A Beginner's Guide to Molecular Identification of Seaweed

Giuseppe C. Zuccarello[1]* and Nicholas A. Paul[2]

[1]School of Biological Sciences, Victoria University of Wellington, PO Box 600, Wellington, 6140, New Zealand
[2]School of Science & Engineering, University of the Sunshine Coast, Maroochydore DC, Queensland, 4558 Australia

Abstract

The most rapidly expanding areas for seaweed production in the world are the tropics, including Indonesia, yet these areas are also where molecular identification of local marine flora has only been sporadically employed. Furthermore, a goal for the Government of Indonesia is to diversify the types of seaweed that are being utilized, targeting valuable products and, hand in hand, to develop aquaculture techniques for these species. Morphological methods for species identification in algae are complex or unreliable, due to simple morphologies and plasticity. Therefore, it is crucial that the correct identification is made for species and varieties of commercial interest so that growth and biochemical results can be compared and contrasted between locations, across environments and over time without taxonomic ambiguity. This guide presents entry level methodologies for sample collection, DNA preservation, DNA extraction, PCR, and analyses of DNA sequence data, as a first step in the genetic characterization of both well-known cultivated species and identification of different species with potential economic properties.

Keywords: DNA barcoding, DNA extraction, molecular identification, polymerase chain reaction, *Kappaphycus*

1. Introduction

Pressure on terrestrial environments has led to an increased dependence on marine organisms for food and biological products. This environmental pressure is further exacerbated by changes in climate that have altered the utility and opportunities for land use. Marine macroalgae (seaweeds) have always played an important part in both human diets and as sources of bioproducts for human use and agriculture applications. Seaweeds have been an important component of diets in many countries, both traditionally and presently, especially, but not exclusively, in Asia. Traditionally, these needs have been fulfilled by wild harvest. As needs have increased, seaweed cultivation has increased to meet demands and reduce the impact on natural ecosystems. This has led to large cultivation areas in many parts of the world, including Indonesia which are now one of the largest global producers (FAO FishStat Plus, 2019).

Like all newly started industries, especially of organisms that are not well studied, the identity and diversity of organisms, both that are presently being cultivated but also that may potentially be beneficial, is not easily determined. This is especially true in seaweeds in which characteristics (morphological characters) are simple, difficult to determine, environmentally plastic or convergent. Accurate identification and naming of the species presently under cultivation, and of those species that have new properties, is needed before other cultivation practices (for example, breeding experiments; searches for sister species with similar or better properties) can be undertaken.

The low morphological differences amongst taxonomically distinct seaweed species (Verbruggen, 2014), plus their ability to change vegetative morphology (called plasticity) in different environments, has led to confusion in the naming of specimens under cultivation and also to a limited understanding of their biodiversity in many areas of the world. Therefore, the biodiversity of any particular area is now commonly determined using molecular methods. This molecular-assisted taxonomy has given us a better understanding of general seaweed biodiversity of many areas of the world (e.g. Cianciola, Popolizio,

*Corresponding author.
E-mail: joe.zuccarello@vuw.ac.nz

Schneider, & Lane, 2010; Dubrasquet, Reyes, Sanchez, Valdivia, & Guillemin, 2018; Freshwater et al., 2017), or elucidated the species diversity within particular groups or genera (e.g. Buchanan & Zuccarello, 2018; Diaz-Martinez , Zuccarello, Chávez, Pedroche, & Avila-Ortiz, 2016; Schneider, 2015).

Molecular methods have also been used to clarify questions of seaweed diversity and identification in cultivation. For example, the first study of cultivated *Eucheuma* and *Kappaphycus* (especially with regard to diversity of *Eucheuma denticulatum* and *Kappaphycus alvarezii*) used a mitochondrial region (the intergenic region between cytochrome oxidase subunit 2 and subunit 3-cox2-3 spacer), plus an intergenic region of the chloroplast (between the large subunit and small subunit of ribulose bisphosphate carboxylase- the Rubisco spacer) to determine the diversity and correct species designation for commercial samples (Zuccarello et al., 2006) (useful marker regions shown in Table 1). This study showed low intra-specific diversity amongst >100 cultivars, with most essentially comprising a single genotype used in most cultivation areas around the world, and with a lower diversity seen with the chloroplast marker compared with the cox2-3 spacer. The study also revealed variants that existed in non-cultivated populations, and indicated that some of the strain names used in cultivation could not be genetically distinguished using these markers. Follow up studies showed that a cultivated strain is displacing the native strains in many areas (Dumilag, Orosco, & Lluisma 2016a, Halling et al., 2013, Tano, Halling, Lind, Buriyo, & Wikström, 2015) and higher levels of variation using other markers (cytochrome oxidase subunit 1) (Dumilag, Salvador, & Halling, 2016b). While these studies are extremely useful, the knowledge base especially for seaweed DNA analysis is very variable between countries. While development of strains can be facilitated by DNA analysis, in some situations just an understanding of the species under cultivation or the biodiversity of the seaweed species in the area is still needed.

For molecular methods to be useful in identification of seaweed species across different studies, several criteria must be met. Homologous DNA regions (genes, partial genes, or non-genic areas) need to be easily accessible (e.g., PCR primers already designed) and variable (e.g., appropriate levels of variation to distinguish species) so that comparisons and relationships can be determined. These regions need to be selected carefully for species identification and/or variation within species. This has led to the present day use of certain DNA regions that can be easily amplified and have been used historically in particular groups. One such region is the so called 'DNA barcode' used extensively in animals (Hebert, Cywinska, Ball, & deWaard, 2003). This is a partial, 5-prime end, of the cytochrome oxidase subunit 1 gene (COI), usually around 600 base pairs in length. While this region has proved useful in animals, its utility in species identification for the diverse species of red algae, brown algae and green algae remains to be fully explored. The diversity of seaweeds means that, while some primers are useful for a particular algal group (e.g., between phyla and occasionally within phyla) there is no universal primer combination that is useful for all algae. Many papers now indicate that the COI region can be amplified in many seaweeds using a standard set of primers in red algae (Saunders & Moore, 2013) and brown algae (Saunders & McDevit, 2012) and is useful in determining the diversity of species. The COI region has not been used for green algae and is considered not to be useful in this group (Saunders & Kucera, 2010). Therefore, most of the successful work on green algae involves other regions such as ribosomal RNA regions and plastid genes.

Because of the diversity of algae, choice of markers for barcoding or other molecular characterization for each algal group is critical. This involves understanding the question you want to address, and/or the particular species or group of organisms you are studying. The methodologies and procedures for molecular work are extremely varied and rely on many factors. These include scientific factors such as the questions addressed and previous knowledge on the study group; economic factors such as equipment available and budget; and the standard procedures that people use in their laboratories. Just because somebody else in your laboratory or a colleague has used one particular primer (and has offered it) does not mean that you can necessarily use the same one for your study, especially if it is on different species.

The purpose of this review is not to summarize all molecular methodologies available for identifying seaweeds but to instead present what can be called "a beginner's guide" to molecular laboratory work with seaweeds, starting with the sampling of seaweed and moving through considerations that are unique for seaweed and providing some trouble shooting options to navigate through unexpected or frustrating results.

2. Sample Collection

The question of sampling design, and numbers of samples to analyse, is very much based on the questions you are trying to address, plus unknown factors, such as population size and distribution of variants. If you want to identify a sample that has a clear provenance, then a single sample is, of course,

adequate. If you are wanting to sample a cultivation area for genetic diversity, and you know that all samples were clonally propagated, then low sampling may also be adequate (3-5 samples). If you are sampling a natural population of unknown diversity or to compare between populations, then for capturing the diversity larger samples sizes are needed, especially if diversity exists in low frequencies (e.g., 0.05). In most seaweed diversity studies using organelle markers (i.e. mitochondria and chloroplasts), approximately 10 samples per population is often used (e.g., Neiva, Assis, Fernandes, Pearson, & Serrão, 2014) but values of 30 are a good rule of thumb to more accurately measure diversity (Krueger-Hadfield & Hoban, 2016; Ward & Jasieniuk, 2009). The sampling design should also take into consideration what is already known of algal population structure and diversity. With poor spore dispersal and settlement close to parents, algal populations are often highly structured, meaning that diversity is concentrated in patches even within populations (Krueger-Hadfield, Roze, Mauger, & Valero, 2013; Zuccarello, Yeates, Wright, & Bartlett, 2001). Therefore, randomly sampling plants at some distance from each other, versus nearest neighbors, is more likely to capture the diversity of a particular population. Structuring sampling procedures and records can be done in a way to capture this information even when it is not explicitly being tested, for example, by labeling samples in a way that indicates their spatial relationships to each other even if this is not formally quantified. For example, it is not necessary in this instance to measure the actual distance in meters between samples but the intent should be to only collect samples that are at least 5 m apart.

3. DNA Preservation

While DNA extraction from fresh algal material is preferable, for the most part DNA cannot be extracted from fresh material, as often laboratory facilities are far away from collection locations. Therefore, several methods are employed for algal preservation for later DNA extraction. The goal of all methods is to reduce the degradation of the DNA by endogenous DNases within the alga or by exogenous DNases from epiphytic bacteria or fungi.

The most common method of DNA preservation is using a desiccant, such as silica gel, to rapidly dry, and keep dry, the algal sample. To do this the proportion of silica gel used must be much greater than the amount of algal material (1:10 ratio or more). It is critical that only a small amount of tissue is used relative to the amount of silica. Using silica gel with a self-indicator will also ensure that the silica gel is not saturated with water. If the silica gel is saturated, it is advisable to replace with fresh silica gel. Saturation is more likely to occur with certain species of seaweed that have a high moisture content, for example, *Caulerpa* and other siphonous algae can have a moisture content greater than 95% (or a fresh:dry ratio of >20:1), and if the ratio of silica gel to alga is low. Another common method of preservation is the use of 95-100% ethanol. Again, the amount of ethanol relative to tissue must be great to not reduce the concentration of ethanol in the final sample (again, more problematic with material that has high moisture content). Both these methods have been used in seaweeds; with the easy of transportation of silica gel over ethanol this is the most commonly used method. Another important consideration is to realize that, in order to save on silica gel, only a small amount of tissue is needed for most PCR applications. For a terete algal branch, 2-3 cm is more than sufficient for several DNA extractions. For an algal blade a piece 4 cm^2 is enough.

A voucher of the species should be kept. This means a sample of the 'species' you are collecting (something that looks typical of what you are collecting, or the plant that you removed a small piece from for DNA analysis) needs to be preserved on a sheet of paper. If you collect 100 samples of the same species, 100 vouchers are excessive, but several representative samples are useful to maintain. This voucher is especially important if you are investigating species diversity in an unexplored region and may discover a species new to science. In this case the voucher sample can be deposited at an internationally recognized herbarium prior to publication.

4. DNA Extraction

There are many ways to extract DNA from organisms, but algae have their own issues. The goal is to extract high molecular weight DNA from algal tissue, or at least DNA that is amplifiable (this is not necessarily the same thing). The first challenge is to break the cells to be able to extract the DNA. Several methods have been used: one is freezing tissue in liquid nitrogen and grinding using a mortar and pestle, while maintaining it frozen; another method is to use a tissue homogenizer to powder the dried tissue; a simple method is to grind the tissue in a lysis buffer in a microfuge tube, using a microfuge tube pestle (e.g., Kontes ® pestles). The addition of sterile sand or small glass beads can help the grinding.

Another issue is the presence of carbohydrates (cell wall material; e.g., carrageenan) and secondary metabolites (e.g., tannins, phenolics) that can co-precipitate with the DNA and interfere with the PCR.

There are methods to remove excess carbohydrates (Liu et al., 1998) or specific extraction procedures that remove carbohydrates (Boom & Sol, 1990). These methods have been successful in some studies but are not a certainty. Metabolites that may interfere with PCR can be usually reduced by the DNA extraction method or additives to the PCR mix (this will be revisited in a later section; see also Table 2).

Beginners feel that the best thing to do is to pay for a commercial extraction kit. There are many manufacturers of DNA extraction kits and some researcher use the "Plant" DNA extraction kit varieties (sometimes kits for soil or even animal DNA extractions). While these kits can work on some types of algal tissue they have not been optimized for seaweed DNA extraction, and can also not work. The classic method for DNA extraction in seaweeds is using a detergent called CTAB (cetyl trimethylammonium bromide) and was first introduced by Doyle and Doyle (1990) for land plants. The method has been modified over time (e.g., Zuccarello & Lokhorst, 2005). This method works for many red, brown and green algae, but does involve several steps and organic solvents. Another common method, that is more convenient and works on some algae, is the Chelex method, first introduced to algae by Goff and Moon (1993). No method works universally, but these simplified methods can be easily modified for your algal group of interest.

The CTAB method starts by adding a small amount of dried, or fresh, tissue (approximately 10 mg) to a microfuge tube containing 500µl of CTAB extraction buffer (2% CTAB, 0.1 M Tris-HCl (pH 8.0), 1.4 M NaCl, 20mM EDTA, 1% polyvinylpyrrolidone -PVP) plus the optional addition of approximately 50 µg RNAse A and 80 µg Proteinase K. These volumes can be reduced for smaller samples (or increased if a larger amount of DNA is needed). Samples are then ground with a microfuge pestle. Samples are then placed at 55-60°C for at least 1 hour, mixing occasionally, longer extraction times (e.g., overnight) are also possible and in some cases aid in extraction. Then an equal volume of chloroform: isoamyl alcohol (at a ratio of 24:1) is added to the samples, thoroughly mixed and spun for 10 min at 12,000 g, or maximum speed of the centrifuge used. This removes non-polar molecules and denatures the proteins that get in the way of downstream DNA use (e.g., PCR). The top aqueous layer, avoiding the lower chloroform:isoamyl alcohol layer and the interface debris, is transferred to a new microfuge tube and again extracted with an equal volume of chloroform:isoamyl alcohol (24:1), spun for 5 min and the aqueous layer is removed to a new centrifuge tube. An equal volume of 100% isopropanol is added to the tube, the tube is inverted several times

(essentially slowly inverting in one's hand) and then placed in a rack at room temperature for 30 min. The sample is spun for 20 min at maximum speed and immediately decanted (i.e. tipping the liquid out of the tube carefully and not dislodging the DNA pellet). To the DNA pellet, approximately 500 µL of 70% ethanol is added, partially to remove the isopropanol and also any remaining salts, and the sample is spun for 5 minutes and decanted. The sample is air-dried for 20 min, or until ethanol is removed, and 50 µL of 0.1 X TE buffer (1 mM Tris pH 8, 0.1 mM EDTA) or sterile pure water added. The sample can then be stored at -20 °C until use. The quality and quantity of high molecular weight DNA can be checked on an 1% agarose gel stained with ethidium bromide (or other DNA stain) and a quantified DNA molecular marker used, or spectrophotometrically (absorbance at 260 nm, Sanbrook, Fritsc h, & Maniatis, 1989).

Another method that has worked very well for select groups of algae is the Chelex® method (Goff & Moon 1993); it has the advantage to be quick, easy to use on small samples and non-toxic. The best way to ensure a small sample is taken is to excise approximately 5 mm of an axis of the thalli (small amounts of algal material are crucial for this method to work). This method has been modified (Zuccarello, West, Kamiya, & King, 1999), but basically: samples are placed in 500 µl of a 10 mM Tris (pH 8.0) solution containing 5% Chelex 100 resin in a microfuge tube. The Chelex solution needs to be shaken vigorously before each aliquot to ensure proper proportions of the resin in the extraction mix, as the resin particles sink quickly. Smaller amounts of tissue (including spores) can be extracted by reducing the amount of Chelex solution used. Tissue is then ground in the microfuge tube using a microfuge tube pestle until the solution become colored (the presence of pigment indicates that the cells are disrupted), and when only small fragments of tissue are visible. The samples need to be kept on ice before and after grinding. After grinding, samples are placed in 95-100 °C water bath or heat block for 10 min and then immediately placed on ice. Samples need to then be centrifuged at top speed for 10 min and the supernatant carefully removed and placed in newly marked, clean microcentrifuge tubes. Care needs to be taken to not transfer any Chelex resin beads nor tissue fragments. The samples can now be frozen at -20 °C. Checking DNA content from Chelex extractions is not useful for this method.

To remove exogenous DNA from the microfuge pestles, so that they can be reused, the pestles can be sterilized by placing them in 10% commercial bleach for an hour or more, and then washing repeatedly in autoclaved double-distilled water until

all bleach odor is eliminated. The pestle can then be dried in a moderately warm oven. This procedure is more effective for DNA removal than autoclaving.

Trouble shooting DNA extractions may need to be done. The most important thing is to check the ability of your DNA to be amplified. It is rare that the amount of DNA is not sufficient for successful PCR, and problems are often DNA quality or that contaminants retard PCR. Degraded DNA (a smear in an agarose gel) would most likely indicate a poorly preserved sample rather than poor extraction, and not much can be done without recollecting the sample. Again, if the DNA is able to be amplified, this is not a concern. There are applications were undegraded high-molecular-weight DNA is essential but not for most PCR applications. If there is a very low concentration of DNA, the most common reasons are the tissue was not ground properly (grind more thoroughly), extraction was inefficient (incubate for longer, or try another extraction method) or the DNA precipitation was inefficient (either speed of centrifuge was too low, or the DNA pellet was dislodged and lost; especially for the CTAB method where the DNA pellet can be difficult to visualise at the base of the microfuge tube).

5. Polymerase Chain Reaction (PCR)

The PCR is a method in which regions of DNA can be copied (amplified) using a set of DNA oligo-nucleotides (primers) and a thermostable DNA polymerase (e.g., *Taq* polymerase). This method can produce billions of copies of a DNA region that can then be sequenced to determine the order of nucleotides in the amplified region. Before PCR reactions are undertaken, the marker or markers to be used need to be considered. Again, this depends on the organism you have, or species/genera/phyla you might have, and the question you are addressing. A rough guide to the molecular markers most commonly used in seaweed studies is presented in Table 1, but note, that these are general DNA marker regions and specific primers for each region, for your particular seaweed, need to be obtained from previous studies. Things to consider are whether there are gene databases that allow you to compare your samples to other published information, especially if you are interested in species identification. For organisms of which you have no idea what they are, 'universal' primers for small subunit ribosomal RNA (SSU) are available and will tell you which algal group you have (e.g., Stoeck et al., 2010). With slightly more taxonomic knowledge (phylum, class or even genus of alga) marker selection is much easier. You need to check for publications that have studied that group, or query Genbank (https://www.ncbi.nlm.nih.gov/) for gene data of that particular group of organisms. A good selection of red seaweed markers, and primers, are presented in Saunders and Moore (2013).

PCR conditions in publications should be used as general guidelines. PCR conditions in your laboratory, using different supplies, often vary. There are several factors that influence PCR reactions and help in having

Table 1. A general selection of the most commonly used molecular markers in the three major groups of seaweeds

Algal Group	Classes/order	Genera	Species
Rhodophyta (red algae)	SSU, *rbc*L	*rbc*L, *psb*A, LSU	COI, *rbc*L, *cox*2-3 spacer
Chlorophyta (green algae)	SSU, *rbc*L	*rbc*L, *tuf*A, LSU	ITS, *rbc*L, *tuf*A
Phaeophyceae (brown algae)	SSU, *rbc*L, *psa*A, *psb*A	*rbc*L, *cox*3	ITS, COI, *cox*3

Note: Many of these markers can be used at multiple taxonomic levels of investigations and/or in combination. Classes/order= for phylogenetic analyses of classes or orders of algae. Genera= determining phylogenetic relationships between genera in a group. Species= often used for species identification, variation within species and relationships within genera. SSU= nuclear small subunit ribosomal RNA gene (18S); LSU= nuclear large subunit ribosomal RNA gene (26S); ITS= Internal transcribed spacer of the ribosomal cistron, divided into ITS1 and ITS2; *rbc*L= plastid-encoded large subunit of ribulose bisphosphate carboxylase/oxygenase; plastid- encoded *psb*A= photosystem II D1 protein; *psa*A= plastid- encoded photosystem I P700 chlorophyll a apoprotein A1; *tuf*A= plastid-encoded elongation factor Tu; *cox*3= mitochondrial-encoded cytochrome c oxidase subunit 3; cox2-3 spacer= mitochondrial spacer between cytochrome c oxidase subunit 2 and 3; COI= mitochondrial-encoded cytochrome c oxidase subunit 1.

Table 2. Additives that have been used in PCR to inhibit contaminants or to provide more specific PCRs

Additive	Final Conc.
Bovine serum albumin (BSA)	10–100 µg/mL
N, N,N-trimethylglycine (betaine)	1–3 M
Dimethyl sulfoxide (DMSO)	1–10%
Glycerol	5–20%
Formamide	1.25–10%
Polyethylene glycol (PEG)	5–15%
Nonionic detergents (e.g., Triton X-100, Tween 20)	0.05–01%

Note: Final conc.= recommended concentration range in the PCR reaction

successful reactions. The main ones are: 1) amount of DNA and contaminants; 2) $MgCl_2$ concentration; and, 3) annealing temperature.

Amount of DNA and contaminants: Successful PCR relies on an optimal amount of DNA and usually this is much lower than beginners may think. The easiest solution is to have multiple dilutions of your DNA template, for CTAB extraction 10- to 1000-times dilution is usually recommended. Contaminants, i.e. chemicals that co-precipitated with your DNA that will inhibit the DNA polymerase, are common in algae. These could be cell wall components, secondary metabolites, or environmental contaminants attached to the alga. Additives are often used to combat these inhibitors (Table 1), most commonly used are bovine serum albumin (BSA), betaine and dimethylsulfoxide (DMSO). Some of these additives are also used to increase the specificity of the PCR reactions, if for example when multiple bands are present (e.g., DMSO).

Mg^{2+} concentration: One method to increase specificity, either to produce a band or remove extra amplification bands, is to change the $MgCl_2$ concentration in the PCR reaction. Increased Mg ion concentrations decrease the specificity of the annealing and elongation from the primers. Mg ion concentrations can vary between 1-5 mM in PCRs, with 2.0-2.5 being an average used. If you get no bands, one possibility is to increase the Mg ion concentration and to decrease it if you have multiple bands.

Annealing temperature. The most common method to optimize PCR reactions is to change the annealing temperature. A decrease in annealing temperature will increase primer binding and an increase in temperature will increase primer binding specificity. If the PCR does not produce bands, one solution to try is to lower the annealing temperature, and if multiple bands are

produced increasing annealing temperature is often done. While annealing temperatures for specific primer and primer combinations are presented as theoretical temperatures when purchased, or can be calculated from the primer sequence, they may also be present in publications that used the primers. However, usually optimization is still needed. Annealing temperatures vary from 45 °C to 65 °C. Some PCR machines allow for a gradient of annealing temperatures across the heating block which can make optimization of PCR less time consuming.

Besides changes in Mg ion concentration and annealing temperature there are other methods to produce acceptable PCR products. These include the length of time at particular temperatures, mostly annealing temperature times but also the number of cycles. There is a whole array of things that can be changed to get usable PCR products (e.g., DNA concentration, PCR reaction additives, Mg concentrations, annealing temperatures) even changes in primer concentration. This leads to many potential combinations of optimization steps that may need to be done to the PCR reaction to produce products. This should be done in as systematic a way as possible, taking plenty of notes in the laboratory books for future reference.

A simple method that is a good first step for many algal studies, and especially for a new project, is as follows: Having a positive control is very useful (i.e. a DNA extraction that you know has worked in previous studies, or a species for which the primers were designed); Use a standard PCR reaction with a final Mg ion concentration of 2.5 mM and 20 µg of BSA, plus 1-2 dilutions of your DNA; use a touchdown PCR – this is a PCR in which the annealing temperature is reduced from a 'high' temperature to a 'lower' temperature per cycle for several cycles, followed by cycles at the lower temperature. For example, a touch

down program could be: an initial denaturing step at 94 °C for 3-5 min; followed by 10 cycles of 94 °C for 30 s, 55 °C for 1 min which is reduced by 1 °C per cycle, and 72 °C for 30 s-1 min (depending on product length); followed by 25 cycles of 94 °C/30 s, 45 °C/30 s, 72 °C/30 s; and a final extension of 72 °C for 5 min. From the results, adjust further PCR conditions. For no bands, DNA dilution and extra BSA are a good first step; for extra bands, a higher annealing temperature and possibly lower concentration of Mg ions are a good start.

PCR products should be checked on a 1-2% agarose gel. If products are of the correct size and of adequate concentration, they can be sequenced. Mostly researchers out-source their Sanger sequencing reactions. A careful search of companies that will fulfill your requirement should be done. The costs will depend on many factors, one is sample preparation and another is the number of samples. PCR products can be purified in your laboratory to cut down costs. A common method is enzymatic purification, using exonuclease I and shrimp alkaline phosphatase (to remove excess single-stranded primers and dNTPs respectively, exo-sap), but other methods include commercial PCR clean-up kits and even by precipitation (Rosenthal, Coutelle, & Craxton, 1993).

It is advisable to sequence PCR products in both directions (forward and reverse), but this will depend on your budget, plus experience in the quality of single direction sequencing of your PCR products, and the question to be addressed. Sometimes with small PCR products a single direction gives good quality signal for a significant portion of the amplicon, which is enough for species identification. While for longer PCR products (for phylogenetic analyses) or intra-population studies (where correct scoring of single base pair mutations are essential for the resolution required), bi-directional sequencing is critical.

6. Sequence Analysis

Once sequences are returned from the sequencing company they need to be edited, and/or assembled if bidirectional. There are many free programs to read Sanger electropherograms ("traces" as .ab1 files) (e.g., BioEdit, MEGA, 4peaks, etc.). For assembling the sequence in both directions, you need a specialised software program that will assemble the sequences, unfortunately these programs are rarely free. One free software program is the Staden package (not widely used). More user-friendly programs are Geneious or Codoncode, but they are not free. Editing of sequences is critical for any further analyses. Even with prefect sequences, the short reads and longer

reads (both ends of your trace) are often full of ambiguous signal (unclear or mixed peaks). These need to be removed and the remaining sequence checked. The key is to only accept basepair calls based on peaks that are unambiguous (only one clear peak per location, with occasional background peaks less than 20% of main peak) and that the peaks are equidistant. Incorrect editing could enter extraneous data that could affect your downstream analyses, so if you have the luxury of long sequences it is more important to be conservative and accept only very clear peaks (substitute an N or IUPAC DNA base codes for the ambiguous region). This is especially critical if you are looking for small variation within species (for example, a haplotype that is associated with a particular strain of commercial alga).

Once you have edited your sequences, there are two simple ways to determine the origin of your sequences and therefore check the identity of your sample (or, more accurately, the identity of your PCR product): Blast searching and phylogenetic analysis. The first is the simplest and quickest to do and involves pasting the edited sequence into the online NCBI nucleotide Blast (blastn) sequence query page *(https://blast.ncbi.nlm.nih.gov/Blast.cgi)*, and search your sequences' closest match in the Genbank database. This closest match indicates what your sequence could be, and percent identity will tell you how close your sequences matches the available sequences. These results will also tell you if your sequence is from the genomic region you targeted (i.e. did you really amplify COI?). If you were targeting COI but amplified another gene (for example, the Blast closest match is to another gene), it is likely you are using the wrong primers. Also, the Blast search will tell you if you have the correct organism, which means that you have to have some prior knowledge of the organism you are trying to identify. If your top Blast hit is from a bacterium or diatom, yet you were trying to identify a *Sargassum* species, one possibility is that you have amplified and sequenced a contaminant rather than the target seaweed. In this case you should use different primers that may be more specific for *Sargassum* species, try another *Sargassum* sample, or alter the PCR conditions to be more specific for your primers (e.g., increase annealing temperature, decrease Mg concentrations). While Blast will give you the closest match, plus many other matches (known as hits), it is not the only way, nor necessarily the best, to determine the species you may want to identify.

Genbank accessions are full of useful information and should be carefully examined. They will tell you the paper the sequences came from, the researchers who did the work, where the sample was collected

and when. Sometimes the records are out of date and the publication needs to be found, using the author names and organisms studied. This is useful information to gain more insight into your sequence. For example, if the top hit is from England, but your sample is from Indonesia, it is unlikely to be the same species, although some species have wide distributions, growing in tropical and cold temperate environments is unlikely for most seaweed species. It is also useful to follow up the research on this group by other researchers and what has been done with the genetic data (phylogenetic analyses), and especially to learn more about the taxonomy from peer-reviewed manuscripts. While Genbank is a useful depository of DNA sequence data, seaweed species names are provided by the submitters of the sequence, they should not be taken as correct taxonomic identifications without further information (peer-reviewed publications, herbarium vouchers, taxonomic expertise of the researchers).

The second method is to do a phylogenetic analysis with your sequence and selected sequences from Genbank that are the best hits from your Blast search, plus species that you think your samples could be. Phylogenetic inference is a complex field and, in many cases, is not straight forward and quite subjective. The following is a relatively easy method for identifying a single sequence from a database that may have similar taxa. Remember Genbank does not contain every organism on earth, just what has been sequenced so far, and of course not with all genes, plus with variable confidence of species identification (see above).

Firstly, you need to produce an aligned data set. There are several steps to do this. First, select sequences of the homologous gene from your sample (e.g., all COI) from your Blast search. Click on the sequences you want, usually the top ten Blast hits and several other named species (many sequences in Blast are not given species names, it is important to select a few sequences that have been named by the researchers that submitted the sequences). Download them from Genbank in a format that can be read by an alignment and analysis program - a common sequence data format is FASTA (.fasta, or .fsa). A free user-friendly entry-level alignment and analysis program is MEGA (www.megasoftware.net). MEGA will import FASTA sequences, plus it has the advantage of a Genbank web browser interface to download sequences directly into a data set.

Second, when the sequences are combined in a single file (your sequences plus downloaded sequences), they then need to be aligned before further analysis. The reason they need to be aligned is that not all sequences, even when using the same primers, are necessarily of equal length (because they start in different positions or due to differences in editing by you or other researchers) and therefore the sequence file will not have homologous bases in columns. Every column is a hypothesized homologous position and this can only be achieved by aligning the file. Many algorithms are available for alignment, common ones are MAFFT, T-COFFEE, MUSCLE and ClustalW, the latter two are available in MEGA. As a first step, use the default parameters for the alignment program you select, but depending on the gene and data set you are analyzing, alignment methods and parameters can give different inference results. Check your aligned dataset for anomalies, including the wrong gene or reversed sequences (usually yours). You should see many homologous bases (columns in the alignment that have the same base in that position) with variable amounts of columns that are different. Most alignment viewers display this well with colored bases, similarity shading or matching bases as dots to top sequence (in MEGA).

Once satisfied with the alignment, a phylogenetic tree can be produced from this aligned dataset. There are many methods for inferring phylogenetic relationships (Hall 2017; Lemey, Salemi, & Vandamme, 2007), and following procedures from a manuscript are a good start. Two starter methods that can be easily implemented in MEGA are a neighbor-joining (NJ) distance tree and a maximum-likelihood (ML) phylogeny. In the analysis interface of MEGA, select 'phylogeny' and 'construct neighbor-joining tree'. You will get an interface that has selectable parameters on it, such as "test of phylogeny", for example to bootstrap the phylogeny (Felsenstein 1985). Bootstrap the phylogeny with 500 replicates. The other important parameter to select is the molecular evolution model ("model/method"). The best fit model needs to be determined first and can be done in MEGA under find 'Models'/'Find Best DNA/Protein model', or online with programs such as jModelTest (Santorum, Darriba, Taboada, & Posada, 2014). After a best fit model has been determined, select that model for your analysis; alternately just select a simple model such as Jukes and Cantor (1969). The NJ tree in MEGA will show the closest sister sequence to your sequence and will also display the support for that relationship as a bootstrap value. A more robust method for phylogenetic reconstruction, also implemented in MEGA, is ML. For an ML analysis a best model of molecular evolution needs to be selected, following the methods stated before. A ML tree can also be bootstrapped.

The different phylogenetic methods, plus the Blast searching, should be congruent with the closest species/sequence to your test sequence being the

same. If this is the case, you have a good basis to propose that the closest sequence is that particular sequence in Genbank. This does not necessarily mean that your species is that species, but if your sequence is 99-100% similar to the sequence on Genbank, you can fairly safely assume, i.e. use as your working hypothesis, that your sequence is of the same species as the species that the Genbank sequence is derived from. There are two things to consider in making a final determination rather than having a working hypothesis. One is species delimitation; if your sequence is 1.8% or 2.2% different from the closest sequence in Genbank do you have the same species or do you have a different species? Genetic species delimitation is a growing field and is summarized for algae in Leliaert et al. (2014). Unfortunately, species status cannot be easily concluded from a phylogenetic tree without further work. The second point is species names in Genbank. These names are submitted by the authors of the sequence, often without voucher information, although this can sometimes be found from the related publication. The reality is that the authors of the sequence data determined the species name, but their taxonomic skills can be variable. Also there are nomenclatural issues, it could be that the species name being used for a particular alga in a particular region does not correspond to the species that the name is linked with, known as the "type" specimen. Therefore, the name is incorrectly applied, even if all the samples are from the same species. These issues are also not easy to solve without further work, for example, examining the specimens used for the sequences in Genbank, consulting taxonomic experts, and preferably sequencing of the type specimen.

7. Conclusions

Samples need to be collected in a way and in sufficient number to address particular questions. It is always safer to over collect and store samples for possible future work, especially if you do not regularly travel to that field site in a remote part of an archipelago. After DNA extraction, PCR needs to be performed. For this to be successful careful consideration of the markers (i.e. genes) and primers used needs to be made, follow these guidelines (Table 1) or find papers addressing similar questions in the same or similar organisms. Remember the methods section of a publication is only a starting guideline and you should adapt your processes to get a successful result (modify annealing temperatures, etc, etc.). Successful PCR amplifications can be sequenced and need to be carefully edited to only use basepairs that are clear and likely to be correct. Both Blast searches and phylogenetic analysis will

identify the sequence to the closest match in Genbank. If this match has a high similarity to your sequence, then it will be a good working hypothesis of the species that you have in hand.

Molecular identification of seaweeds is an excellent way to understand the species that you have in hand and is an entry into many important aspects of seaweed biology that can be answered using molecular tools. For example, understanding variation within a cultivated species, the identification of introduced seaweeds, the distribution of genetic diversity and areas of high diversity, more systematic selection of varieties with useful properties or in breeding studies, plus a better overall understanding of seaweed diversity in a particular region. For Indonesia, seaweed farming has fundamentally changed the economic and social status of coastal communities, but this growth has primarily been based on culture of a few species of red seaweed and, because of vegetative propagation, only a few varieties of each. Through the molecular identification of new species and strains of commercial importance, the Indonesian seaweed industry will be able to diversify and grow from this strong foundation.

Acknowledgements

References

Boom, R. & Sol, C. (1990). Rapid and simple method for purification of nuclei acids. *Methods*, 28, 495–503.

Buchanan, J. & Zuccarello, G. C. (2018) Utility of molecular-assisted alpha taxonomy of the genus *Cystophora* (Fucales, Phaeophyceae) from New Zealand and Australia. *Phycologia*, 57, 374–384.

Cianciola, E., Popolizio, T., Schneider, C., & Lane, C. (2010). Using Molecular-Assisted Alpha Taxonomy to Better Understand Red Algal Biodiversity in Bermuda. *Diversity*, 2, 946–958.

Díaz-Martínez, S., Zuccarello, G. C., Chávez, G. A. S., Pedroche, F. F., & Avila-Ortiz, A. G. (2016). Species of Padina (Dictyotales, Phaeophyceae) in tropical Mexican waters based on molecular-assisted taxonomy . *Phycologia*, 55(6), 673–687. https://doi.org/10.2216/16-15.1 .

Doyle, J. J. & Doyle, J. L. D. (1990). Isolation of plant DNA from fresh tissue. *Focus.* 12, 13-15.

Dubrasquet, H., Reyes, J., Sanchez, R. P., Valdivia, N., & Guillemin, M.-L. (2018). Molecular-Assisted Revision

of Red Macroalgal Diversity and Distribution along the Western Antarctic Peninsula and South Shetland Islands. Cryptogamie, Algologie, 39(4), 409. https://doi.org/10.7872/crya/v39.iss4.2018.409 .

Dumilag, R. V., Orosco, F. L., & Lluisma, A. O. (2016a). Genetic diversity of *Kappaphycus* species (Gigartinales, Rhodophyta) in the Philippines. *Systematics and Biodiversity*. doi: 10.1080/14772000.2016.1157643

Dumilag, R. V., Salvador, R. C., & Halling, C. (2016b). Genotype introduction affects population composition of native Philippine *Kappaphycus* (Gigartinales, Rhodophyta). Conservation Genetics Resources 1–3.

FAO FishStat Plus. (2019). Universal software for fishery statistical time series. Fisheries and aquaculture software. [Accessed May 2019].

Felsenstein, J. (1985). Confidence limits of phylogenies: An approach using the Bootstrap. *Evolution,* 39, 783–791.

Freshwater, D. W., Idol, J. N., Parham, S. L., Fernández-García, C., León, N., Gabrielson, P. W., & Wysor, B. (2017). Molecular assisted identification reveals hidden red algae diversity from the Burica Peninsula, Pacific Panama. Diversity, 9(2). https://doi.org/10.3390/d9020019 .

Goff, L. J., & Moon, D. A. (1993). PCR amplification of nuclear and plastid genes from algal herbarium specimens and algal spores. *J Phycol* 29:381–384.

Hall, B. G. (2017). *Phylogenetic Trees Made Easy* (5th Edition). Oxford University Press, Oxford. pp. 385.

Halling, C., Wikström, S. A., Lilliesköld-Sjöö, G., Mörk, E., Lundsør, E., & Zuccarello, G. C. (2013). Introduction of Asian strains and low genetic variation in farmed seaweeds: Indications for new management practices. Journal of Applied Phycology, 25(1), 89–95. https://doi.org/10.1007/s10811-012-9842-0 .

Hebert, P. D. N., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences,* 270, 313–321.

Jukes, T. H. & Cantor, C. R. (1969). *Evolution of protein molecules*. Pp. 21–132 *in* H. N. Munro, ed. Mammalian protein metabolism. Academic Press, New York.

Krueger-Hadfield, S. A. & Hoban, S. M. (2016). The importance of effective sampling for exploring the population dynamics of haploid-diploid seaweeds. *J. Phycol.,* 52, 1–9.

Krueger-Hadfield, S. A., Roze, D., Mauger, S., & Valero, M. (2013). Inter-gametophytic selfing and microgeographic genetic structure shape populations of the intertidal red seaweed *Chondrus crispus*. *Molecular Ecology,* 22, 3242–3260.

Leliaert, F., Verbruggen, H., Vanormelingen, P., Steen, F., López-Bautista, J. M., Zuccarello, G. C., & De Clerck, O. (2014). DNA-based species delimitation in algae. European Journal of Phycology, 49(2), 179–196. https://doi.org/10.1080/09670262.2014.904524

Liu, J., Goh, C., Loh, C., & Liu, P. E. I. (1998). A Method for Isolation of Total RNA from Fruit Tissues of Banana. *Plant Molecular Biology Reporter*, 1–6. .

Lemey, P., Salemi, M., & Vandamme, A. M. (eds). (2007). *The Phylogenetic Handbook*. Cambridge University Press, Cambridge. pp.750

Neiva, J., Assis, J., Fernandes, F., Pearson, G. A., & Serrão, E. A. (2014). Species distribution models and mitochondrial DNA phylogeography suggest an extensive biogeographical shift in the high-intertidal seaweed Pelvetia canaliculata. Journal of Biogeography, 41(6), 1137–1148. https://doi.org/10.1111/jbi.12278 .

Rosenthal, A., Coutelle, O., & Craxton, M. (1993). Large-scale production of DNA sequencing templates by microtitre format PCR. *Nucleic Acids Research.* 21, 173–174.

Sanbrook, J., Fritsch, E. F., & Maniatis, T. (1989). Molecular cloning: a laboratory manual. *Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.*

Santorum, J. M., Darriba, D., Taboada, G. L., & Posada, D. (2014). jmodeltest.org: selection of nucleotide substitution models on the cloud. *Bioinformatics*, 30, 1310–1311.

Saunders, G. W. & Kucera, H. (2010). An evaluation of rbcL, tufA, 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). Methods for DNA Barcoding Photosynthetic Protists Emphasizing the Macroalgae and Diatoms. In: DNA Barcodes: Methods and Protocols. Humana Press, Totowa, NJ, pp 207–222.

Saunders, G. W. & Moore, T. E. (2013). Refinements for the amplification and sequencing of red algal DNA barcode and RedToL phylogenetic markers: a summary of current primers, profiles and strategies. *Algae*, 28, 31-43.

Schneider, C. W. (2015). A molecular-assisted alpha taxonomic study of the genus *Centroceras* (Ceramiaceae, Rhodophyta) in Bermuda reveals two novel species. *Algae*. 30, 15-33

Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M. D. M., Breiner, H. W., & Richards, T. A. (2010). Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Molecular Ecology*, 19(SUPPL. 1), 21–31. https://doi.org/10.1111/j.1365-294X.2009.04480.x.

Tano, S. A., Halling, C., Lind, E., Buriyo, A., & Wikström, S. A. (2015). Extensive spread of farmed seaweeds causes a shift from native to non-native haplotypes in natural seaweed beds. Marine Biology, 162(10), 1983–1992. https://doi.org/10.1007/s00227-015-2724-7.

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

Ward, S. M. & Jasieniuk, M. (2009). Review: Sampling weedy and invasive plant populations for genetic diversity analysis. *Weed Science,* 57, 593–602.

Zuccarello, G. C., Critchley, A. T., Smith, J., Sieber, V., Lhonneur, G. B., & West, J. A. (2006). Systematics and genetic variation in commercial shape *Kappaphycus* and shape *Eucheuma (Solieriaceae*, Rhodophyta). *Journal of Applied Phycology*, 18(3–5), 643–651. https://doi.org/10.1007/s10811-006-9066-2

Zuccarello, G., Yeates, P., Wright, J., & Bartlett, J. (2001) Population structure and physiological differentiation of haplotypes of Caloglossa leprieurii (Rhodophyta) in a mangrove intertidal zone. *J Phycol,*37:235–244.

Zuccarello, G. C. & Lokhorst, G. M. (2005). Molecular phylogeny of the genus *Tribonema* (Xanthophyceae) using rbc L gene sequence data: monophyly of morphologically simple algal species. *Phycologia*, 44, 384–392.

Zuccarello, G. C., West, J. A., Kamiya, M. & King, R. J. (1999). A rapid method to score plastid haplotypes in red seaweeds and its use in determining parental inheritance of plastids in the red alga *Bostrychia* (Ceramiales). *Hydrobiologia*, 401, 207–214.