Pearman et al. BMC Bioinformatics (2020) 21:220
https://doi.org/10.1186/s12859-020-3528-4

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Testing the advantages and disadvantages of short- and long- read eukaryotic metagenomics using simulated reads

William S. Pearman*, Nikki E. Freed and Olin K. Silander*

* Correspondence: wpearman1996@gmail.com; olinsilander@gmail.com
School of Natural and Computational Sciences, Massey University, Private Bag 102904, North Shore, Auckland 0745, New Zealand

Abstract

Background: The first step in understanding ecological community diversity and dynamics is quantifying community membership. An increasingly common method for doing so is through metagenomics. Because of the rapidly increasing popularity of this approach, a large number of computational tools and pipelines are available for analysing metagenomic data. However, the majority of these tools have been designed and benchmarked using highly accurate short read data (i.e. Illumina), with few studies benchmarking classification accuracy for long error-prone reads (PacBio or Oxford Nanopore). In addition, few tools have been benchmarked for non-microbial communities.

Results: Here we compare simulated long reads from Oxford Nanopore and Pacific Biosciences (PacBio) with high accuracy Illumina read sets to systematically investigate the effects of sequence length and taxon type on classification accuracy for metagenomic data from both microbial and non-microbial communities. We show that very generally, classification accuracy is far lower for non-microbial communities, even at low taxonomic resolution (e.g. family rather than genus). We then show that for two popular taxonomic classifiers, long reads can significantly increase classification accuracy, and this is most pronounced for non-microbial communities.

Conclusions: This work provides insight on the expected accuracy for metagenomic analyses for different taxonomic groups, and establishes the point at which read length becomes more important than error rate for assigning the correct taxon.

Keywords: Metagenomics, Nanopore, Illumina, Long read, Community composition

Background

Applying metagenomic methods to quantify community composition

To understand ecological community diversity, it is essential to quantify taxon frequency. The most common method of quantifying taxa frequencies is through metabarcoding [1]. In this method, conserved genomic regions (often 16S rRNA in the case of bacterial and archaeal species; 18S rRNA or Cytochrome c oxidase I for eukaryotic species) are amplified from the sample of interest, sequenced (most often using high-

throughput methods such as Illumina), and then classified using one of several available pipelines (e.g. QIIME, MEGAN, Mothur) [2–4]. Many of these pipelines have been designed around the analysis of bacterial datasets.

In contrast to metabarcoding, metagenomic approaches do not rely on the amplification of specific genomic sequences, which can introduce bias. Instead, they aim to quantify community composition based on the recovery and sequencing of all DNA from community samples. Metagenomic methods limit biases that can occur during the amplification steps of metabarcoding, and yield insight into the functional diversity present in ecosystems [5, 6].

While metabarcoding approaches have been widely applied to both microbial and eukaryotic taxa, the vast majority of metagenomic studies have focused only on microbial communities. Unsurprisingly, the various advantages and disadvantages of using metagenomic analyses for microbial communities are well-documented [7–9]. There are likely several factors driving this microbe-centric application of metagenomics, including (1) the greater level of diversity of microbial taxa; (2) the considerable number of microbial taxa that are "unculturable," making it difficult to collect the requisite amount of DNA for genomic sequencing; (3) the availability of a multitude of non-molecular methods for quantifying multicellular taxa; and (4) the relative paucity of genomic sequence for multicellular organisms in databases [10] (Supp. Figure S1). This latter factor is perhaps the single largest factor in driving the bias toward microbial metagenomics.

However, the amount and diversity of multicellular genomic sequence data is rapidly increasing. Although multicellular metabarcode databases are currently far more complete relative to genomic databases, this gap is closing quickly. For example, the Earth BioGenome project aims to sequence the genomes of upwards of one million eukaryotic species within the next decade [11]. Regardless of the success of this effort, there are a host of ongoing eukaryotic sequencing projects, including Bat 1 K [12], Bird 10 K (10,000 bird genomes [13]), G10K (10,000 vertebrate genomes [14]), and i5K (5000 arthropod genomes [15]), among others. This suggests that within the next 5 years, most multicellular organisms will have at least one member of their family present in genomic databases, with some groups of multicellular organisms being completely represented at the genus level.

This would increase the utility of metagenomics for assessing membership in plant and animal communities, especially for cases in which organisms are difficult to observe or degraded. This is frequently the case for diet studies [16], many invertebrate communities such as in treeholes [17] or algal holdfasts [18].

Analysis of short-read metagenomic data

Many metagenomic classification analyses rely on first pass classifiers to assign reads to one or more taxa, followed by second pass classifiers that can improve on the initial classification by taking into account the number and relationship of taxa identified in the first pass. This second step often relies on a lowest common ancestor algorithm [3, 19, 20], or by refining taxonomic representation by examining the results from the first pass classifier [21].

The most widely used first pass classifier is BLAST (basic local alignment search tool), and it is considered the gold standard [22]. However, BLAST is not

computationally efficient enough to deal with tens or hundreds of millions of reads. Thus, algorithms for fast metagenomic classification have been the subject of intense research over the last few years, and include k-mer based approaches such as CLARK [23], Kraken and related tools (Kraken, Kraken2, and KrakenUniq) [19], Centrifuge [20], EnSVMB [24], and Kaiju [25], as well as reduced alphabet amino acid based approaches such as DIAMOND [26]. In almost all cases these have been designed and benchmarked using short read data [22].

Analysis of long-read data metagenomic data

The advent of "third generation" single molecule long read technologies (PacBio and Oxford Nanopore) has significant implications for metagenomic analyses, most notably for genome assembly [27, 28]. These technologies allow read lengths of 10 kilobase pairs (Kbp) and beyond, in strong contrast with the approximately 300 base pairs (bp) limit of Illumina. However, both PacBio and Nanopore technologies have far higher error rates (88–94% accuracy for Nanopore [29] and 85–87% for PacBio [30]). The lower accuracy of Nanopore and PacBio (non-circular consensus) sequence reads may affect the success of current classification methods, and there are few algorithms designed to exploit long-read data.

As a first approach toward determining the use of long-read technologies for metagenomic applications, we would like to understand the relative advantages and disadvantages of using short accurate reads versus long error-prone reads. Recent work has shown that relatively high genus level classifications of approximately 93% have been achieved using Nanopore-based metagenomic analyses of a mock bacterial community [31]. Here we expand this analysis to allow direct comparison between short and long read approaches. In addition, we compare metagenomic classification success in microbial communities as compared to communities of multicellular organisms. We find that longer reads, despite their higher error rate, can considerably improve classification accuracy compared to shorter reads, and that this is especially true for specific taxa.

Results

We first looked only at short read lengths to quantify the effects of sequencing technology and classifier (BLAST or Kraken2) on recall at the level of genus. For both bacteria and fungi, we found that recall was at or above 99.9% for Illumina reads of any length (100 bp, 150 bp, or 300 bp), for both BLAST and Kraken2 (Fig. 1). Similarly, PacBio had very high recall for these groups at approximately 97% for 100 bp reads, and 99.98% for 300 bp reads. In strong contrast, for Nanopore data the recall was far lower; approximately 25% for 100 bp reads and increasing to 75% at 300 bp. In general, Kraken2 had slightly lower recall than BLAST.

However, for plants and animals, average recall was low regardless of sequencing technology. Average recall for Illumina reads peaked at approximately 55 and 75% for animals and plants, respectively (Fig. 1, black). Nanopore recall rates peaked at just over 20 and 35% for animals and plants, respectively, while for PacBio these values were 33 and 62% respectively. However, this was highly taxon-dependent, with some taxa consistently having recall near 100%, while others remained close to 0% regardless of sequencing technology or read length. Perhaps surprisingly, on average Kraken2

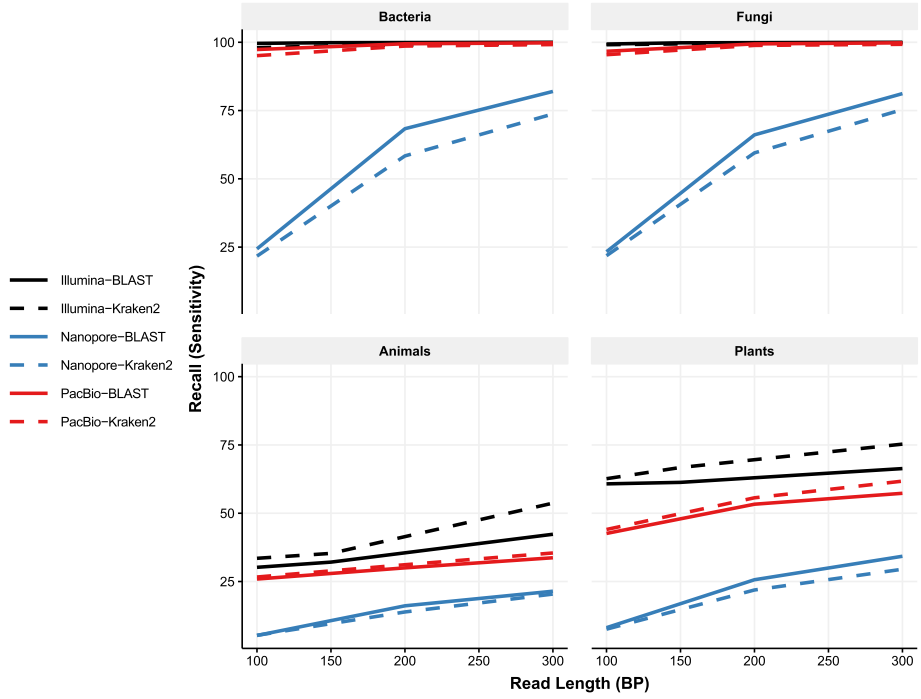

Fig. 1 Recall is consistently higher in bacteria and fungi than plants or animals for both short Illumina and Nanopore reads. Each panel shows recall for the different kingdoms. Median recall at the genus level across each panel is shown by blue (Nanopore) or red (PacBio), and black (Illumina). Dashed lines indicate recall rates for reads classified using Kraken2; solid for reads classified using BLAST. Illumina reads exhibit consistently higher recall, followed by PacBio; bacteria and fungi exhibit higher recall than plants or animals

outperformed BLAST for Illumina reads for both plant and animal taxa. This difference may be due to the default BLAST parameters being sub-optimal for short reads.

We next quantified differences in classification success (the proportion of all classified reads that were correctly classified), again considering only short read lengths. For bacteria and fungi, all three sequencing methods exhibited high classification success, with the exception of Kraken2 classification of Nanopore reads (Fig. 2). For each sequencing method and classifier, classification success for plants and animals was low relative to bacteria and fungi. For both Illumina and Nanopore, BLAST resulted in approximately 87 and 97% of reads being correctly classified, for animals and plants respectively. While for PacBio these values were slightly lower at 82 and 95%. Kraken2 success was far lower, especially for Nanopore reads, peaking at 54% in animals (Fig. 2). PacBio reads exhibited slightly lower classification success for Kraken2 at 70 and 85% for animals and plants, Illumina reads were only slightly better for these taxa at 72 and 90%. Over this range of read lengths, we found only a weak relationship between read length and classification success, in contrast to the results for recall.

It is perhaps expected that highly accurate Illumina reads would result in more accurate taxonomic classification, followed by slightly lower accuracy PacBio circular consensus sequence (CCS) reads, and more error-prone Nanopore reads. However, it is possible to obtain PacBio and Nanopore reads far in excess of 300 bp (single Nanopore reads of up to 2 megabase pairs have been sequenced), so we next quantified recall and classification success for reads with lengths up to 4000 bp. Because such read lengths

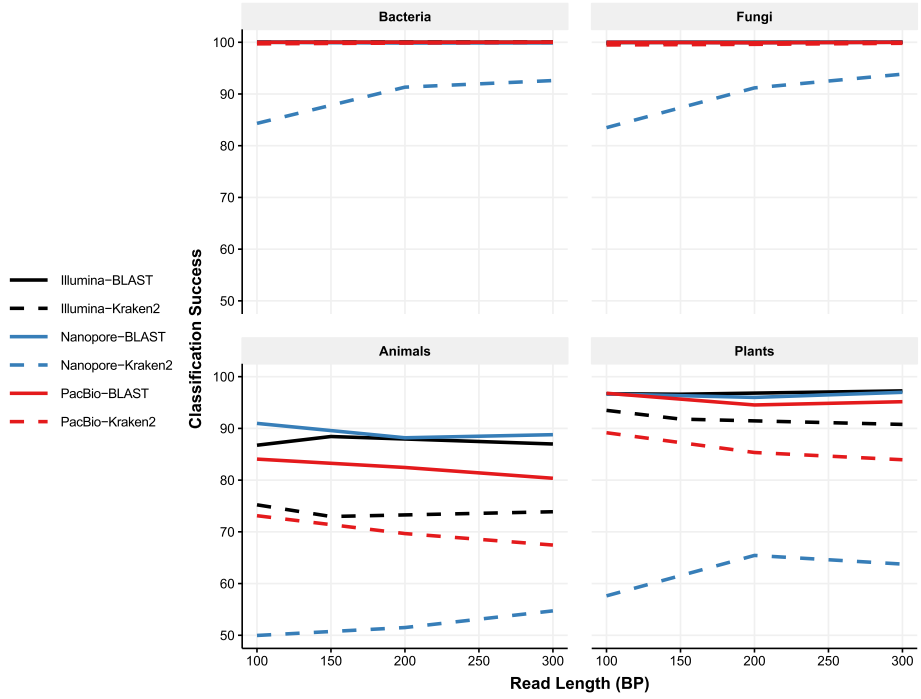

Fig. 2 Classification success for short reads is weakly related to read length and strongly dependent on classification method. Each panel shows recall for the different kingdoms. Median classification success at the genus level across each panel is shown by blue (Nanopore) or red (PacBio), and black (Illumina). Dashed lines indicate recall rates for reads classified using Kraken2; solid for reads classified using BLAST. For bacteria and fungi, median classification rates of of all pairs except for Nanopore-Kraken2 are almost exactly 100% for all read lengths in Bacteria and Fungi, resulting in obscured lines at these higher levels

are not currently possible to obtain using Illumina technology, we did not measure recall and classification success for Illumina reads of similar lengths.

We observed a similar relationship between read length and recall for both BLAST and Kraken2. For bacteria and fungi, nanopore recall increased from ~ 20% using 100 bp reads to almost 100% when using 1500 bp reads, while PacBio recall increased from ~ 95% to ~ 100% for these taxa. For animals and plants we observed similar trends, although at no point did recall approach 100%. However, long Nanopore reads surpassed the recall of even the longest Illumina reads (300 bp) classified with Kraken2, with crossover points at approximately 3000 bp for animals and 2500 bp for plants regardless of classifier (Fig. 3, black and blue solid lines). The crossover points between PacBio read lengths with Illumina 300 bp read lengths were approximately 700 bp (BLAST) and 900 bp (Kraken2) for animals, while for plants these values were 600 bp (BLAST) and 800 bp (Kraken2).

We also considered this metric at the level of family. In this case found that for animals, Nanopore reads surpassed Illumina reads only at lengths close to 4000 bp, reaching approximately 70% recall at this point (Supp. Figure S2). However, for plants Nanopore recall surpassed Illumina recall at 2500 bp, with 4000 bp reads yielding a recall of approximately 90%. When using PacBio reads at the family level, recall exceeded that of Illumina in all instances regardless of classifier (crossover points at 1500 bp for Kraken2 with animals, and 2500 bp for BLAST with animals). For plants the crossover points were 600 bp and 800 bp for animals and plants respectively.

We next examined classification success at longer read lengths. For BLAST we observed no relationship between classification success and read length for both PacBio and Nanopore (Fig. 4.). Bacteria and fungi both had consistently high classification success (median 100%), while animals and plants had lower classification success (median 82 and 96%, respectively). Interestingly, classification success in animals for PacBio reads was lower than that of Nanopore reads, at medians of 77 and 86% respectively. This is likely explained by Nanopore reads having more failed queries overall (lower recall). In all other taxa, classification success was virtually the same for both sequencing technologies.

In contrast to BLAST, for Kraken2 we observed a consistent increase in classification success as read length increased (Fig. 4). This was the case for both PacBio and Nanopore reads. However, for Nanopore reads, Kraken2 classification success never exceeded that which was observed for BLAST. This was not the case for PacBio reads, where Kraken2 classification success exceeded that of BLAST at 2600 bp (animals) and 2800 bp (plants).

Finally, we tested classification success at the level of Family. In this case, we observed that for BLAST, the classification success for plants was approximately 99% over all read lengths regardless of sequencing technology, while for Kraken2 only 4000 bp reads reached this level. For animals, BLAST classification success of nanopore reads was approximately 95% over all read lengths while for PacBio reads this value

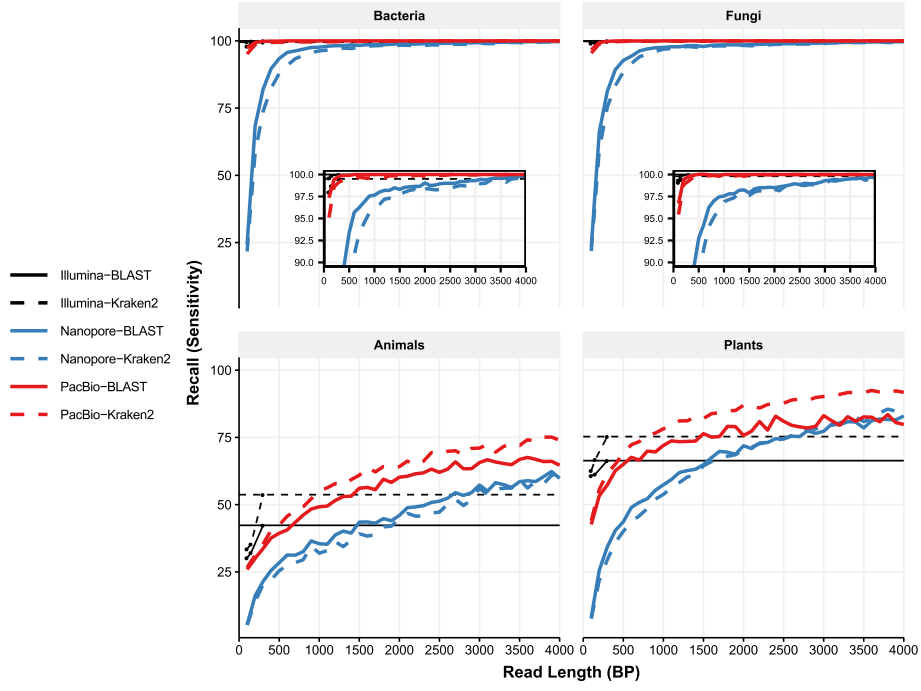

Fig. 3 Long PacBio and Nanopore reads equal or surpass the recall of the longest Illumina reads for both BLAST and Kraken2. Each panel shows recall for the different kingdoms. Median recall at the genus level is indicated by blue lines for Nanopore or red lines for PacBio, either dashed (Kraken2) or solid (BLAST). The recall rates for 300 bp Illumina reads are shown as thin solid black lines, again either dashed (Kraken2) or solid (BLAST). Black points show the recall for the short Illumina reads at lengths of 100 bp, 150 bp, and 300 bp. Insets show the upper end of the y-axis in greater detail for both bacteria and fungi

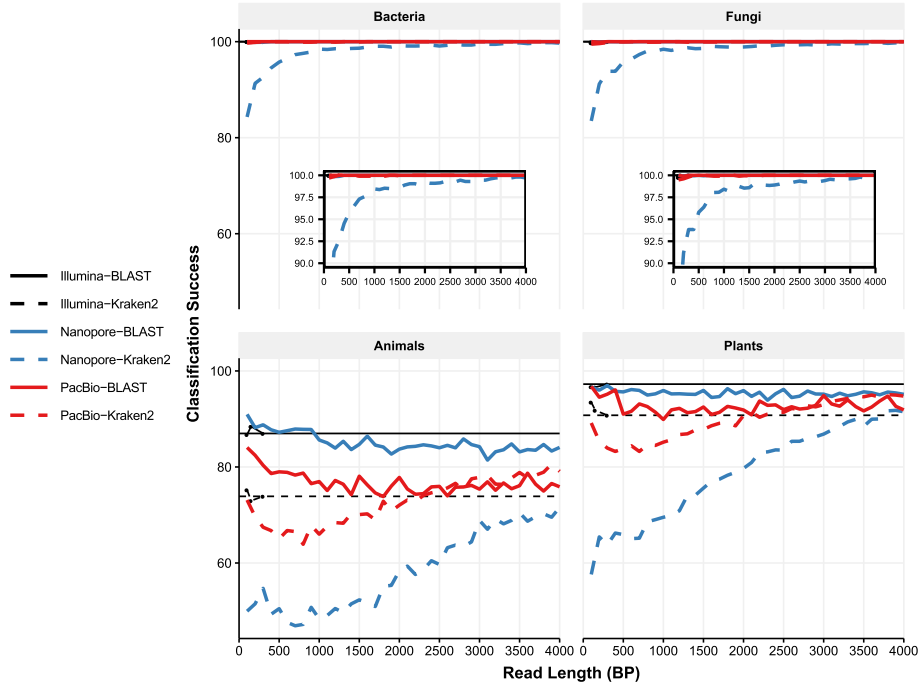

Fig. 4 Classification success for long reads is dependent on read length only for Kraken2 classification. Each panel shows classification success for the different kingdoms. Median classification success at the genus level is indicated by blue lines for Nanopore or red lines for PacBio, either dashed (Kraken2) or solid (BLAST). The recall rates for 300 bp Illumina reads are shown as thin solid black lines, again either dashed (Kraken2) or solid (BLAST). Black points show the classification success for all Illumina reads of all lengths (100 bp, 150 bp, and 300 bp). For animals and plants, the classification success of Kraken2 depends strongly on read length, and only surpasses BLAST for PacBio at approximately 3500 bp for both animals and plants. Insets show the upper end of the y-axis in greater detail for both bacteria and fungi

was ~ 87%. For Kraken2, the highest classification success was 85% for nanopore reads and 91% for PacBio (Supp. Figure S3).

Discussion

Here we have compared the relative accuracy of taxon classification using simulated short accurate reads (Illumina) and long reads (Nanopore and PacBio) with known ground truth. We have used two simple metrics of success: recall (the ratio of correctly classified reads to all reads) and classification success (the ratio of correctly classified reads to all classified reads). We have tested taxon classification using a broad range of taxa, including bacteria, fungi, animals, and plants.

Recall for both BLAST and Kraken2 was improved by the use of long reads, especially in the case of animals and plants, for which recall improved almost three-fold as read length increased from 300 bp to 4000 bp. Generally, both Kraken2 and BLAST achieved similar levels of recall. The exception was for short reads for animals and plants, for which Kraken2 was more accurate than BLAST.

We found no relationship between classification success and read length for BLAST. This implies that when a read is classified as belonging to a taxa, the likelihood it was correctly classified remains relatively constant over different read lengths. However, the number of reads that are classified *at all* increases with read length (causing an increase in recall). The exception to this lack of relationship between classification

success and read length was for Kraken2, for which the proportion of correctly classified reads increases with read length by more than 50% for both plants and animals. While relatively few studies have analysed the role of read length in classification succession, at shorter read lengths (such as those achieved through pyrosequencing) [32] or for metagenomic assembled fragments (those constructed through Illumina sequencing with an assembly step prior to classification) [33], similar results have been achieved. Our results are unique in that we show a similarly important role of read length, but for long reads without additional steps to improve read quality.

Our results also indicate that recall for long PacBio and Nanopore reads was equal to or higher than even the longest Illumina reads (300 bp). This was true regardless of kingdom, or classification method (Fig. 3). One implication of this is that a simple way to improve Nanopore classification accuracy is to impose minimum read lengths. This can be achieved by performing size selection during library preparation or during computational analyses. Additionally, these results indicate that recall for PacBio reads is consistently higher than Nanopore, and also rapidly surpasses that of Illumina. Like Nanopore reads, this was true regardless of kingdom, or classification methods.

At first glance, then, there appears to be a clear trade-off between short read Illumina and long read sequencing for metagenomic analyses. While both PacBio and Nanopore allow higher recall at long read lengths, this advantage is offset by the fact that Illumina generally provides more reads per run. This consideration is most significant for PacBio sequencing, which (while having greater accuracy than Nanopore due to the vast majority of reads being CCS), provides considerably less sequencing throughput (at most, approximately two million short reads short (e.g. < 4 Kbp) per 8 M SMRT cell). This contrasts with Nanopore PromethION throughput, which can yield well in excess of 20 million 4 Kbp reads per run. Even MinION throughput is easily in excess of 2 million reads per run for short 2Kbp reads.

The more critical question is whether the higher recall makes up for this deficit in read number. At most, recall for Nanopore improves recall 50% beyond 300 bp Illumina reads, while classification success is similar (using BLAST). Thus, if the read capacity of Illumina runs is 50% or more than Nanopore (or Pacbio), the number of correctly classified reads will be maximised using Illumina technology - on a per sequencing run basis. However, for many researchers the more relevant metric is cost per read, or in many cases, time, which may change this cost calculation. Given this, we find no clear advantage in using Illumina over Nanopore given the observed classification accuracy for long inaccurate Nanopore reads. However, the low number of PacBio reads, despite their high accuracy, may limit the effectiveness of this platform for metagenomic analyses.

Differences in accuracy between bacteria, fungi, animals, and plants

We find very large differences in classification accuracy (mostly in terms of recall) for bacteria and fungi versus plants and animals. The discrepancy between taxonomic groups likely arises from a variety of factors. Among these are the higher degree of divergence between bacterial species relative to animal and plant species, and the complexity of bacterial genomes compared to eukaryotic genomes. We discuss these factors below.

Bacterial taxa are often considered separate species once they have diverged by 6% ANI (Average Nucleotide Identity) on a genomic level [34, 35]. The degree of nucleotide divergence between eukaryotic species is not standardised [36], and species are generally designated as such based on the biological species concept put forward by Mayr [37]. Additionally, divergence levels differ substantially between loci (as for bacteria). However, for some loci general ranges for eukaryotic species have emerged. For example, for mitochondrial COI, between-species divergence is usually greater than 3% [38, 39]. These loci are among the fastest diverging loci in plant and animal genomes, and many other loci may differ by far less than 1% between species. Due to this low level of divergence, metagenomic classifiers may frequently classify animal and plant genera with lower accuracy than bacterial genera.

A second explanation for the increased classification success in bacteria and fungi is that these genomes contain fewer repetitive elements than animals or plants [40]. Although such repetitive regions are usually masked from classifiers (including BLAST and Kraken2), this masking may not be complete.

A third reason is that the genomic databases for plants and animals are far less complete than for bacteria and fungi. There is a large difference in the number of genomes and sequences available for different Kingdoms, with bacteria having significantly more species present than the next closest kingdom (See Supp. Figure S1). However, we expect this factor will be mitigated in the future as genomic databases continue to expand and computational search methods continue to improve.

Differences in accuracy between Kraken2 and BLAST

We observed similar levels of recall for BLAST and Kraken2 over most read lengths for Nanopore reads, while for PacBio reads Kraken2 appeared to perform better than BLAST in animals and plants. However, there were strong differences in classification success. For short reads, Kraken2 classification success was far lower than BLAST. As read lengths increased, Kraken2 classification success approached those of BLAST for Nanopore reads. Classification success for PacBio reads appeared to be lower than that of Nanopore reads when using BLAST, while when using Kraken2 the opposite was the case, and PacBio had consistently higher classification success than Nanopore. Part of this is likely due to longer reads allowing multiple k-mer matches, decreasing the probability of a false positive classification. One perhaps underappreciated advantage of Kraken2 over BLAST is that Kraken2 has reduced sensitivity to structural variation within reads. As Kraken2 allows multiple k-mers to match within a read, structural changes (e.g. inversions) are less likely to influence the outcome of Kraken2 matching. Such structural changes may influence BLAST due to the matching and extend algorithm. Thus for long reads, classifiers that are insensitive to synteny may be more successful, especially for taxa in which structural rearrangements are common.

Conclusions

Here we have shown both PacBio and Nanopore reads, despite being more error-prone than Illumina reads, are useful for metagenomic classification due to their increased length. For plant and animal communities, the classification accuracy of long Pacbio and Nanopore reads exceeds that of Illumina. We found that classification accuracy is

more dependent on the set of taxa being considered than on the metagenomic classifier being used (Kraken2 or BLAST), and that this was true for both short accurate (Illumina and long (Nanopore and PacBio) sequence data). Together these data suggest that one consideration in selecting a metagenomic sequencing method (i.e. long or short read) is the taxonomic group of interest.

Methods

Genomic data

For each of four major taxonomic divisions (bacteria, fungi, animals, and plants), we downloaded 20 genomes from GenBank [41]. Within each of these divisions, we included genomes from a total of 22 classes, 46 orders, and 58 families (Fig. 5; details in Supplementary Table S1). To select these genomes, we first constructed a filtered list of genomes that were represented at the chromosome level or greater in GenBank. Within divisions we then selected genomes randomly from this list (without replacement, such that no species was represented more than once).

Read simulation

We simulated Nanopore reads using NanoSim 2.0.0 [42] with the default error parameters for *E. coli* R9 1D data. This method uses a mixture model to produce simulated reads with indel and error rates similar to real datasets. The error model is applied equally to all parts of a read, and the read lengths are drawn from a distribution approximating real data. To create simulated read data of specific lengths, we truncated the simulated reads after the relevant number of basepairs using a custom bash script (i.e. to simulate 100 bp Nanopore reads, we truncated all reads in a simulated dataset to 100 bp (see example command below).

```
simulator.py linear –r Reference –c ecoli –n 1000 –o Output  ––min_len Length  ––max_len 8000
cat Output.fasta | awk–v'RS => "NR > 1{print" > "1;printf(",2)}' > Output_trimmed.fasta
```

We simulated PacBio reads using SimLORD. This method assumes that the number of passes along a read is chi-squared distributed and dependent on read length, as would occur in a standard SMRT run. Thus, shorter reads have more passes, with very short reads (e.g. < 1000 bp in length) having well over 10 passes on average; reads 2000 basepairs in length having a median of over four passes; and reads 3000 basepairs in length having a median of more than three passes. Thus, short reads are almost exclusively CCS ("HiFi") reads. For both Nanopore and PacBio, we simulated reads for lengths varying from 100 bp to 4000 bp at 100 bp intervals, simulating 1000 reads per interval for all taxa (a total of 40,000 reads for each taxon per read simulation program, and 3.2 million reads for all taxa and read lengths per read simulation program). (See example command below).

```
simlord ––read–reference Reference –n 1000 –f1 Length Output
```

We simulated Illumina data using dwgsim 0.1.12 [43] with the following options:

```
dwgsim –e 0.0001 –E 0.0001 –N 2000 –1 100 –2 100 –r 0.0001 –R 0.01 –y 0.000 –c 0
```

This implements errors to mirror those in Illumina data, with constant error rates of 1e-4 and no indels (which are extremely rare in Illumina data). We generated 1000

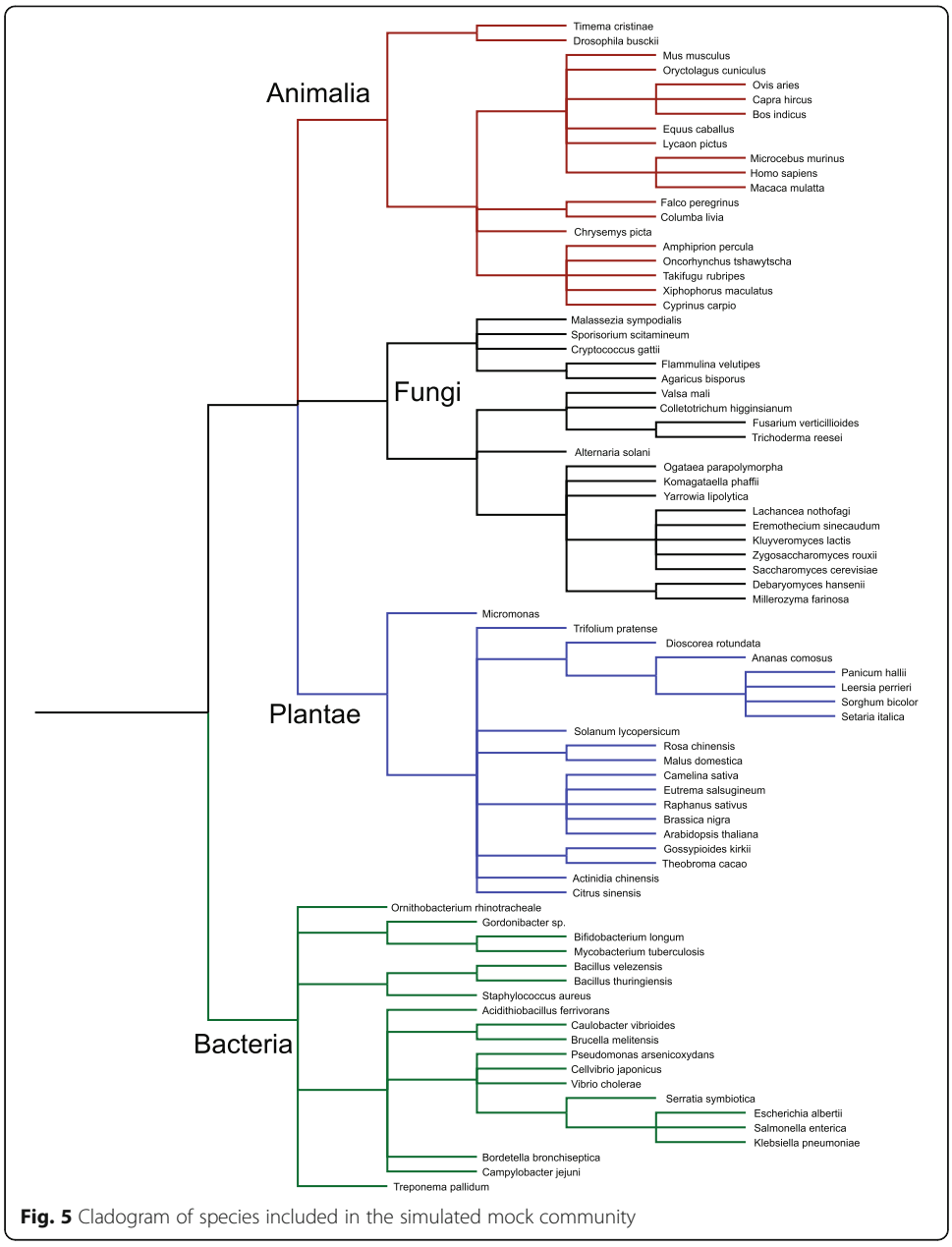

Fig. 5 Cladogram of species included in the simulated mock community

reads for each genome, at three read lengths: 100 bp, 150 bp, and 300 bp (a total of 240, 000 reads across all taxa and lengths), and used only single end reads for all analyses.

Sequence classification

We used BLAST 2.7.1 [44] and Kraken2 [19] for sequence classification. We created a local custom database consisting of the NCBI nt database (downloaded on Feb 8, 2019) and the genomes of the 80 taxa that we used to test classification success. We used the default alignment parameters for BLAST (word_size = 11, match/mismatches scores = 2,-3, gap costs – existence = 5, extension =2, filter = low complexity regions), except for implementing a maximum e-value of 0.1. We used the match with the highest bit score

for all downstream analyses. For Kraken2 analyses we used the default parameters (in which the k-mer length is 35 bp and default minimiser length is 31 bp). For Kraken2 we used the taxon assigned by the lowest common ancestor (LCA) algorithm employed in Kraken2 for downstream analyses.

Accuracy metrics

To assess the effects of read length on classification accuracy we focus our analysis only on how often a read is assigned to the correct taxon. For our simulated reads there are three possible outcomes when querying a database (Table 1).

We expect that taxa that are well represented in the database, and which have few closely related taxa, will have high rates of true matches. Taxa with many close relatives in the database will have many false matches. Taxa that are poorly represented in the database will have high rates of failed queries. Both of these latter results are in a class usually referred to as false negatives: we falsely infer taxon A is absent. However, they largely arise from different mechanisms. Importantly, as genomic databases become more complete, we expect the fraction of failed queries will decrease. At the same time we expect that the fraction of false matches may increase, as more and more closely related taxa become present in the database. The exact nature of this tradeoff is not well explored. Novel statistical approaches, such as Bayesian re-estimation of species frequencies, may mitigate the problem [21]; however, improved methods are required to address this problem [45].

There are other aspects of classification success that we do not focus on here. The first of these is the notion of a true negative: a sequence that is known to *not* arise from any taxa, should not return a match to any taxa. This is not a biologically realistic situation (all sequences arise from a taxon), although this aspect is useful when trying to assess the performance of different classifiers [46] and presenting the full truth table. The second aspect we do not consider here are false positives: if a read query matches taxon A, but does not arise from taxon A. We would thus falsely interpret taxon A as being present in a community. This metric is intrinsic to the composition of the community rather than just each taxon and the database. For example, if taxon A dominates the community, then it cannot have a high fraction of false positives relative to true positives simply because the vast majority of read queries from the community will be from taxon A and thus true positives. Conversely if taxon B is extremely rare, there will be a large number of false positives relative to true positives, as very few read queries will be from taxon B, resulting in a very small fraction of true positives.

Thus, we use a simplified set of metrics (see Table 1) that are not intrinsically related to community composition: true matches, false matches, and failed queries. We used

Table 1 Description of outcomes for database queries

Description of outcome	Metric	Notation
A read query from taxon A returns a match from taxon A	True match (we correctly infer taxon A is present)	M_{true}
A read query from taxon A returns a match from a taxon that is not A	False match (we infer taxon A is absent due to a secondary match)	M_{false}
A read query from taxon A returns no hit at all	Failed query (we infer taxon A is absent due to database paucity).	M_{fail}

our simulated genomic sequence reads from 80 taxa to quantify these three outcomes at both the genus and family level. To assign genus and family from species, we used the NCBI taxonomy database [47] (which is used by BLAST as the default taxon classifier).

We calculate two ratios from the three metrics in Table 1. The first is the fraction of true positives classified correctly (i.e. recall):

$$\text{Recall} = \mathrm{M_{true}}/(\mathrm{M_{true}} + \mathrm{M_{false}} + \mathrm{M_{fail}})$$

The second is the ratio of true matches to false matches. This simply excludes failed queries from the equation. We term this second metric classification success.

$$\text{Classification Success} = \mathrm{M_{true}}/(\mathrm{M_{true}} + \mathrm{M_{false}})$$

The critical difference between these metrics is that taxa which are poorly represented in the database may nevertheless have high rates of classification success, although recall will necessarily be low. However, as the fraction of failed queries approaches zero (which we expect as genomic databases grow), these two metrics become equivalent.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-3528-4.

Additional file 1: Figure S1. The number of species present in the NCBI RefSeq database has grown roughly exponentially over time. **Figure S2.** Recall at the family level. **Figure S3.** Classification success at the family level. **Table S1.** List of species include in the *in silico* mock community, with associated Kingdom and NCBI.

Abbreviations

BLAST: Basic Local Alignment Search Tool; BP: Basepairs; CCS: Circular Consensus Sequence; LCA: Lowest Common Ancestor; ONT: Oxford Nanopore Technologies; PacBio: Pacific Biosciences

Acknowledgements

Thanks to Paul Gardner for his helpful and insightful comments on the manuscript. Thank you to two anonymous reviewers for their comments on the manuscript.

Authors' contributions

WP, NF, and OS conceived the project, WP and OS simulated and generated the data. WP and OS analysed the data. WP, NF, and OS wrote the paper, and have read and approved the manuscript.

Funding

Some of this work was supported by a Massey University Strategic Research Excellence Fund awarded to NF and a Marsden Fund grant (MAU1703) to OS. The funding bodies had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

Data for this manuscript is available on the OpenScienceFramework: https://osf.io/2gw8d/

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors have no competing interests to declare.

Received: 5 June 2019 Accepted: 30 April 2020
Published online: 29 May 2020

References

1. Ji Y, Ashton L, Pedley SM, Edwards DP, Tang Y, Nakamura A, et al. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Ecol Lett. 2013;16(10):1245–57. https://doi.org/10.1111/ele.12162.

2. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7(5):335–6 Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3156573.
3. Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, et al. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLoS Comput Biol. 2016;12(6):e1004957. https://doi.org/10.1371/journal.pcbi.1004957.
4. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75(23):7537–41. https://doi.org/10.1128/AEM.01541-09.
5. Schloss PD, Handelsman J. Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. Genome Biol. 2005;6(8):229. https://doi.org/10.1186/gb-2005-6-8-229.
6. Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, et al. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol. 2014;12(6):e1001889 Available from: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001889.
7. Roumpeka DD, Wallace RJ, Escalettes F, Fotheringham I, Watson M. A Review of Bioinformatics Tools for Bio-Prospecting from Metagenomic Sequence Data. Front Genet. 2017;8:23. https://doi.org/10.3389/fgene.2017.00023.
8. Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. Microb Inform Exp. 2012;2(1):3. https://doi.org/10.1186/2042-5783-2-3.
9. Temperton B, Giovannoni SJ. Metagenomics: microbial diversity through a scratched lens. Curr Opin Microbiol. 2012;15(5):605–12. https://doi.org/10.1016/j.mib.2012.07.001.
10. Escobar-Zepeda A, Vera-Ponce de León A, Sanchez-Flores A. The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics. Front Genet. 2015;6:348. https://doi.org/10.3389/fgene.2015.00348.
11. Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, et al. Earth BioGenome project: sequencing life for the future of life. Proc Natl Acad Sci U S A. 2018;115(17):4325–33. https://doi.org/10.1073/pnas.1720115115.
12. Teeling EC, Vernes SC, Dávalos LM, Ray DA, Gilbert MTP, Myers E, et al. Bat biology, genomes, and the Bat1K project: to generate chromosome-level genomes for all living bat species. Annu Rev Anim Biosci. 2018;6:23–46. https://doi.org/10.1146/annurev-animal-022516-022811.
13. OBrien SJ, Haussler D, Ryder O. The birds of Genome10K. Gigascience. 2014;3(1):32. https://doi.org/10.1186/2047-217X-3-32.
14. 10K Community of Scientists G. Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. J Hered. 2009; Available from: https://academic.oup.com/jhered/article-abstract/100/6/659/839176.
15. Robinson GE, Hackett KJ, Purcell-Miramontes M, Brown SJ, Evans JD, Goldsmith MR, et al. Creating a buzz about insect genomes. Science. 2011;331(6023):1386. https://doi.org/10.1126/science.331.6023.1386.
16. Pearman W, Smith ANH, Breckell G, Dale J, Freed NE, Silander OK. New tools for diet analyses: nanopore sequencing of metagenomic DNA from stomach contents to quantify diet in an invasive population of rats. bioRxiv. 2018:363622 [Cited 2018 Aug 8]. Available from: https://www.biorxiv.org/content/early/2018/07/06/363622.
17. Gossner MM, Lade P, Rohland A, Sichardt N, Kahl T, Bauhus J, et al. Effects of management on aquatic tree-hole communities in temperate forests are mediated by detritus amount and water chemistry. J Anim Ecol. 2016 Jan;85(1):213–26. https://doi.org/10.1111/1365-2656.12437.
18. Ojeda FP, Santelices B. Invertebrate communities in holdfasts of the kelp macrocystic pyrifera from southern Chile. Mar Ecol Prog Ser Oldendorf. 1984;16(1):65–73 Available from: http://www.int-res.com/articles/meps/16/m016p065.pdf.
19. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15(3):R46. https://doi.org/10.1186/gb-2014-15-3-r46.
20. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 2016;26(12):1721–9. https://doi.org/10.1101/gr.210641.116.
21. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data; 2016. https://doi.org/10.1101/051813.
22. McIntyre ABR, Ounit R, Afshinnekoo E, Prill RJ, Hénaff E, Alexander N, et al. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biol. 2017;18(1):182 Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1299-7.
23. Ounit R, Wanamaker S, Close TJ, Lonardi S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics. 2015;16:236. https://doi.org/10.1186/s12864-015-1419-2.
24. Jiang Y, Wang J, Xia D, Yu G. EnSVMB: Metagenomics Fragments Classification using Ensemble SVM and BLAST. Sci Rep. 2017;7(1):9440. https://doi.org/10.1038/s41598-017-09947-y.
25. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun. 2016;7:11257. https://doi.org/10.1038/ncomms11257.
26. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59–60. https://doi.org/10.1038/nmeth.3176.
27. Frank JA, Pan Y, Tooming-Klunderud A, Eijsink VGH, McHardy AC, Nederbragt AJ, et al. Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. Sci Rep. 2016;6:25373. https://doi.org/10.1038/srep25373.
28. Nicholls SM, Quick JC, Tang S, Loman NJ. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. Gigascience. 2019;8(5). https://doi.org/10.1093/gigascience/giz043.
29. Wick R, Judd LM, Holt KE. Comparison of Oxford Nanopore basecalling tools. 2018. Available from: https://zenodo.org/record/1188469.
30. Ardui S, Ameur A, Vermeesch JR, Hestand MS. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res. 2018;46(5):2159–68. https://doi.org/10.1093/nar/gky066.
31. Brown BL, Watson M, Minot SS, Rivera MC, Franklin RB. MinION™nanopore sequencing of environmental metagenomes: a synthetic approach. Gigascience. 2017;6(3):1–10. https://doi.org/10.1093/gigascience/gix007.

32. Wommack KE, Bhavsar J, Ravel J. Metagenomics: read length matters. Appl Environ Microbiol. 2008;74(5):1453–63. https://doi.org/10.1128/AEM.02181-07.
33. McHardy AC, Martín HG, Tsirigos A, Hugenholtz P, Rigoutsos I. Accurate phylogenetic classification of variable-length DNA fragments. Nat Methods. 2007;4(1):63–72. https://doi.org/10.1038/nmeth976.
34. Stackebrandt E, Goebel BM. Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. Int J Syst Evol Microbiol. 1994;44(4):846–9. [Cited 2018 Jul 7]. Available from. https://doi.org/10.1099/00207713-44-4-846.
35. Konstantinidis KT, Tiedje JM. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A. 2005;102(7):2567–72. https://doi.org/10.1073/pnas.0409727102.
36. Cognato AI. Standard percent DNA sequence difference for insects does not predict species boundaries. J Econ Entomol. 2006;99(4):1037–45 Available from: https://www.ncbi.nlm.nih.gov/pubmed/16937653.
37. Mayr E. Systematics and the origin of species, from the viewpoint of a zoologist: Harvard University Press; 1999. p. 334. Available from: https://market.android.com/details?id=book-mAljnLp6r_MC.
38. Song H, Buhay JE, Whiting MF, Crandall KA. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. Proc Natl Acad Sci U S A. 2008;105(36):13486–91. https://doi.org/10.1073/pnas.0803076105.
39. Lefébure T, Douady CJ, Gouy M, Gibert J. Relationship between morphological taxonomy and molecular divergence within Crustacea: proposal of a molecular threshold to help species delimitation. Mol Phylogenet Evol. 2006;40(2):435–47. https://doi.org/10.1016/j.ympev.2006.03.014.
40. Treangen TJ, Abraham A-L, Touchon M, EPC R. Genesis, effects and fates of repeats in prokaryotic genomes. FEMS Microbiol Rev. 2009;33(3):539–71 Available from: https://www.ncbi.nlm.nih.gov/pubmed/19396957.
41. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Res. 2013; 41(Database issue):D36–42. https://doi.org/10.1093/nar/gks1195.
42. Yang C, Chu J, Warren RL, Birol I. NanoSim: Nanopore sequence read simulator based on statistical characterization. Gigascience. 2017;6(4):1–6. https://doi.org/10.1093/gigascience/gix010.
43. Homer N. DWGSIM. Github; 2017 [cited 2018 Sep 5]. Available from: https://github.com/nh13/DWGSIM.
44. Madden T. The BLAST Sequence Analysis Tool. In: National Center for Biotechnology Information (US); 2013. [Cited 2018 Sep 5]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK153387/.
45. Nasko DJ, Koren S, Phillippy AM, Treangen TJ. RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. Genome Biol. 2018;19(1):165. https://doi.org/10.1186/s13059-018-1554-6.
46. Lindgreen S, Adair KL, Gardner PP. An evaluation of the accuracy and speed of metagenome analysis tools. Sci Rep. 2016;6:1–14. https://doi.org/10.1038/srep19233.
47. Federhen S. The NCBI taxonomy database. Nucleic Acids Res. 2012;40(Database issue):D136–43. https://doi.org/10.1093/nar/gkr1178.

Publisher's Note