baseLess

lightweight detection of sequences in raw MinION data

Noordijk, Ben; Nijland, Reindert; Carrion, Victor J.; Raaijmakers, Jos M.; De Ridder, Dick; De Lannoy, Carlos

DOI
10.1093/bioadv/vbad017

Publication date
2023

Document Version
Final published version

Published in
Bioinformatics Advances

Citation (APA)
Noordijk, B., Nijland, R., Carrion, V. J., Raaijmakers, J. M., De Ridder, D., & De Lannoy, C. (2023). baseLess: lightweight detection of sequences in raw MinION data. *Bioinformatics Advances*, *3*(1), Article vbad017. https://doi.org/10.1093/bioadv/vbad017

Bioinformatics Advances, 2023, vbad017
https://doi.org/10.1093/bioadv/vbad017
Advance Access Publication Date: 15 February 2023
Original Paper

Sequence analysis

baseLess: lightweight detection of sequences in raw MinION data

Ben Noordijk [1], Reindert Nijland[2], Victor J. Carrion[3,4,5], Jos M. Raaijmakers[3,4], Dick de Ridder [1] and Carlos de Lannoy [1,6,*]

[1]Bioinformatics Group, Wageningen University, Wageningen 6700AH, The Netherlands, [2]Marine Animal Ecology, Wageningen University, Wageningen 6700AP, The Netherlands, [3]Institute of Biology, Leiden University, Leiden 2300RA, The Netherlands, [4]Department of Microbial Ecology, Netherlands Institute of Ecology, Wageningen 6700AB, The Netherlands, [5]Departamento de Microbiología, Instituto de Hortofruticultura Subtropical y Mediterránea 'La Mayora', Universidad de Málaga-Consejo Superior de Investigaciones Científicas (IHSM-UMA-CSIC), Málaga 29010, Spain and [6]Department of Bionanoscience, Delft University of Technology, Delft 2600GA, The Netherlands

*To whom correspondence should be addressed.

Associate Editor: Alex Bateman

Received on January 4, 2023; revised on January 27, 2023; editorial decision on February 4, 2023; accepted on February 12, 2023

Abstract

Summary: With its candybar form factor and low initial investment cost, the MinION brought affordable portable nucleic acid analysis within reach. However, translating the electrical signal it outputs into a sequence of bases still requires mid-tier computer hardware, which remains a caveat when aiming for deployment of many devices at once or usage in remote areas. For applications focusing on detection of a target sequence, such as infectious disease monitoring or species identification, the computational cost of analysis may be reduced by directly detecting the target sequence in the electrical signal instead. Here, we present baseLess, a computational tool that enables such target-detection-only analysis. BaseLess makes use of an array of small neural networks, each of which efficiently detects a fixed-size subsequence of the target sequence directly from the electrical signal. We show that baseLess can accurately determine the identity of reads between three closely related fish species and can classify sequences in mixtures of 20 bacterial species, on an inexpensive single-board computer.

Availability and implementation: baseLess and all code used in data preparation and validation are available on Github at https://github.com/cvdelannoy/baseLess, under an MIT license. Used validation data and scripts can be found at https://doi.org/10.4121/20261392, under an MIT license.

Contact: c.v.delannoy@tudelft.nl

Supplementary information: Supplementary data are available at *Bioinformatics Advances* online.

1 Introduction

DNA sequencing is no longer the costly endeavor it once was; while two decades ago analysis of a single genome could occupy multiple labs over several years (Green *et al.*, 2015), technological innovations have now driven the per-base cost down sufficiently to allow routine sequencing for other purposes than scientific discovery, including forensics (Bruijns *et al.*, 2018) and clinical diagnoses (Cristiano *et al.*, 2019; Newman *et al.*, 2014; Normand *et al.*, 2018). The case for such usage was strengthened further with the introduction of Oxford Nanopore Technology (ONT)'s MinION, a low-cost, small-size sequencing device. No longer inhibited by high initial investment costs or poor portability, small laboratories and individual users may now opt for in-house sequencing and on-site analysis in remote locations (Faria *et al.*, 2016; Goordial *et al.*, 2017; Pomerantz *et al.*, 2018).

This development was possible due to the introduction of a new sequencing mechanism; rather than the fluorescence-based sequencing-by-synthesis approach employed by previous devices, the MinION sequences DNA strands of arbitrary length by ratcheting them through a nanopore while reading out the electric current (de Lannoy *et al.*, 2017). This readout is colloquially referred to as a 'squiggle'. As the nucleotide combination residing in the nanopore at a given moment influences the electrical resistance, the squiggle carries information on the sequence. In a process termed 'basecalling', the nucleotide sequence is deduced from the squiggle.

Although the MinION itself is an inexpensive sequencer, real-time data analysis currently still requires at least a mid-tier laptop,

outfitted with a GPU with sufficient memory (upward of 4 GB). For some applications, for example, the distribution of thousands of devices for infectious disease screening, this may bring along prohibitively high additional costs. It would therefore be beneficial if inexpensive computing hardware could be used instead. Depending on the intended purpose, a computationally lighter analysis pipeline may be a solution. As fast computing hardware is mainly required for basecalling, some basecallers have been developed that trade off lower resource requirements against a decreased basecalling accuracy. DeepNano-blitz (Boža *et al.*, 2020) is the most recent open-source example of such an implementation, while ONT's proprietary basecaller guppy has a 'fast' running mode for this purpose.

Not all applications require information on the full read sequence however. If only detection of a set of known sequences is required, these sequences could be detected directly in the squiggle instead, potentially reducing the computational load even further. Several direct-from-squiggle sequence detection methods have been proposed. Kovaka *et al.* (2021) developed UNCALLED, which assigns a probability for each 5-mer potentially matching to each squiggle segment and then compares probable series of 5-mers to a pre-indexed genome to quickly map the read to its likely location. Its original purpose is to facilitate 'adaptive sampling', that is, to rapidly detect the likely origin of a read while the strand is still being sequenced, so that sequencing of strands from non-target sources may be terminated early (Loose *et al.*, 2016). UNCALLED can easily be repurposed to perform general sequence detection; however, the index-based approach carries several disadvantages. Efficiency decreases for larger and more repetitive genomes and re-indexing is required to attune the tool to a new target sequence. Moreover, accuracy was found to be low for short sequences (Bao *et al.*, 2021). Similarly to UNCALLED, SquiggleNet was designed for adaptive sampling (Bao *et al.*, 2021). Following a more straightforward approach, it uses a neural network trained for the recognition of a given genome to decide whether squiggles belong to a species or not. Previously, SquiggleNet was found to outperform UNCALLED in terms of both accuracy and processing speed, but the required retraining of SquiggleNet for a given species is a highly resource- and time-consuming process.

Here, we introduce baseLess, a computationally efficient and flexible approach for direct sequence detection (Fig. 1). Using an array of small neural networks, each pre-trained to recognize a single *k*-mer, baseLess can determine whether a read can be mapped to a given sequence or not. Configuring our tool to detect a sequence requires only the selection of target *k*-mers and their associated pre-trained neural networks. We show that baseLess can perform species detection on eukaryotic whole-genome sequencing data against a background of similar species, as well as 16S-based species detection of prokaryotes agnostic of background sequences. BaseLess is more accurate than direct sequence detection pipelines, but currently outperformed in speed and accuracy by basecalling-and-mapping. Nevertheless, the baseLess pipeline uses a smaller analysis model than basecallers and may run more efficiently after software optimizations, thus making an initial step toward species detection on more affordable (~$100) computational analysis hardware. Further development could remove an important economical bottleneck for highly distributed and remote field analysis using the MinION.

2 Results

2.1 Tool structure

BaseLess deduces the presence of a target sequence by detecting squiggle segments corresponding to salient short sequences, *k*-mers, using an array of convolutional neural networks (CNNs) (Fig. 1A). Each CNN detects a single *k*-mer, a relatively simple task, thus the network complexity can be kept low. This divide-and-conquer strategy has several advantages. All CNNs can process a read in parallel, which makes baseLess computationally efficient. Furthermore, given a library of pre-trained CNNs, baseLess can easily be reconfigured to detect a different target sequence by combining a different set of CNNs. Finally, sufficient data to train the CNNs are usually available; shorter sequences generally occur more often than longer sequences, thus a read set of any source, once corrected for basecalling errors (see Section 4), provides sufficient data to train for a wide range of *k*-mers.

To complete the baseLess network, the outputs of the CNN array are combined using one of two aggregation rules. If configured in 'abundance mode', baseLess returns the number of occurrences found for each *k*-mer, which may then be compared with abundance estimates derived from a target genome (Fig. 1B). In 'read detection mode', the network is configured to decide whether a sufficiently large fraction of its *k*-mers has been found in a given read to conclude that it contained the target sequence (Fig. 1C). These modes are explained and evaluated in more detail below.

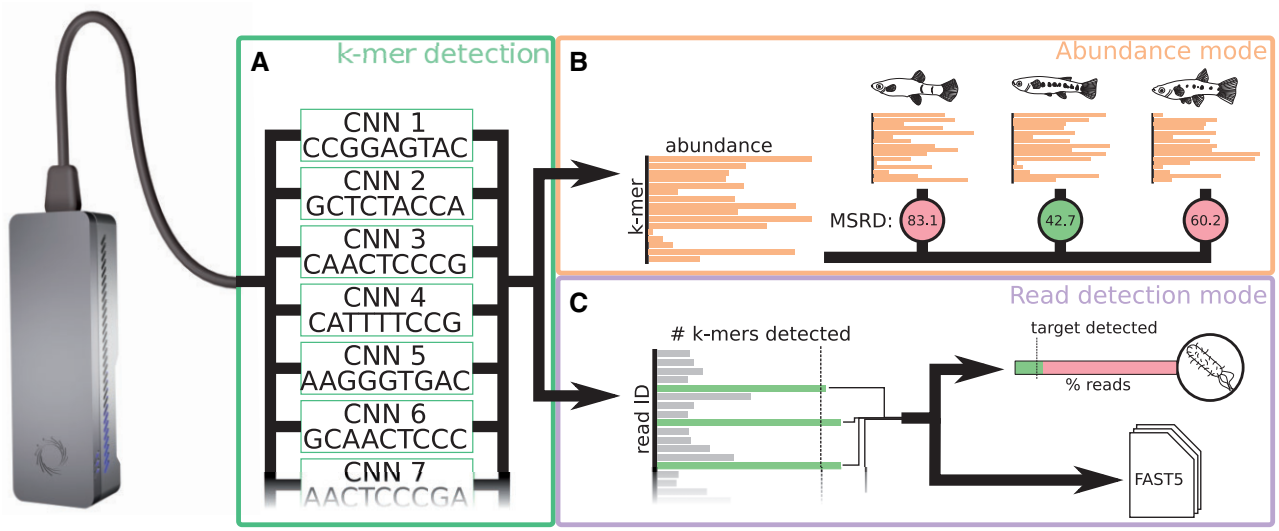

Fig. 1. Schematic overview of the baseLess sequence detection tool. (**A**) baseLess detects sequences using an array of pre-trained interchangeable neural networks, each of which detects a specific *k*-mer. These *k*-mers have been specifically selected to allow discrimination of a target sequence. (**B**) In abundance mode, network outputs are summed over all reads and presented as an estimate of *k*-mer abundance. This estimate is compared against genome-based estimates for several closely related species by calculating the MSRD. The species for which the MSRD is lowest is the most likely source of the reads. (**C**) In read-based detection mode, a target sequence is sought in each individual read. A minimum fraction of *k*-mers needs to be detected in a read before it is classified as a target read. A target species is detected if a minimum fraction of analyzed reads can be ascribed to it. Reads ascribed to the target species are also stored in a FAST5 file for further analysis

2.2 Abundance-based species detection

As baseLess provides fast and accurate inference on low-cost hardware, it is highly suited to determine the species or strain to which a given individual belongs at remote sampling locations, or at many locations simultaneously. One practical application of such usage may be found in ecological monitoring of visually similar species. For such tasks, baseLess should be configured in abundance mode, which requires the target species' genome and a set of background genomes—genomes of species from which the target species must be discerned. The *k*-mer set used for discrimination is then found by combining *k*-mers that are highly abundant in the target genome yet found less than average in the background genomes, or vice versa. To determine the origin of a sample, baseLess ranks *k*-mers by abundance as measured in the reads and compares it to their abundance ranking in the target and background genomes, using the mean-squared rank difference (MSRD):

$$\mathrm{MSRD} = \frac{1}{N}\sum_{n=1}^{N}(m_{b,n} - m_{r,n})^2.$$

Here, $m_{b,n}$ and $m_{r,n}$ are the rank for *k*-mer m_n based on abundances in analyzed reads and in a reference, respectively. N is the total number of *k*-mers analyzed in the reads.

To test baseLess' performance in this scenario, we analyzed unamplified whole-genome MinION reads from three related guppy species: *Phalloptychus januarius*, *Poeciliopsis gracilis* and *Poeciliopsis turneri* (van Kruistum *et al.*, 2021). In three separate analyses, we configured our tool for detection of one of the species against the other two, using Illumina short-read assemblies of the same individuals as target and background genomes to avoid the risk of detecting species based on MinION-specific sequencing errors. We then analyzed a set of 2000 MinION reads originating from the target species. We found that baseLess consistently calls the correct species for each analyzed readset (Fig. 2A–C). Moreover, baseLess did not need the full 2000 reads for any classification; stable MSRD values were attained after 52, 352 and 84 reads for *P.gracilis*, *P.januarius* and *P.turneri*, respectively. It should be noted that MSRD scores cannot directly be compared between the three different models used here. This is because each includes different *k*-mer detecting submodels, which are marked by different error models and abundances in individuals. Basecalling followed by mapping gives accurate results within five reads (Supplementary Fig. S1), which still makes it a viable alternative compared with baseLess at the moment.

To verify whether baseLess indeed detects differences between species and not between individuals, we also ran classification on samples of a family of four *P.gracilis* individuals, using a *k*-mer set selected using the genome of an unrelated *P.gracilis* individual (Fig. 2D). BaseLess consistently called the correct species while requiring less than a hundred reads. Notably, MSRD values for the correct class were consistently lower for this family than for the individual of which an Illumina assembly was used to compose the model (Fig. 2C), which may be a result of the higher read quality obtained for the family. This also indicates that our model was not overfitting to the *k*-mer profile of a single individual.

Interestingly, the *k*-mer rankings also followed the phylogenetic relation between the species; in all detection experiments, MSRD values for *P.gracilis* and *P.turneri* were consistently closer to each other than to *P.januarius*, which is indeed of a different genus. This implies that, even if the genome of the correct species is not included, the relative identity of a sample may be inferred by comparing measured abundances to several related species.

2.3 Read-based species detection

In specific applications, a sample may contain a mixture of DNA of many species, from which a species of interest must be detected. Possible scenarios include the screening for infectious disease agents at events or at national borders, or detection of indicator species for environmental health. In 16S cDNA samples, baseLess may be configured to detect such a species of interest by selecting a combination of *k*-mers unique to the target's 16S sequence, and running it in read detection mode. In this configuration, baseLess detects each *k*-mer on a per-read basis, rather than summing occurrences over all reads as is done in abundance mode. If a minimum fraction of target *k*-mers is found in a read, it is attributed to the target species. The raw squiggle of found target sequences is stored to allow more in-depth analysis at a later stage, while non-target reads can be discarded to decrease data storage footprint. To allow reliable detection of a wide range of species against an arbitrary genomic background, we composed a list of *k*-mers which both varied in sequence composition and produced easily differentiable squiggle segments. This list was further filtered to only contain *k*-mers that are present in NCBI 16S sequences, yet sufficiently rare to allow for species discrimination (see Section 4). We find that, on average, this subset of *k*-mers suffices to uniquely identify the majority of species in samples containing up to 10 000 different constituents (Supplementary Fig. S2). This indicates that, in the context of a given microbiome analysis, this *k*-mer set should often provide sufficient resolution, as the pool of present species and/or the number of species of interest among them is typically smaller [e.g. the human gut microbiome contains ∼1000 species (Yang *et al.*, 2020)]. Additionally, the discriminative power of baseLess can be further improved by training additional *k*-mer models, which takes a few hours at most (see Section 4).

To test this approach, we amplified and sequenced the 16S rRNA regions of an artificial microbial community of 21 known species on the MinION (Supplementary Table S1). In total, 400 000 reads were fully basecalled and mapped to the 21 genomes of the species to determine their likely origin. No reads were mapped to the *Porphyromonas gingivalis* genome, thus this species was left out of subsequent analysis. Read numbers for other species varied between 11 and 51 040.

We reconfigured baseLess and ran inference for each of the species in a 5-fold cross-validation scheme, to determine how well it could identify the origin of reads. Running speeds were benchmarked on two different classes of hardware; the Nvidia Jetson Nano (2 GB), a ∼$100 single-board computer with dedicated GPU

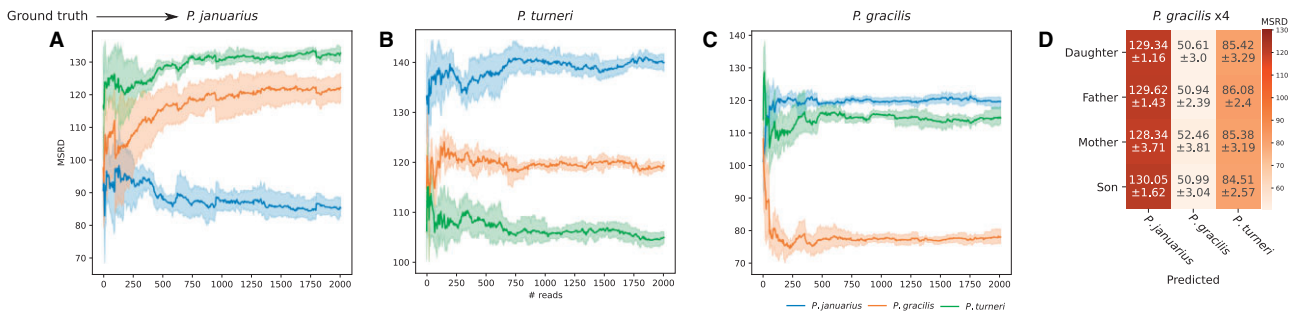

Fig. 2. MSRDs based on a comparison of *k*-mer abundances estimated from reads by baseLess and the abundances in genomes of three closely related fish species. A low MSRD indicates that *k*-mer abundances in sample and genome are alike and that reads are thus more likely derived from that genome. Results are presented for (**A**) *P.januarius*, (**B**) *P.turneri*, (**C**) *P.gracilis* and (**D**) a family of four *P.gracilis* individuals of which no assembled genome was used in the configuration of baseLess. In A, B and C, colored areas denote the 95% confidence interval. In D, numbers are formatted as mean ± standard deviation over 2000 reads

(Nvidia Maxwell, 128 cores@921MHz) and a high-end desktop computer with dedicated GPU (Nvidia GeForce RTX 3070, 5888 cores@173GHz). To allow straightforward comparison, all tools were given access to 3 CPU cores and the GPU if required. We compared the performance of baseLess on 16S read classification with that of four other pipelines: full basecalling by either DeepNano-blitz (Boža *et al.*, 2020) or guppy in 'fast' mode, followed by mapping using minimap2 (Li, 2018) ('DeepNano+minimap2' and 'guppy+minimap2', respectively); UNCALLED (Kovaka *et al.*, 2021) and SquiggleNet (Bao *et al.*, 2021).

BaseLess consistently identified its target reads with more than 95% accuracy and an F_1 score of 0.54 on average (Fig. 3A), with the exception of *Helicobacter pylori*. Compared with DeepNano+minimap2 and UNCALLED, baseLess yielded a higher accuracy and a higher F_1 score for all species. Nevertheless, Guppy+minimap2 consistently outperformed baseLess and all other pipelines. Under default settings, SquiggleNet only had sufficient data to classify reads of the three species for which the most reads were available: *Escherichia coli*, *H.pylori* and *Listeria monocytogenes*. On these species, baseLess performed similar to, or better than SquiggleNet.

In the speed benchmark on the Jetson Nano, baseLess processed 5.0 kb per second (kbps) which is more than twice as fast as Guppy+minimap2 (2.0 kbps). Despite its lower speed, Guppy is still a viable alternative to baseLess as its higher per-read accuracy compensates for its lower speed (Supplementary Fig. S3). SquiggleNet ran the fastest at 17 kbps. None of the tools were able to match the theoretical maximum throughput of the MinION (230 kbps). We were unable to install DeepNano-blitz and UNCALLED on this hardware, possibly due to incompatibility with the energy-efficient AARCH64 CPU architecture used in the Jetson Nano and most other single-board computers. As expected, processing speeds were much higher on high-end desktop hardware with the three GPU-accelerated tools—baseLess, guppy+minimap2 and SquiggleNet—performing best. At 1.3 megabase per second (Mbps), guppy+minimap2 was faster than all other tools. SquiggleNet (420 kbps) was again faster than baseLess (210 kbps), which in turn out-competed DeepNano+minimap2 (120 kbps) and UNCALLED (84 kbps).

3 Discussion

In this work, we proposed a method to identify whole genomes or amplified sequences in nanopore reads by detecting salient *k*-mers using an array of individual, interchangeable neural networks. We show that baseLess, our implementation of this method, is capable of correctly classifying single-species whole-genome sequencing samples, given the target species' genome and a set of off-target genomes. This is useful for species determination of larger organisms, though not for environmental samples of microbes, which contain many species of which most may be unknown. We therefore also implemented an alternative running mode, which allows microbial species detection against an unknown background, suitable for smaller genomes or PCR-amplified samples.

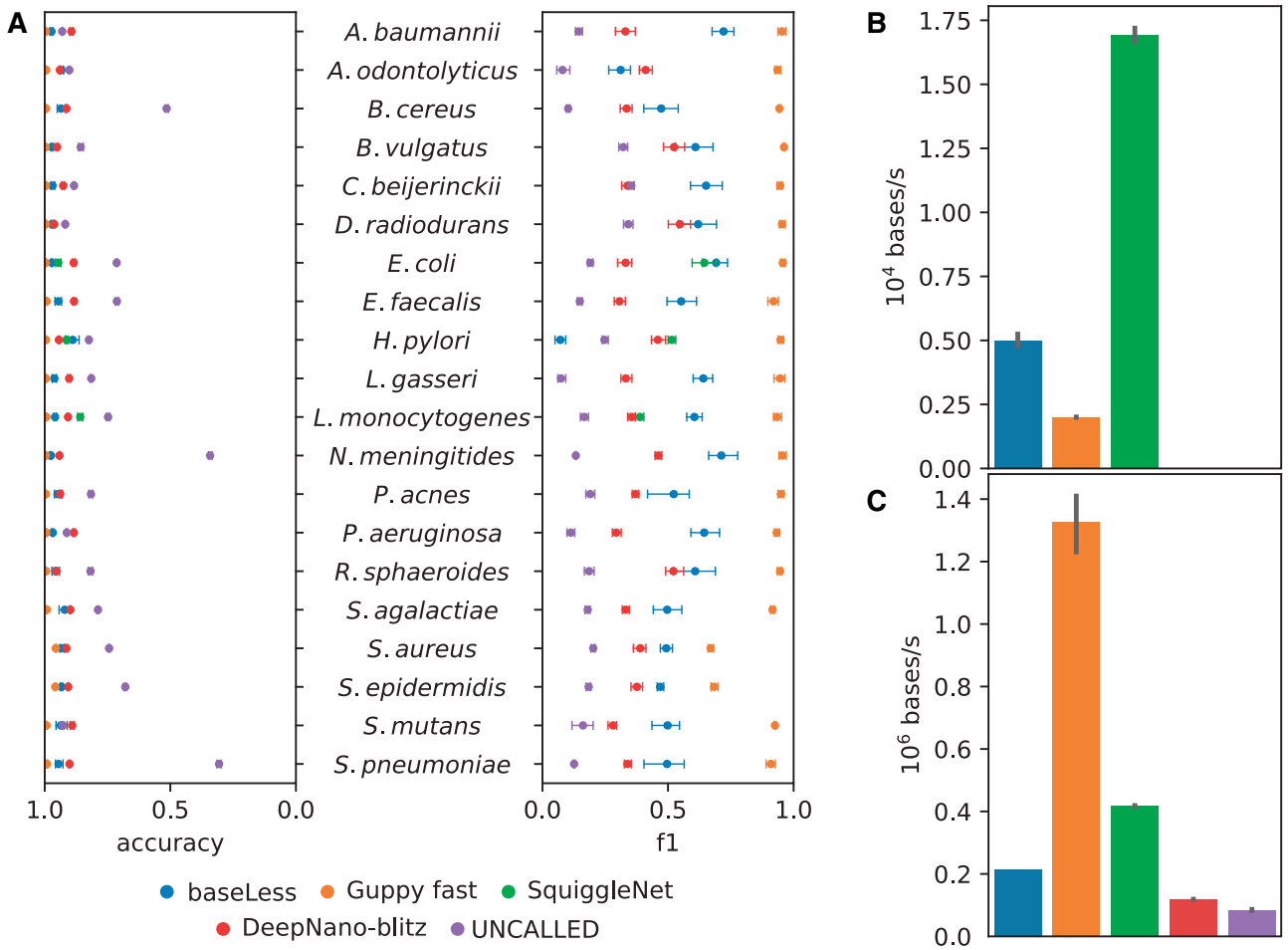

Fig. 3. Performance in 16S-based species detection for a community of 20 equally abundant species, of baseLess and 4 competing analysis tools; DeepNano-blitz; Guppy (fast mode); UNCALLED and SquiggleNet. As DeepNano-blitz and Guppy are basecallers, not read mapping tools, minimap2 is used to obtain the mapping. (**A**) Accuracy and F_1 score per species for all four tools. Ground truth is determined through Guppy high-accuracy basecalling followed by mapping using BLASTN. Black bars denote standard deviation over five cross-validation folds. (**B**) Analysis speed for each tool on the Nvidia Jetson Nano (2 GB) single-board computer and (**C**) on a high-end desktop computer. Black bars denote standard deviation over 10 analysis runs on 1000 reads

The world-wide demand for microbe screening, most prominently for infectious disease agents, is currently filled mostly by lateral-flow antibody and qPCR tests. Antibody tests have a turnaround time of minutes, require little training to use and can be mass-produced at low expense, but require a redesign and subsequent re-distribution for the detection of different targets. Moreover, detection is not as reliable as that of nucleic acid analysis (Mistry *et al.*, 2021). qPCR is generally more reliable, but also requires newly designed primers for different targets. We propose that baseLess could allow for rapid detection of novel targets without the requirement to synthesize and distribute novel primers, because reconfiguring baseLess for the detection of a new agent or variant only requires loading the networks for a different set of *k*-mers. Additionally, as found target reads are stored, these may be analyzed in depth afterward, giving researchers an unprecedented wealth of information on mutations from each detected occurrence of the agent. Especially relevant in this context, though left unexplored here, would be microbe detection using direct DNA or RNA sequencing, as omission of PCR steps would bring down turnaround time even further.

We compared baseLess with two fast full basecalling-and-mapping pipelines and two adaptive sequencing tools. Of these competitors, only the guppy+minimap2 pipeline could consistently classify reads with a higher accuracy than baseLess. SquiggleNet performed similarly to baseLess in terms of accuracy and was more than three times faster on the Jetson Nano. However, due to its high training data requirements, it could only be evaluated on 3 of the 20 species tested here. BaseLess did not suffer from this disadvantage as it only needs examples of *k*-mers to train, which may be obtained from any source. Furthermore, SquiggleNet requires retraining to detect new species, while baseLess only needs reconfiguration for a different set of *k*-mers. We thus argue that baseLess has more potential to be developed into an accurate yet flexible sequence detection tool than its competitors, especially after it receives further optimizations.

Several venues may be explored to further optimize our workflow. Importantly, baseLess' computational efficiency can be further increased; we ran our tool using Tensorflow, a fully equipped deep learning library, however to run inference on low-powered hardware more efficiently, light-weight frameworks such as Tensorflow-lite and TensorRT may be employed. We expect that further optimization would allow baseLess to analyze reads faster than Guppy+minimap2, as baseless is conceptually more lightweight; its model size (∼1 MB for a single species-detection model) is only a fraction of that of Guppy+minimap2 (∼300 MB). Furthermore, we note that the amplification of 16S sequences used in our 16S performance evaluation remains a bottleneck in sequence detection. Instead, the MinION may also be used to directly sequence RNA. As ribosomal RNA makes up a large part of the total RNA content of prokaryotes (Pust *et al.*, 2021), it would be interesting to evaluate classification based on unamplified RNA content instead.

In summary, the results obtained inspire confidence that using computationally light direct analysis of squiggles, the MinION can be turned into a mobile species detector for under $1000, thus paving the way for nucleic acid-based detection of biological agents.

4 Materials and methods

4.1 Network design procedure

Individual *k*-mers are recognized using 1D CNNs implemented in Tensorflow 2.3 (Abadi *et al.*, 2015). We optimized hyperparameters through 100 rounds of training and evaluation on 33 549 and 3241 held-out training and test reads, respectively, to obtain the final network architecture (Supplementary Fig. S4). After each round of training and evaluation, the next hyperparameter set was selected using a tree-structured Parzen estimator implemented in hyperopt (Bergstra *et al.*, 2013). The objective function was designed to increase the F_1 score while decreasing network size:

$$L = (1 - F_1) + \lambda \cdot \frac{p_c}{p_{\max}}.$$

Here, L denotes the loss to be minimized, p_c denotes the number of parameters in the current iteration of the network and $p_{\max}$ denotes the maximum number of parameters attainable given the boundaries of the parameter search space. The parameter λ controls the trade-off between accuracy and network size and was set to 0.01.

Networks output the posterior probability of their target *k*-mers being present in a squiggle segment. The threshold above which this posterior probability is considered sufficiently high to detect the presence of a *k*-mer was chosen to maximize the F_1 score, using a grid search on training data for probabilities between 0.75 and 0.999 with a step size of 0.001. For read detection mode, the fraction of *k*-mers to be detected before the target sequence is considered present must be set as well. This parameter was optimized simultaneously with the posterior probability threshold.

4.2 False positive rate simulation

An optimal choice for the value of *k* should balance the abundance of a *k*-mer, such that it is rare enough to discriminate sequences, yet not so rare that it never occurs at all. We approximate this optimal value by considering the probability of detecting target sequences in random sequences by chance.

Assuming all canonical *k*-mers are equally represented, the expected number of *k*-mer occurrences in a read of length L_{read} is $L_{\text{read}} \cdot 2 \cdot 4^{-k}$ and the probability of a *k*-mer occurring in the sequence at least once can be estimated using a Poisson distribution. Assuming we draw *k*-mer-detecting networks from a library of perfectly accurate pre-generated networks A, the expected number of *k*-mers found in a target sequence at least once can be calculated:

$$E = P(X_{\text{target}} \geq 1) \cdot |A|.$$

Here, X_{target} is the number of occurrences of a *k*-mer in the target sequence, $|A|$ is the size of the *k*-mer network library and E is the expected number of *k*-mers in A found in the target sequence. A false positive occurs when a read that does not contain the target sequence contains all the selected *k*-mers of the target read by chance. The rate at which this occurs can be estimated as follows:

$$\text{FPR} = P(X_{\text{non-target}} \geq 1)^E.$$

Here, $X_{\text{non-target}}$ denotes the number of occurrences of a *k*-mer in the non-target sequence and FPR denotes the false positive rate. We performed FPR simulations for different values of *k*, representative values for non-target sequence lengths—30 and 50 kb, representing full nanopore read lengths—and target sequence lengths—0.6, 1.5 and 30 kb, representing BOLD barcodes (Ratnasingham and Hebert, 2007), 16S sequences and whole coronavirus genomes, respectively—and selected the value for *k* that minimized FPR. Both $k = 8$ and $k = 9$ returned good FPR values, thus we included *k*-mers of both sizes in subsequent steps.

4.3 k-mer library design

For 16S sequence detection, we composed a library of 1500 suitable *k*-mers, which should allow detection of a wide range of species. Similar to Doroschak *et al.* (2020), we used an evolutionary algorithm to select *k*-mers that are dissimilar in sequence and produce easily distinguishable squiggles. To enforce sequence dissimilarity, only *k*-mers with a maximum Smith–Waterman score of 6 (assuming gap penalty, match score and mismatch score of −4, 1 and −1, respectively) to other selected *k*-mers are allowed, while squiggle dissimilarity is enforced by comparing simulated squiggles as produced by guppy (v. 5.0.11 + 2b6dbff). That is, we only accept modifications made to *k*-mers by the evolutionary algorithm if both the minimum and the average dynamic timewarping score between its squiggle and the other squiggles in the set increase. Furthermore, for the bacterial case study, we remove the outer 10 percentiles of most abundant *k*-mers based on 20 959 16S rRNA sequences obtained from NCBI (Bioproject:PRJNA33175) because these *k*-mers are

excessively rare or ubiquitous. Additionally, *k*-mers containing four or more of G/C or 5 or more of A/T in a row are rejected as the length of homopolymer stretches can be difficult to detect in squiggles. Starting with a set of random sequences, we ran the evolutionary algorithm for 10 rounds of decreasing numbers of proposed mutations per sequence; the initial two rounds applied five mutations in each sequence, after which the number of mutations decreased by one for each two rounds.

4.4 Nanopore sequencing

Poeciliidae reads were obtained from a previous study and have been obtained as described in van Kruistum *et al.* (2021). For 16S reads, we sequenced pre-made DNA isolate of microbial mock community A (v3.1, HM-278D, BEI resources) on a MinION (Mk1B, Oxford Nanopore plc.) using accompanying flowcell (FLO-MIN106) and 16S sequencing kit (SQK-RAB204). In this sequencing run we saved the fast5 files, which contain the raw data. Subsequently, we used these raw data to perform the speed and accuracy benchmarks on the Jetson Nano, high-end desktop and server (see Section 4.6).

4.5 Data preparation

To obtain a ground truth species assignment, all reads were basecalled using guppy (v5.0.11 + 2b6dbff) in high-accuracy mode. For 16S reads, we mapped them using BLASTN (v2.9.0+) to the expected 21 bacterial GenBank genomes (Supplementary Table S1). The species to which the sequence identity was highest was selected as the ground truth species for that read. To correct sequencing errors and assign individual bases to each squiggle segment, we aligned reads to reference genomes using tombo (v1.5.1). *Poeciliopsis gracilis*, *P.januarius* and *P.turneri* reads were aligned to genomes constructed from the nanopore reads, while 16S reads were aligned to their respective GenBank genomes. These genomes were also used for salient *k*-mer detection in the evaluation of read detection mode. For abundance mode validation, *k*-mers were selected from GenBank short read genomes, built from Illumina reads of the same three individuals (GCA_903067085.1, GCA_902982915.1 and GCA_903068135.1 for *P.gracilis*, *P.januarius* and *P.turneri*, respectively).

4.6 Benchmarking

We compared baseLess performance on 16S reads to four other tools; UNCALLED (v2.0-127-g0fc1cab), SquiggleNet (v1.0), DeepNano-blitz (v1.0) and Guppy (v5.0.11 + 2b6dbff, 'fast' mode). As the latter two tools are basecallers and not mapping tools, the basecalled reads returned by these were mapped to target genomes using minimap2 (2.17-r941) to produce the final prediction.

We performed accuracy and F_1 score benchmarks in stratified 5-fold cross-validation on 335 000 reads. Tools were run on a PowerEdge R740 server (Dell), on three Xeon Gold 6242 CPUs @2.80GHz (Intel). As Guppy and SquiggleNet were optimized for GPU usage, they were run on a Tesla T4 GPU (NVIDIA). We ran all tools in a Snakemake (Köster and Rahmann, 2012) workflow. Training 1500 baseLess *k*-mer models on the server took around 7 h.

Speed benchmarks were performed on two systems, an Nvidia Jetson Nano System-on-Module (2 GB RAM, ARM CPU, 4 cores@1.43GHz, Nvidia Maxwell GPU, 128 cores@921MHz) and a high-end desktop computer (32 GB RAM, AMD Ryzen 3700× CPU, 16 cores@3.6GHz, Nvidia GeForce RTX 3070, 5888 cores@173GHz). Inference was performed 10 times per tool over 1000 reads. Tools were given access to 3 CPU cores and the GPU if they were configured to use it. On the Jetson Nano, baseLess' maximum memory usage was set to 512 MB, which theoretically ensures sufficient overhead to run MinKNOW.

Acknowledgements

We thank Elio Schijlen and Bas te Lintel Hekkert for help with nanopore sequencing. We also thank Henri van Kruistum for the provision of raw nanopore reads and nanopore assemblies for *P.gracilis*, *P.januarius* and *P.turneri*.

Author contributions

Ben Noordijk (Formal analysis [equal], Investigation [equal], Methodology [equal], Software [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review and editing [equal]), Rijndert Nijland (Methodology [equal], Resources [equal], Supervision [equal], Validation [equal], Writing—review and editing [equal]), Victor J. Carrion (Resources [equal], Supervision [equal], Writing—review and editing [equal]), Jos M. Raaijmakers (Resources [equal], Supervision [equal], Writing—review and editing [equal]), Dick de Ridder (Conceptualization [equal], Funding acquisition [equal], Methodology [equal], Resources [equal], Software [equal], Supervision [equal], Writing—review and editing [equal]), Carlos de Lannoy (Conceptualization [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Resources [equal], Software [equal], Supervision [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review and editing [equal]).

Funding

This work was supported by the Netherlands Foundation of Scientific Research Institutes (NWO-I, formerly FOM) [SMPS to D.d.R.].

Conflict of Interest: none declared.

References

Abadi,M. *et al.* (2015) TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. https://www.tensorflow.org/

Bao,Y. *et al.* (2021) Squigglenet: Real-time, direct classification of nanopore signals. *Genome Biol.*, **22**, 1–16.

Bergstra,J. *et al.* (2013) Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In: *International Conference on Machine Learning*, pp. 115–123. PMLR.

Boža,V. *et al.* (2020) Deepnano-blitz: A fast base caller for minion nanopore sequencers. *Bioinformatics*, **36**, 4191–4192.

Bruijns,B. *et al.* (2018) Massively parallel sequencing techniques for forensics: A review. *Electrophoresis*, **39**, 2642–2654.

Cristiano,S. *et al.* (2019) Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature*, **570**, 385–389.

de Lannoy,C. *et al.* (2017) The long reads ahead: De novo genome assembly using the minion. *F1000Research*, **6**, 1083.

Doroschak,K. *et al.* (2020) Rapid and robust assembly and decoding of molecular tags with DNA-based nanopore signatures. *Nat. Commun.*, **11**, 1–8.

Faria,N.R. *et al.* (2016) Mobile real-time surveillance of Zika virus in Brazil. *Genome Med.*, **8**, 1–4.

Goordial,J. *et al.* (2017) *In situ* field sequencing and life detection in remote (79°26′N) Canadian high arctic permafrost ice wedge microbial communities. *Front. Microbiol.*, **8**, 2594.

Green,E.D. *et al.* (2015) Human genome project: Twenty-five years of big biology. *Nature*, **526**, 29–31.

Köster,J. and Rahmann,S. (2012) Snakemake—A scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.

Kovaka,S. *et al.* (2021) Targeted nanopore sequencing by real-time mapping of raw electrical signal with uncalled. *Nat. Biotechnol.*, **39**, 431–441.

Li,H. (2018) Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.

Loose,M. *et al.* (2016) Real-time selective sequencing using nanopore technology. *Nat. Methods*, **13**, 751–754.

Mistry,D.A., *et al.* (2021) A systematic review of the sensitivity and specificity of lateral flow devices in the detection of SARS-COV-2. *BMC Infect. Dis.*, **21**, 1–14.

Newman,A.M. *et al.* (2014) An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat. Med.*, **20**, 548–554.
Normand,E.A. *et al.* (2018) Clinical exome sequencing for fetuses with ultrasound abnormalities and a suspected Mendelian disorder. *Genome Med.*, **10**, 1–14.
Pomerantz,A. *et al.* (2018) Real-time DNA barcoding in a rainforest using nanopore sequencing: Opportunities for rapid biodiversity assessments and local capacity building. *GigaScience*, **7**, giy033.
Pust,M.-M. *et al.* (2021) Direct RNA nanopore sequencing of *Pseudomonas aeruginosa* clone c transcriptomes. *J. Bacteriol.*, **204**, e00418–21.
Ratnasingham,S. and Hebert,P.D. (2007) Bold: The barcode of life data system (http://www. barcodinglife. org). *Mol. Ecol. Notes*, **7**, 355–364.
van Kruistum,H. *et al.* (2021) Parallel genomic changes drive repeated evolution of placentas in live-bearing fish. *Mol. Biol. Evol.*, **38**, 2627–2638.
Yang,J. *et al.* (2020) Species-level analysis of human gut microbiota with metataxonomics. *Front. Microbiol.*, **11**, 2029.