The advantages and disadvantages of short- and long-read metagenomics to infer bacterial and eukaryotic community composition

William Pearman[1], Nikki Freed[1], Olin Silander[1]

1 School of Natural and Computational Sciences, Massey University, Auckland, New Zealand

Corresponding Authors:

William Pearman

Olin Silander

Private Bag 102904, North Shore, 0745 Auckland, New Zealand

Email addresses: wpearman1996@gmail.com; olinsilander@gmail.com

Abstract

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. Here we use simulated error prone Oxford Nanopore and 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 error-prone reads can significantly increase classification accuracy, and this is most pronounced for non-microbial communities. 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.

Introduction

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 (Ji et al. 2013). 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) (Caporaso et al. 2010; Huson et al. 2016; Schloss et al. 2009). 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. Not only do metagenomic methods profile taxon composition in a less biased way than metabarcoding, but they can also yield insight into the functional diversity present in ecosystems (Schloss and Handelsman 2005; Keeling et al. 2014)

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 (Roumpeka et al. 2017; Thomas, Gilbert, and Meyer 2012; Temperton and Giovannoni 2012). 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 (Escobar-Zepeda, Vera-Ponce de León, and Sanchez-Flores 2015) (Supp. Fig.1). This latter factor is perhaps the single largest factor in driving the bias toward microbial metagenomics.

However, the amount and diversity of eukaryotic 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 (Lewin et al. 2018). Regardless of the success of this effort, there are a host of ongoing eukaryotic sequencing projects, including Bat 1K (100 bat genomes (Teeling et al. 2018)), Bird 10K (10,000 bird genomes (OBrien, Haussler, and Ryder 2014)), G10K (10,000 vertebrate genomes (10K Community of Scientists 2009)), and i5K (5000 arthropod genomes (Robinson et al. 2011)), among others. This suggests that within the next five 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.

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 (Wood and Salzberg 2014; Kim et al. 2016; Huson et al. 2016), or by refining taxonomic representation by examining the results from the first pass classifier (Lu et al. 2016).

The most widely used first pass classifier is BLAST, and it is considered gold standard (McIntyre et al. 2017). 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 (Ounit et al. 2015), Kraken and related tools (Kraken, Kraken2, and KrakenUniq) (Wood and Salzberg 2014), Centrifuge (Kim et al. 2016), EnSVMB (Jiang et al. 2017), and Kaiju (Menzel, Ng, and Krogh 2016), as well as reduced alphabet amino acid based approaches such as DIAMOND (Buchfink, Xie, and Huson 2015). In almost all cases these have been designed and benchmarked using short read data (McIntyre et al. 2017).

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 (Frank et al. 2016; Nicholls et al. 2019). 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 (Wick, Judd, and Holt 2018) and 85-87% for PacBio (Ardui et al. 2018)). 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 (Brown et al. 2017). 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.

Methods

Genomic data

For each of four major taxonomic divisions (bacteria, fungi, animals, and plants), we downloaded 20 genomes from GenBank (Benson et al. 2013). Within each of these divisions, we included genomes from a total of 22 classes, 46 orders, and 58 families (**Figure 1**).

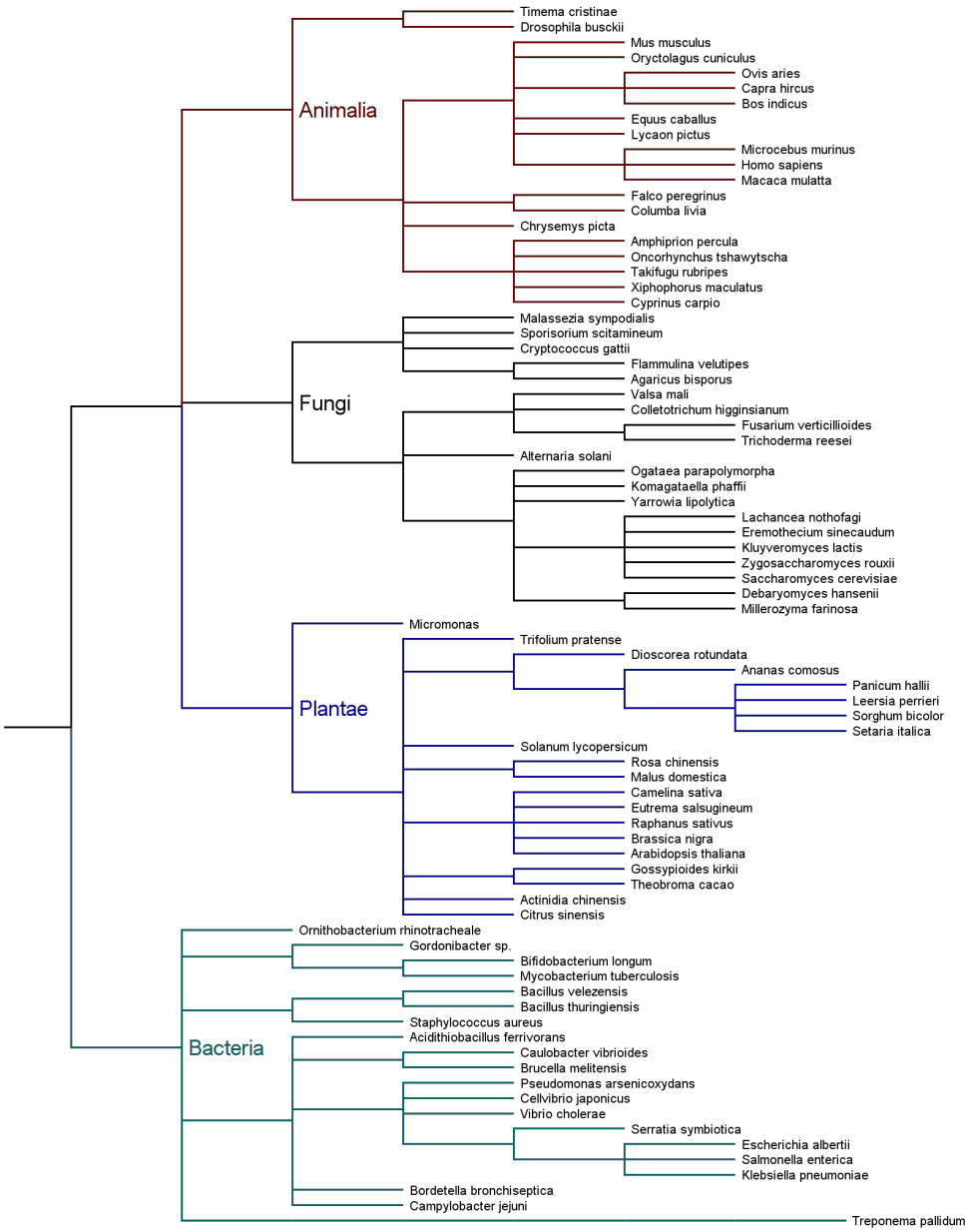


Figure 1. Cladogram of species included in the simulated mock community based on the NCBI taxonomy. We included taxa from four major divisions (bacteria, fungi, animals, and plants).

Read simulation

We simulated Nanopore reads using NanoSim 2.0.0 (Yang et al. 2017) 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 perl script (i.e. to simulate 100bp Nanopore reads, we truncated all reads in a simulated dataset to 100bp). We did this for read lengths varying from 100 bp to 4,000 bp at 100 bp intervals, simulating 1,000 reads per interval for all taxa (a total of 40,000 reads for each taxon, and 3.2 million reads for all taxa and read lengths).

We simulated Illumina data using dwgsim 0.1.12 (Homer 2017) 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 1,000 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 (Madden 2013) and Kraken2 (Wood and Salzberg 2014) 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, 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.

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**).

150 **Table 1**. Description, metrics, and notations of the results of a database query.

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}

151

152

153 We expect that taxa that are well represented in the database, and which have few closely
154 related taxa, will have high rates of true matches. Taxa with many close relatives in the
155 database will have many false matches. Taxa that are poorly represented in the database
156 will have high rates of failed queries. Both of these latter results are in a class usually
157 referred to as false negatives: we falsely infer taxon A is absent. However, they largely arise
158 from different mechanisms. Importantly, as genomic databases become more complete, we
159 expect the fraction of failed queries will decrease. At the same time we expect that the
160 fraction of false matches may increase, as more and more closely related taxa become
161 present in the database. The exact nature of this tradeoff is not well explored. Novel
162 statistical approaches, such as Bayesian re-estimation of species frequencies, may mitigate
163 the problem (Lu et al. 2016); however, improved methods are required address this problem
164 (Nasko et al. 2018).

165 There are other aspects of classification success that we do not focus on here. The first of
166 these is the notion of a true negative: a sequence that is known to *not* arise from any taxa,
167 should not return a match to any taxa. This is not a biologically realistic situation (all
168 sequences arise from a taxon), although this aspect is useful when trying to assess the
169 performance of different classifiers (Lindgreen et al. 2016) and presenting the full truth table.
170 The second aspect we do not consider here are false positives: if a read query matches
171 taxon A, but does not arise from taxon A. We would thus falsely interpret taxon A as being
172 present in a community. This metric is intrinsic to the composition of the community rather
173 than just each taxon and the database. For example, if taxon A dominates the community,
174 then it cannot have high rates of false positives relative to true positives simply because the
175 vast majority of read queries from the community will be from taxon A and thus true
176 positives. Conversely if taxon B is extremely rare, there will be a large number of false

177 positives relative to true positives, as very few read queries will be from taxon B, resulting in
178 a very small number of true positives.

179 Thus, we use a simplified set of metrics (see **Table 1**) that are not intrinsically related to
180 community composition: true matches, false matches, and failed queries. We used our
181 simulated genomic sequence reads from 80 taxa to quantify these three outcomes at both
182 the genus and family level. To assign genus and family from species, we used the NCBI
183 taxonomy database (Federhen 2012) (which is used by BLAST as the default taxon
184 classifier).

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

187 Recall = $M_{true}/(M_{true} + M_{false} + M_{fail})$

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

190 Classification Success = $M_{true}/(M_{true} + M_{false})$

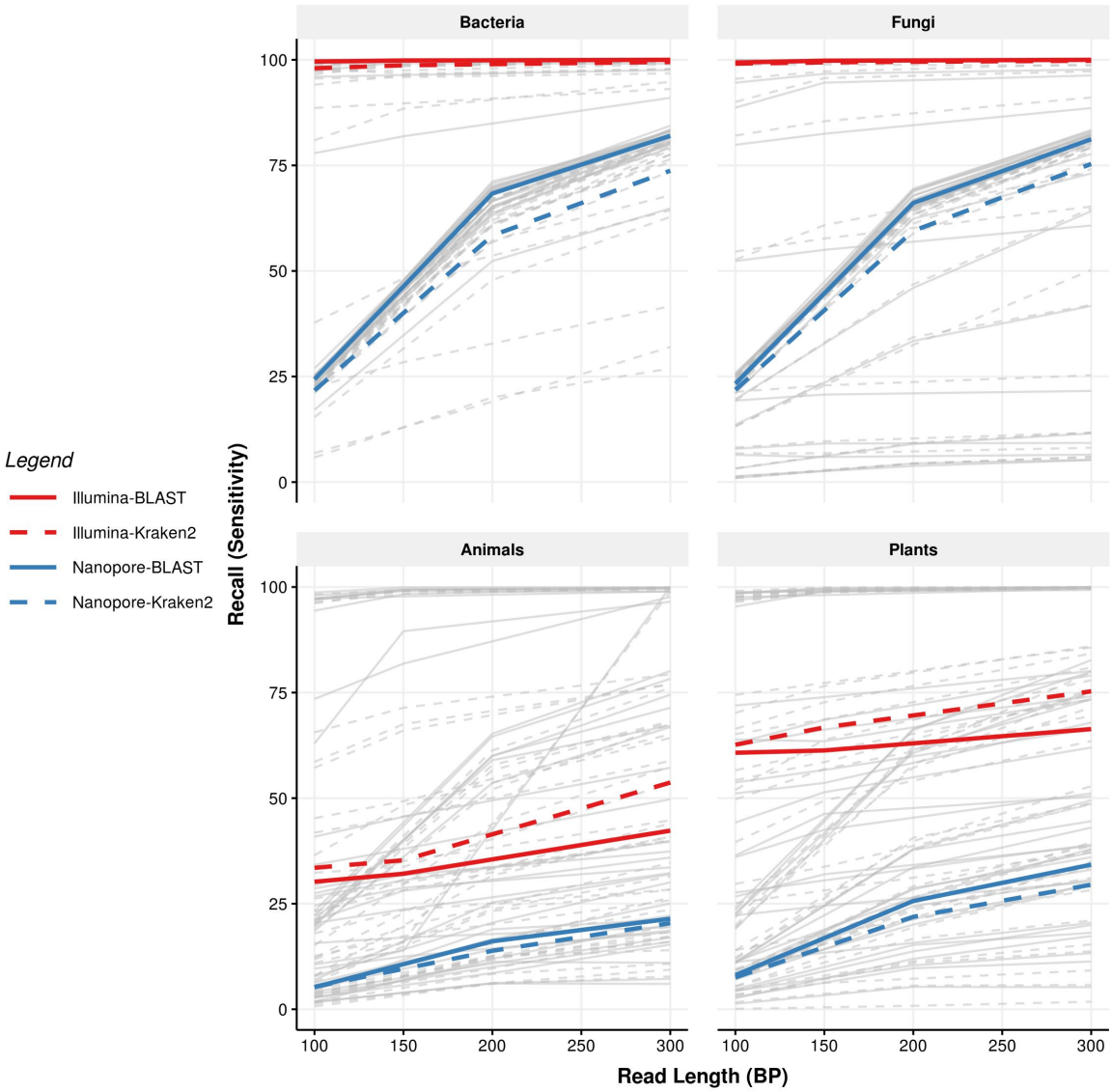
191 The critical difference between these metrics is that taxa which are poorly represented in
192 the database may nevertheless have high rates of classification success, although recall will
193 necessarily be low. However, as the fraction of failed queries approaches zero (which we
194 expect as genomic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

195 Results

196 We first looked only at short read lengths to quantify the effects of sequencing technology
197 and classifier (BLAST or Kraken2) on recall at the level of genus. For both bacteria and
198 fungi, we found that recall was at or above 99.9% for Illumina reads of any length (100bp,
199 150bp, or 300bp), for both BLAST and Kraken2 (**Fig. 2**). In strong contrast, for Nanopore
200 data, recall was far lower; approximately 25% for 100bp reads and increasing to 75% at
201 300bp. In general, Kraken2 had slightly lower recall than BLAST.

202 However, for plants and animals, average recall was low regardless of sequencing
203 technology. Average recall for Illumina reads peaked at approximately 55% and 75% for
204 animals and plants, respectively (**Fig 2,** light blue lines). Nanopore recall rates peaked at
205 just over 20% and 35% for animals and plants, respectively. However, this was highly
206 taxon-dependent, with some taxa consistently having recall near 100%, while others
207 remained close to 0% regardless of sequencing technology or read length (**Fig. 2,** grey

208 lines). Perhaps surprisingly, on average Kraken2 outperformed BLAST for Illumina reads for
209 both plant and animal taxa.



210

211 **Figure 2. Recall is consistently higher in bacteria and fungi than plants or animals for**
212 **both short Illumina and Nanopore reads.** Each panel shows recall for the different
213 kingdoms. Recall for individual taxa is indicated in grey, with median recall shown by
214 dashed (Nanopore) or solid (Illumina). Blue lines indicate recall rates for reads classified
215 using Kraken2; red for reads classified using BLAST. Illumina reads exhibit consistently
216 higher recall; bacteria and fungi exhibit higher recall than plants or animals.

217

218 We next quantified differences in classification success (the proportion of all classified
219 reads that were correctly classified), again considering only short read lengths. For bacteria
220 and fungi, both Illumina and Nanopore reads exhibited high classification success, with the
221 exception of Kraken2 classification of Nanopore reads (**Fig. 3**). 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. However, Kraken2 success was far lower, especially for Nanopore reads, peaking at 54% in animals (**Fig. 3**). 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.

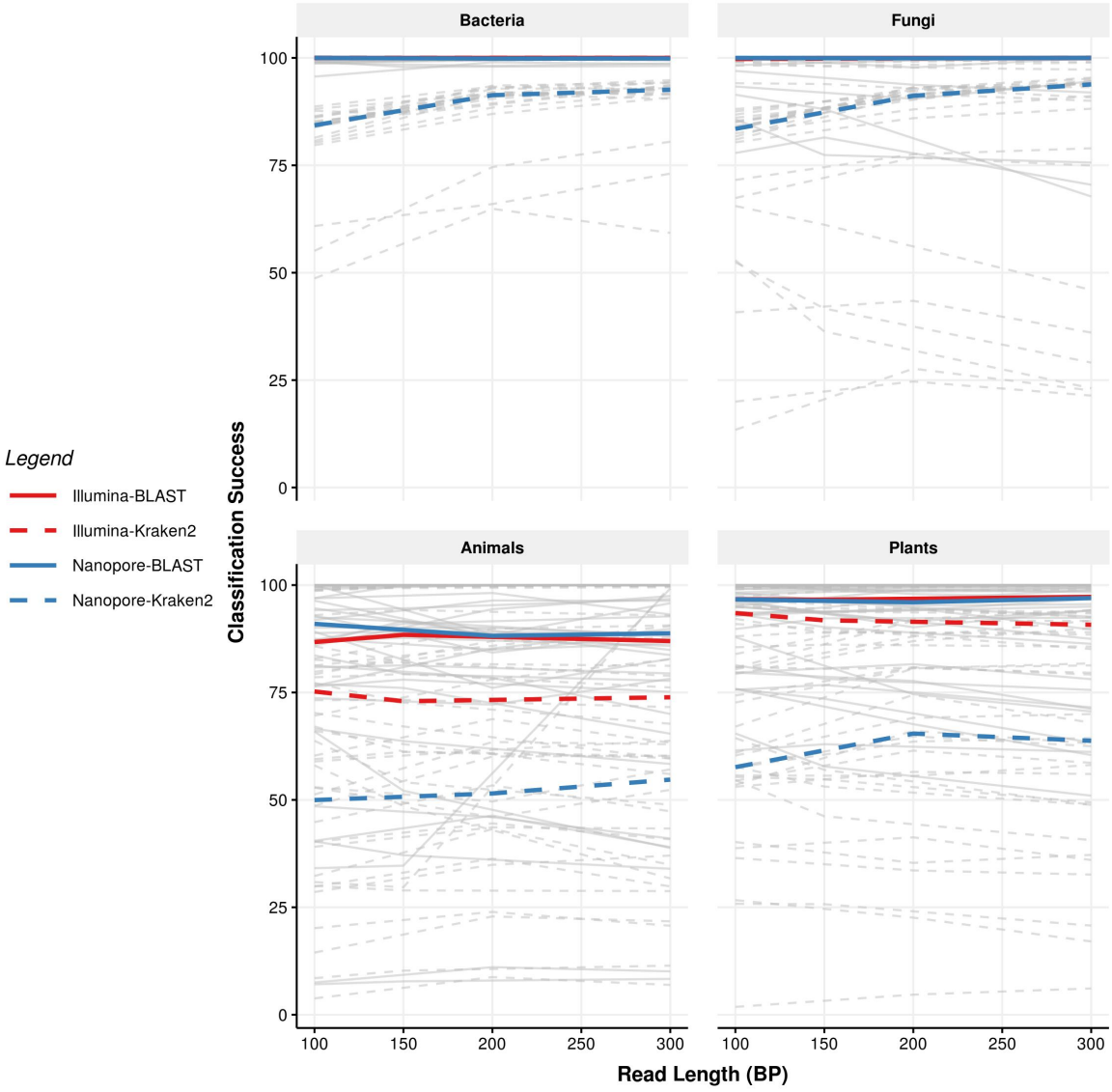
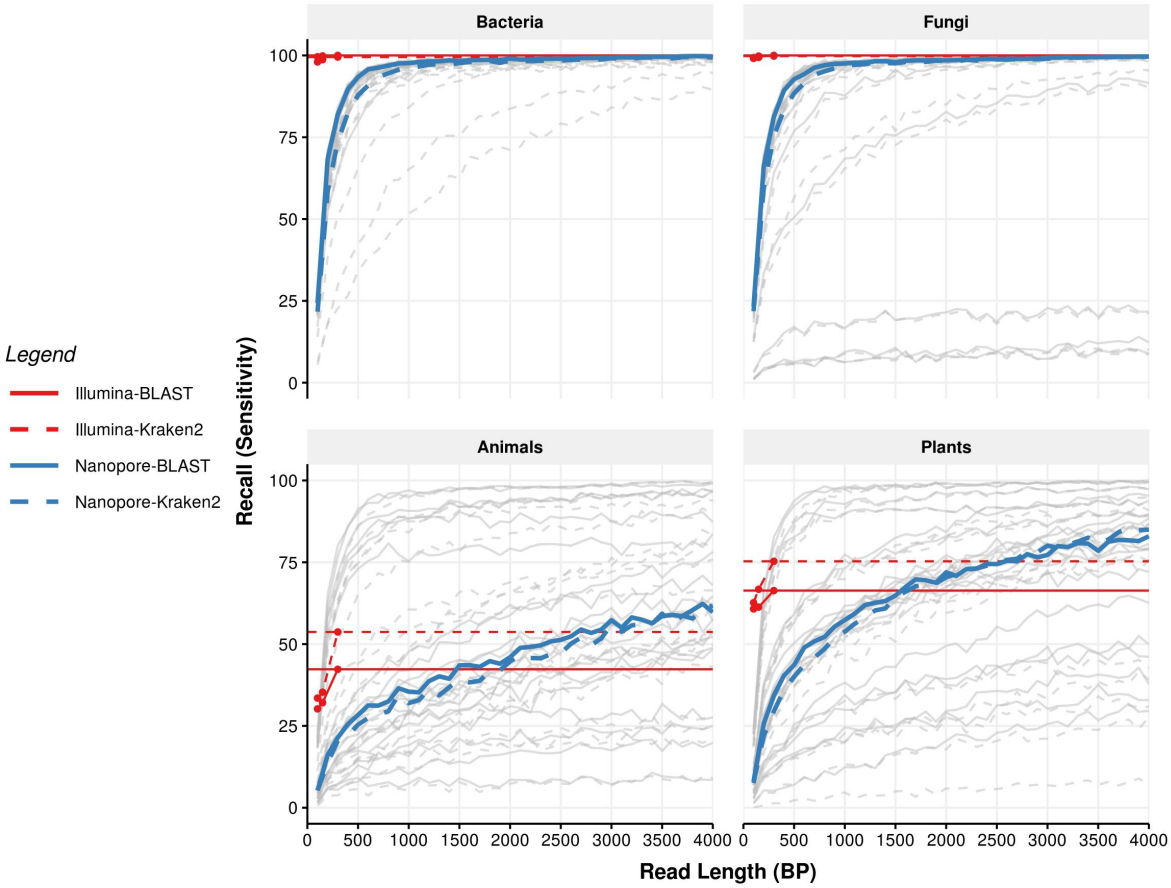


Figure 3. Classification success for short reads is weakly related to read length and strongly dependent on classification method. Each panel shows recall for the different kingdoms. Classification success for individual taxa is indicated in grey, with median classification success shown by solid lines (Illumina) or dashed lines (Nanopore). Blue lines indicate recall rates for reads classified using Kraken2; red for reads classified using BLAST. For bacteria and fungi, median classification rates of Illumina-BLAST, Illumina-Kraken2, and Nanopore-BLAST are almost exactly 100% for all read lengths.

236 It is perhaps expected that highly accurate Illumina reads would result in more accurate
237 taxonomic classification than long error-prone Nanopore reads. However, it is possible to
238 obtain Nanopore reads far in excess of 300bp (reads up to 2 megabase pairs have been
239 sequenced), so we next quantified recall and classification success for reads with lengths
240 up to 4,000 bp. Because such read lengths are not currently possible to obtain using
241 Illumina technology, we did not measure recall and classification success for Illumina reads
242 of similar lengths.



243

244 **Figure 4. Long Nanopore reads surpass the recall of the longest Illumina reads for**
245 **both BLAST and Kraken2.** Each panel shows recall for the different kingdoms. Recall for
246 Nanopore reads for individual taxa is indicated in grey, with median recall indicated by
247 dashed lines, either blue (Kraken2) or red (BLAST). The recall rates for 300 bp Illumina reads
248 are shown as thin solid lines, again either blue (Kraken2) or red (BLAST). Coloured points
249 show the recall for all Illumina reads of all lengths (100 bp, 150 bp, and 300 bp).

250

251 We observed similar relationship between read length and recall for both BLAST and
252 Kraken2. For bacteria and fungi, recall increased from ~20% using 100 bp reads to almost
253 100% when using 1500 bp reads. For animals and plants we observed similar trends,
254 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 Kraken, with crossover points at approximately 3000 bp for animals and 2500 bp for plants **(Fig. 4**, red and blue solid lines).

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 Fig. 2**). However, for plants Nanopore recall surpassed Illumina recall at 2500 bp, with 4000 bp reads yielding a recall of approximately 90%. We found again that for both animals and plants, Kraken2 recall surpassed BLAST when relying on Illumina reads.

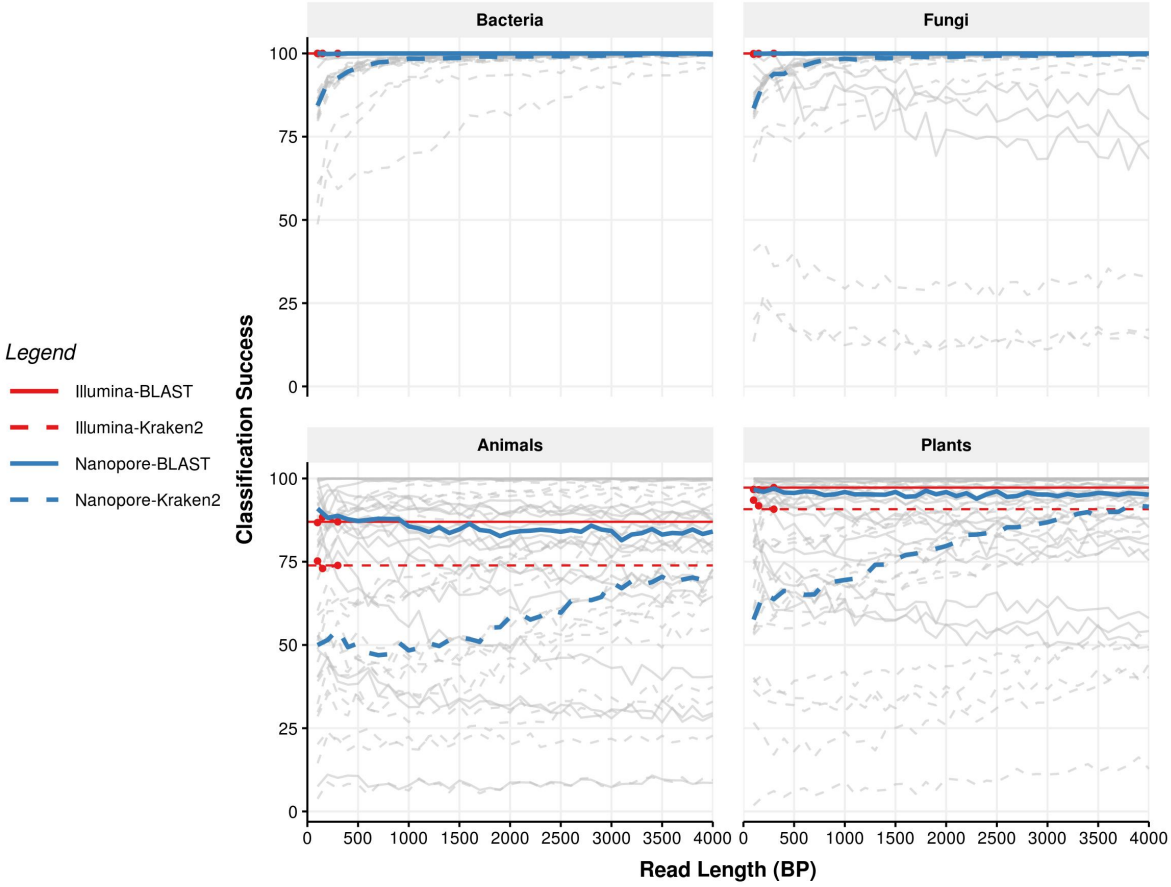


Figure 5. Classification success for long reads is dependent on read length only for Kraken2 classification. Each panel shows classification success for the different kingdoms. Classification success for individual taxa is indicated in grey, with median classification success shown by solid lines (Illumina) or dashed lines (Nanopore). Blue indicates classification success rates for reads classified using Kraken2, while red indicates those classified using BLAST. For animal and plants, the classification success of Kraken2 depends strongly on read length, and never surpasses BLAST or Illumina at any length.

273 We next examined classification success at longer read lengths. For BLAST we observed
274 no relationship between classification success and read length for any taxon (**Fig 5**.).
275 Bacteria and fungi both had consistently high classification success (median 100%), while
276 animals and plants had lower classification success (median 82% and 96%, respectively).
277 However, for Kraken2 we observed a consistent increase in classification success as read
278 length increased. However, this never exceeded the classification success we observed for
279 BLAST, nor did it succeed the classification success we observed for short accurate
280 Illumina reads.

281 Finally, we tested classification success at the level of Family. In this case, we observed
282 that for BLAST, the classification success for plants was approximately 99% overall read
283 lengths, while for Kraken2 only 4000 bp reads reached this level. For animals, BLAST
284 classification success was approximately 95% over all read lengths, but for Kraken2
285 reached a maximum of 85% only at the longest read lengths (**Supp. Fig. 3**).

286 Discussion

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

293 Recall for both BLAST and Kraken2 was improved by the use of long reads, especially in
294 the case of animals and plants, for which recall improved almost three-fold as read length
295 increased from 300 bp to 4,000 bp. Generally both Kraken2 and BLAST achieved similar
296 levels of recall. The exception was for short reads for animals and plants, for which Kraken2
297 was more accurate than BLAST.

298 We found no relationship between classification success and read length for BLAST. This
299 implies that the ratio of correctly classified reads to all classified reads remains relatively
300 constant over different read lengths. However, the number of reads that are classified *at all*
301 increases with read length (causing an increase in recall). These observations are in line
302 with what has been observed by others (McHardy et al. 2007). The exception to this lack of
303 relationship between classification success and read length was for Kraken2, for which the
304 proportion of correctly classified reads increases with read length by more than 50% for
305 both plants and animals.

306 Our results also indicate that recall for long Nanopore reads was as high or higher than
307 short Illumina reads. This was true regardless of kingdom, or classification method, with
308 Nanopore surpassing 300 bp Illumina reads at approximately 1500 bp (**Fig. 4**). These
309 results do suggest that one approach to improve classification accuracy is to impose
310 minimum read lengths. This can be achieved by performing size selection during library
311 preparation or during computational analyses.

312 At first glance, then, there appears to be a clear trade-off between short read Illumina and
313 long read Nanopore sequencing for metagenomic analyses. While Nanopore allows higher
314 recall at long read lengths, this advantage is offset by the fact that Illlumina generally
315 provides more reads per run. At most, recall for Nanopore improves 50% beyond 300 bp
316 Illumina reads, while classification success is similar (using BLAST). Thus, if the read
317 capacity of Illumina runs is 50% or more than Nanopore, the number of classified reads will
318 be maximised using Illumina technology - on a per sequencing run basis. However, for
319 many researchers the more relevant metric is cost per read. In this case, MinION read
320 yields are approximately equal to MiSeq, and only HiSeq or NovaSeq provides a clear cost
321 advantage over Nanopore MinION. On the other hand, cost per read for PromethION are
322 not far from NovaSeq. Thus, we find no clear advantage in using Illumina over Nanopore
323 given the observed classification accuracy for long inaccurate Nanopore reads.

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

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

330 Bacterial taxa are often considered separate species once they have diverged by 6% ANI
331 (Average Nucleotide Identity) on a genomic level (Stackebrandt and Goebel 1994;
332 Konstantinidis and Tiedje 2005). The degree of nucleotide divergence between eukaryotic
333 species is not standardised (Cognato 2006), and species are generally designated as such
334 based on the biological species concept put forward by Mayr (Mayr 1999). Although
335 divergence levels differ substantially between loci (as for bacteria), for some loci general
336 ranges for eukaryotic species have emerged. For example, for mitochondrial COI, between-
337 species divergence is usually greater than 3% (Song et al. 2008; Lefébure et al. 2006).
338 These loci are among the fastest diverging loci in plant and animal genomes, and many

339 other loci may differ by far less than 1% between species. Due to this low level of
340 divergence, metagenomic classifiers may frequently classify animal and plant genera with
341 lower accuracy than bacterial genera.

342 A second explanation for the increased classification success in bacteria and fungi is that
343 these genomes contain fewer repetitive elements than animals or plants (Treangen et al.
344 2009). Although such repetitive regions are usually masked from classifiers (including
345 BLAST and Kraken2), this masking may not be complete.

346 A third reason is that the genomic databases for plants and animals are far less complete
347 than for bacteria and fungi. There is a large difference in the number of genomes and
348 sequences available for different Kingdoms, with bacteria having significantly more species
349 present than the next closest kingdom (See Supp. Fig.1). However, we expect this factor
350 will be mitigated in the future as genomic databases continue to expand and computational
351 search methods continue to improve.

352 Differences in accuracy between Kraken2 and BLAST

353 We observed similar levels of recall for BLAST and Kraken2 over most reads lengths.
354 However, there were strong differences in classification success. For short reads, Kraken2
355 classification success was far lower than BLAST. As read lengths increased, Kraken2
356 classification success approached BLAST. Part of this is likely due to longer reads allowing
357 multiple k-mer matches, decreasing the probability of a false positive classification. One
358 perhaps underappreciated advantage of Kraken2 over BLAST is that Kraken2 has reduced
359 sensitivity to structural variation within reads. As Kraken2 allows multiple k-mers to match
360 within a read, structural changes (e.g. inversions) are less likely to influence the outcome of
361 Kraken2 matching. Such structural changes may influence BLAST due to the matching and
362 extend algorithm. Thus for long reads, classifiers that are insensitive to synteny may be
363 more successful, especially for taxa in which structural rearrangements are common.

364 Conclusion

365 Here we have shown despite being error-prone, Nanopore reads are useful for
366 metagenomic classification due to their increased length, and that for plant and animal
367 communities, the classification accuracy of long Nanopore reads exceeds that of Illumina.
368 We found that classification accuracy is more dependent on the set of taxa being
369 considered than on the metagenomic classifier being used (Kraken2 or BLAST), and that
370 this was true for both short accurate (Illumina) and long error-prone (Nanopore) 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.

Acknowledgments

Some of this work was supported by a Massey University Strategic Research Excellence Fund awarded to NF.

References

10K Community of Scientists, Genome. 2009. “Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species.” *The Journal of Heredity*. https://academic.oup.com/jhered/article-abstract/100/6/659/839176.

Ardui, Simon, Adam Ameur, Joris R. Vermeesch, and Matthew S. Hestand. 2018. “Single Molecule Real-Time (SMRT) Sequencing Comes of Age: Applications and Utilities for Medical Diagnostics.” *Nucleic Acids Research* 46 (5): 2159–68.

Benson, Dennis A., Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. 2013. “GenBank.” *Nucleic Acids Research* 41 (Database issue): D36–42.

Brown, Bonnie L., Mick Watson, Samuel S. Minot, Maria C. Rivera, and Rima B. Franklin. 2017. “MinIONTMnanopore Sequencing of Environmental Metagenomes: A Synthetic Approach.” *GigaScience* 6 (3): 1–10.

Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2015. “Fast and Sensitive Protein Alignment Using DIAMOND.” *Nature Methods* 12 (1): 59–60.

Caporaso, J. Gregory, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D. Bushman, Elizabeth K. Costello, Noah Fierer, et al. 2010. “QIIME Allows Analysis of High-Throughput Community Sequencing Data.” *Nature Methods* 7 (5): 335–36.

Cognato, Anthony I. 2006. “Standard Percent DNA Sequence Difference for Insects Does Not Predict Species Boundaries.” *Journal of Economic Entomology* 99 (4): 1037–45.

Escobar-Zepeda, Alejandra, Arturo Vera-Ponce de León, and Alejandro Sanchez-Flores. 2015. “The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics.” *Frontiers in Genetics* 6 (December): 348.

Federhen, Scott. 2012. “The NCBI Taxonomy Database.” *Nucleic Acids Research* 40

(Database issue): D136–43.

Frank, J. A., Y. Pan, A. Tooming-Klunderud, V. G. H. Eijsink, A. C. McHardy, A. J. Nederbragt, and P. B. Pope. 2016. “Improved Metagenome Assemblies and Taxonomic Binning Using Long-Read Circular Consensus Sequence Data.” *Scientific Reports* 6 (May): 25373.

Homer, Nils. 2017. *DWGSIM* (version 0.1.12). Github. https://github.com/nh13/DWGSIM.

Huson, Daniel H., Sina Beier, Isabell Flade, Anna Górska, Mohamed El-Hadidi, Suparna Mitra, Hans-Joachim Ruscheweyh, and Rewati Tappu. 2016. “MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data.” *PLoS Computational Biology* 12 (6): e1004957.

Jiang, Yuan, Jun Wang, Dawen Xia, and Guoxian Yu. 2017. “EnSVMB: Metagenomics Fragments Classification Using Ensemble SVM and BLAST.” *Scientific Reports* 7 (1): 9440.

Ji, Yinqiu, Louise Ashton, Scott M. Pedley, David P. Edwards, Yong Tang, Akihiro Nakamura, Roger Kitching, et al. 2013. “Reliable, Verifiable and Efficient Monitoring of Biodiversity via Metabarcoding.” *Ecology Letters* 16 (10): 1245–57.

Keeling, Patrick J., Fabien Burki, Heather M. Wilcox, Bassem Allam, Eric E. Allen, Linda A. Amaral-Zettler, E. Virginia Armbrust, et al. 2014. “The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing.” *PLoS Biology* 12 (6): e1001889.

Kim, Daehwan, Li Song, Florian P. Breitwieser, and Steven L. Salzberg. 2016. “Centrifuge: Rapid and Sensitive Classification of Metagenomic Sequences.” *Genome Research* 26 (12): 1721–29.

Konstantinidis, Konstantinos T., and James M. Tiedje. 2005. “Genomic Insights That Advance the Species Definition for Prokaryotes.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (7): 2567–72.

Lefébure, T., C. J. Douady, M. Gouy, and J. Gibert. 2006. “Relationship between Morphological Taxonomy and Molecular Divergence within Crustacea: Proposal of a Molecular Threshold to Help Species Delimitation.” *Molecular Phylogenetics and Evolution* 40 (2): 435–47.

Lewin, Harris A., Gene E. Robinson, W. John Kress, William J. Baker, Jonathan Coddington,

Keith A. Crandall, Richard Durbin, et al. 2018. “Earth BioGenome Project: Sequencing Life for the Future of Life.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (17): 4325–33.

Lindgreen, Stinus, Karen L. Adair, and Paul Gardner. 2016. “An evaluation of the accuracy and speed of metagenome analysis tools.” *Scientific Reports* 6 (19): doi:10.1038/srep19233.

Lu, Jennifer, Florian P. Breitwieser, Peter Thielen, and Steven L. Salzberg. 2016. “Bracken: Estimating Species Abundance in Metagenomics Data.” https://doi.org/10.1101/051813.

Madden, Thomas. 2013. *The BLAST Sequence Analysis Tool*. National Center for Biotechnology Information (US).

Mayr, Ernst. 1999. *Systematics and the Origin of Species, from the Viewpoint of a Zoologist*. Harvard University Press.

McHardy, Alice Carolyn, Héctor García Martín, Aristotelis Tsirigos, Philip Hugenholtz, and Isidore Rigoutsos. 2007. “Accurate Phylogenetic Classification of Variable-Length DNA Fragments.” *Nature Methods* 4 (1): 63–72.

McIntyre, Alexa B. R., Rachid Ounit, Ebrahim Afshinnekoo, Robert J. Prill, Elizabeth Hénaff, Noah Alexander, Samuel S. Minot, et al. 2017. “Comprehensive Benchmarking and Ensemble Approaches for Metagenomic Classifiers.” *Genome Biology* 18 (1): 182.

Menzel, Peter, Kim Lee Ng, and Anders Krogh. 2016. “Fast and Sensitive Taxonomic Classification for Metagenomics with Kaiju.” *Nature Communications* 7 (April): 11257.

Nasko, Daniel J., Sergey Koren, Adam M. Phillippy, and Todd J. Treangen. 2018. “RefSeq Database Growth Influences the Accuracy of K-Mer-Based Lowest Common Ancestor Species Identification.” *Genome Biology* 19 (1): 165.

Nicholls, Samuel M., Joshua C. Quick, Shuiquan Tang, and Nicholas J. Loman. 2019. “Ultra-Deep, Long-Read Nanopore Sequencing of Mock Microbial Community Standards.” *GigaScience* 8 (5). https://doi.org/10.1093/gigascience/giz043.

OBrien, Stephen J., David Haussler, and Oliver Ryder. 2014. “The Birds of Genome10K.” *GigaScience* 3 (1): 32.

O’Leary, Nuala A., Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, et al. 2016. “Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation.” *Nucleic Acids*

464 *Research* 44 (D1): D733–45.

465 Ounit, Rachid, Steve Wanamaker, Timothy J. Close, and Stefano Lonardi. 2015. "CLARK:
466 Fast and Accurate Classification of Metagenomic and Genomic Sequences Using
467 Discriminative K-Mers." *BMC Genomics* 16 (March): 236.

468 Robinson, Gene E., Kevin J. Hackett, Mary Purcell-Miramontes, Susan J. Brown, Jay D.
469 Evans, Marian R. Goldsmith, Daniel Lawson, Jack Okamuro, Hugh M. Robertson, and
470 David J. Schneider. 2011. "Creating a Buzz about Insect Genomes." *Science* 331
471 (6023): 1386.

472 Roumpeka, Despoina D., R. John Wallace, Frank Escalettes, Ian Fotheringham, and Mick
473 Watson. 2017. "A Review of Bioinformatics Tools for Bio-Prospecting from
474 Metagenomic Sequence Data." *Frontiers in Genetics* 8 (March): 23.

475 Schloss, Patrick D., and Jo Handelsman. 2005. "Metagenomics for Studying Unculturable
476 Microorganisms: Cutting the Gordian Knot." *Genome Biology* 6 (8): 229.

477 Schloss, Patrick D., Sarah L. Westcott, Thomas Ryabin, Justine R. Hall, Martin Hartmann,
478 Emily B. Hollister, Ryan A. Lesniewski, et al. 2009. "Introducing Mothur: Open-Source,
479 Platform-Independent, Community-Supported Software for Describing and Comparing
480 Microbial Communities." *Applied and Environmental Microbiology* 75 (23): 7537–41.

481 Song, Hojun, Jennifer E. Buhay, Michael F. Whiting, and Keith A. Crandall. 2008. "Many
482 Species in One: DNA Barcoding Overestimates the Number of Species When Nuclear
483 Mitochondrial Pseudogenes Are Coamplified." *Proceedings of the National Academy of*
484 *Sciences of the United States of America* 105 (36): 13486–91.

485 Stackebrandt, E., and B. M. Goebel. 1994. "Taxonomic Note: A Place for DNA-DNA
486 Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in
487 Bacteriology." *International Journal of Systematic and Evolutionary Microbiology* 44 (4):
488 846–49.

489 Teeling, Emma C., Sonja C. Vernes, Liliana M. Dávalos, David A. Ray, M. Thomas P. Gilbert,
490 Eugene Myers, and Bat1K Consortium. 2018. "Bat Biology, Genomes, and the Bat1K
491 Project: To Generate Chromosome-Level Genomes for All Living Bat Species." *Annual*
492 *Review of Animal Biosciences* 6 (February): 23–46.

493 Temperton, Ben, and Stephen J. Giovannoni. 2012. "Metagenomics: Microbial Diversity
494 through a Scratched Lens." *Current Opinion in Microbiology* 15 (5): 605–12.

495 Thomas, Torsten, Jack Gilbert, and Folker Meyer. 2012. "Metagenomics - a Guide from

496 Sampling to Data Analysis." *Microbial Informatics and Experimentation* 2 (1): 3.

497 Treangen, Todd J., Anne-Laure Abraham, Marie Touchon, and Eduardo P. C. Rocha. 2009.
498 "Genesis, Effects and Fates of Repeats in Prokaryotic Genomes." *FEMS Microbiology*
499 *Reviews* 33 (3): 539–71.

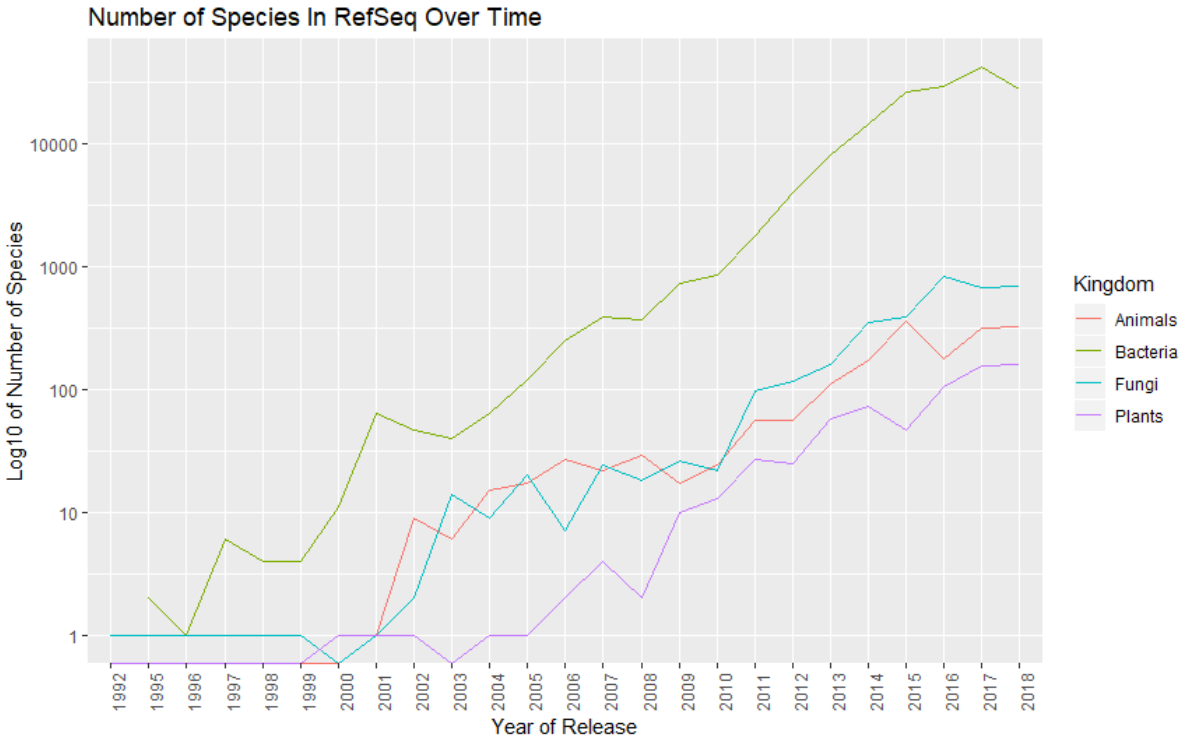
500 Wick, Ryan, Louise M. Judd, and Kathryn E. Holt. 2018. *Comparison of Oxford Nanopore*
501 *Basecalling Tools*. https://doi.org/10.5281/zenodo.1188469.

502 Wood, Derrick E., and Steven L. Salzberg. 2014. "Kraken: Ultrafast Metagenomic Sequence
503 Classification Using Exact Alignments." *Genome Biology* 15 (3): R46.

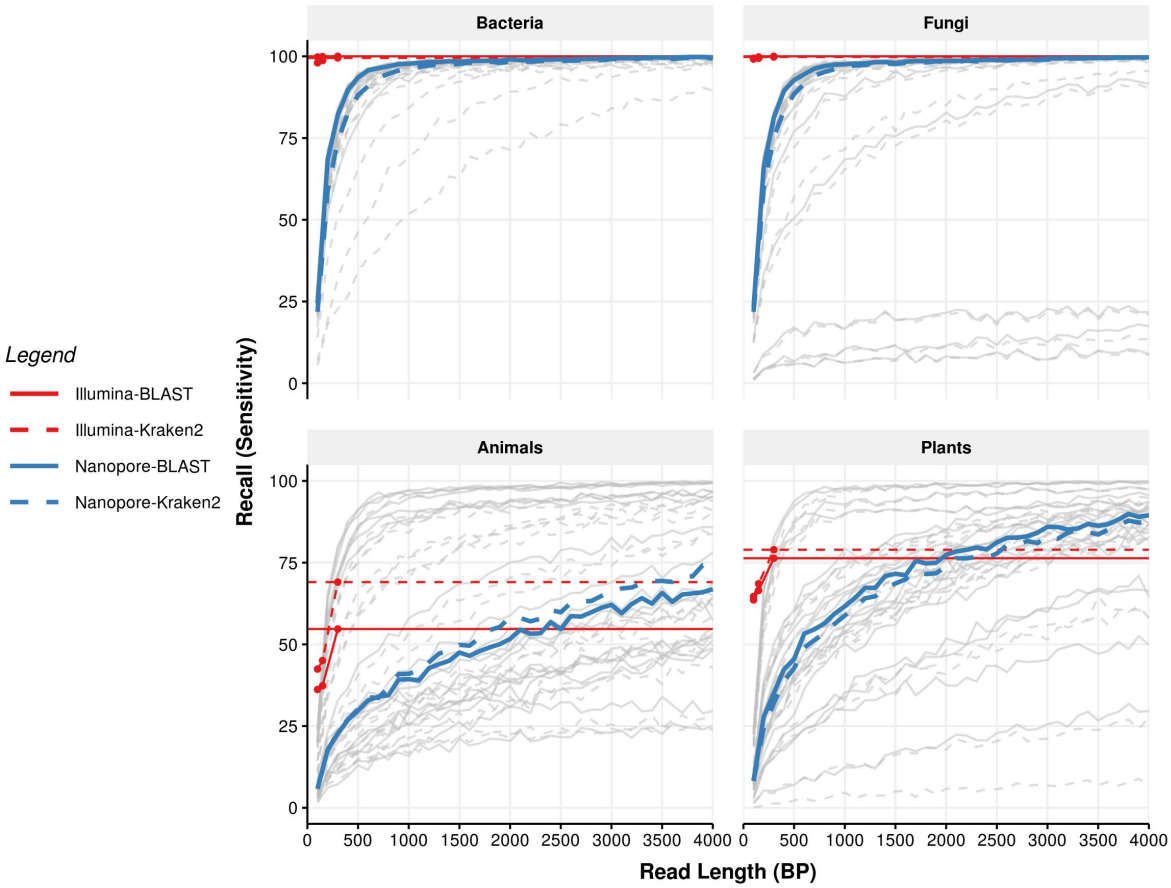
504 Yang, Chen, Justin Chu, René L. Warren, and Inanç Birol. 2017. "NanoSim: Nanopore
505 Sequence Read Simulator Based on Statistical Characterization." *GigaScience* 6 (4): 1–
506 6.

507

Supplementary Figures



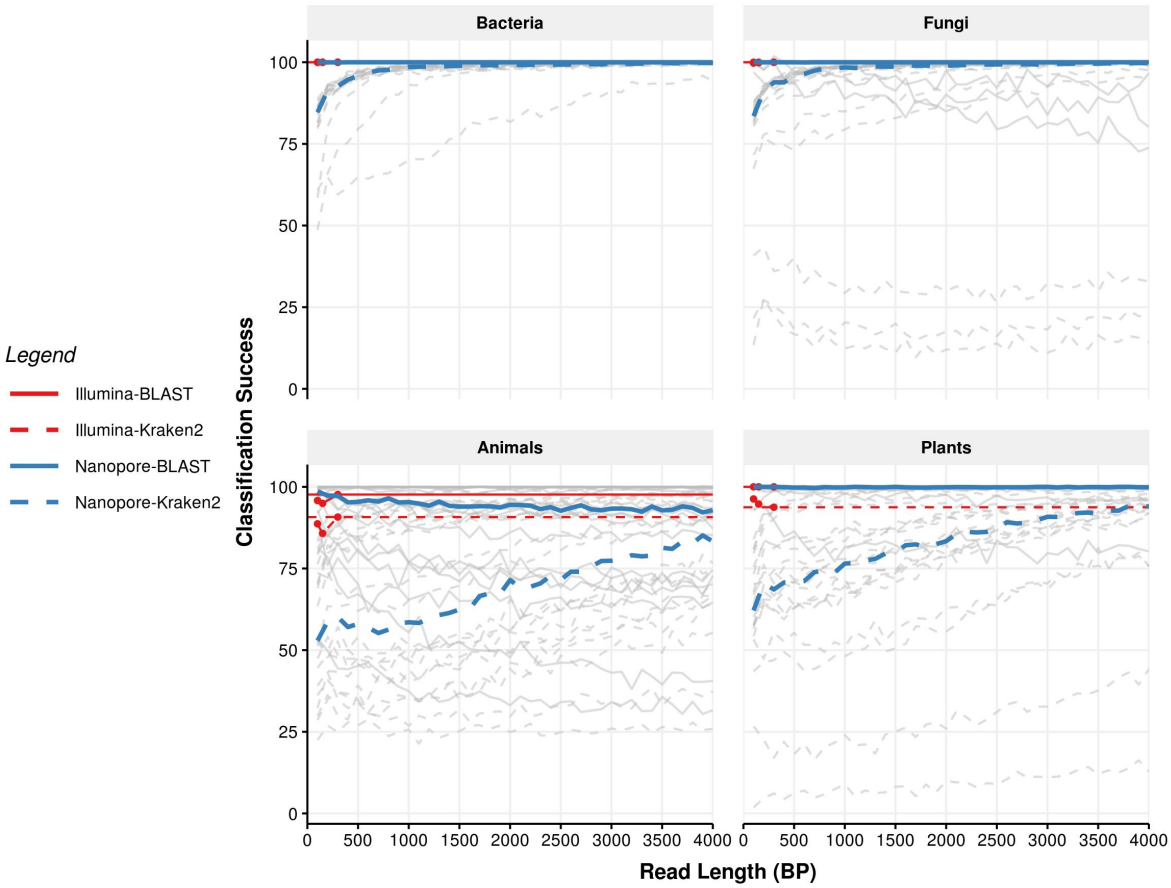
Supplementary Figure 1. **Number of species present in the NCBI RefSeq database has grown roughly exponentially over time**. The y-axis is plotted on a log scale. Data were retrieved from the RefSeq database (O'Leary et al. 2016): https://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/eukaryotes.txt and https://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/prokaryotes.txt



516

517 **Supplementary Figure 2 Recall at the family level.** Each panel shows recall for the
518 different kingdoms. Recall for Nanopore reads for individual taxa is indicated in grey, with
519 median recall indicated by dashed lines, either blue (Kraken2) or red (BLAST). The recall
520 rates for 300 bp Illumina reads are shown as thin solid lines, again either blue (Kraken2) or
521 red (BLAST). Coloured points show the recall for all Illumina reads of all lengths (100 bp,
522 150 bp, and 300 bp).

523



Supplementary Figure 3 Classification success at the family level. Each panel shows classification success for the different kingdoms. Classification success for individual taxa is indicated in grey, with median classification success shown by solid lines (Illumina) or dashed lines (Nanopore). Blue indicates classification success rates for reads classified using Kraken2, while red indicates those classified using BLAST. For animal and plants, the classification success of Kraken2 depends strongly on read length, and never surpasses BLAST or Illumina at any length.

Supplementary Table 1. List of species include in the *in silico* mock community, with associated Kingdom and NCBI Accession details.

Species	Kingdom	NCBI Accession
Actinidia chinensis	Plantae	CM009654.1
Ananas comosus	Plantae	CM003813.1
Arabidopsis thaliana	Plantae	CP002684.1
Brassica nigra	Plantae	CM004491.1
Camelina sativa	Plantae	CM002729.1

Citrus sinensis	Plantae	CM001701.1
Dioscorea rotundata	Plantae	BDMI01000001.1
Eutrema salsugineum	Plantae	CM001778.1
Gossypioides kirkii	Plantae	CM008980.1
Leersia perrieri	Plantae	CM002476.1
Malus domestica	Plantae	CM007867.1
Micromonas sp.	Plantae	CP001574.1
Panicum hallii	Plantae	CM008046.2
Raphanus sativus	Plantae	CM007999.1
Rosa chinensis	Plantae	CM009582.1
Setaria italica	Plantae	CM004364.1
Solanum lycopersicum	Plantae	CM001064.3
Sorghum bicolor	Plantae	CM000760.3
Theobroma cacao	Plantae	LT594788.1
Trifolium pratense	Plantae	LT555306.1
Amphiprion percula	Animalia	CM009708.1
Bos indicus	Animalia	CM003021.1
Capra hircus	Animalia	CM001710.2
Chrysemys picta	Animalia	CM002655.1
Columba livia	Animalia	CM007525.1
Cyprinus carpio	Animalia	LN590701.1
Drosophila busckii	Animalia	CP012523.1
Equus caballus	Animalia	CM000377.2
Falco peregrinus	Animalia	CM007505.1
Homo sapiens	Animalia	CM004593.1
Lycaon pictus	Animalia	CM007565.1
Macaca mulatta	Animalia	CM000308.1
Microcebus murinus	Animalia	CM007661.1

Mus musculus	Animalia	CM004154.1
Oncorhynchus tshawytscha	Animalia	CM009202.1
Oryctolagus cuniculus	Animalia	CM000790.1
Ovis aries	Animalia	CM008472.1
Takifugu rubripes	Animalia	HE602535.1
Timema cristinae	Animalia	CM007794.2
Xiphophorus maculatus	Animalia	CM008938.1
Agaricus bisporus	Fungi	CP015470.1
Alternaria solani	Fungi	CP022024.1
Colletotrichum higginsianum	Fungi	CM004455.1
Cryptococcus gattii	Fungi	CP025759.1
Debaryomyces hansenii	Fungi	CR382133.2
Eremothecium sinecaudum	Fungi	CP014242.1
Flammulina velutipes	Fungi	CM002695.1
Fusarium verticillioides	Fungi	CM000578.1
Kluyveromyces lactis	Fungi	CR382121.1
Komagataella phaffii	Fungi	LT962476.1
Lachancea nothofagi	Fungi	LT598449.1
Malassezia sympodialis	Fungi	LT671813.1
Millerozyma farinosa	Fungi	FO082059.1
Ogataea parapolymorpha	Fungi	CM002300.1
Saccharomyces cerevisiae	Fungi	BK006935.2
Sporisorium scitamineum	Fungi	CP010913.1
Trichoderma reesei	Fungi	CP016232.1
Valsa mali	Fungi	CM003098.1
Yarrowia lipolytica	Fungi	HG934059.1
Zygosaccharomyces rouxii	Fungi	CU928173.1
Acidithiobacillus ferrivorans	Bacteria	LT841305.1

Bacillus thuringiensis	Bacteria	CP015250.1
Bacillus velezensis	Bacteria	CP025939.1
Bifidobacterium longum	Bacteria	CP013673.1
Bordetella bronchiseptica	Bacteria	CM002881.1
Brucella melitensis	Bacteria	CP018494.1
Campylobacter jejuni	Bacteria	CP012689.1
Caulobacter crescentus	Bacteria	AE005673.1
Cellvibrio japonicus	Bacteria	CP000934.1
Escherichia albertii	Bacteria	AP014855.1
Gordonibacter sp.	Bacteria	LT827128.1
Klebsiella pneumoniae	Bacteria	CP025088.1
Mycobacterium tuberculosis	Bacteria	CP023640.1
Ornithobacterium rhinotracheale	Bacteria	CP006828.1
Pseudomonas arsenicoxydans	Bacteria	LT629705.1
Salmonella enterica	Bacteria	CP007400.2
Serratia symbiotica	Bacteria	LN890288.1
Staphylococcus aureus	Bacteria	CP012974.1
Treponema pallidum	Bacteria	CP020366.1
Vibrio cholerae	Bacteria	LT907989.1

535