In deposition environments with low metabolic activity (i.e., frozen, anoxic, or dry sediments), DNA molecules may be preserved for thousands of years. The ubiquity of plant root and leaf fragments and pollen trapped in soils and shallow aquatic sediments makes these materials excellent choices for the molecular characterization of past floras. However, one obstacle in this area of study is the limiting bias in the bioinformatic classification of short fragments of degraded DNA from the large, complex genomes of plants. State-of-the-art analysis of environmental DNA (eDNA), sedimentary DNA, and ancient DNA (aDNA) for the identification of plants typically relies on the amplification of specific barcoding loci that are well represented in databases and contain sufficient information for discrimination between taxa (Parducci et al., 2017; Sjögren et al., 2017). However, due to locus specificity and primer mismatch (Hollingsworth et al., 2011) and the mechanics of PCR amplification, the reliable detection of rare molecules can be low and may bias the reconstruction of communities even when repetition and verification steps are built into the experiment (Ficetola et al., 2014).
Furthermore, in many cases, individual barcode loci are known to lack discriminatory power at the species level (Hollingsworth et al., 2016). For aDNA from well-preserved sources, fragments may be long enough for PCR amplification approaches (e.g., Zimmermann et al., 2017). In more highly degraded samples, DNA molecules may be too short for most barcode target amplification (e.g., <200 bp), but still may be identified using broader reference databases and massively parallel short-read sequencing technology.

Unbiased and complete molecular reconstruction of mixed DNA samples (i.e., microbiomes, eDNA, or from ancient and museum specimens) can now be performed using short-read (<200 bp) shotgun sequencing (e.g., Pedersen et al., 2016). Although shotgun sequencing of mixed samples of plant DNA is not likely to be efficient with today's technology (i.e., a massive sequencing effort will be required for a relatively small amount of classifiable sequence data), improvements in DNA sequencing technology suggest that in the future it will be time- and cost-effective to produce sufficiently large sequencing data sets for this process (Mardis, 2017). The discriminatory power of such a genome-skimming protocol is expected to be substantially higher than traditional barcode loci due to the greater information content of whole-genome reference data sets. These methods will also be backward-compatible with existing databases (e.g., National Center for Biotechnology Information [NCBI], Barcode of Life Database) of barcode and genomic sequence data (Hollingsworth et al., 2016; Parducci et al., 2017).

Of particular interest is complete plastid genome sequencing or a target-capture method that concentrates fragments originating from the chloroplast. These data would be broadly alignable across virtually all of the Viridiplantae without large insertions, duplications, or deletions (Ruhfel et al., 2014), and would therefore provide significant discriminatory power with relatively lower potential for bias in read classification. Furthermore, chloroplast DNA has a higher copy number than nuclear DNA and therefore should be more complete in degraded environmental samples.

Metagenomic methods

Because of the massive volume and fragmented nature of shotgun metagenomic data, careful bioinformatic processing is required to yield reliable taxonomic classification. The methods to accomplish this have not been fully explored in plants, and plant genomes tend to be larger and more complex (Wendel et al., 2016) than those typically targeted by available metagenomic methods (i.e., bacterial communities).

There are two main classes of very-high-throughput read classification programs currently available that are capable of efficient processing against very large databases. These are: (1) assembly or mapping approaches using a Burrows–Wheeler transform (Burrows and Wheeler, 1994; e.g., Centrifuge [Kim et al., 2016] and HOLLI [Pedersen et al., 2016]), and (2) ultrafast k-mer “pseudoalignment” algorithms that require exact matches with short oligonucleotide sequences (e.g., Kraken [Wood and Salzberg, 2014] and CLARK [Ounit et al., 2015]). Classic alignment-based search tools like BLAST and the faster, less-stringent version, MegaBLAST (Zhang et al., 2000), have been shown to yield higher sensitivity when classifying longer reads or contigs, but at the cost of several orders of magnitude more computational time to perform searches (Wood and Salzberg, 2014). Other phylogenetic and machine learning methods are available for microbial metagenomics, but again, low throughput prevents these methods from being computationally cost-effective for analyzing hundreds of thousands of next-generation sequencing (NGS) reads (i.e., MetaPhAn [Segata et al., 2012] and Naive Bayes Classifier [Rosen et al., 2011]). These methods also have not been applied to plant community reconstruction but may be suitable in some cases with high-quality data.

aDNA analysis with shotgun sequencing

To date, the only analysis of NGS short-read shotgun sequence data from ancient samples that can be verified to contain plant DNA was done using the HOLI pipeline (a holistic pipeline for processing high-throughput metagenomic data from environmental samples; https://github.com/ancient-eDNA/Holi/tree/a8fdd3638b98729b4b1b12da6cabdcf8ea61b, 9 March 2016 version used by Pedersen et al., 2016), a UNIX shell script that classifies sequence reads using local alignment mapping via Bowtie2 against the entire NCBI Nucleotide (nt) database. However, the reference database choice and parameters (i.e., stringency of classification criteria) of the analysis have not been explicitly explored. The Centrifuge metagenomics program (Kim et al., 2016) will be substituted for the HOLI pipeline in this study. Centrifuge also performs alignment of reads via the Burrows–Wheeler transform implemented in Bowtie2 (Langmead and Salzberg, 2012), but is well documented, properly versioned, and includes more tools for results processing and filtering.

For this study, we proposed to more thoroughly explore the performance of classification pipelines for short-read shotgun metagenomic sequence data, focusing on the performance of Kraken (Wood and Salzberg, 2014), Centrifuge (Kim et al., 2016), and MegaBLAST (Zhang et al., 2000). Kraken was chosen for this study as the k-mer classifier over CLARK (Ounit et al., 2015), primarily because Kraken is more widely used and cited at this point and performance is similar depending on settings (Ounit et al., 2015). Performance will be tested on a simulated short-read data set and the Pedersen et al. (2016) aDNA soil metagenome data against a single, plant-targeted reference database. The goals of the proposed study were to: (1) demonstrate the use of high-throughput sequence read classification programs using reference databases customized for plant genomic data, (2) quantify false-positive rates under different database/algorithm schemes, (3) provide examples and documentation of how these methods can be implemented, and (4) demonstrate and discuss the relative merits of alignment-based and k-mer-based classification algorithms in the context of plant community metagenomics and previous findings from sedimentary aDNA sequencing studies (Pedersen et al., 2016). The results of this study will inform future studies of plant aDNA as well as studies using eDNA for the biomonitoring of plant community composition in recent time for the presence of spatially or temporally rare plants (Sjögren et al., 2017) and studies establishing conservation baselines from pre-colonization vegetation (Wilmshurst et al., 2014).

METHODS

Code to reproduce all analyses in this study is publicly available at https://github.com/rsh249/ISOETES1.git. Details and instructions on the methods presented here can be found by viewing the README file associated with the ISOETES1 repository (https://github.com/rsh249/ISOETES1/blob/master/README.md). This repository includes scripts to download and convert sequence data.
to build the reference database targeting plants and bash scripts to build reference indices, run read classification programs, and summarize results.

Reference sequence data from all plant (Viridiplantae) taxa in GenBank (Wheeler et al., 2007) and the RefSeq chloroplast sequences (Pruitt et al., 2006) were downloaded from the NCBI repositories. This strategy for the reference database was applied over a total evidence approach of using the entire NCBI nt database for two reasons: (1) the custom database built here is feasible to work with on most modern consumer computers with at least 16 GB of RAM, and (2) the time to classify against the reduced, plant-specific custom database was low for all methods and this made it possible to run several iterations of the classification experiment on simulated data. MegaBLAST (Zhang et al., 2000) can use these data directly, whereas Centrifuge (Kim et al., 2016) builds a reference database using the Burrows–Wheeler transform (Burrows and Wheeler, 1994) and Ferragina–Manzini index (Ferragina and Manzini, 2000) and Kraken (Kim et al., 2016) builds an ordered k-mer hash table in Jellyfish 1.x (Marçais and Kingsford, 2011).

The same preprocessing steps and postanalysis filtering were performed in all analyses. The genome assembly program String Graph Assembler (Simpson and Durbin, 2012; Simpson, 2014) was used to clean raw read data to remove duplicate, repetitive, and low-quality reads (for details, see: https://github.com/rsh249/ISOETES1/blob/master/reproduce/kraken_search_fastq). Cleaned and quality-controlled reads were analyzed using Kraken 0.10.6-released (Wood and Salzberg, 2014), Centrifuge 1.0.3-beta (Kim et al., 2016), and MegaBLAST 2.2.26 (Zhang et al., 2000) to test three different approaches to short-read alignment and classification. Kraken and Centrifuge were run using all default settings (Wood and Salzberg, 2014; Kim et al., 2016). MegaBLAST was run with a minimum word size of 11 bp. All read classifications were referenced against the NCBI taxonomy (downloaded August 2016; Federhen, 2011) to identify the lowest common ancestor associated for each classified read.

Metagenome simulation data were generated using wgsim (https://github.com/lh3/wgsim) to generate 100,000 reads using the default settings of 70-bp read-length average and a base mutation rate of 0.001 and a read error rate of 0. Soil metagenome sequence data from Pedersen et al. (2016) were downloaded from the European Nucleotide Archive (project number PRJEB14494; available at http://www.ebi.ac.uk/ena/data/view/PRJEB14494).

All analyses were performed on a compute cluster hosted by the Sackler Institute for Comparative Genomics at the American Museum of Natural History (New York, New York, USA). The cluster comprises 16 identical compute nodes, each with twin Intel Xeon E5-2697A v4 (2.6 GHz, 16-core) central processing units (CPUs) and 256 GB of RAM. Data were served to compute nodes from a 40-TB local network attached storage device that could adversely affect performance for disk-intensive processes.

RESULTS

One consideration in classification algorithm choice is limitations in available computational resources. RAM requirements varied depending on the nature of reference database indexing and handling. For example, MegaBLAST does not read the reference database used here (all plants in GenBank plus RefSeq plastids) into RAM in its native FASTA format and requires 338 MB of memory.

Kraken's k-mer database required significantly more memory at 5.4 GB and Centrifuge fell in the middle, with its indexed Burrows–Wheeler transform database requiring 1.04 GB in practice. Any of these methods would be suitable for analysis with this reference database and conditions. Kraken and Centrifuge were run using all default settings (Wood and Salzberg, 2014; Kim et al., 2016). MegaBLAST was run with a minimum word size of 11 bp. All read classifications were referenced against the NCBI taxonomy (downloaded August 2016; Federhen, 2011) to identify the lowest common ancestor associated for each classified read.

Metagenome simulation data were generated using wgsim (https://github.com/lh3/wgsim) to generate 100,000 reads using the default settings of 70-bp read-length average and a base mutation rate of 0.001 and a read error rate of 0. Soil metagenome sequence data from Pedersen et al. (2016) were downloaded from the European Nucleotide Archive (project number PRJEB14494; available at http://www.ebi.ac.uk/ena/data/view/PRJEB14494).

All analyses were performed on a compute cluster hosted by the Sackler Institute for Comparative Genomics at the American Museum of Natural History (New York, New York, USA). The cluster comprises 16 identical compute nodes, each with twin Intel Xeon E5-2697A v4 (2.6 GHz, 16-core) central processing units (CPUs) and 256 GB of RAM. Data were served to compute nodes from a 40-TB local network attached storage device that could adversely affect performance for disk-intensive processes.

RESULTS

One consideration in classification algorithm choice is limitations in available computational resources. RAM requirements varied depending on the nature of reference database indexing and handling. For example, MegaBLAST does not read the reference database used here (all plants in GenBank plus RefSeq plastids) into RAM in its native FASTA format and requires 338 MB of memory.
database on a modern laptop computer with at least 16 GB of available memory.

Kraken and Centrifuge are both designed to be very fast search algorithms (Wood and Salzberg, 2014; Kim et al., 2016). Time to classify the soil metagenome data from Pedersen et al. (2016) varied by total amount of sequence data (Fig. 1). Kraken and Centrifuge required comparable amounts of computation time, generally requiring less than 7 min of CPU time (less actual time if multithreading was employed). MegaBLAST required two to four orders of magnitude more time on the same data sets (Fig. 1). The extended time required for MegaBLAST does not correspond to more reads classified. On the contrary, Kraken and Centrifuge generally classified more reads than MegaBLAST (Fig. 2).

To compare the classification performance of each of these methods, short-read data were simulated from the reference database using wgsim (https://github.com/lh3/wgsim) for 10, 50, or 100 randomly selected taxa with five rounds of repetition. Reads were classified to lowest common genus and order. In general, Centrifuge has a higher true-positive rate or sensitivity than the other two methods (Fig. 3). That is, Centrifuge correctly identified more of the known taxa in nearly every case than did Kraken or MegaBLAST. However, Kraken appears to be generally more conservative (Fig. 3), having the highest precision (fewest false positives relative to total classifications made). MegaBLAST ranked lowest in all of these tests, with relatively high rates of false positives and low rates of true positives.

Soil metagenomic data from the North American Ice-Free Corridor (IFC) were reanalyzed with Kraken, Centrifuge, and MegaBLAST with the plant-specific reference database described here, and genera were classified as "present" if more than 0.1% of all classified reads were unambiguously classified to that genus. In general, the total number of genera identified by MegaBLAST was much higher than for Kraken and Centrifuge (Fig. 4). For Kraken and Centrifuge, there was no clear pattern as to which method routinely predicted higher or lower generic diversity (Fig. 4). No trend in generic diversity through time was observed for either lake series (Fig. 4).

DISCUSSION

Reanalysis of the North American IFC soil metagenome sequence data performed with Kraken, Centrifuge, and MegaBLAST yielded broadly comparable results to those published using HOLI (Fig. 5; Pedersen et al., 2016). Soil shotgun metagenome sequence data dating back to 12,810 yr before present (BP) (Pedersen et al., 2016) yielded reads that could be classified to plants in our reference database at a rate of less than 1.5% of the total unique, quality-controlled reads (Fig. 2).

The number of genera identified by each of the three methods tested supports the idea that MegaBLAST has issues with high false-positive rates when classifying short-read data (Figs. 3 and 4). This is not surprising as MegaBLAST was not designed for classifying very short molecules (Zhang et al., 2000) and was included in this study as a "straw-man" baseline method. There are many plant taxa identified by the three methods tested here and HOLI (Pedersen et al., 2016) that are ecologically important in the successional vegetation of the IFC (Fig. 5). However, the occurrence pattern through time of some of these taxa depends on which classification method is employed (Fig. 5).

The read classification protocol used can drastically alter any conclusions drawn about paleoecology and the successional processes taking place in the IFC through this time series, although there are places where the methods all generally converge. In every sample in both the Charlie Lake and Spring Lake series, *Populus* L. was identified as present by HOLI, and in all but the oldest two samples, *Salix* L. was as well (Fig. 5). These results suggest that early vegetation in the IFC included a poplar–willow woodland community. Kraken, Centrifuge, and MegaBLAST generally agree with this conclusion (Fig. 5), but Kraken and Centrifuge

---

**FIGURE 3.** Plant community metagenomics simulation taxonomic classification performance. Metagenomic data simulated from data for 10, 50, or 100 species (×5 repetitions each) were analyzed by Kraken, Centrifuge, and MegaBLAST. Performance was measured by the sensitivity (true-positive rate [TPR = TP/(TP + FN)]) for classifications made to genera (A) and orders (B), and by precision (positive predictive value [PPV = TP/(TP + FP)]) for classifications made to genera (C) and orders (D). Accuracy for all instances is close to 1 and is not shown here because there are >100,000 taxa represented in the reference database and the vast majority are predicted to be absent in all samples. TP = true positive; FP = false positive; FN = false negative.
infer notably patchier occurrence of these genera in layers older than 9300 yr BP. Reads mapping to grasses generally (Pedersen et al., 2016), and Hordeum L. more specifically (Fig. 5), prior to ~11,500 yr BP are interpreted as suggesting that the poplar–willow woodland was either sparse or interrupted by areas of grassland. However, grass and Salicaceae pollen types are identified in low levels in all layers (see Pedersen et al., 2016, Figs. S4 and S5), suggesting that the differences observed in the molecular evidence are likely due to detection issues.

In one case of notable congruence, the occurrence of Ceratophyllum L. in the IFC lakes was inferred by all methods in the same samples, all dating to the past 10,000 yr, suggesting that the aquatic vegetation of the glacial lakes shifted as the lakes became shallower and warmer as the glaciers receded from the region (Pedersen et al., 2016). However, despite the agreement on the molecular evidence across methods, the pollen and macrofossil assemblages analyzed by Pedersen et al. (2016) disagree. Pollen and macrofossil evidence (see Pedersen et al., 2016, Fig. S4 and associated data) does not show presence of Ceratophyllum older than 7000 yr BP, whereas the molecular evidence suggests its presence between 7700 and 9300 yr BP (Fig. 5).

Notable differences between methods were observed in Picea A. Dietr. and Pinus L. In the original study, HOLI classified reads from all samples less than 10,000 years old as including Picea (Pedersen et al., 2016). Reanalysis with Kraken and MegaBLAST infer the presence of Picea in some of the samples between 11,500 and 12,800 yr BP (Fig. 5), which would suggest that at least patches of boreal forest were present in the vicinity of the lakes nearly back to the opening of the IFC. The presence of Pinus inferred by HOLI suggests that in the past 10,000 years, Pinus was present as the modern boreal forest ecosystem developed. Kraken does not identify the presence of Pinus until 1444 yr BP, and Centrifuge has a gap in samples containing Pinus between 9300 and 7001 yr BP (Fig. 5). Interestingly, none of the metagenomic methods match the pollen record, which indicates the presence of Pinus-type pollen dating back to at least 12,000 yr BP (Pedersen et al., 2016), potentially suggesting that pollen is not a good vector for DNA preservation in lake sediments.

The presence of Artemisia L. pollen in the lake sediments suggests that elements of dry steppe vegetation have been present in the area throughout this time period. However, HOLI only inferred the presence of Artemisia DNA in samples dating between 11,200 and 12,500 yr BP. The other methods tested here are even more conservative, with the extreme case once again being Kraken, which only identifies reads matching Artemisia in samples older than 12,300 yr (Fig. 5). It is important to note that discrepancies between the pollen record and the aDNA record could be a result of pollen import from extra-local areas and/or molecular taphonomic preservation biases against Artemisia.

Why are there differences between methods?

The reanalysis of IFC soil metagenome data supports the assertion that Kraken is the most conservative of the methods tested here. Kraken often predicts the presence of a genus in fewer samples than all of the other methods (Fig. 5). MegaBLAST has difficulty classifying short-read data (Fig. 3), but it does infer relatively similar trends in the common and ecologically important IFC genera (Fig. 5).

Although Centrifuge and HOLI use alignment against a Burrows–Wheeler transform via Bowtie2 (Langmead and Salzberg, 2012) to establish read classification, the reanalysis of the IFC soil metagenome data from Pedersen et al. (2016) revealed potentially important differences between classifications made by HOLI and Centrifuge. It is important to note why the HOLI pipeline was not tested in this study. First, the HOLI pipeline as referenced by Pedersen et al. (2016) was incomplete and lacked the last common ancestor designation step as well as any user guidance or documentation in the code repository (https://github.com/ancient-eDNA/Holi/commit/a8fd3638b98729b4b1b12a23da6cabdf8e6a1b; 9 March 2016 version). Furthermore, since that time, the HOLI pipeline has continued to be developed for a new study that is not yet published (https://github.com/ancient-eDNA/Holi), and it may therefore be impossible to reproduce the results of Pedersen et al. (2016) using the current repository. This mismanagement of

![FIGURE 4. Generic diversity identified by Centrifuge, Kraken, and MegaBLAST in ancient DNA soil metagenomic data. Total genera counts are the number of genera with greater than 0.1% of classified reads mapping to a genus by each of the three methods tested here for sedimentary ancient DNA samples from the Charlie Lake and Spring Lake series from Pedersen et al. (2016).](http://www.wileyonlinelibrary.com/journal/AppsPlantSci)
Higher misclassification rates, particularly when restricted reference data sets are used (Schmieder and Edwards, 2011; Breitwieser et al., 2017). To avoid issues with spurious classifications of short- and low-complexity reads, all metagenomic data analyzed here were first pre-processed with tools implemented in String Graph Assembler (Simpson and Durbin, 2012) to remove duplicate and low-complexity reads using the subprograms ‘preprocess’ and ‘filter’ (https://github.com/jits/sga/tree/master/src#readme). A second and possibly likelier cause of differing results, Centrifuge performs a rapid gapped alignment (Kim et al., 2016), whereas HOLI does not allow gaps in read alignments under 50 bp (Pedersen et al., 2016). Gapped alignment may allow reads to map more ambiguously to similar sequences across broader taxonomic scales, causing fewer genera to be unambiguously classified by Centrifuge versus HOLI and overall more conservative inferences made by Centrifuge (Fig. 5).

The issue of missing data

One major issue with all uses of available genetic and genomic data is the colossal amount of missing data. This problem is not limited to the use of these data for metagenomic classification; it also affects other fields like phylogenetics and functional or evolutionary genomics. The NCBI houses a massive database of nucleotide sequence data. As of August 2017, publicly available sequence data from NCBI covers more than 159,012 species of Viridiplantae, of which 149,652 are in the Embryophyta. However, there are only 267 whole genomes and ~1900 chloroplast genomes for all plants, showing that the vast majority of taxa are represented by only a few sequences. Therefore, in a random sample of DNA fragments from a mixed sample, it is expected that many of those fragments will most closely match to a taxon for which we have a large amount of sequence data (i.e., whole genome) for the sole reason that there are no other homologous sequences in the database for that part of the genome. For this reason, it may be advisable to design databases that optimize taxonomic coverage and minimize missing data by targeting genes for which there are the most available data even if sequencing effort is not necessarily available sequence data from NCBI covers more than 159,012 species of Viridiplantae, of which 149,652 are in the Embryophyta. However, there are only 267 whole genomes and ~1900 chloroplast genomes for all plants, showing that the vast majority of taxa are represented by only a few sequences. Therefore, in a random sample of DNA fragments from a mixed sample, it is expected that many of those fragments will most closely match to a taxon for which we have a large amount of sequence data (i.e., whole genome) for the sole reason that there are no other homologous sequences in the database for that part of the genome. For this reason, it may be advisable to design databases that optimize taxonomic coverage and minimize missing data by targeting genes for which there are the most available data even if sequencing effort is not targeting these loci specifically. Furthermore, if whole genome data are included in the reference database, it may be informative to consider where reads are mapping and whether or not there are homologous sequences to that region in other taxa.

Short-read shotgun metagenomics holds exciting potential for the study of plant communities at various spatial and temporal scales through the analysis of eDNA and aDNA. However, the potential of this technology is complicated by the complexity of plant genomes, missing data for most plant taxa, and the ability to accurately classify short (less than 200 bp) reads from today’s massively parallel DNA sequencing platforms. Search algorithms...
now exist (Wood and Salzberg, 2014; Kim et al., 2016) that can rapidly classify reads against very large reference databases an order of magnitude faster than MegaBLAST (Fig. 1). In this study, we show that bioinformatic choices for read classification algorithms can make a difference in terms of conclusions drawn from the analysis of aDNA from the IFC of North America during the Pleistocene–Holocene transition (Fig. 5). These differences may be a result of the classification performance characteristics of each of the tested methods (Kraken, Centrifuge, and MegaBLAST) illuminated through a series of in silico simulation experiments (Fig. 3). Further testing of all aspects of these bioinformatic pipelines is essential, including testing more classification algorithms and upstream and downstream processing. Although they are designed for microbial metagenomics, Kraken and Centrifuge rank at the top of the methods discussed here because of their superior performance (Fig. 3) and their user support and development (see https://github.com/DerrickWood/kraken and https://github.com/infphilo/centrifuge). Users seeking to perform shotgun metagenomic characterization of plant communities from environmental or ancient samples should consider these tools for future analyses.

ACKNOWLEDGMENTS

The author would like to thank the Gerstner Family Foundation for their generous support. All computational work was carried out with resources made available by the Sackler Institute for Comparative Genomics at the American Museum of Natural History.

LITERATURE CITED

Breitwieser, F. P., J. Lu, and S. L. Salzberg. 2017. A review of methods and databases for metagenomic classification and assembly. Briefings in Bioinformatics https://doi.org/10.1093/bib/bbx120.

Burrows, M., and D. J. Wheeler. 1994. A block-sorting lossless data compression algorithm. Technical Report 124. Digital Equipment Corporation, Palo Alto, California, USA.

Federhen, S. 2011. The NCBI taxonomy database. Nucleic Acids Research 40: D136–D143.

Ferragina, P., and G. Manzini. 2000. Opportunistic data structures with applications. In Proceedings of the 41st IEEE Symposium on Foundations of Computer Science, Redondo Beach, California, USA.

Ficetola, G. F., J. Pansu, A. Bonin, E. Coissac, C. Gigué-Covex, M. De Barba, L. Gielly, et al. 2014. Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. Molecular Ecology Resources 15: 543–556.

Hollingsworth, P. M., S. W. Graham, and D. P. Little. 2011. Choosing and using a plant DNA barcode. PLoS ONE 6: e19254.

Hollingsworth, P. M., D. Li, M. van der Bank, and A. D. Twyford. 2016. Telling plant species apart with DNA: From barcode to genomes. Philosophical Transactions of the Royal Society B 371: 20150338.

Kim, D., L. Song, F. P. Breitwieser, and S. L. Salzberg. 2016. Centrifuge: Rapid and sensitive classification of metagenomic sequences. Genome Research 26: 1721–1729.

Langmead, B., and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9: 357–359.

Marçais, G., and C. Kingsford. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27: 764–770.

Mardis, E. R. 2017. DNA sequencing technologies: 2006–2016. Nature Protocols 12: 213–218.

Ounit, R., S. Wamamaker, T. J. Close, and S. Lonardi. 2015. CLARK: Fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics 16: 236.

Parducci, L., K. D. Bennett, G. F. Ficetola, I. G. Alisos, Y. Suyama, J. R. Wood, and M. W. Pedersen. 2017. Ancient plant DNA in lake sediments. New Phytologist 214: 924–942.

Pedersen, M. W., A. Ruter, C. Schweger, H. Friebe, R. A. Staff, K. K. Kjeldsøe, M. L. Mendoza, et al. 2016. Postglacial viability and colonization in North America’s ice-free corridor. Nature 537: 45–49.

Pruitt, K. D., T. Tatusova, and D. R. Maglott. 2006. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research 35: D61–D65.

Rosen, G. L., E. R. Reichenberger, and A. M. Rosenfeld. 2011. NBC: The Naïve Bayes Classification tool webserver for taxonomic classification of metagenomic reads. Bioinformatics 27: 127–129.

Ruhfel, B. R., M. A. Gitzendanner, P. S. Soltis, D. E. Soltis, and J. G. Burleigh. 2014. From algae to angiosperms—Inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. BMC Evolutionary Biology 14(1): 23.

Schmieder, R., and R. Edwards. 2011. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS ONE 6: e17288.

Segata, N., L. Waldron, A. Ballarini, V. Narasimha, O. Jousso, and C. Huttenhower. 2012. Metagenomic microbial community profiling using unique clade specific marker genes. Nature Methods 9: 811–814.

Simpson, J. T. 2014. Exploring genome characteristics and sequence quality without a reference. Bioinformatics 30: 1228–1235.

Simpson, J. T., and R. Durbin. 2012. Efficient de novo assembly of large genomes using compressed data structures. Genome Research 22: 549–556.

Sjögren, P., M. E. Edwards, L. Gielly, C. T. Langdon, I. W. Croudace, M. K. F. Merkel, T. Fonville, and I. G. Alisos. 2017. Lake sedimentary DNA accurately records 20th century introductions of exotic conifers in Scotland. New Phytologist 213: 929–941.

Wendel, J. F., S. A. Jackson, B. C. Meyers, and R. A. Wing. 2016. Evolution of plant genome architecture. Genome Biology 17(1): 37.

Wheeler, D. L., T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, et al. 2007. Database resources of the national center for biotechnology information. Nucleic Acids Research 36: D13–D21.

Wilmshurst, J. M., N. T. Moar, J. R. Wood, P. J. Bellingham, A. M. Findlater, J. Robinson, and C. Stone. 2014. Use of pollen and ancient DNA as conservation baselines for offshore islands in New Zealand. Conservation Biology 28: 202–212.

Wood, D. E., and S. L. Salzberg. 2014. Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biology 15: R46.

Zhang, Z., S. Schwartz, L. Wagner, and W. Miller. 2000. A greedy algorithm for aligning DNA sequences. Journal of Computational Biology 7: 203–214.

Zimmermann, H. H., E. Raschke, L. S. Epp, K. R. Stooft-Leichsenring, G. Schwamborn, L. Schirrmeister, P. P. Overduin, and U. Herzschuh. 2017. Sedimentary ancient DNA and pollen reveal the composition of plant organic matter in Late Quaternary permafrost sediments of the Buur Khaya Peninsula (north-eastern Siberia). Biogeosciences 14: 575–596.