High-throughput sequencing for algal systematics

Mariana C. Oliveira[a], Sonja I. Repetti[b], Cintia Iha[ab], Christopher J. Jackson[b], Pilar Díaz-Tapia[bc], Karoline Magalhães Ferreira Lubiana[a], Valéria Cassano[a], Joana F. Costa[b], Ma. Chiela M. Cremen[b], Vanessa R. Marcelino[bde], Heroen Verbruggen[b]

[a] Department of Botany, Biosciences Institute, University of São Paulo, São Paulo 05508-090, Brazil

[b] School of BioSciences, University of Melbourne, Victoria 3010, Australia

[c] Coastal Biology Research Group, Faculty of Sciences and Centre for Advanced Scientific Research (CICA), University of A Coruña, 15071, A Coruña, Spain

[d] Marie Bashir Institute for Infectious Diseases and Biosecurity and Sydney Medical School, University of Sydney, New South Wales 2006, Australia

[e] Westmead Institute for Medical Research, Westmead, New South Wales 2145, Australia

Abstract

In recent years, the use of molecular data in algal systematics has increased as high-throughput sequencing (HTS) has become more accessible, generating very large data sets at a reasonable cost. In this perspectives paper, our goal is to describe how HTS technologies can advance algal systematics. Following an introduction to some common HTS technologies, we discuss how metabarcoding can accelerate algal species discovery. We show how various HTS methods can be applied to generate datasets for accurate species delimitation, and how HTS can be applied to historical type specimens to assist the nomenclature process. Finally, we discuss how HTS data such as organellar genomes and transcriptomes can be used to construct well resolved phylogenies, leading to a stable and natural classification of algal groups. We include examples of bioinformatic workflows that may be applied to process data for each purpose, along with common programs used to achieve each step. We also discuss possible strategies and the new skill set that will be required to fully embrace HTS as a part of algal systematics, along with considerations of cost and experimental design. HTS technology has revolutionized many fields in biology, and will certainly do the same in algal systematics.

Key words: bioinformatic workflows, classification, high-throughput sequencing, molecular data, multiplexing, nomenclature, species boundaries, species discovery, systematics.

Introduction

In recent decades, the use of molecular data has become gradually more dominant in algal systematics. These data have diverse applications including for species discovery, defining species boundaries, nomenclatural decisions and higher-level classification. The current trend of increasing use of molecular data can be expected to persist into the future, running in parallel with developments in DNA sequencing technology and its affordability.

In recent years, high-throughput sequencing (HTS) has become more accessible and this technology has the ability to generate very large data sets at a reasonable cost (e.g., Song *et al.,* 2016). New HTS techniques along with improvements in algorithms for handling/analysing HTS data and falling costs have allowed the development of a wide range of applications and led to a huge increase in molecular sequence data (Reuter *et al.*, 2015; Mardis, 2017).

Previous reviews have described HTS technologies and their application in systematics. A comparison of platforms and discussion of emerging applications is given in, e.g., Glenn (2011) and Goodwin *et al.* (2016). The use of different HTS sequencing platforms and their applicability in plant systematics is reviewed by Harrison & Kidner (2011), and Lemmon & Lemmon (2013) review methods for obtaining HTS data and discuss their use in phylogenetic analyses.

In this paper, we describe the various ways current HTS technologies can advance the field of algal systematics. We present some common and readily available HTS technologies and illustrate their application in species discovery, species delimitation, nomenclature and classification. We also discuss the new skill set that will be required to fully embrace HTS as a part of algal systematics, along with considerations of cost and experimental design.

HTS technology

A range of technologies are now available for use in high-throughput sequencing. We limit ourselves to an overview of those most commonly used in systematics at the moment (Table 1). Among these are several short-read sequencing platforms, which produce much shorter reads than traditional Sanger sequencers, as well as some long-read platforms. For short-read sequencing, Illumina is the dominant system with several well-established platforms (Reuter *et al.,* 2015; Goodwin *et al.*, 2016). Illumina MiSeq and NextSeq are benchtop sequencers with varying levels of throughput. Moving up in throughput even further, the Illumina HiSeq is a large, production-scale sequencer capable of producing 1,000 Gb per run. Error rates are very low for all these platforms, compared to other technologies. DNA sequencing is a fast moving field, and new higher-throughput technologies are already becoming available (e.g. the NovaSeq 6000 is a new benchtop sequencer with higher throughput of up to 6,000 Gb per run) or are under development.

PacBio and MinION technologies can produce long sequence reads, but produce lower yield and generally have a higher error rate than Illumina. They are typically used to complement the more accurate and higher-throughput Illumina methods for assembly and genome finishing. MinION is a small and inexpensive handheld device that can run via USB connection to a personal computer (Dijk *et al.*, 2014; Goodwin *et al.*, 2016), and recently has been used in the field for plant species identification (Parker *et al.* 2017).

HTS platforms require that a DNA library be constructed for each sample before sequencing. A DNA library is a collection of DNA fragments each containing oligonucleotide adapters at their ends that are compatible with the sequencing system (Head *et al.,* 2015). Broadly speaking, library protocols consist of three main steps: DNA fragmentation, size selection and adapter ligation (Head *et al.* 2015).

One of the biggest advantages of HTS for systematics is that multiple samples can be sequenced in the same sequencing run, resulting in cost and time savings. This process is called multiplexing, and involves pooling DNA libraries together. Sequences from each library are identified by a short unique DNA sequence, also known as index sequence or sequencing barcode, which is ligated to each fragment during library preparation (Harrison & Kidner, 2011; McCormack *et al.*, 2013). After sequencing, these barcodes can be used to identify and sort reads corresponding to each library using bioinformatic analyses. All commonly used platforms have multiplexing options. Off-the-shelf library preparation kits allow multiplexing of up to 384 samples (Illumina Nextera XT DNA Library Prep), and more than 1,000 samples have been pooled with user-designed barcodes (e.g. Shokralla *et al.*, 2015).

Different types of input DNA/RNA can be sequenced in various ways to suit diverse applications in systematics. In whole genome shotgun (WGS) sequencing, total genomic DNA is extracted from the organism and sequenced. This approach yields reads belonging to all genomes present in the organism, but chloroplast and mitochondrial genomes are usually dominant in the data (in terms of coverage) because they are present in multiple copies in the cell. RNA sequencing (also called RNAseq and transcriptome sequencing) involves the extraction of total RNA from an organism and enrichment of the mRNA by poly(A)-tail selection or ribosomal RNA depletion, followed by high-throughput sequencing (Mortazavi *et al.*, 2008).

Instead of sequencing complete genomes or transcriptomes, one can also use a variety of methods to selectively enrich particular genomic regions. RADseq enables broad sampling of many independent regions across the genome and requires no prior information about the genome. With this method, DNA is fragmented by restriction enzymes and, following a size selection step, the resulting fragments are sequenced (McCormack *et al.*, 2013). Because the position of restriction sites in the genome is fairly conserved within species and among closely related species, this method can yield data for the same genomic regions across many samples, which can be used to infer species boundaries. RADseq is not

generally useful when studying more distantly related species because the position of restriction sites may not be conserved among them.

Amplicon sequencing is another enrichment technique that targets specific genomic regions (the amplicons). It is often performed as a two-step process, involving a first round of PCR amplification using amplicon-specific primers flanked by tails, to enrich for one or multiple parts of the genome. The primer tails allow for a second round of PCR to add Illumina adapter sequences, and barcodes if multiplexing is desired. This enables the simultaneous sequencing of a specific set of genes for multiple samples (Cruaud *et al.* 2017). In this case, a priori information about the genome is needed for primer design, and in order for the amplicons to be fully sequenced, they need to be designed to fit within the read length of the technology. For example, when using MiSeq with 300 nt-long reads, amplicons should not exceed 500 nt if they are to be completely sequenced (with some overlap) by the 300 nt paired-end reads.

Target capture is another method to enrich particular areas of the genome. This method relies on selective enrichment of chosen genomic regions by hybridisation to probes. These probes are short DNA or RNA molecules (ca. 70-150 nt) designed to match particular regions of the genome (usually exons or other conserved regions of the genome). The DNA of the sample to be sequenced is then hybridised to the probes which bind to complementary DNA. Unbound DNA is washed away, and the bound DNA is subsequently sequenced. In the case of target capture, a priori genomic information is required for probe design (McCormack *et al.*, 2013).

Box 1: Glossary

Amplicon sequencing: Sequencing of specific genomic regions amplified by PCR using locus-specific primers.

Annotation: The process of identifying regions of significance or interest in an assembled genome, including coding and non-coding regions (together with putative functions), and other attributes such as structural features and repeat regions.

Assembly: Process by which overlapping short reads are combined into large contiguous segments of DNA (contigs). Assembly can be performed by mapping shorts reads to an existing reference such as the genome of a closely related species (mapping assembly), or reads can be combined without a reference sequence (*de novo* assembly). A variety of algorithms are available to perform assembly.

Contig: Contiguous segment of DNA resulting from assembly of shorter HTS reads.

Coverage: Number of sequencing reads that support the occurrence of a nucleotide in a given position in a contig.

DNA barcode: DNA sequence used as a taxonomic tool to identify a species. These are short DNA fragments of selected molecular markers. Usually different markers are used for different taxonomic groups. Not to be confused with a sequencing barcode.

Demultiplexing: A bioinformatic procedure that uses short index sequences (see “index sequence”) from pooled samples (see "multiplexing") to sort sequence reads into separate files, each containing data from a single sample.

Exome sequencing: Sequencing of the exome (i.e. all exon sequences) of an organism using target capture.

Genome finishing: Contiguous segments of a given genome are ordered and joined; ambiguities or gaps between them are resolved to complete the whole-genome. See text for strategies.

Genomic partitioning: Different methods can be used to enrich a sequence library for specific target regions of a genome, as an alternative to whole genome sequencing.

High-throughput sequencing: Category of sequencing technologies that parallelize the sequencing process. Sequencing costs are reduced because they can generate millions of sequences from thousands of DNA templates at once. Many HTS platforms are available, differing in their sequencing chemistry and sequencing instrument used.

Index sequence: Also known as barcode tag or sequencing barcode. Short sequence ligated to each DNA fragment of a sample during library preparation. See "multiplexing" for more information. Not to be confused with a DNA barcode.

Multiplexing: The process of pooling several samples together in the same run. This can be achieved by ligating unique short sequence tags (called index sequences or barcodes) to the DNA fragments of each sample. HTS reads can be identified by using these tags during bioinformatic analyses.

Next generation sequencing: Alternative term for high-throughput sequencing.

Paired-end sequencing: DNA fragments are sequenced from both ends.

Workflow: Also referred to as a pipeline. A succession of bioinformatic procedures that transform raw data into interpretable results. These processes are usually automated by programs/scripts written in a scripting language (e.g. Perl, Python) and include tasks such as quality control and data analysis.

RADseq: Sequencing of restriction site-associated DNA (see text for full explanation).

RNAseq: RNA sequencing, typically of mRNA.

Read mapping: Aligning (mapping) of HTS short reads to a reference sequence (e.g. contig, genome).

Scaffold: Construct consisting of non-overlapping contigs where the contig order is known, but the sequence between them is not (i.e. there is a sequencing gap). Such contigs are usually linked together by paired-end sequences where each read maps to the end of a different contig.

Target capture: A strategy to enrich for specific regions of the genome using hybridisation to probes. Target capture requires prior knowledge of the genome of interest (or a closely related genome) for probe design. The probes are then hybridized with the DNA sample(s) of interest and the enriched DNA is sequenced.

Two-step PCR: A cost-effective approach to prepare libraries of amplicons for high-throughput sequencing. One or more target genomic fragments (amplicons) are amplified during the first PCR using amplicon-specific primers containing an additional stretch of nucleotides (called a tail or adapter). These tails allow a second PCR round to add sequencing barcodes and platform adapters (e.g. for Illumina) to the amplicons.

Whole genome shotgun sequencing: Approach used to sequence genomic DNA, involving fragmentation of total DNA into smaller fragments that are then sequenced. The resulting sequence reads are assembled into longer contigs.

HTS technologies have created high demand for bioinformatic tools capable of sorting, analysing and managing large amounts of data (Dijk *et al.*, 2014). HTS platforms produce millions of sequence reads that require specific analysis workflows to produce interpretable results. Individual steps in a workflow can be carried out by a number of different programs, each with their own pros and cons and customisable sets of features. Because of this, it is important to set aside enough time to experiment with programs in the workflow to determine what is most suitable for the dataset being analysed.

In the following sections addressing the application of HTS in different aspects of systematics, we will illustrate examples of bioinformatic workflows that may be applied to process data for each purpose, along with common programs used to achieve each step. These workflows are meant merely as an illustration of what can be achieved and will differ depending on the exact question being answered and sequencing strategies chosen.

HTS for new species discovery

While the rate at which new species are described has been constant or decreasing for well-known groups such as mammals and birds, more algal species are being discovered than ever before (Costello & Wilson, 2011; De Clerck *et al.*, 2013). The development of molecular techniques has positively contributed to this trend as these can allow identification of cryptic diversity (e.g. Beszteri *et al.,* 2007, Payo *et al.,* 2013), which is especially prevalent in morphologically simple algal lineages (Verbruggen, 2014). Working with microalgae typically relies on isolation and culturing, which can be difficult to achieve for algal species that are rare or unculturable. Indeed, only a small fraction of wild cells from the environment can be cultured with conventional techniques (DeLong, 2009), and consequently the algae strains deposited in culture collections represent less than 10% of natural diversity (Kim *et al.*, 2014). HTS techniques are providing opportunities to study undiscovered species without the need for culturing.

Metabarcoding is routinely used to characterise microbial communities and can contribute to uncovering unknown algal diversity. This technique consists of extracting DNA directly from environmental samples and amplifying DNA metabarcodes (e.g. 16S, 18S, *cox*1, *rbc*L, *tuf*A), followed by HTS and identification of the species present in the bulk environmental sample (Shokralla *et al.*, 2012). This approach has allowed the discovery of a massive number of microorganisms that had not previously been isolated (e.g., Rappe & Giovannoni, 2003).

The Tara Oceans expeditions obtained a very large metabarcoding dataset across tropical and temperate oceans (de Vargas *et al.*, 2015; Malviya *et al.*, 2016). Using the eukaryotic 18S rRNA gene, the authors found ~110,000 operational taxonomic units (OTUs), whereas only ~11,200 species had previously been described for marine eukaryotic plankton (de Vargas *et al.*, 2015). Not all these newly discovered OTUs are microalgae, but it is a striking example of how metabarcoding can reveal new species that might otherwise go unnoticed. The highest diversity and number of sequences of undescribed species was found among the picoplanktonic cell size group, organisms that are more difficult to characterise using morphology (de Vargas *et al.*, 2015). The pico/nano-planktonic Prasinophyte Clade VII, for example, was poorly characterised before the Tara Oceans survey, and now it is known to be a highly diverse group (Lopes dos Santos *et al.* 2016). Even in better-studied phytoplankton groups, many sequences of unknown species were obtained. For example, ~8,000 dinoflagellate OTUs were found even

though only ~2,000 species have been described, and for both the Dictyochophyceae and Chlorarachniophyta, the number of OTUs recovered is more than 60 times higher than the number of described species (de Vargas *et al.*, 2015).

Smaller-scale metabarcoding surveys have also revealed previously uncharted diversity of microalgae. Studies targeting snow algae have found that uncultured *Chlamydomonadaceae* compose most of the species diversity in this habitat (Lutz *et al.* 2015, Lutz *et al.* 2016). Further, studies on haptophytes from Norway and Naples found that the majority of the sequences identified were from undescribed species (Bittner *et al.*, 2013; Egge *et al.*, 2015). The biodiversity of haptophytes has also been surveyed with the 28S rRNA metabarcode, but the low number of reference sequences from known species hinders an accurate assessment of the species identity with this marker (Gran-Stadniczeñko *et al.* 2017). The *rbc*L metabarcode is more informative to identify diatom species (Kermarrec *et al.* 2013), and to our knowledge has not been broadly applied in species discovery yet.

Metabarcoding is also useful for microalgae discovery in benthic habitats, which have received little attention when compared to planktonic environments (Forster *et al.* 2016). A high diversity of undescribed siphonous green algae was identified in coral samples surveyed with the 16S rRNA gene, which is commonly used to characterize bacterial communities but can also detect chloroplast DNA (del Campo *et al.*, 2017). Metabarcoding studies using the *tuf*A gene (commonly used as a DNA barcode in green algae; Saunders & Kucera, 2010) to target endolithic algae in tropical marine limestone samples revealed over 100 undescribed algal OTUs near the species level, including several new lineages of green algae (Marcelino & Verbruggen, 2016; Sauvage *et al.*, 2016). We expect that the application of metabarcoding in algal turfs and other habitats with small and/or morphologically feature-poor multicellular algae will yield similar discoveries.

In addition to the 16S rRNA, *rbc*L and *tuf*A genes, the universal plastid amplicon (UPA, a fragment of the plastid 23S rRNA gene) is also suitable for environmental sequencing with the ability to identify both eukaryotic algae and cyanobacteria (Sherwood & Presting, 2007). Biodiversity assays using this metabarcode retrieved a large number of algal lineages, many of which may constitute new species (Steven *et al.,* 2012; Marcelino & Verbruggen, 2016; Sherwood *et al.*, 2016). As with 16S rRNA, a downside to using UPA is that it is very conserved and does not always permit distinction between closely related species (Saunders & Kucera, 2010). All potential metabarcoding loci will have advantages and disadvantages, and it has been shown that a combination of multiple metabarcodes, which is achievable with amplicon HTS, gives the most comprehensive insights into algal diversity (Marcelino & Verbruggen, 2016).

It is possible to sequence environmental DNA without amplifying specific markers and to identify species based on *de novo* genome assemblies, a technique known as metagenomics. In a recent study,

over 7,900 uncultivated prokaryotic species have been discovered using this technique (Parks *et al.*, 2017). Likewise, chloroplast and mitochondrial genomes can be retrieved from metagenomic data (e.g. Worden *et al.*, 2012). Linking metagenomic sequences to eukaryotic species is still challenging and therefore this technique has not been broadly applied to algal species discovery yet.

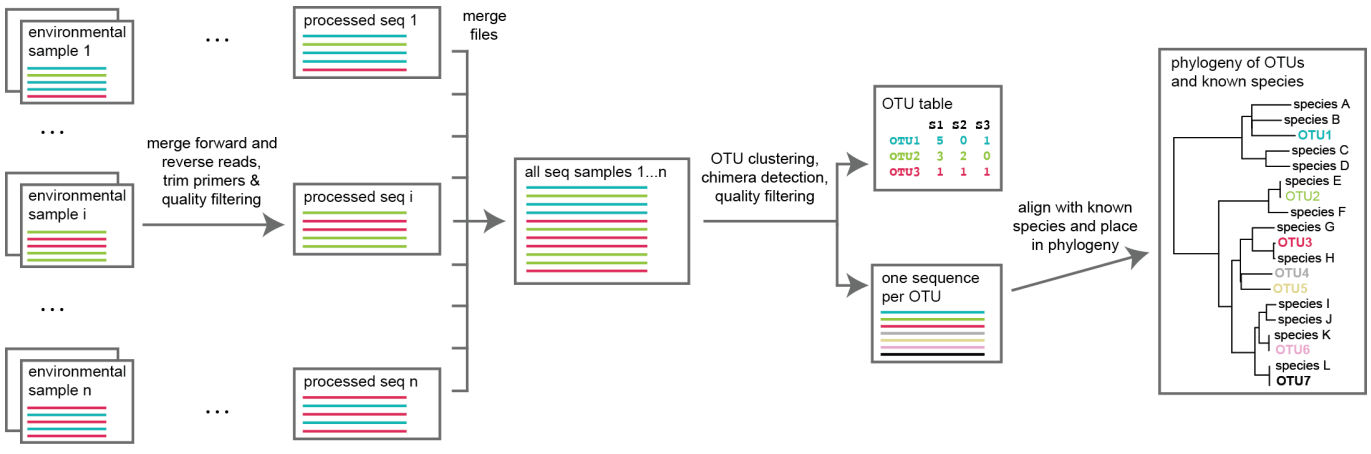
It is important to note that metabarcoding and metagenomics do not characterise newly discovered species beyond the obtained sequence, and that much more work will be needed if these species are to be named formally. The great strength of metabarcoding lies in its potential to give us a better understanding of the true magnitude of algal biodiversity and the distribution of species, whether they have been described or not, in different habitats across the planet.

Box 2: Bioinformatics for Species discovery

The concepts and analysis workflow for HTS-based species discovery are quite different from other applications described in this paper. In this case the samples are environmental samples, and the HTS data from each sample contains the PCR-amplified marker sequences of the multitude of organisms in that environmental sample. After the forward and reverse paired-end reads have been overlapped and merged with one another, forming a consensus sequence, the PCR primers are trimmed off and quality filtering is carried out, which includes verification of the amplicon length, sequence quality and whether the sequence is in fact the target gene. The processed sequences from different samples are then pooled, keeping track of which sample they belong to.

The next step involves clustering of the sequences into bins based on an identity threshold. These bins correspond to operational taxonomic units (OTUs). The clustering methods typically start with an empty database of OTU sequences, and then add sequences one at a time. If the sequence to be added matches with one already in the OTU database within the predefined identity threshold (often 97%) it is added to that OTU, otherwise it is used to define a new OTU. Good OTU clustering programs will include algorithms to detect possible chimeras formed during the PCR and a number of other quality control steps. The output from the program includes an alignment with a representative sequence for each OTU as well as an OTU table specifying the abundance of each OTU in each of the environmental samples.

For species discovery, the OTU sequences can then be aligned with sequences of known species. The phylogeny inferred from this alignment allows the assessment of which OTUs match with known species (OTUs 2,3,6,7 in figure) and which could be newly discovered diversity (OTUs 1,4,5 in figure).



If a combination of different markers is used, a few extra steps are needed to sort reads by marker, and the OTU clustering and phylogenetic inference are performed for each marker separately (e.g. Marcelino & Verbruggen, 2016).

Several established workflows are capable of performing all fundamental bioinformatic steps, taking raw data through to taxonomic assignment and alignment generation. These include QIIME (pronounced "chime"; Caporaso *et al.*, 2010 b), MOTHUR (Schloss *et al.*, 2009), BioMaS (Fosso *et al.*, 2015) and UPARSE (Edgar, 2013). Stand-alone programs also exist for different steps in the workflow. Quality control and preparation of OTUs for clustering can be achieved using FLASH (Magoč & Salzberg, 2011), PRINSEQ (Schmieder & Edwards, 2011) and PyNAST (Caporaso *et al.*, 2010 a). UCHIME in USEARCH (Edgar *et al.*, 2011) can perform de novo chimera detection and removal. The UPARSE workflow (Edgar, 2013), among others, can be used to construct OTUs de novo from sequence reads and cluster OTUs based on a defined similarity threshold, as well as to perform chimera filtering. PyNAST, MAFFT (Katoh *et al.*, 2012), USEARCH and TANGO (Clemente *et al.*, 2010; Clemente *et al.*, 2011) can be used for alignment and taxonomic assignment of OTUs. For diversity and phylogenetic analyses, OTUs can be aligned with reference sequences of known species using alignment programs such as MUSCLE (Edgar, 2004), MAFFT and Geneious (http://www.geneious.com, Kearse *et al.* 2012).

HTS for species delimitation

Defining species boundaries is a major goal of systematic biology, but it can be excessively difficult for taxa lacking morphological diagnostic characters or for groups that have recently diverged and have yet to accumulate diagnostic features. Algae are known to be taxonomically challenging because of their structural simplicity (i.e. lack of morphological characters), morphological plasticity and high levels of cryptic diversity (Verbruggen, 2014). Because of these issues, it has become common practice to define algal species boundaries with molecular data, with the identification of morphological diagnostic features required to achieve an integrative taxonomy generally occurring afterwards. We refer to Leliaert *et al.* (2014) for a more in-depth review of DNA-based species delimitation and its application to algae.

Sanger-sequenced single-locus datasets still dominate studies of algal species delimitation, despite multi-locus methods being superior in sensitivity and accuracy (Dupuis *et al.*, 2012; Leliaert *et al.*, 2014). One HTS technique that can target amplicons commonly used in single-locus studies is two-step PCR (see Box 1), where multiple amplicons can be generated simultaneously for a large number of specimens (Cruaud *et al.*, 2017; Gohl *et al.,* 2016). This approach allows one to take advantage of amplicon sequences in current databases (e.g. Genbank, BOLD), and along with the use of familiar PCR methods this arguably makes two-step PCR an attractive option. On the downside, it requires reference data to design primers, and significant time investment is sometimes required to optimise and perform PCRs. Moreover, the two-step PCR approach yields short amplicons (corresponding to twice HTS sequencing read lengths, 500nt maximum with current MiSeq technology). Together with the vastly reduced number of loci that can be practically recovered in comparison to RADseq and target-capture (see below), this

means that the amount of data recovered is relatively small. Consequently, two-step PCR does not leverage the ability of HTS to produce large-scale multi-locus data.

Simulations have shown that species delimitation using a multispecies coalescent model which accounts for incomplete lineage sorting increases in accuracy with the number of unlinked loci (Jones *et al.*, 2015). Plastid data provide linked loci, as do mitochondrial data, as these genomes are inherited without undergoing genetic recombination. Unlinked loci from the nuclear genome are therefore best for species delimitation, and HTS offers great potential to generate such multi-locus datasets in a fast and cost-effective manner. Importantly, HTS can also recover data from multiple genomic compartments (nucleus, plastid, mitochondrion) simultaneously (Amaral-Zettler *et al.* 2016), without many of the technical challenges of classical PCR and Sanger sequencing methods.

Although standard shotgun HTS can be employed to generate multi-locus HTS data for species delimitation, several modified HTS methods are arguably better-suited for this purpose. One such approach is RADseq, which has been used recently to infer robust species boundaries in corals (Herrera & Shank, 2016), skinks (Rittmeyer & Austin, 2015), alpine plants (Boucher *et al.*, 2016) and macroalgae (Fraser *et al.*, 2016; Montecinos, 2016). RADseq is an attractive method for non-model taxa, and hence useful for algae, because it does not require prior genomic information. Additionally, owing to the fact that sequences recovered from RADseq are pre-selected via restriction sites, the coverage per locus is greater than whole-genome sequencing as well as more cost-effective. RADseq loci include both coding and non-coding regions from diverse genomic contexts and histories. However, the same RADseq loci dataset often cannot be fully recovered across divergent species due to mutations at restriction sites (Rubin *et al.*, 2012; Huang & Knowles, 2016) or variation in sequence coverage. DNA requirements (quality and quantity) are another drawback of this approach which may limit its applicability.

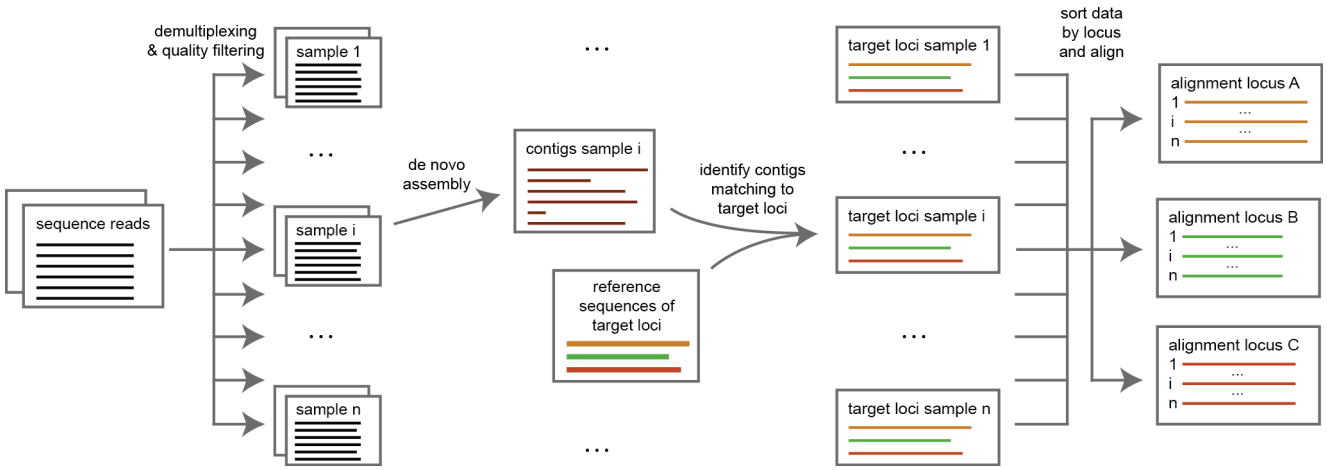
Like RADseq, target capture also has the potential to yield many unlinked loci, and it has been used in species delimitation of, for example, fish (Song *et al.*, 2017) and ants (Ješovnik *et al.*, 2017). A major difference between these approaches is that target capture produces data for known selected markers, yielding large datasets for comparative studies. However, the characteristics that make target capture appealing also have their disadvantages. In order to design capture oligonucleotides prior knowledge of the genome is required, which is not always possible. This can be obtained from a transcriptome, or probes of highly conserved genomic regions can be used. The more complex library preparation and the production of the probes also make this approach more expensive.

In the field of algal species delimitation, HTS has not yet been widely adopted. Sanger sequencing of standard molecular markers have been used efficiently for this purpose (e.g. Guillemin *et al.*, 2016; Montecinos *et al.*, 2017). However, several phycological studies have already proven its utility in cases of recent or incipient speciation. Fraser *et al.* (2016) used a type of RADseq to identify recent speciation in

Durvillaea (bull kelp), whereas Amaral-Zettler *et al.* (2016) used HTS sequencing of mitochondrial and plastid genomes to delineate species of the brown macroalga *Sargassum*. Similarly, a RADseq approach has been used for the model brown-algal genus *Ectocarpus* (Montecinos, 2016). Even though very few studies have used HTS-based approaches for species delimitation in algae, its success in other groups of organisms suggests that it could become the new standard method in algal species delimitation.

Box 3: Bioinformatics for species delimitation

The data analysis workflow for species delimitation converts the raw HTS data, which comprises sequence reads of multiple samples and multiple loci, into alignments of individual loci. There are some differences depending on which type of approach is used. The figure shows an analysis workflow that could work with target capture, multi-locus PCR amplified loci and transcriptome data.



The sequence reads are first demultiplexed, with reads belonging to different samples separated into files based on their index sequence. Each file undergoes quality filtering to remove low quality reads and trim off the lower-quality ends of reads. Reads for each sample are then assembled into longer contigs or scaffolds. The contigs matching the target loci are identified based on similarity searches using a set of previously published reference sequences from those target loci. Once the target loci have been determined for every sample, the data get reshuffled, with all sequences of any given locus from different samples put together in locus-specific files, which are then aligned. These alignments are the end product of the workflow and can be used as the input for a range of species delimitation tools.

Many widely available programs can be used to perform the steps in this workflow, including CLC Genomics Workbench (www.qiagenbioinformatics.com/) and Geneious, which are capable of performing de novo assembly and alignment functions. For a comprehensive list of freely available assembly and read mapping programs please refer to *Table 2* in Song *et al.* (2016). BLAST, which is available in both Geneious and CLC Genomics Workbench as well as stand-alone software that can be downloaded and run on a local computer, can be used for similarity searches to compare contigs and reference sequences.

For RADseq data, the HTS reads are demultiplexed and filtered for quality as above. Because these loci are anonymous and no reference sequences are thus available, the remainder of the workflow differs. For RADseq, PyRAD (Eaton, 2014) and Stacks (Catchen *et al.*, 2013) are the two most common programs that can implement the entire workflow from start to finish. For each sample, highly similar sequences sharing a given sequence similarity are grouped (clustered) into different loci called "stacks". Each stack is summarised into a consensus sequence, and for heterozygous loci both alleles are reported in the result.

The resulting consensus sequences are clustered between individuals to generate one data matrix per locus, which is then aligned. This is the same output as for the workflow described above, and the resulting alignments can be fed into species delimitation software.

BPP (Yang & Rannala, 2014), the DISSECT package (Jones *et al.*, 2015) for BEAST (Drummond *et al.*, 2012), and recently the package STACEY (Jones, 2017) for BEAST2 (Bouckaert *et al.*, 2014) are examples of programs that make use of the unlinked multi-locus data produced to establish species delimitation. BEAST2 also has useful packages SNAPP (Bryant *et al.*, 2012) and *BEAST (Heled & Drummond, 2010) to infer species phylogeny and population sizes.

HTS of historical type specimens to assist nomenclature

Once DNA-based species boundaries are established, appropriate species names have to be selected. Linking species recovered from molecular studies to type specimens – the samples on which species descriptions are based and where the taxonomic name is anchored --– can be a difficult task when only morphological and/or anatomical features were used in the original species description. Accessing the DNA of historical type specimens, including those of morphologically similar old species names now considered synonyms, is the most obvious solution to link the various DNA-based entities to named species (e.g. Hughey *et al.*, 2001, 2002; Hayden *et al.*, 2003; Gabrielson, 2008 a, 2008 b; Saunders & McDevit, 2012; Hind *et al.*, 2014; Sissini *et al.*, 2014; Hernandez-Kantun *et al.*, 2015; Lindstrom *et al.*, 2015; Vieira *et al.*, 2016). The biggest challenges with this approach are to deal with the highly fragmented and tiny quantities of DNA preserved in type specimens, and to avoid contamination from more recent collections with abundant DNA used in the same laboratory.

The time since collection and the preservation method significantly affect DNA degradation and the quantity of DNA that can be recovered (Särkinen *et al.*, 2012; Choi *et al.*, 2015). Amplification of historical DNA via PCR is likely to pick up even minute contaminations of contemporary undegraded DNA (Saunders & McDevit, 2012). Measures commonly used in handling historical DNA can be adopted to avoid contamination of type specimens with DNA of contemporary samples. Minimum standards to work with DNA of historical samples include: i) perform DNA extractions and PCR preparation in a dedicated work area where no modern samples are handled; ii) use multiple mock extractions and PCR negative controls, while positive controls should be avoided as they provide an additional source of contamination; iii) sterilize tools (e.g. pipettes) and working space, using bleach and/or UV light (Cooper & Poinar, 2000; Brown & Brown, 2011; Llamas *et al.*, 2017).

PCR is the Achilles' heel of this process, and therefore amplification-free approaches facilitated by HTS significantly reduce the risks of contamination (Saunders & McDevit, 2012; Bakker, 2017). DNA barcodes and even complete organellar genomes can be obtained from whole-genome shotgun sequencing of type specimens (Box 4, Hughey & Gabrielson, 2012; Staats *et al.*, 2013; Prosser *et al.*, 2016; Yeates *et al.*, 2016). High-throughput sequencing data has been obtained from historical specimens (type,

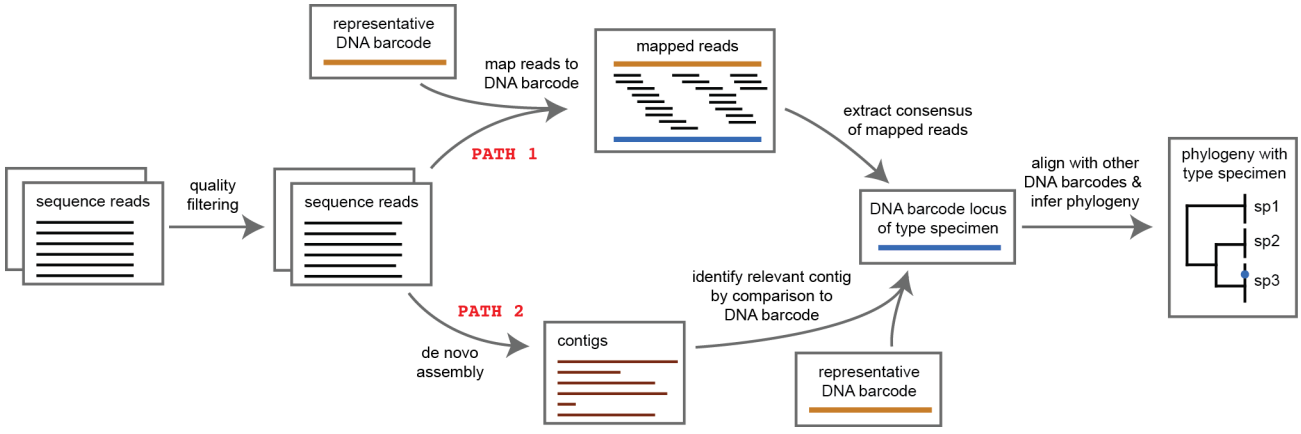
holotype, topotype) to reassess the taxonomic classification of *Pyropia perforata* (Hughey *et al*., 2014), *Otohimella* gen. nov. (Suzuki *et al.,* 2016), three *Membranoptera* species (Hughey *et al*., 2017) and ten species in the genera *Gelidium* and *Pterocladiella* (Boo *et al.*, 2016).

Whether we should aim to provide every species with a Latin binomial name is up for debate (De Clerck *et al.,* 2013; Casiraghi *et al.,* 2016). A scientific name and the correct assignment of cryptic species to type specimens are useful, for example, to coordinate conservation efforts and track invasive species (e.g. Payo *et al.,* 2013; Belton *et al.,* 2014). However, there is uncertainty in the formal name of many old algal species, and the pace at which new species-level lineages are discovered largely exceeds the rates at which species can be formally described. Consequently, there is a tendency in phycology and other disciplines to use strain or voucher names to transmit taxonomic information (De Clerck *et al.,* 2013; Page, 2016).

Box 4: Bioinformatics for HTS of historical type specimens to assist the nomenclature process

The data generated for historical specimens will usually be whole genome shotgun (WGS) sequencing reads. The goal of the workflow in this case is to obtain the sequence for a DNA barcode locus from these WGS reads.

Following quality filtering, the corrected reads can be processed in two different ways. The first option (Path 1 in figure) maps the reads to a pre-determined DNA barcode from the same or a related species based on sequence similarity, and the consensus sequence of the mapped reads is then extracted and used to infer its phylogenetic position among a set of other DNA barcodes. The second option (Path 2) is to assemble the reads *de novo* and use a representative pre-determined DNA barcode to identify the contig that corresponds to the DNA barcode locus. The relevant part of the contig is then extracted, aligned with other DNA barcodes and used to determine which species-level cluster it belongs to.



CLC Genomics Workbench and Geneious as well as some freely available programs summarised in Song *et al.* (2016) can be used for assembly, read mapping and alignment. BLAST is typically used to identify target contigs by comparison to reference sequences.

HTS for higher-level classification

Molecular data have revolutionized our understanding of algal systematics and are now essential to the proposal of any classification scheme. Most algal studies use only one or a few molecular markers, although phylogenies based on more extensive datasets have been published (e.g., Verbruggen *et al.*, 2009; Silberfeld *et al.*, 2010; Nozaki *et al.*, 2014; Ruck *et al.*, 2016; Yang *et al.*, 2016). A common problem in phylogenies based on one or a few genes is the poor resolution of phylogenetic relationships, which causes uncertainty about the monophyly of some taxa and limits our ability to advance towards a natural classification of algal lineages. Low branch support is often explained by the scarcity of molecular data and conflicting signals among markers (the so-called soft polytomies) but could in some cases be hard polytomies representing rapid radiation of lineages (e.g., Reviers & Rousseau, 1999; Verbruggen *et al.,* 2010). HTS allows us to overcome limitations of data quantity by sequencing organellar genomes and/or transcriptomes for a large number of taxa, producing large-scale molecular datasets in a rapid and cost-effective manner. HTS data has resolved challenging phylogenies and classification issues for plants (Ma *et al.,* 2014; Lu *et al.,* 2015), animals (Finstermeier *et al*., 2012; Prum *et al*. 2015) and protists (Cavalier-Smith *et al*., 2015; Kang *et al*., 2017).

The number of nuclear and organellar genomes and the amount of transcriptome data available for algae are still low compared to plants and animals. The few algal studies employing HTS mostly focus on describing the structure of organellar genomes, and phylogenies, if present, contain relatively few species (e.g. Jeong *et al.*, 2014; Zhang *et al.,* 2015; An *et al.*, 2016; Sěvčıková *et al.*, 2016; McManus *et al.,* 2017). Nonetheless, these studies demonstrate the high potential of HTS data to produce well resolved algal phylogenies (Box 5, Janouškovec *et al.* 2013; Wang *et al.,* 2013; Lemieux *et al.*, 2014, 2015; Yang *et al.*, 2015; Villain *et al.*, 2017).

The number of studies applying HTS data to address algae classification is steadily growing. Muñoz-Gómez *et al.* (2017) sequenced six plastid genomes for mesophilic non-seaweed red algae with uncertain relationships. The deeper nodes in the phylogeny of Rhodophyta were resolved based on chloroplast genomes and this group of six species formed a monophyletic clade, for which they proposed the new subphylum Proteorhodophytina. Leliaert *et al.* (2016) used chloroplast genome data to establish a new class Palmophyllophyceae (Chlorophyta) for a green algal lineage whose affinities remained uncertain. Verbruggen *et al.* (2017) used chloroplast phylogenomics to clarify the relationships of *Ostreobium* within the siphonous green algae (Ulvophyceae), for which a new suborder was proposed. Costa *et al.* (2016) sequenced the chloroplast genomes of 22 species of the red algal order Nemaliales, resolving the placement of several previously contradictory clades and recognizing two new suborders and six families. Further, Díaz-Tapia *et al.* (2017) sequenced 52 chloroplast genomes from

Rhodomelaceae (Rhodophyta) resolving a well supported phylogeny where five new tribes were recognized.

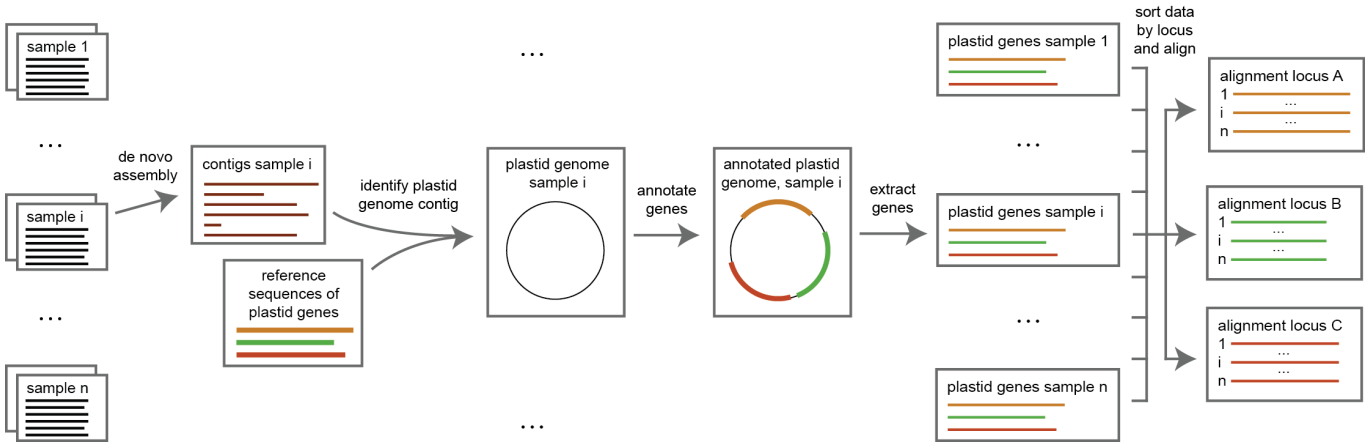
HTS transcriptome data have also been used to address phylogenetic questions (e.g. Jeong *et al.*, 2014; Sun *et al.*, 2014). Derelle *et al.* (2016) combined transcriptome and genome data to resolve phylogenetic relationships among lineages of the Stramenopiles, confirming the monophyly of the two major groups Bigyra and Gyrista and proposing a new classification within Chyrista and Diatomista. Bachvaroff *et al.* (2014) constructed a phylogeny for the alveolates, resolving uncertainties and conflicts regarding basal lineages of the dinoflagellates. Janouškovec *et al.* (2017) analyzed the core dinoflagellates producing a well-resolved phylogeny that demonstrates the monophyly of thecate taxa and notices the paraphyly of Gymnodiniales, providing the basis for future classification decisions. Fučíková *et al.* (2014) combined chloroplast genes, whole genome shotgun sequencing and transcriptome data to address the phylogeny of classes Trebouxiophyceae and Ulvophyceae. Moreover, analyses by Jackson *et al.* (2017) demonstrated that for brown algae, large-scale transcriptome data had superior resolving power compared to the previous most marker-rich study involving an eight-gene alignment, leading to changes in the classification of genera.

These examples demonstrate that HTS provides an invaluable source of data for producing well resolved algal trees in a variety of phyla and at different taxonomic levels, often in situations where smaller datasets were inconclusive. These phylogenetic trees have been key to assessing the monophyly of certain taxa and resolving phylogenetic relationships, leading to solutions to persistent issues in classification and the description of new higher taxa at many taxonomic levels.

Box 5: Bioinformatics for higher-level classification

The input data for this type of application consist of sequence reads for a range of samples. The output is a set of alignments, one per locus, each containing the sequences for all the samples. Depending on the type of sequencing used for this purpose, the workflow will vary somewhat. For DNA capture and transcriptome data, the workflow described in Box 3 can be used to obtain the alignments.

Very often, however, whole genome shotgun sequencing is performed with the goal to obtain data from plastid and mitochondrial genomes for phylogenetic analysis. The figure below shows a typical data analysis workflow for this scenario, using the plastid genome as an example.



Starting from sequence reads that, for the sake of the illustration, have already been demultiplexed and quality checked (see Box 3), *de novo* assembly is performed for every sample. Using a set of predetermined plastid genes (e.g. from a related species downloaded from GenBank), the contig corresponding to the plastid genome is identified in the assembly file. The position of genes is subsequently annotated on that sequence, and the sequences of all the genes get stored into one dataset per species. For the sake of the illustration we have assumed that the assembly yields a complete genome that is depicted as circular. But even if it is fragmented, the contigs can still be identified, annotated and the genes from different contigs combined downstream.

In the final step, the sample-specific datasets get reshuffled to yield one dataset per gene, each containing the relevant sequence of all samples, and those datasets are aligned. These aligned datasets are the end product of the workflow, as the next steps are variable. For instance, one could infer phylogenies from each gene separately, or concatenate the gene alignments to infer a phylogeny from that.

The programs listed in Boxes 3 and 4 also apply for this workflow. Table 3 from Song *et al.* (2016) has a summary of programs for gene prediction and annotation.

Strategies

Many strategic choices need to be made in the application of HTS to questions in systematics. How much of the work will be outsourced to the sequencing provider and what can be done in-house? How much data should be obtained from each sample? What is the most suitable platform for the different applications?

Using HTS in systematics is not excessively expensive, especially when considering how much data it delivers and how much more conclusive the results typically are. It is difficult to be specific about prices because they differ widely between providers, change with time, and depend on the needs of the project (platform, amount of data, etc.). At the time of writing, one could get a gigabase of Illumina HiSeq data for about USD $20. The library preparation that has to happen prior to sequencing is usually more expensive than the sequencing itself. Outsourcing it to a sequencing provider costs ca. USD $120 per library, but price reductions can often be negotiated for larger numbers of samples. For example, sequencing 2.5 Gb of WGS data for each of 96 samples can currently be done for under USD $10,000,

and this would yield more than enough data to assemble organelle genomes as well as other high-abundance DNA such as nuclear ribosomal RNA genes.

A very important consideration is whether to outsource aspects of the work or perform them in-house. For example, it is possible to save on library preparation costs by carrying out the procedure in-house. This is quite cheap in terms of consumables (ca. USD $30 per sample), but comes at a significant cost in terms of labour and equipment (e.g. sonicator, real-time PCR, fragment analyser), so it may not be a suitable choice for all labs.

Using HTS also requires more computational skills and resources than pre-HTS molecular taxonomic practices. From the equipment side of things, a multi-core server is very useful to analyse HTS datasets. Some laboratories will be able to access such computational resources through their institution or from government-provided supercomputer facilities, or one can purchase access to cloud computing (e.g. Amazon EC2). Most labs will also want to invest in network-attached storage to keep their data safe. As with any other skill, learning HTS analysis takes time, but many excellent intensive courses are offered, and there is a vast amount of information available online.

Similar to data generation, it is also possible to outsource parts of the analysis workflow. For example, it is not uncommon for sequencing providers to offer de novo assembly of WGS data or QIIME analysis of 16S amplicon data. While this can help to avoid some costs (e.g. server purchase), at the time of writing one should not expect that computational aspects can be completely outsourced. Collaboration with laboratories who already have experience with HTS data and have access to the computational resources is probably one of the best strategies available to groups that want to start using HTS in their work.

The computational skill set needed for HTS-centred systematics is also quite different from that required when only one or a few genes need to be analysed. Because of the massively large data sets generated, it is not possible to apply the same level of manual curation and tuning of the dataset, and one relies much more on automated methods. To illustrate the difference between them: the experience of manually editing a single-gene alignment in an alignment editor is quite different from the equivalent process with HTS data, which would involve writing a script that loops through hundreds of datasets and, for each of them, calls a program to align the sequences, another program that checks the quality of the alignment, and possibly a program that removes unreliably aligned positions based on some predefined thresholds of sequence conservation.

The developments in the methodology, the reduction of cost and the use of standardized workflows will enable the use of HTS as a standard tool for algal systematics, just as the use of Sanger sequencing and phylogenetic methods became the standard in this field two decades ago. HTS technology has revolutionized many fields in biology, and will certainly do the same in algal systematics.

Acknowledgments

This work resulted from a SPRINT bilateral grant between FAPESP (2015/50078-1) and the University of Melbourne. MCO is grateful to CNPq (301491/2013-5; 406351/2016-3). PDT acknowledges support by the postdoctoral programmes Axudas de apoio á etapa inicial de formación posdoutoral do Plan I2C (Xunta de Galicia). The Verbruggen lab received funding from the Australian Research Council (DP150100705) and the Australian Biological Resources Study (RFL213-08).

References

Amaral-Zettler, L.A., Dragone, N.B., Schell, J., Slikas, B., Murphy, L.G., Morrall, C.E. & Zettler, E. R. (2016). Comparative mitochondrial and chloroplast genomics of a genetically distinct form of *Sargassum* contributing to recent "Golden Tides" in the Western Atlantic. *Ecology and Evolution*, **7**: 516–525.

An, S.M., Noh, J.H., Lee, H.R., Choi, D.H., Lee, J.H. & Yang, E.C. (2016). Complete mitochondrial genome of biraphid benthic diatom, *Navicula ramosissima* (Naviculales, Bacillariophyceae). *Mitochondrial DNA Part B*, **1**: 549-550.

Bachvaroff, T.R., Gornik, S.G., Concepcion, G.T., Waller, R.F., Mendez, G.S., Lippmeier, J.C. & Delwiche, C.F. (2014). Dinoflagellate phylogeny revisited: Using ribosomal proteins to resolve deep branching dinoflagellate clades. *Molecular Phylogenetics and Evolution*, 70: 314–322.

Bakker, F.T. (2017). Herbarium genomics: skimming and plastomics from archival specimens. *Webbia: Journal of Plant Taxonomy and Geography*, **72**: 35–45.

Belton, G.S., Prud'homme van Reine, W.F., Huisman, J.M., Draisma, S.G.A. & Gurgel, C.F.D. (2014). Resolving phenotypic plasticity and species designation in the morphologically challenging *Caulerpa racemosa–peltata* complex (Chlorophyta, Caulerpaceae). *Journal of Phycology*, **50**: 32-54.

Beszteri, B., John, U. & Medlin, L.K. (2007). An assessment of cryptic genetic diversity within the *Cyclotella meneghiniana* species complex (Bacillariophyta) based on nuclear and plastid genes, and amplified fragment length polymorphisms. *European Journal of Phycology*, **42**: 47–60.

Bittner, L., Gobet, A., Audic, S., Romac, S., Egge, E.S., Santini, S., Ogata, H., Probert, I., Edvardsen, B. & de Vargas, C. (2013). Diversity patterns of uncultured Haptophytes unravelled by pyrosequencing in Naples Bay. *Molecular Ecology*, **22**: 87–101.

Boo, G.H., Hughey, J.R., Miller, K.A. & Boo, S.M. (2016). Mitogenomes from type specimens, a genotyping tool for morphologically simple species: ten genomes of agar-producing red algae. *Scientific Reports*, **6**: 35337.

Boucher, F.C., Casazza, G., Szövényi, P. & Conti, E. (2016). Sequence capture using RAD probes clarifies phylogenetic relationships and species boundaries in *Primula* sect. *Auricula*. *Molecular Phylogenetics and Evolution*, **104**: 60–72.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M., Rambaut, A. & Drummond, A.J. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLOS Computational Biology*, **10**: e1003537.

Brown, T. A. & Brown, K. (2011). The Technical Challenges of Biomolecular Archaeology. In *Biomolecular Archaeology: An Introduction* (Brown, T. A. & Brown, K., editors), 136-148. Wiley-Blackwell, Oxford, UK.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. & RoyChoudhury, A. (2012). Inferring Species Trees Directly from Biallelic Genetic Markers: Bypassing Gene Trees in a Full Coalescent Analysis. *Molecular Biology and Evolution*, **29**: 1917–1932.

Caporaso, J., Bittinger, K., Bushman, F., DeSantis, T., Andersen, G. & Knight, R. (2010a). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics*, **26**: 266–267.

Caporaso, J., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F., Costello, E., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J. & Knight, R. (2010b) QIIME allows analysis of high-throughput community sequencing data. *Natute Methods,* **7**: 335–336.

Casiraghi, M., Galimberti, A., Sandionigi, A., Bruno, A. & Labra, M. (2016). Life with or without names. *Evolutionary Biology*, **43**: 582–595.

Catchen, J., Hohenlohe, P., Bassham, S., Amores, A. & Cresko, W. (2013). Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22**: 3124–3140.

Cavalier-Smith, T., Chao, E.E. & Lewis, R. (2015). Multiple origins of Heliozoa from flagellate ancestors: New cryptist subphylum Corbihelia, superclass Corbistoma, and monophyly of Haptista, Cryptista, Hacrobia and Chromista. *Molecular Phylogenetics and Evolution,* **93**: 331–362.

Choi, J., Lee, H. & Shipunov, A. (2015). All that is gold does not glitter? Age, taxonomy, and ancient plant DNA quality. *PeerJ,* **3**: e1087.

Clemente, J., Jansson, J. & Valiente, G. (2011). Flexible taxonomic assignment of ambiguous sequencing reads. *BMC Bioinformatics*, **12**: 8.

Clemente, J.C., Jansson, J. & Valiente, G. (2010). Accurate taxonomic assignment of short pyrosequencing reads. *Pacific Symposium on Biocomputing,* **15**: 3–9.

Cooper, A. & Poinar, H.N. (2000). Ancient DNA: Do it right or not at all. *Science*, **289**: 1139.

Costa, J.F., Lin, S.M., Macaya, E.C., Fernández-García, C. & Verbruggen, H. (2016). Chloroplast genomes as a tool to resolve red algal phylogenies: a case study in the Nemaliales. *BMC Evolutionary Biology,* **16**: 205.

Costello, M.J. & Wilson, S.P. (2011). Predicting the number of known and unknown species in European seas using rates of description. *Global Ecology and Biogeography* **20**: 319–330.

Cruaud, P., Rasplus, J.-Y., Rodriguez, L.J. & Cruaud, A. (2017). High-throughput sequencing of multiple amplicons for barcoding and integrative taxonomy. *Scientific Reports*, **7**: 41948.

De Clerck, O., Guiry, M.D., Leliaert, F., Samyn, Y., Verbruggen, H. (2013) Algal taxonomy: a road to nowhere? *Journal of Phycology,* **49**: 215–225.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.-M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., Flegontova, O., Guidi, L., Horak, A., Jaillon, O., Lima-Mendez, G., Luke, J., Malviya, S., Morard, R., Mulot, M., Scalco, E., Siano, R., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Acinas, S.G., Bork, P., Bowler, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Raes, J., Sieracki, M.E., Speich, S., Stemmann, L., Sunagawa, S., Weissenbach, J., Wincker, P. & Karsenti, E. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, **348**: 1261605-1/11.

del Campo, J., Pombert, J.F., Slapeta, J., Larkum, A. & Keeling, P.J. (2017). The 'other' coral symbiont: *Ostreobium* diversity and distribution. *The ISME Journal*, **11**: 296–299.

DeLong, E.F. (2009). The microbial ocean from genomes to biomes. *Nature,* **459**: 200–206.

Derelle, R., López-García, P., Timpano, H., Moreira, D. (2016). A phylogenomic framework to study the diversity and evolution of Stramenopiles (=Heterokonts). *Molecular Biology and Evolution*, **33**: 2890–2898.

Díaz-Tapia, P., Maggs, C.A., West, J.A. & Verbruggen, H. (2017). Analysis of chloroplast genomes and a supermatrix inform reclassification of the Rhodomelaceae (Rhodophyta). *Journal of Phycology*, XXXXX.

Dijk, E., Auger, H., Jaszczyszyn, Y. & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics*, **30**: 418–426.

Drummond, A., Suchard, M., Xie, D. & Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, **29**: 1969–1973.

Dupuis, J.R., Roe, A.D. & Sperling, F.A.H. (2012) Multi-locus species delimitation in closely related animals and fungi: One marker is not enough. *Molecular Ecology*, **21**: 4422–4436.

Eaton, D. (2014). PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics*, **30**: 1844–1849.

Edgar, R. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, **32**:1792–1797.

Edgar, R. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, **10**: 996–998.

Edgar, R., Haas, B., Clemente, J., Quince, C. & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, **27**: 2194–2200.

Egge, E.S., Johannessen, T.V., Andersen, T., Eikrem, W., Bittner, L., Larsen, A., Sandaa, R.A. & Edvardsen, B. (2015). Seasonal diversity and dynamics of haptophytes in the Skagerrak, Norway, explored by high-throughput sequencing. *Molecular Ecology,* **24**: 3026–3042.

English, A., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D.M., Reid, J.G., Worley, K.C. & Gibbs, R.A. (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE*, **7**: e47768.

Finstermeier, K., Zinner, D., Brameier, M., Meyer, M., Kreuz, E., Hofreiter, M. & Roos, C. (2012). A mitogenomic phylogeny of living primates. *PLoS ONE*, **8**: e69504.

Forster, D., Micah, D., Mahé, F., Dolan, J.R., Audic, S., Bass, D., Bittner, L., Boutte, C., Christen, R., Claverie, J.-M., Decelle, J., Edvardsen, B., Egge, E., Eikrem, W., Gobet, A., Kooistra, W.H.C.F., Logares, R., Massana, R., Montresor, M., Not, F., Ogata, H., Pawlowski, J., Pernice, M.C., Romac, S., Shalchian-Tabrizi K., Simon, N., Richards, T.A., Santini, S., Sarno, D., Siano, R., Vaulot, D., Wincker, P., Zingone, A., de Vargas, C. & Stoeck, T. (2016). Benthic protists: the under-charted majority. *FEMS Microbiology Ecology*, **92**: fiw120.

Fosso, B., Santamaria, M., Marzano, M., Alonso-Alemany, D., Valiente, G., Donvito, G., Monaco, A., Notarangelo, P. & Pesole, G. (2015). BioMaS: a modular pipeline for Bioinformatic analysis of Metagenomic AmpliconS. *BMC Bioinformatics*, **16**: 203.

Fraser, C.I., McGaughran, A., Chuah, A. & Waters, J.M. (2016). The importance of replicating genomic analyses to verify phylogenetic signal for recently evolved lineages. *Molecular Ecology*, **25**: 3683–3695.

Fučíková, K., Lewis, P.O. & Lewis, L. A. (2014). Putting *incertae sedis* taxa in their place: a proposal for ten new families and three new genera in Sphaeropleales (Chlorophyceae, Chlorophyta). *Journal of Phycology*, **50**: 14–25.

Gabrielson, P.W. (2008a). Molecular sequencing of Northeast Pacific type material reveals two earlier names for *Prionitis lyallii*, *Prionitis jubata* and *Prionitis sternbergii*, with brief comments on *Grateloupia versicolor* (Halymeniaceae, Rhodophyta). *Phycologia,* **47**: 89–97.

Gabrielson, P.W. (2008b). On the absence of previously reported Japanese and Peruvian species of *Prionitis* (Halymeniaceae, Rhodophyta) in the northeast Pacific. *Phycological Research,* **56**: 105–114.

Glenn, T. (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, **11**: 759–769.

Gohl, D.M., Vangay, P., Garbe, J., MacLean, A., Hauge, A., Becker, A., Gould, T.J., Clayton, J.B., Johnson, T.J., Hunter, R., Knights, D. & Beckman, K.B. (2016) Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nature Biotechnology*, **34**: 942–949.

Goodwin, S., McPherson, J.D. & McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, **17**: 333–351.

Gran-Stadniczeñko, S., Šupraha, L., Egge, E.D. & Edvardsen, B. (2017) Haptophyte Diversity and Vertical Distribution Explored by 18S and 28S Ribosomal RNA Gene Metabarcoding and Scanning Electron Microscopy. *Journal of Eukaryotic Microbiology,* **64**: 514–532.

Guillemin, M.-L., Contreras-Porcia L., Ramírez M.E., Macaya, E.C., Contador, C.B., Woods, H. Wyatt, C. & Brodie, J. (2016). The bladed Bangiales (Rhodophyta) of the South Eastern Pacific: Molecular species delimitation reveals extensive diversity. *Molecular Phylogenetics and Evolution*, **94**: 814–826.

Harrison, N. & Kidner, C.A. (2011). Next-generation sequencing and systematics: what can a billion base pairs of DNA sequence data do for you?. *Taxon*, *60*: 1552–1566.

Hayden, H.S., Blomster, J., Maggs, C.A., Silva, P.C., Stanhope, M.J. & Waaland, R. (2003). Linnaeus was right all along: *Ulva* and *Enteromorpha* are not distinct genera. *European Journal of Phycology*, **38**: 277–294.

Head, S.R., Komori, H.K., LaMere, S.A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D.R. & Ordoukhanian, P. (2014). Library construction for next-generation sequencing: Overviews and challenges. *Biotechniques*, **56**: 61-passim.

Heled, J. & Drummond, A. (2010). Bayesian Inference of Species Trees from Multilocus Data. *Molecular Phylogenetics and Evolution*, **27**: 570–580.

Hernandez-Kantun, J.J., Rindi, F., Adey, W.H., Heesch, S., Peña, V., Le Gall, L. & Gabrielson, P.W. (2015). Sequencing type material resolves the identity and distribution of the generitype *Lithophyllum incrustans*, and related european species *L. hibernicum* and *L. Bathyporum* (Corallinales, Rhodophyta). *Journal of Phycology*, **51**: 791–807.

Herrera, S. & Shank, T.M. (2016). RAD sequencing enables unprecedented phylogenetic resolution and objective species delimitation in recalcitrant divergent taxa. *Molecular Phylogenetics and Evolution*, **100**: 70–79.

Hind, K.R., Gabrielson, P.W., Lindstrom, S.C. & Martone, P.T. (2014). Misleading morphologies and the importance of sequencing type specimens for resolving coralline taxonomy (Corallinales, Rhodophyta): *Pachyarthron cretaceum* is *Corallina officinalis*. *Journal of Phycology*, **50**: 760–764.

Huang, H. & Knowles, L.L. (2016). Unforeseen consequences of excluding missing data from next-generation sequences: Simulation study of rad sequences. *Systematic Biology*, **65**: 357–365.

Hughey, J.R. & Gabrielson, P.W. (2012). Comment on "Acquiring DNA sequence data from dried archival red algae (Florideophyceae) for the purpose of applying variable names to contemporary genetic species: a critical assessment. *Botany,* **90**: 1191–1194.

Hughey, J.R., Gabrielson, P.W., Rohmer, L., Tortolani, J., Silva, M, Miller, K.A., Young, J.D., Martell, C. & Ruediger, E. (2014). Minimally destructive sampling of type specimens of *Pyropia* (Bangiales, Rhodophyta) recovers complete plastid and mitochondrial genomes. *Scientific Reports* **4**: 5113.

Hughey, J.R., Hommersand, M.H., Gabrielson, P.W., Miller, K.A. & Fuller, T. (2017). Analysis of the complete plastomes of three species of *Membranoptera* (Ceramiales, Rhodophyta) from Pacific North America. *Journal of Phycology,* **53**: 32–43.

Hughey, J.R., Silva, P.C. & Hommersand, M.H. (2001). Solving taxonomic and nomenclatural problems in Pacific Gigartinaceae (Rhodophyta) using DNA from type material. *Journal of Phycology,* **37**: 1091–1109.

Hughey, J.R., Silva, P.C. & Hommersand, M.H. (2002). ITS1 sequences of type specimens of *Gigartina* and *Sarcothalia* and their significance for the classification of South African Gigartinaceae (Gigartinales, Rhodophyta). *European Journal of Phycology*, **37**: 209–216.

Jackson, C., Salomaki, E.D., Lane, C.E. & Saunders G.W. (2017). Kelp transcriptomes provide robust support for interfamilial relationships and revision of the little known Arthrothamnaceae (Laminariales). *Journal of Phycology*, **53**: 1–6.

Jain, M., Olsen, H.E., Paten, B. & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, **17**: 239.

Janouškovec J., Gavelis G.S., Burki, F., Dinh, D., Bachvaroff, T.R., Gornik, S.G., Bright, K.J., Imanian,B., Strom, S.L., Delwiche, C.F., Waller, R.F., Fensome, R.A., Leander, B.S., Rohwer, F.L. & Saldarriaga, J.F. (2017). Major transitions in dinoflagellate evolution unveiled by phylotranscriptomics. *PNAS*, **114**: E171-E180.

Janouškovec, J., Liu, S.-L., Martone, P.T., Carré, W., Leblanc, C., Collén, J. & Keeling, P.J. (2013). Evolution of red algal plastid genomes: ancient architectures, introns, horizontal gene transfer, and taxonomic utility of plastid markers. *PLoS ONE*, **8**: e59001.

Jeong, H., Lim, J.-M., Park, J., Sim, Y.M., Choi, H.-G., Lee, J. & Jeong, W.-J. (2014). Plastid and mitochondrion genomic sequences from Arctic *Chlorella* sp. ArM0029B. *BMC Genomics*, **15**: 286.

Ješovnik, A., Sosa-Calvo, J., Lloyd, M.W., Branstetter, M.G., Fernández, F. & Shultz, T.R. (2017). Phylogenomic species delimitation and host-symbiont coevolution in the fungus-farming ant genus *Sericomyrmex* Mayr (Hymenoptera: Formicidae): ultraconserved elements (UCEs) resolve a recent radiation. *Systematic Entomology*, **42**: 523–542.

Jones, G. (2017). Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. *Journal of Mathematical Biology,* **74**: 447–67.

Jones, G., Aydin, Z. & Oxelman, B. (2015). DISSECT: an assignment-free Bayesian discovery method for species delimitation under the multispecies coalescent. *Bioinformatics*, **31**: 991–998.

Kang, S.A., Tice, A.K., Spiegel, F.W., Silberman, J.D., Pánek, T., Čepička, I., Kostka, M., Kosakyan, A., Alcântara, D.M., Roger, A.J., Shadwick, L.L., Smirnov, A., Kudryavstev, A., Lahr, D.J.G. & Brown, M.W. (2017). Between a pod and a hard test: the deep evolution of amoebae. *Molecular Biology and Evolution*, **34**: 2258-2270.

Karamitros, T. & Magiorkinis, G. (2015). A novel method for the multiplexed target enrichment of MinION next generation sequencing libraries using PCR-generated baits. *Nucleic acids research*, **43**: e152-e152.

Katoh, K., Misawa, K., Kuma, K. & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research,* **30**: 3059–3066.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Mentjies, P., & Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**: 1647–1649.

Kermarrec, L., Franc, A., Rimet, F., Chaumeil, P., Humbert, J.F. & Bouchez, A. (2013). Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: a test for freshwater diatoms. *Molecular ecology resources,* **13**: 607–619.

Kim, K., Park, J.-H., Bhattacharya, D. & Yoon, H. (2014). Applications of next-generation sequencing to unravelling the evolutionary history of algae. *International Journal of Systematic and Evolutionary Microbiology*, **64**: 333–345.

Koren, S., Schatz, M.C., Walenz, B.P., Martin, J., Howard, J.T., Ganapathy, G., Wang, Z., Rasko, D.A., McCombie, W.R., Jarvis, E.D. & Phillippy, A.M. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*, **30**: 693–700.

Leliaert, F., Tronholm, A., Lemieux, C., Turmel, M., DePriest, M.S., Bhattacharya, D., Karol, K.G., Fredericq, S., Zechman, F.W. & Lopez-Bautista, J.M. (2016). Chloroplast phylogenomic analyses reveal the deepest-branching lineage of the Chlorophyta, Palmophyllophyceae class. nov. *Scientific Reports,* **6**: 25367.

Leliaert, F., Verbruggen, H., Vanormelingen, P., Steen, F., Zuccarello, G.C. & De Clerck, O. (2014). DNA-based species delimitation in algae. *European Journal of Phycology*, **49**: 179–196.

Lemieux, C., Otis, C. & Turmel, M. (2014). Chloroplast phylogenomic analysis resolves deep-level relationships within the green algal class Trebouxiophyceae. *BMC Evolutionary Biology*, **14**: 211.

Lemieux, C., Vincent, A.T., Labarre, A., Otis, C. & Turmel, M. (2015). Chloroplast phylogenomic analysis of chlorophyte green algae identifies a novel lineage sister to the Sphaeropleales (Chlorophyceae). *BMC Evolutionary Biology*, **15**: 264.

Lemmon, E.M. & Lemmon, A.R. (2013). High-Throughput Genomic Data in Systematics and Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, **44**: 99–121.

Lindstrom, S.C., Gabrielson, P.W., Hughey, J.R., Macaya, E.C. & Nelson, W.A. (2015). Sequencing of historic and modern specimens reveals cryptic diversity in *Nothogenia* (Scinaiaceae, Rhodophyta). *Phycologia,* **54**: 97–108.

Llamas, B., Valverde, G., Fehren-Schmitz, L., Weyrich, L.S., Cooper, A. & Haak, W. 2017. From the field to the laboratory: controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era. *STAR: Science & Technology of Archaeological Research*, **3**: 1–14.

Lopes dos Santos, A.L., Gourvil, P., Tragin, M., Noël, M.H., Decelle, J., Romac, S. & Vaulot, D. (2016). Diversity and oceanic distribution of prasinophytes clade VII, the dominant group of green algae in oceanic waters. *The ISME journal*, **11**: 512–528.

Lu, J.-M., Zhang, N., Du, X.-Y., Wen, J. & Li, D.-Z. (2015). Chloroplast phylogenomics resolves key relationships in ferns. *Journal of Systematics and Evolution,* **53**: 448–457.

Lutz, S., Anesio, A. M., Raiswell, R., Edwards, A., Newton, R. J., Gill, F., & Benning, L. G. (2016). The biogeography of red snow microbiomes and their role in melting arctic glaciers. *Nature communications*, **7**: 11968.

Lutz, S., Anesio, A.M., Field, K. & Benning, L.G. (2015). Integrated 'omics', targeted metabolite and single-cell analyses of Arctic snow algae functionality and adaptability. *Frontiers in Microbiology*, **6**: 1323.

Ma, P.F., Zhang, Y.X., Zeng, C.X., Guo, Z.H. & Li, D.Z. (2014). Chloroplast phylogenomic analyses resolve deep level relationships of an intractable bamboo tribe Arundinarieae (Poaceae). *Systematic Biology,* **63**: 933–950.

Magoč, T. & Salzberg, S. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, **27**: 2957–2963.

Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., Wincker, P., Iudicone, D., de Vargas, C., Bittner, L., Zingone, A. & Bowler, C. (2016). Insights into global diatom distribution and diversity in the world's ocean. *Proceedings of the National Academy of Sciences of the United States of America*, **113**: E1516–E1525.

Marcelino, V.R. & Verbruggen, H. (2016). Multi-marker metabarcoding of coral skeletons reveals a rich microbiome and diverse evolutionary origins of endolithic algae. *Scientific Reports,* **6**: 31508.

Mardis, E. (2017). DNA sequencing technologies: 2006-2016. *Nature Protocols*, **12**: 213–218.

McCormack, J., Hird, S., Zellmer, A., Carstens, B. & Brumfield, R. (2013). Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, **66**: 526–538.

McManus, H.A., Sanchez, D.J. & Karol, K.G. (2017). Plastomes of the green algae *Hydrodictyon reticulatum* and *Pediastrum duplex* (Sphaeropleales, Chlorophyceae). *PeerJ*, **17**: e3325.

Montecinos, A. (2016). *Species delineation and hybridization in the brown seaweed* Ectocarpus *complex*. Université Pierre et Marie Curie, Universidad Austral de Chile Sciences, Roscoff.

Montecinos, A.E., Couceiro, L., Peters, A.F., Desrut, A., Valero, M. & Guillemin, M.-L. (2017). Species delimitation and phylogeographic analyses in the *Ectocarpus* subgroup *siliculosi* (Ectocarpales, Phaeophyceae). *Journal of Phycology*, **53**: 17–31.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods*, **5**: 621-628.

Muñoz-Gómez, S.A., Mejía-Franco, F.G., Durnin, K., Colp, M., Grisdale, C.J., Archibald, J.M. & Slamovits, C.H. (2017). The new red algal subphylum Proteorhodophytina comprises the largest and most divergent plastid genomes known. *Current Biology*, **27**: 1-8.

Nozaki, H., Yamada, T. K., Takahashi, F., Matsuzaki, R., & Nakada, T. (2014). New "missing link" genus of the colonial volvocine green algae gives insights into the evolution of oogamy. *BMC Evolutionary Biology*, **14**: 37.

Page, R.D.M. (2016). DNA barcoding and taxonomy: dark taxa and dark texts. *Philosophical Transactions of the Royal Society B*, **371**: 20150334.

Parker, J., Helmstetter, A.J., Devey, D., Wilkinson, T. & Papadoupulos, A.S.T. (2017). Field-based species identification of closely-related plants using real-time nanopore sequencing. *Scientific Reports*, 7: 8345.

Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., & Tyson, G.W. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the Tree of Life. *Nature Microbiology*, **2:**1533-1542.

Payo, D.A., Leliaert, F., Verbruggen, H., D'hondt, S., Calumpong, H. P. & De Clerck, O. (2013). Extensive cryptic species diversity and fine-scale endemism in the marine red alga *Portieria* in the Philippines. *Proceedings of the Royal Society B: Biological Sciences,* **280**: 20122660.

Prosser, S.W.J., deWaard, J.R., Miller, S.E. & Hebert, P.D.N. (2016). DNA barcodes from century-old type specimens using next-generation sequencing. *Molecular Ecology Resources*, **16**: 487–497.

Prum, R.O., Berv, J.S., Dornburg, A., Field, D.J., Townsend, J.P., Lemmon, E.M. & Lemmon, A.R. (2015). A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature*, **526**: 569–573.

Rappe, M.S., Giovannoni, S.J. (2003). The uncultured microbial majority. *Annual Review of Microbiology,* **57**: 369–394.

Reuter, J.A., Spacek, D & Snyder, M.P. (2015). High-throughput sequencing technologies. *Molecular Cell*, **58**: 586–597.

Reviers, B. & Rousseau, F. (1999). Towards a new classification of the brown algae. In *Progress in Phycology Research* (Round, F.E. & Chapman, D.J., editors), **13**: 107–201. Biopress Ltd., Bristol.

Rhoads, A. & Au, K.F. (2015). PacBio sequencing and its applications. *Genomics, Proteomics & Bioinformatics*, **13**: 278–289.

Rittmeyer, E.N. & Austin, C.C. (2015). Combined next-generation sequencing and morphology reveal fine-scale speciation in Crocodile Skinks (Squamata: Scincidae: Tribolonotus). *Molecular Ecology*, **24**: 466–483.

Rubin B.E.R., Ree R.H. & Moreau C.S. (2012). Inferring phylogenies from RAD sequence data. *PLoS ONE*, **7**: 1–12.

Ruck, E.C., Nakov, T., Alverson, A.J. & Theriot, E.C. (2016). Phylogeny, ecology, morphological evolution, and reclassification of the diatom orders Surirellales and Rhopalodiales. *Molecular Phylogenetics and Evolution*, **103**: 155–171.

Särkinen, T., Staats, M., Richardson, J.E., Cowan, R.S. & Bakker, F.T. (2012). How to open the treasure chest? Optimising DNA extraction from herbarium specimens. *PLoS ONE,* **7**: e43808.

Saunders, G.W. & Kucera, H. (2010). An evaluation of *rbc*L, *tuf*A, UPA, LSU and ITS as DNA barcode markers for the marine green macroalgae. *Cryptogamie, Algologie*, **31**: 487-528.

Saunders, G.W. & McDevit, D.C. (2012). Acquiring DNA sequence data from dried archival red algae (Florideophyceae) for the purpose of applying available names to contemporary genetic species: a critical assessment. *Botany*, **90**:191–203.

Sauvage, T., Schmidt, W.E., Suda, S., Fredericq, S. (2016). A metabarcoding framework for facilitated survey of endolithic phototrophs with *tuf*A. *BMC Ecology*, **16**: 8.

Schloss, P., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., Van Horn, D.J. & Weber, C.F. (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied Environmental Microbiology*, **75**: 7537–7541.

Schmieder, R. & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**: 863–864.

Sěvčıková, T., V. Klimeš, V., Zbránková, V., Strnad, H., Hroudová, M., Vlček, C. & Eliáš, M. (2016). A comparative analysis of mitochondrial genomes in Eustigmatophyte algae. *Genome Biology and Evolution*, **8**: 705–722.

Sherwood, A.R. & Presting, G.G. (2007). Universal primers amplify a 23S rDNA plastid marker in eukaryotic algae and cyanobacteria. *Journal of Phycology*, **43**: 605–608.

Sherwood, A.R., Dittbern, M.N., Johnston, E.T. & Conklin, K.Y. (2016). A metabarcoding comparison of windward and leeward airborne algal diversity across the Ko'olau mountain range on the island of O'ahu, Hawai'i. *Journal of Phycology*, **53**: 437–445.

Shokralla, S., Porter, T.M., Gibson, J.F., Dobosz, R., Janzen, D.H., Hallwachs, W., Golding, G.B. & Hajibabaei, M. (2015). Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific Reports*, **5**: 9687.

Shokralla, S., Spall, J.L., Gibson, J.F. & Hajibabaei, M. (2012). Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology*, **21**: 1794–1805.

Silberfeld, T., Leigh, J.W., Verbruggen, H., Cruaud, C., de Reviers, B. & Rousseau, F. (2010). A multi-locus time-calibrated phylogeny of the brown algae (Heterokonta, Ochrophyta, Phaeophyceae): Investigating the evolutionary nature of the ‘‘brown algal crown radiation”. *Molecular Phylogenetics and Evolution*, **56**: 659–674.

Sissini, M.N., Oliveira, M.C., Gabrielson, P.W., Robinson, N.M., Okolodkov, Y.B., Riosmena-Rodriguez, R. & Horta, P. A. (2014). *Mesophyllum erubescens* (Corallinales, Rhodophyta) –so many species in one epithet. *Phytotaxa,* **190**: 299–319.

Song, H., Lee, J., Graf, L., Rho, M., Qiu, H., Bhattacharya, D. & Yoon, H. (2016). A novice’s guide to analyzing NGS-derived organelle and metagenome data. *Algae*, **31**: 137–154.

Song, S., Zhao, J. & Li, C. (2017). Species delimitation and phylogenetic reconstruction of the sinipercids (Perciformes: Sinipercidae) based on target enrichment of thousands of nuclear coding sequences. *Molecular Phylogenetics and Evolution*, **111**: 44–55.

Staats, M., Erkens, R.H.J., van de Vossenberg, B., Wieringa, J.J., Kraaijeveld, K., Stielow, B., Geml, J., Richardson, J.E. & Bakker, F.T. (2013). Genomic Treasure Troves: Complete Genome Sequencing of Herbarium and Insect Museum Specimens. *PLoS ONE,* **8**: e69189.

Steven, B., McCann, S. & Ward, N.L. (2012). Pyrosequencing of plastid 23S rRNA genes reveals diverse and dynamic cyanobacterial and algal populations in two eutrophic lakes. *FEMS Microbiology Ecology,* **82**: 607–615.

Sun, J., Wang, L., Wu, S., Wang, X., Xiao, J., Chi, S., Liu, C., Ren, L., Zhao, Y., Liu, T. & Yu, J. (2014). Transcriptome-wide evolutionary analysis on essential brown algae (*Phaeophyceae*) in China. *Acta Oceanologica Sinica,* **33**: 13–19.

Suzuki, M., Segawa, T., Mori, H., Akiyoshi, A., Ootsuki, R., Kurihara, A., Sakayama, H., Kitayama, T., Abe, T., Kogame, K., Kawai, H. & Nozaki, H. (2016). Next-generation sequencing of an 88-year-old specimen of the poorly known species *Liagora japonica* (Nemaliales, Rhodophyta) supports the recognition of *Otohimella* gen. nov. *PLoS ONE,* **11**: 1–18.

Verbruggen, H. (2014). Morphological complexity, plasticity, and species diagnosability in the application of old species names in DNA-based taxonomies. *Journal of Phycology*, **50**: 26–31.

Verbruggen, H., Ashworth, M., LoDuca, S.T., Vlaeminck, C., Cocquyt, E., Sauvage, T., Zechman, F.W., Littler, D.S., Littler, M.M., Leliaert, F. & De Clerck, O. (2009). A multi-locus time-calibrated phylogeny of the siphonous green algae. *Molecular Phylogenetics and Evolution,* **50**: 642–653.

Verbruggen, H., Maggs, C.A, Saunders, G.W., Le Gall, L., Yoon, H. S. & De Clerck, O. (2010). Data mining approach identifies research priorities and data requirements for resolving the red algal tree of life. *BMC Evolutionary Biology,* **10**: 16.

Verbruggen, H., Marcelino, V. R., Guiry, M. D., Cremen, Ma. C. M. & Jackson, C. J. (2017). Phylogenetic position of the coral symbiont *Ostreobium* (Ulvophyceae) inferred from chloroplast genome data. *Journal of Phycology*, **53**: 790–803.

Vieira, C., Camacho, O., Wynne, M.J., Mattio, L., Anderson R.J., Bolton, J.J., Sanson, M., D'Hondt, S., Leluaert, F., Fredericq, S., Payri, C. & De Clerck, O. (2016). Shedding new light on old algae: Matching names and sequences in the brown algal genus *Lobophora* (Dictyotales, Phaeophyceae). *Taxon*, **65**: 689–707.

Villain, A., Kojadinovic, M., Puppo, C., Prioretti, L., Hubert, P., Zhang, Y., Grégori, G., Roulet, A., Roques, C., Claverie, J.-M., Gontero, B. & Blanc, G. (2017). Complete mitochondrial genome sequence of the freshwater diatom *Asterionella formosa*. *Mitochondrial DNA Part B*, **2:** 97–98.

Wang, L., Mao, Y., Kong, F., Li, G., Ma, F., Zhang, B., Sun, P., Bi, G., Zhang, F., Xue, H. & Cao, M. (2013). Complete sequence and analysis of plastid genomes of two economically important red algae: *Pyropia haitanensis* and *Pyropia yezoensis*. *PLoS ONE* **8**: e65902.

Worden, A.Z., Janouskovec, J., McRose, D., Engman, A., Welsh, R.M., Malfatti, S., Tringe, S.G. & Keeling, P.J. (2012). Global distribution of a wild alga revealed by targeted metagenomics. *Current Biology*, **22**:675-677.

Yang, E.C., Boo, S.M., Bhattacharya, D., Saunders, G.W., Knoll, A.H., Fredericq, S., Graf, L. & Yoon, H.S. (2016). Divergence time estimates and the evolution of major lineages in the florideophyte red algae. *Scientific Reports,* **6**: 2136.

Yang, E.C., Kim, K.M., Kim, S.Y., Lee, J.M., Boo, G.H., Lee, J.-H., Nelson, W.A., Yi, G., Schmidt, W.E., Fredericq, S., Boo, S.M., Bhattacharya, D. & Yoon, H.S. (2015). Highly conserved mitochondrial genomes among multicellular red algae of the Florideophyceae. *Genome Biology Evolution*, **7**: 2394–2406.

Yang, Z. & Rannala, B. (2014). Unguided Species Delimitation Using DNA Sequence Data from Multiple Loci. *Molecular Phylogenetics and Evolution*, **31**: 3125–3135.

Yeates, D.K., Zwick, A. & Mikheyev, A.S. (2016). Museums are biobanks: Unlocking the genetic potential of the three billion specimens in the world's biological collections. *Current Opinion in Insect Science,* **18**: 83–88.

Zhang, L., Wang, X., Liu, T., Wang, H., Wang, G., Chi, S.& Liu, C. (2015). Complete Plastid Genome of the Brown Alga *Costaria costata* (Laminariales, Phaeophyceae). *PLoS ONE,* **10**: e0140144.

Table 1. HTS technologies commonly used at the time of writing, with their key features and potential applications in systematics. Unless specified, information is from manufacturers.

Methods	Illumina MiSeq	Illumina NextSeq 500	Illumina HiSeq 2500 (Rapid Run Mode)	Illumina HiSeq 2500 (High-Output Mode)	Pacific Biosciences PACBIO RS II	Oxford Nanopore Technologies MinION
Max output	15 Gb	120 Gb	300 Gb	1000 Gb	5–8 Gb (per SMRT cell)	20 Gb
Max read length	2 × 300 bp	2 × 150 bp	2 × 250 bp	2 × 125 bp	60 kb	> 250 kb
Max reads per run	12–25 million	400 million	600 million	4 billion	~360 000	Up to 4.4 million
Max time per run	~56 hours	29 hours	60 hours	6 days	6 hours per SMRT cell	48 hours
Error rates[1]	0.1%	0.1%	0.1%	0.1%	~13% single-pass <1% (CCS)	8% (Jain *et al.*, 2016)
Required DNA	1 ng – 1 µg	1 ng – 1 µg	25 ng – 1 µg	25 ng – 1 µg	5 µg	10 pg – 1 µg
Multiplexing	Yes, 384 barcodes per lane	Yes, 384 barcodes per lane	Yes, 384 barcodes per lane	Yes, 384 barcodes per lane	Yes, 384 barcodes	Yes (See e.g. Karamitros & Magiorkinis, 2015)
Costs per Gb (USD)[2]	$100–$1000	$30–$50	$40–$100	$30–$80	~$300	~$150
Applications in systematics	• Small whole genome sequencing • Targeted gene sequencing (amplicon) • 16S metagenomic sequencing	• Exome sequencing • Targeted gene sequencing • Whole transcriptome sequencing	• Exome sequencing • Whole transcriptome sequencing • Whole genome sequencing	• Exome sequencing • Whole transcriptome sequencing • Whole genome sequencing	• Useful for *de novo* genome assembly as long reads provide larger scaffolds (Rhoads & Au, 2015) • Whole genome sequencing	• Useful for *de novo* assembly of small genomes and finishing of larger genomes (English *et al.*, 2012)

[1]Error rates are not exactly comparable. The error rates in Illumina platforms applies to >85% of reads (Glenn, 2011). CCS (circular consensus read) is a Pacific Biosciences technique which permits the reading of a circularized molecule multiple times, improving accuracy. However this approach reduces the read length (Koren et al., 2012).

[2]Rounded from Field Guide to next-generation DNA sequencers (Glenn, 2011) and 2016 update. Cost refers to reagent cost. Cost of library preparation and equipment purchase not included.

Author/s:
Oliveira, MC; Repetti, SI; Iha, C; Jackson, CJ; Diaz-Tapia, P; Lubiana, KMF; Cassano, V; Costa, JF; Cremen, MCM; Marcelino, VR; Verbruggen, H

Title:
High-throughput sequencing for algal systematics

Date:
2018-01-01

Citation:
Oliveira, M. C., Repetti, S. I., Iha, C., Jackson, C. J., Diaz-Tapia, P., Lubiana, K. M. F., Cassano, V., Costa, J. F., Cremen, M. C. M., Marcelino, V. R. & Verbruggen, H. (2018). High-throughput sequencing for algal systematics. EUROPEAN JOURNAL OF PHYCOLOGY, 53 (3), pp.256-272. https://doi.org/10.1080/09670262.2018.1441446.