
9. Roland Faure, Dominique Lavenier, and Jean-François Flot. Hairsplitter: haplotype assembly from long, noisy reads. *Peer Community Journal*, 4, 10 2024.
10. Erik Garrison and Gabor Marth. Haplotype-based variant detection from short-read sequencing. *arXiv*, 1207, 07 2012.
11. Daniel C Koboldt, Ken Chen, Todd Wylie, David E Larson, Michael D McLellan, Elaine R Mardis, George M Weinstock, Richard K Wilson, and Li Ding. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, 25(17):2283–2285, September 2009.
12. Mikhail Kolmogorov, Derek M Bickhart, Bahar Behsaz, Alexey Gurevich, Mikhail Rayko, Sung Bong Shin, Kristen Kuhn, Jeffrey Yuan, Evgeny Polevikov, Timothy P L Smith, and Pavel A Pevzner. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods*, 17(11):1103–1110, October 2020.
13. Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.
14. Ruibang Luo, Chak-Lim Wong, Yat-Sing Wong, Chi-Ian Tang, Chi-Man Liu, Chi-Ming Leung, and Tak-Wah Lam. Exploring the limit of using a deep neural network on pileup data for germline variant calling. *Nature Machine Intelligence*, 2:1–8, 04 2020.
15. Iain Milne, Micha Bayer, Linda Cardle, Paul Shaw, Gordon Stephen, Frank Wright, and David Marshall. Tablet - next generation sequence assembly visualization. *Bioinformatics*, 26:401–2, 12 2009.
16. Matt Olm, Alexander Crits-Christoph, Keith Bouma-Gregson, Brian Firek, Michael Morowitz, and Jillian Banfield. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nature Biotechnology*, 39, 06 2021.
17. Ryan Poplin, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T Afshar, Sam S Gross, Lizzie Dorfman, Cory Y McLean, and Mark A DePristo. A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnologies*, 36(10):983–987, November 2018.
18. Christopher Quince, Sergey Nurk, Sebastien Raguideau, Robert James, Orkun S Soyer, J Kimberly Summers, Antoine Limasset, A Murat Eren, Rayan Chikhi, and Aaron E Darling. STRONG: metagenomics strain resolution on assembly graphs. *Genome Biol.*, 22(1):214, July 2021.
19. Christopher Quince, Alan W Walker, Jared T Simpson, Nicholas J Loman, and Nicola Segata. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.*, 35(9):833–844, September 2017.
20. Mantas Sereika, Aaron James Mussig, Chenjing Jiang, Kalinka Sand Knudsen, Thomas Bygh Nymann Jensen, Francesca Petriglieri, Yu Yang, Vibeke Rudkjøbing Jørgensen, Francesco Delogu, Emil Aarre Sørensen, Per Halkjær Nielsen, Caitlin Margaret Singleton, Philip Hugenholtz, and Mads Albertsen. Genome-resolved long-read sequencing expands known microbial diversity across terrestrial habitats. *Nat. Microbiol.*, 10(8):2018–2030, August 2025.
21. Jim Shaw, Jean-Sebastien Gounot, Hanrong Chen, Niranjan Nagarajan, and Yun William Yu. Floria: fast and accurate strain haplotyping in metagenomes. *Bioinformatics*, 40:i30–i38, 06 2024.
22. Leo Speidel, Marie Forest, Sinan Shi, and Simon R Myers. A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.*, 51(9):1321–1329, September 2019.
23. Duy Tin Truong, Adrian Tett, Edoardo Pasolli, Curtis Huttenhower, and Nicola Segata. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.*, 27(4):626–638, April 2017.
24. Tam Truong, Roland Faure, and Rumen Andonov. Assembling close strains in metagenome assemblies using discrete optimization. In *Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies - BIOINFORMATICS*, pages 347–356. INSTICC, SciTePress, 2024.
25. Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of GWAS discovery: Biology, function, and translation. *Am.*

J. Hum. Genet., 101(1):5–22, July 2017.
26. Ryan Wick. Badread: simulation of error-prone long reads. *Journal of Open Source Software*, 4(36):1316, 2019.
27. Taedong Yun, Cory McLean, Pi-Chuan Chang, and Andrew Carroll. Improved non-human variant calling using species-specific deepvariant models, December 2018. Accessed: 2025-10-21.
28. Zhenxian Zheng, Shumin Li, Junhao Su, Amy Leung, Tak-Wah Lam, and Ruibang Luo. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Nature Computational Science*, 2:797–803, 12 2022.

Appendix

		strains	coverage	error rate (%)	reference
Number of strains	2	Y5 H5	20x	2	Flye assembly
	4	Y5 H5 AMSCJX03 EC590	20x	2	Flye assembly
	6	Y5 H5 AMSCJX03 EC590 K12 LD27-1	20x	2	Flye assembly
	8	Y5 H5 AMSCJX03 EC590 K12 LD27-1 ME8067 RM14721	20x	2	Flye assembly
	10	Y5 H5 AMSCJX03 EC590 K12 LD27-1 ME8067 RM14721 SE15 UMN026	20x	2	Flye assembly
Error rate (%)	0.1	Y5 H5 AMSCJX03 EC590	20x	0.1	Flye assembly
	0.5	Y5 H5 AMSCJX03 EC590	20x	0.5	Flye assembly
	1	Y5 H5 AMSCJX03 EC590	20x	1	Flye assembly
	2	Y5 H5 AMSCJX03 EC590	20x	2	Flye assembly
	3	Y5 H5 AMSCJX03 EC590	20x	3	Flye assembly
	5	Y5 H5 AMSCJX03 EC590	20x	5	Flye assembly
Even coverage	20	Y5	20x	2	H5
	10	Y5	10x	2	H5
	5	Y5	5x	2	H5
	3	Y5	3x	2	H5
	2	Y5	2x	2	H5
	1	Y5	1x	2	H5
Uneven coverage	20	Y5 H5 AMSCJX03 EC590	20x, 20x, 20x, 20x	2	Flye assembly
	10	Y5 H5 AMSCJX03 EC590	20x, 20x, 20x, 10x	2	Flye assembly
	5	Y5 H5 AMSCJX03 EC590	20x, 20x, 20x, 5x	2	Flye assembly
	3	Y5 H5 AMSCJX03 EC590	20x, 20x, 20x, 3x	2	Flye assembly
	2	Y5 H5 AMSCJX03 EC590	20x, 20x, 20x, 2x	2	Flye assembly
	1	Y5 H5 AMSCJX03 EC590	20x, 20x, 20x, 1x	2	Flye assembly

Table 1. Description of the experiments run with the simulated datasets on *E. coli*