Darwin Tree of Life DNA barcoding of vascular and non-vascular plants and of lichens - Standard Operating Procedure

Version 1.6
Published date: November 2020
Revision date: May 2022
Authors: Laura Forrest and Michelle Hart

Purpose: This SOP outlines the method used to DNA barcode plants collected for the Darwin Tree of Life project. The DNA barcodes will be generated for all land plant material being collected by the Genome Acquisition Labs (GALs) and sent to Sanger for sequencing, in order to verify morphological identifications and to provide markers to track the final plant genomes back to the original samples.

Document history

Version	Date	Major changes	Contributors
1.0	August 2020	Draft version	Laura Forrest, Michelle Hart
1.1	November 2020	Revisions	Laura Forrest
1.2	December 2020	Revisions	Laura Forrest
1.3	January 2021	Addition of primer table and revisions	Laura Forrest
1.4	April 2021	Removal of sample submission information into stand-alone SOP, and revisions	Laura Forrest
1.5	June 2021	Lichen protocols added and general revision	Laura Forrest
1.6	May 2022	Revisions	Laura Forrest

Plant material

Plant material (typically leaf or thallus material, but may be other plant tissue such as flowers) or lichen thallus tissue will be collected and submitted by the GALs following the **RBGE DToL Sample collection Standard Operating Procedure_Vascular Plants** [link] or the **RBGE DToL Sample collection Standard Operating Procedure_Bryophytes v1.1** [link] and the **SOP RBGE Plant DNA Barcoding sample submission 1.1.docx** [link]; the GAL will also email a completed **DToL SAMPLE MANIFEST V2.3** for the collections that are to be barcoded [link]. The collection information on this will be used to generate an Edinburgh DNA (**EDNA**) number for each collection, so that they can be incorporated in our in-house DNA database and our DNA bank.

The **DToL sample manifests** are modified to only contain one row per specimen collection, with the **DToL number** field containing a comma-separated list of all the DToL numbers that apply to each specimen collection. This **barcode-adapted sample manifest** is saved in the DToL folder in the RBGE DNA drive.

EDNA numbers: An **EDNA upload form** (Excel spreadsheet available from the RBGE DNA server/Molecular lab registration forms) will be completed at RBGE for each batch of extractions, and uploaded to our in-house DNA database, **EDNA**. This spreadsheet comprises a mixture of compulsory and optional cells for each extraction, many of which are defined in some way.

Most of the EDNA information can be obtained from the **barcode-adapted sample manifest**. The changes to the way that this form is usually filled in that are specific for DToL samples are:

1. That the ‘Sample Note’ column (M) is filled in with all the **DToL numbers** that apply to each DNA extraction, separated by commas (no spaces). Unfortunately a limited number of characters are allowed in each field, so where the list is too long to upload to the database, the “DToL” is omitted from the later numbers in the series; for excessively long series, ranges have to be hyphenated. This can be cut and pasted from the **barcode-adapted sample manifest**, but will usually then need to be condensed as described.
2. That the **DNA extraction numbers** follow the DToL letter series described below.
3. That “Darwin Tree of Life sample” is added in the note column (T).
4. That the samples on the spreadsheet are uploaded to EDNA under the Darwin Tree of Life project.
5. The EDNA upload form is saved in the DToL barcode folder on the RBGE DNA server.

Once **EDNA numbers** have been assigned, an **EDNA.csv** file can be downloaded from the **EDNA database**. This will contain the **EDNA numbers** and **collector numbers** for each collection, associated with the **DToL numbers** for each sample. This file is saved in the DToL barcode folder on the RBGE DNA server, and is used to update the autofill lists with the new extraction details in the **GenePool form** and the **DToL GenePool form** (Molecular Lab Registration Forms folder on the RBGE DNA server; right click to open editable file), also to update the **Barcode lab book spreadsheet** (Data folder / DToL folder / barcode folder / lab books on the RBGE DNA server) and the **EDNA bank position forms** (Raw folder / EDNA folder on the RBGE DNA server: the **plate list**, the **plate positions list**, and the **EDbank numbers list**).

BOLD: The **barcode-adapted DToL sample manifest** and the **EDNA.csv** download are used to populate a **BOLD Version 3.1 Spreadsheet** (download from BOLD each time as the versions can change) to create **BOLD specimens**. The **EDNA number** is used as the **BOLD sample ID number**, while the complete **DToL specimen ID number/string of numbers** from the barcode-adapted manifest is added into the Notes field. The sample's major lineage is included in the Extra Info column, so that the project and datasets can be sorted into broad taxonomic groups (angiosperms, conifers, ferns, lycophytes, mosses, hornworts, liverworts, lichens).

Once the samples have been uploaded to the **BOLD EDToL project**, they are added to either the **RBGE or RBGK datasets (DS-DTOLRBGE; DS-DTOLRBGK)**, which team members from the GALs have access to, as well as to the relevant taxon datasets (**DS-FERN**, **DS-LICHEN**, **DS-LIVERWRT**, **DS-LYCOPHYT**, **DS-MOSS**, **DS-SEEDPLNT**).

The samples are downloaded from BOLD in order to get the **BOLD ProcessID number** that will be needed to upload the raw trace files to the **BOLD EDTOL project**. BOLD downloads can be saved to the DToL barcode folder on the RBGE DNA server.

All the relevant numbers for each sample should be pasted into a shared google sheet, **RBGE DToL Barcodin**g https://docs.google.com/spreadsheets/d/1pewankT7B0kEuEzwJqtOLLyuYhdJsDQdlGAjNvX_i_0/edit?usp=sharing . This must NOT be sorted, as it contains several linked sheets. (To sort, copy and paste the relevant columns into a new spreadsheet).

After **BOLD specimens** have been created, photographs of the plants can be uploaded to BOLD in a zipped folder including the **Specimen Image form Version 3.0** (download an empty form from BOLD) and the images. This has to be completed by the submitting GAL (or by the Barcode Hub if the GAL submits photos to them along with a **Specimen Image form**). Images should be named in a standard manner, preferably by the collector number as it appears in the

manifest, a one-word description, and a number if there are several images in a series, e.g. MR204_flower_1.jpg, MR204_flower_2.jpg; MR204_leaf.jpg, MR204_habitat.jpg. Copies of the submitted zipped folder should be saved, as if there are any problems in the future, e.g. a set of images needing deleted, BOLD will remove the entire submitted batch; it will be important to know exactly what was in the batch in order to resubmit it. In order to upload images to BOLD, the submitter needs edit rights to the BOLD EDTOL project, not to the individual GAL's datasets.

Reference databases

The **UK seed plants** have been DNA barcoded for two plastid markers (*rbc*L, *mat*K) and one nuclear marker (ITS2). This data is on BOLD and has been released to NCBI. However, there are some gaps in the dataset, most of which can be filled by a BLASTn search of GenBank.

The **UK lycophytes, liverworts and hornworts** have been DNA barcoded for three plastid markers (*rbc*L, *mat*K, *psb*A-trnH) and one nuclear marker (ITS2). This data is on private RBGE servers or private projects in BOLD. For several lineages good quality data is available on GenBank (i.e., projects involving well respected taxonomists; projects involving multiple samples per species).

A limited number of **UK mosses** (e.g. Bryaceae) have been DNA barcoded for three plastid markers (*rbc*L, *mat*K, *psb*A-trnH) and one nuclear marker (ITS2). This data is on private RBGE servers. However, there are no UK reference barcode libraries for most **UK mosses**. In these groups, the well-verified DTOL samples, alongside a small amount of additional sampling, will be used to populate barcode reference libraries, as all DToL material is expert-verified to the highest standards based on morphology. For several lineages good quality data is available on GenBank (i.e., projects involving well respected taxonomists; projects involving multiple samples per species).

There is potentially data for the **UK ferns** stored at the Natural History Museum, for plastid markers *rbc*L, *mat*K and trnL-trnF (pers. commun. Fred Rumsey). We do not currently have access to this data. For several lineages good quality data is available on GenBank (i.e., projects involving well respected taxonomists; projects involving multiple samples per species).

For the **UK lichen reference database** - check with Becky Yahr.

WET LAB MATERIALS

In the following list, general equipment and consumables that are typically available in molecular biology laboratories, such as benchtop centrifuges, water baths, heating blocks, orbital shakers, vortexers, thermocyclers, gel tanks and gel trays, agarose, UV or blue-light trans-illuminators, laminar flow hoods, fume hoods, water purification systems, autoclaves, micropipettes, tips, microcentrifuge tubes and tube racks, are generally omitted.

Plant and lichen DNA Extraction (Lab 31)

1. **Mixer Mill** (Retsch) or TissueLyser (QIAGEN), with 4 mm flattened "flying saucer" tungsten carbide or stainless steel beads, or 5 mm round tungsten carbide or stainless steel beads for single tube extractions. To avoid problems with tubes shattering, we do not use frequencies over 20 Hz.
   Tungsten beads can be cleaned for reuse by rinsing them clean of plant debris in water, sonicating to remove small fragments, and washing for c. 1 min in either 0.4 M HCl or a bleach solution to remove any remaining DNA, rinsing well with water and (optionally) autoclaving.
2. **Wide-based microtubes**
   (e.g. Axygen 2.0 mL microtubes, cat. no. MCT-200-C) (stops tissue and beads getting stuck in the end of the tube).
3. **Mini-pestles** and **acid-washed sand**
4. **Narrow-based microtubes**
5. **DNeasy® Plant Kit** (QIAGEN) – mini-columns and plates.
6. Other relevant plastics, including Rainen filter tips.

PCR (Lab 30)

1. **dNTPS**:
   combine stock solutions of dATP, dCTP, dGTP and dTTP into 500 µl aliquots of all four dNTPs at 2 mM each dNTP (e.g. if each stock dNTP is 100 mM, make a 1:50 dilution by combining 920 µL water and 20 µL each dNTP to make 1 ml working solution).
2. **Primers**, desalted:
   make up 100 µM stock solution. Working aliquots are a 1:10 dilution to 10 µM. Avoid excessive freeze-thaw cycles by making several aliquots of 200-500 µL, and keeping frequently-used working aliquots in the fridge in lab 30.
3. **Combined PCR additives**:
   a. TBT-PAR, 5x: 750 mM trehalose, 1 mg/mL nonacetylated bovine serum albumin (BSA), 1% Tween-20, 8.5 mM Tris hydrochloride, pH 8.0 (Samarakoon *et al,*. 2013).
   To make 10 mL of 5x TBT-PAR (Samarakoon *et al.,* 2013), make up a 20 mg/mL BSA solution (store in 0.5 mL aliquots in the freezer). Also make (and freeze 1 ml aliquots of) 10% Tween-20 detergent. Lastly, prepare 8.5 mL of 10 mM Tris-HCl buffer (8.415 ml water plus 85 µl 1 M Tris-HCl). Dissolve 2.835 g trehalose in 6 mL 10 mM Tris-HCl buffer; adjust the total volume to 8.5 ml, and add one 0.5 ml aliquot of BSA, and one 1 ml aliquot of Tween-20. Aliquot and freeze.
   b. Combinatorial Enhancer Solution (CES), 5x: 0.54 M betaine, 1.34% dimethyl sulphoxide (DMSO), 10 µg/µl BSA (Ralser *et al*., 2006).

To make 50 ml of 5x CES (modified from Ralser *et al.,* 2006), combine 27 ml of 5 M betaine, 3.35 ml of DMSO and 125 µl of 20 mg/ml BSA with 19.525 ml water. Aliquot and freeze.

4. **Single ingredient PCR additives**:
   a. **Betaine** (5 M):
      27g Betaine inner salt monohydrate made up to 40 ml with molecular biology grade Sigma water or Millipore filtered water.
   b. **Trehalose** (2 M):
      3.78g D-(+)-Trehalose dihydrate made up to 5 ml with molecular biology grade Sigma water or Millipore filtered water.
   c. **DMSO**: as supplied.
   d. **BSA**: stock at 20 mg/mL.
5. **Water:** Molecular biology grade Sigma water or Millipore filtered water.
6. **Polymerase**:
   a. Low-cost unspecialized taq polymerase (e.g. **BIOTAQ™ DNA polymerase** from Bioline) for *rbc*L, *psb*A-trnH and ITS, with buffer and $MgCl_2$.
   b. Specialized polymerases for *mat*K amplification and other problematic PCRs, e.g. **Platinum® Hot-Start Taq DNA polymerase** (Invitrogen™), **Phusion® High Fidelity DNA polymerase** (New England Biolabs/ Thermo Scientific®), with buffers and $MgCl_2$.

Template quantification (Lab 33)

1. **DNA gel stain**: SYBR®Safe (Invitrogen).
2. **Tris-Borate-EDTA** (TBE) buffer (kept at 10x or 5x stock; 1x working solution) and agarose.
3. **Loading solution**:
   30% glycerol, 0.25% bromophenol blue, Millipore water.
4. **DNA ladder**: 1 Kb Plus DNA ladder (Invitrogen):
   to make working solution aliquots, combine 100 µL stock, 250 µL Sigma gel loading buffer and 650 µL Sigma water.

Template clean-up (Lab 33)

ExoSap-IT (GE Healthcare).

DNA sequencing (Lab 33)

1. **ABI BigDye® Terminator Cycle Sequencing kit** (either version 1.1 or 3.1) plus BigDye® Terminator v1.1, v3.1 5x sequencing buffer.
2. **BigDye® Enhancing Buffer BDX64** (MCLAB): big dye dilution buffer.
3. **Primers**, desalted:
   make up 100 µM stock solution. Working aliquots are a 1:10 dilution to 10 µM. Avoid excessive freeze-thaw cycles by making several aliquots of 200-500 µL, and keeping frequently-used working aliquots in the fridge in lab 33.

Methods

DNA extraction – Lab 31

1. Sampling. All lab samples for extraction are renumbered in consecutive order, with a short simple numbering scheme, rather than using the **DToL sample numbers**, as the numbers have to be transcribed onto several plastic tubes during the extraction process and this is least prone to error. Note the numbers both in a lab book and by annotating the individual silica gel packages on their top right hand corners. Current **RBGE DToL DNA extraction numbering series** for barcoding are prefaced with **B** (mosses), **L** (liverworts), **F** (ferns and lycophytes), **A** (seed plants) and **X** (lichens). These temporary tube numbers must be entered into the **EDNA submission spreadsheet** and will be recorded in the **EDNA database**.

In order to best troubleshoot contamination issues downstream, where possible avoid placing samples from the same genus, and particularly from the same species, into consecutive tubes - although current extraction numbering usually follows the collector numbers.

2. Homogenization.

Plants. On a vented bench, using clean forceps, put around 1-2 cm^2 dry plant material into a labelled open-based Axygen 2 mL microcentrifuge tube, along with a flattened tungsten bead. The plant tissue should be slightly broken up prior to its homogenization in the tissue lyser. If it ends up lying vertically in the tube the bead may avoid hitting it and the sample will not consistently grind to a fine powder. Distribute the tubes or plates between two balanced adaptors. Homogenize for 2 mins at 20 Hz, rotate the adaptor 180º, then homogenize for a further 2 mins at 20 Hz. If the sample has ground poorly, add a second flattened bead and repeat.

Lichens. Some of the lichen barcoding samples are stored in a freezer in lab 30, in 2 ml tubes in TE buffer. Remove the samples from the freezer, allowing plenty of time for the buffer to defrost prior to homogenization. The lichen thallus will have to be removed from the tube with forceps (do not use No. 5 forceps as these do not reach into the bottom of the tube); the forceps must be cleaned with alcohol between samples. A small amount of lichen tissue can be transferred to a wide-based 2 ml tube with 2 flattened tungsten beads for homogenization using a tissue lyser (see above), or into a narrow-based 1.5 ml tube with a small quantity of acid-washed sand to be hand-ground using a disposable plastic mini-pestle. If using a mini-pestle, adding the first 200 µL of pre-heated 65°C QIAGEN lysis buffer can help the sample grinding. Other lichen barcode samples are air-dried, and can be treated the same way as dry plant samples.

3. DNA extraction using DNeasy® Plant kits (QIAGEN): follow manufacturer's protocols, but with the incubation period in lysis buffer extended to an hour at 65°C in a thermomixer block set to 800 RPM. The on-ice incubation can also be extended (so is a good point for a lunch break). Avoid using too much starting material as this can block the DNeasy column membrane, and always centrifuge for 5 minutes at full speed after the first incubation (this is an optional step in the DNeasy manufacturer's protocol) to pellet the homogenized material. It is particularly important not to overdo starting material from mucous-rich plants. Use filter tips whenever transferring or mixing liquids with DNA in them.

Following the lysis step, and depending on sample numbers, the DNA extraction can either continue into individual QIAGEN plant DNeasy mini-columns, or be transferred into a QIAGEN DNeasy extraction plate for bulk processing. However due to the use of different marker sets and primers, samples from different lineages should never be mixed throughout a plate. If a plate is mixed at all, it should be in such a way that DNA from different lineages forms complete strips of 8 that can be rearranged into lineage specific sets prior to PCR.

A double elution, with 75 µL pre-heated 65°C QIAGEN elution buffer at the first step and 50 µL pre-heated 65°C QIAGEN elution buffer at the second step, is recommended, giving a final extraction volume of c. 100 µL.

As of autumn 2021, we are also extracting plant barcoding samples using MagAttract Plant kits, following the manufacturer's protocols.

4. Storage

Silica gel dried plant tissue – short-term this is stored in plastic bags in closed plastic DToL labelled boxes in lab 31, in case extractions have to be repeated. Once successful sequences have been obtained for each batch, the silica gel dried samples and the relevant **barcode-adjusted sample manifest** and **EDNA accession numbers** will be passed on to Herbarium staff for long term curation in the **RBGE silica-dried tissue store**.

TE-preserved lichen material – if any lichen thallus is left, short-term the tubes are returned to the freezer drawer for re-freezing, in case extractions have to be repeated.

DNA: Short-term, DNA is stored in labelled 1.7 mL elution tubes, in the fridge in lab 30, or in the DToL drawer of the freezer in lab 30. Tubes are arranged taxonomically in racks (i.e. by the temporary extraction numbers: A = seed plants; F = ferns and lycophytes; B = mosses; L = liverworts, X = lichens, each in different racks). Once good quality barcode DNA sequences have been obtained for all samples within a set of tubes, the DNA will be transferred to the **RBGE DNA bank**, in barcoded fluidX tubes and 96-tube racks stored in the -80 freezer in the

canteen corridor. The tube and plate barcodes will be scanned and their details entered into the relevant database.

Standard barcoding PCR reactions. Lab 30

Marker choice

For plant accessions that are required for DToL identification/sample tracking purposes, liverworts and hornworts will be amplified for *rbc*L, ITS2 and *psb*A-trnH, mosses will be amplified for *rbc*L and ITS2, ferns and lycophytes for *rbc*L, and seed plants for *rbc*L and ITS2. The lichen marker is ITS.

Table 1. Land Plant barcoding primers for DToL. Preferred primers are highlighted in bold font.

Locus	Primer name (as given in BOLD)	RBGE Primer name/ code	Taxon	Primer direction	Primer sequence	Primer reference
***rbc*L**	**rbcLa_f**		**Land plants**	**Forward**	**ATGTCACCACAAACAGAGACTAAAG**	Kress WJ, Erickson DL (2007) A Two-Locus Global DNA Barcode for Land Plants: The Coding rbcL Gene Complements the Non-Coding trnH-psbA Spacer Region. PLOS ONE 2(6): e508. https://doi.org/10.1371/journal.pone.0000508
***rbc*L**	**rbcLajf634R**		**Land plants**	**Reverse**	**GAAACGGTCTCTCCAACGCAT**	Fazekas, A.J. et al. 2008. Multiple Multilocus DNA Barcodes from the Plastid Genome Discriminate Plant Species Equally Well. PLOS One 7 e2802
*rbc*L	M745R		Bryophytes	Reverse	CTTCACAWGTACCTGCRGTAGC	Lewis, Mishler & Vilgalys. 1997. Phylogenetic Relationships of the Liverworts (Hepaticae), a Basal Embryophyte Lineage, Inferred From Nucleotide Sequence Data of the Chloroplast Gene rbcL. Mol Phylogenet Evol. 1997 Jun;7(3):377-93. doi: 10.1006/mpev.1996.0395
*rbc*L	rbcL-aar		Land plants	Reverse	CTTCTGCTACAAATAAGAATCGATCTC	Kress WJ, Erickson DL (2007) A Two-Locus Global DNA Barcode for Land Plants: The Coding rbcL Gene Complements the Non-Coding trnH-psbA Spacer Region. PLOS ONE 2(6): e508. https://doi.org/10.1371/journal.pone.0000508
***psb*A-trnH**	**psbA501F**		**Non-seed land plants**	**Forward**	**TTTCTCAGACGGTATGCC**	Cymon Cox in Forrest LL, Crandall-Stotler BJ (2004) A phylogeny of the simple thalloid liverworts (Jungermanniopsida, subclass Metzgeriidae) as inferred from five chloroplast genes. In Goffinet B, Hollowell V, Magill R (eds.) Molecular Systematics of Bryophytes. Monographs in Systematic Botany, Missouri Botanical Garden, 98, 119-140
***psb*A-trnH**	**trnHR**		**Land plants**	**Reverse**	**GAACGACGGGAATTGAAC**	Sang et al. 1997. Chloroplast DNA Phylogeny,Reticulate Evolution, and Biogeography of Paeonia (Paeoniaceae). Amer. J. Bot. 84(9): 1120-1136
ITS2	**ITS2.seqF**		**Bryophytes**	**Forward**	**AACAACTCTCAGCAACGG**	Olsson et al (2009) Bryologist 112: 447-466

ITS2	**ITS.4bryo**		**Bryophytes**	**Reverse**	**TCCTCCGCTTAGTGATATGC**	Stech et al. (2003) Australian Syst. Bot. 16: 561-568
ITS2		**ITS2For or S2F (Primer box 003:H9**	**Seed plants**	**Forward**	**ATGCGATACTTGGTGTGAAT**	Chen S., Yao H, Han J, Liu C, Song J, Shi L, Zhu Y, Ma X, Gao T, Pang X, Luo K, Li Y, Li X, Jia X, Lin Y, Leon C, 2010. Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. PLoS ONE 5, 1–8. Chiou SJ, Yen JH, Fang CL, Chen HL, Lin TY, 2007. Authentication of medicinal herbs using PCR-amplified ITS2 with specific primers. Planta Medica 73, 1421–1426
ITS2		**ITS3R or S3R (Primer box 003:H10)**	**Seed plants**	**Reverse**	**GACGCTTCTCCAGACTACAAT**	Chen S., Yao H, Han J, Liu C, Song J, Shi L, Zhu Y, Ma X, Gao T, Pang X, Luo K, Li Y, Li X, Jia X, Lin Y, Leon C, 2010. Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. PLoS ONE 5, 1–8. Chiou SJ, Yen JH, Fang CL, Chen HL, Lin TY, 2007. Authentication of medicinal herbs using PCR-amplified ITS2 with specific primers. Planta Medica 73, 1421–1426
ITS		**ITS1F**	**lichens**	**forward**	**CTTGGTCATTTAGAGGAAGTAA**	
ITS		**ITS4**	**lichens**	**reverse**	**TCCTCCGCTTATTGATATGC**	White et al. 1990
trnL		trnL_C	ferns	forward		
trnL		trnL_F	ferns	reverse		

PCR amplification. Routinely use PCR enhancers TBT-PAR (Samarakoon *et al*., 2013) or CES (Ralser *et al*., 2006). Set up PCRs in strips rather than plates, due to a high rate of failure of samples around the margins of plates (probably through poor lid seals).

For plants, in 20 µl reactions: 2 µl PCR buffer (at 10x), 2 µl dNTPs (each at 2 mM), 0.6 µl $MgCl_2$ (at 50 mM), 4 µl TBT-PAR or CES additive (both at 5x), 2 µl each primer (at 10 µM), 0.25 µl BIOTAQ (at 5 units/µl) and 1 µl template DNA, made up to 20 µl with Sigma PCR-grade water. (Per reaction, this consists of 1x PCR buffer, 0.2 mM each dNTPs, 1.5 mM $MgCl_2$, 1x additive, 1 µM of each primer, 2.5 U taq polymerase and c. 1 ng template DNA).

For lichens, in 12.5 µl reactions: 1.25 µl PCR buffer (at 10x), 1.25 µl dNTPs (each at 2mM), 0.6 µl $MgCl_2$ (at 50 mM), 2.5 µl TBT-PAR additive (at 5x), 0.5 µl each primer (at 10 µM), 0.13 µl BIOTAQ (at 5 units/µl) and 1 µl template DNA (2-5 ng/reaction), made up to 12.5 µl with Sigma PCR-grade water. (Per reaction, this consists of 1x PCR buffer, 0.2 mM each dNTPs, 2.4 mM $MgCl_2$, 1x additive, 0.4 M of each primer, 0.05 U/µl taq polymerase and c. 2-5 ng template DNA)

Standard PCR parameters (plants):

***rbc*L:** 2 min initial denature at 94°C followed by 35-40 cycles with a 1 min denature at 94°C, 1 min anneal at 51°C and 90 sec extension at 72°C, with a final extension of 72°C for 5 min.

***psb*A-trnH:** 3 min initial denature at 94°C followed by 2 cycles with a 45 sec denature at 94°C, 45 sec annealing at 50°C and 1 min extension at 72°C, then 30 cycles with a 45 sec denature at 94°C, 45 sec anneal at 45°C and 1 min extension at 72°C, with a final extension of 72°C for 5 min.

ITS2: 4 min initial denature at 95°C followed by 30 cycles with a 1 min denature at 94°C, 1 min annealing at 55°C and 45 sec extension at 72°C, with a final extension of 72°C for 5 min.

Standard PCR parameters (lichens):

ITS: 4 min initial denature at 94°C followed by 30 cycles with a 45 sec denature at 94°C, 1 min 30 sec anneal at 55°C and 90 sec extension at 72°C, with a final extension of 72°C for 10 min.

PCR products are stored short-term in labelled racks (project/ date/ locus) in the fridge in lab 33. After visualization, and after successful sequencing if PCR has been successful, tubes are discarded into the plastic recycling bin.

PCR visualization - Lab 33

Approximately 3.5 µl of the PCR product should be added to 1.5 µl of a glycerol/bromophenol blue loading dye in a plastic plate, using the electric Rainin multichannels, and run, for 40 mins to an hour, on a 1% agarose 1x TBE-buffer gel stained using SYBR Safe, with 3.5 µl of 1 kb ladder every third row. (Tips for aliquoting the loading dye are in a labelled red Rainin box, and can be reused multiple times).

There is a Rainin manual multichannel pipette and two flat 96-sample gel rigs (use 100 ml agarose-TBE gel and 5 µl SYBR Safe per rig). Gels can be melted down and re-poured twice. The **gel image jpg** should be labeled with the locus name and date, and saved in the DToL Gel folder on the DNA server, from where it can be copied into the DToL DNA barcoding **lab book PCR spreadsheet**.

Short term, store the PCR products back in the fridge in lab 33. If it is likely to take more than a couple of days to get around to cleaning and processing them, they can be stored in the freezer in lab 33 to reduce the chance of the reactions drying out.

PCR failure

If there is no band on the gel for a given amplification, the PCR will be repeated later.

If PCR of an extraction fails across both/all loci, DNA may be quantified on an agarose gel and by fluorometry or just discarded, and a repeat DNA extraction made from the silica-gel dried tissue sample stored in lab 31. If PCR just fails for one or two loci, it may be repeated using an alternative additive (TBT-PAR or CES) as well as with a 1:25 dilution of the original DNA.

If there are multiple bands, particularly for variable-length markers (usually ITS2 and *psb*A-trnH), PCR may be repeated with alternative primer pairs (where available), or at a higher annealing temperature. If this happens repeatedly then the DNA can be re-extracted from the silica-gel dried tissue, to rule out sample contamination, or the sample may pass barcoding with just a single DNA barcoding locus.

At least one repeat PCR should be carried out for all poor / failed amplifications in a batch.

If over 20% of samples have failed for a marker that is usually successful (e.g. *rbc*L), the entire batch of PCRs are discarded and the PCR is repeated, to save time cherry-picking successful samples.

PCR clean-up - Lab 33

Sample dilution: Depending on the brightness of the PCR band on an agarose gel when compared to a standard, e.g. 3.5 µl of NEB's 1kb DNA ladder – successful barcode PCRs following the above protocols usually have VERY bright bands. For a good PCR product, dilute with 30 µl of Sigma water per well (using the repeat 1000 µl multichannel). It may be necessary to add 45 µl water to extremely bright products, conversely only 10-15µl to weaker products, and no dilution for very faint bands. (**This step is important!** Our current sequencing service has far more sensitive machines than the old service, and failure to dilute strong PCR product will cause signal leakage between capillaries, making the sequences unusable.)

ExoSAP-IT: After the samples are diluted, clean them using the commercial ExoSap-IT (Affymetrix) mix.

- The addition of 1-2 µl of ExoSap-IT to a 10-20 µl PCR is generally more than sufficient (manufacturer's recommendations are addition of 2 µl ExoSap-IT to 5 µl PCR product).
- PCR clean-up incubation parameters: 37°C for 30 min for optimal enzymatic activity, followed by 80°C for 15 min to deactivate the enzymes, then storage at 8-12°C.

Aliquot the ExoSap-IT into a strip, and add 2 µl to each PCR product using a 10 µl repeater multichannel pipette, putting the ExoSap-IT near the top of each well to reduce the risk of carrying over contamination. Reseal lids. Spin down strips, and run the incubation in one of the PCR machines in lab 33.

Cleaned PCR product is stored short-term in labelled racks (project / date/ locus/ cleaned) in the fridge in lab 33. The strips are subsequently discarded, following successful sequencing, into the plastic recycling bin.

Sequencing reactions - Lab 33

We routinely use BDX-64 cycle sequencing additive to reduce the amount of BigDye in each 10 µl reaction to 0.125 µl and to cut the sequencing PCR time; further dilutions of BigDye may be possible for material that routinely gives high quality DNA sequences.

Sequencing PCR (in a total volume of 10 µl): 1.5 µl BigDye® Terminator sequencing buffer, 0.125 µl BigDye®, 0.32 µl each primer, 0.875 µl BDX-64 cycle sequencing additive, 1.5 µl template, made up to 10 µl with Sigma water. Distribute 8.5 µl master mix using a repeater pipette to the bottoms of the wells. Add PCR product near to the top of wells, using a multichannel pipette set to deliver two loads of 1.5 ul (for the forward and reverse reactions). Seal with a labelled foil lid (in the format RBGE_dateletter, e.g. RBGE_210628A, etc, and with a note to say "sample for clean-up and sequencing, lane names in attachment"), and transfer to one of the BioRad PCR machines in lab 32 for cycling using the BDXsq programme.

Sequencing thermocycling parameters, following BDX-64 manufacturer's instructions: a 3 min initial denature at 96°C then 30 cycles with 10 sec denature at 96°C, 5 sec annealing at 50°C and 2 min extension at 60°C.

(Sequencing thermocycling parameters for samples without BDX-64, following BigDye manufacturer's instructions: 25 cycles with a 30 sec denature at 95°C, 20 sec annealing at 50°C and 5 min extension at 60°C.)

A completed **RBGE-DToL GenePool form** (available from the User folder on the RBGE DNA server) for each batch of sequencing must be sent to sequencing@rbge.org.uk, and sequencing strips or plates placed in the left hand drawer of the fridge in lab 33. Samples are sent first class post to DNA Sequencing & Services at the University of Dundee (as either half or full plates; it is not possible to send one or two samples) for clean-up and Sanger sequencing on an ABI 3730. (Instructions for sending sequencing to Dundee are available in the Raw folder of the RBGE DNA server.)

Sequence reads are usually returned within c. 2-3 days from posting date, and are downloaded as zipped folders by plate from the sequencing provider web site. The folders are unzipped, the sequence text files deleted, and the ABI read files are moved into subfolders by locus and by lineage. These are then added to the relevant Sequencher file (see below).

When the returned sequencing reactions fail, first the reads are checked to rule out microsatellites close to the primers causing the failure, and if that is not the reason, the PCR gel images are checked to estimate the quality of the PCR product (band tightness and brightness).

- If band quality is high and microsatellites were not the problem, sequencing can be reattempted from the stored exosapped PCR product.
- If band quality is low, PCR will be repeated as above (PCR failure section).
- If microsatellites are causing the problem, unidirectional sequence will be accepted if it is of high quality, or alternative primers may be tried if available.
- Unreadable sequences for ITS2 can also be due to co-amplifying endogenous fungi, particularly in liverworts, so failures in these cases are to be expected and will not be rectified by redoing the extraction or the PCR; occasionally PCR with alternative primers rectify this but time constraints will generally not make this feasible.
- If sequence reads contain more than one peak at any position (except for where wobbles are likely to be real i.e. a few wobbles in diploid or concatenated markers, in this case the nuclear ribosomal repeat region ITS), indicating contamination, PCR will be repeated if only a single marker is affected (therefore likely due to contamination during PCR or sequencing), or the extraction will be repeated if all markers are similarly affected.

Where a high quality unidirectional sequence read exists, the read will usually be accepted for DToL purposes and uploaded to BOLD, as waiting for the missing sequence to get a bidirectional read could cause a month or more delay to the sample. However, the sample may still be rerun at a later date and the read data on BOLD subsequently updated.

Bioinformatics - Office

An **RBGE DToL GenePool form** (a modification of our existing sequencing **GenePool form**) must be completed prior to sequencing being sent to Dundee. This form concatenates the morphological identification (genus and specific epithet), the **EDNA numbe**r (which is also the **BOLD sample ID**), the BOLD primer name and the submitting GAL in a standardized format, to be incorporated into the sequence read file name (simplifying the deposition process on BOLD). The ABI file names must be less than 100 characters to upload successfully onto BOLD; the **RBGE DToL GenePool form** has a column that gives the total predicted character length of the ABI file name. Please highlight or adjust any file names that will be over 99 characters.

When filling out the **RBGE DToL GenePool form,** the **User Name** for any DToL samples should be "**DToL**". This form is emailed to sequencing@rbge.ac.uk, while the plates or strips for sequencing are left in the Sequencing drawer of the fridge-freezer in lab 33.

One sequencing has been completed, reads from each marker and each major lineage (liverwort, moss, hornwort, fern+lycophyte, seed plant) are uploaded to a separate **Sequencher file** for assembly and editing. The sequencing service alters the dash "-" character in the EDNA number to an underscore, "_". These have to be changed back to dashes to match the **BOLD sample_ID number** (i.e. the **EDNA number**).

Bidirectional reads are assembled by name into contigs that are labelled with the morphological identification and the **BOLD sample_ID number** (i.e. the **EDNA number**), and the base calls manually checked using the Sequencher software.

Once sequences have been manually edited in Sequencher, the contigs are converted to text, aligned (if coding), checked for indels (if coding), and exported from Sequencher in either aligned (*rbc*L) or unaligned (ITS, ITS2, *psb*A-trnH, trnL-trnF) FASTA format. The **FASTA files** are checked for laboratory mistakes, contamination issues and major errors in reading the sequence data using the Basic Local Alignment Search Tool for nucleotide sequence data, **BLASTn**, against the NCBI database, as well as by comparison to any suitable private databases. Where there are problems (e.g. fungal sequences generated for ITS2), any trace files in BOLD can be flagged as contamination. **The submitting GAL and/or plant collector is notified, by**

email, of any identification issues that arise at this point, and any identification queries are entered into the DToL DNA barcoding ID verifications Google sheet https://docs.google.com/spreadsheets/d/1QUaxQzYWBDbeFOesrM7GsrUvctHw-XVqgfFYaQinjdE/edit?usp=sharing.

The **BLAST-verified FASTA text files** are modified to remove the taxon name and replace the underscore after the **EDNA number** with a pipe character ("|"), saved in the DToL folder with "BOLD" appended to their filename, then uploaded to the **BOLD EDTOL project** under the appropriate barcoding marker (sequencing centre = Royal Botanic Garden, Edinburgh; copy and paste the FASTA text).
Basic searches can be run within BOLD using *rbc*L, but not currently for any other plant barcoding marker used for DToL.

Passing / failing samples

Species that do not pass barcoding may need recollected from the wild, with new DNA barcoding, genome sizing, herbarium vouchering, etc, as well as a repeat of the cold chain for the high molecular weight sample - i.e. a barcode fail is expensive and time-consuming, so needs more consideration than one simple BLASTn search.

- **If only one marker has been successfully sequenced:**
    - In most cases the ITS2 marker will be the one which fails (for many reasons including fungal coamplification and microsatellites), and it may not be possible to get a good read for it even if it is re-amplified and re-sequenced.
    - When the successfully sequenced marker is adequate to confirm the identity of the taxon (e.g. 100% match to that taxon; other highest hits not found in the UK flora, etc) then the sample can still be marked as passing barcoding.
- **If the taxon is missing from the reference database:**
    - An informed decision must be made - e.g. does it BLAST to something that we would expect to be a close relative? Using the tree-building function in BOLD, does the taxon sit roughly where expected? On the contrary, if the sample is 100% or close match to something that is not a close relative of its expected taxon, it should fail barcoding despite the gap in the reference database.
    - Longer term, a second sample of the taxon of interest, preferably identified by a different taxonomist than identified the original sample, should be added to the DToL barcode sequencing to act as a reference. This reference sample can come from verified silica dried tissue or a herbarium specimen.
- **If there is a 100% match to multiple species:**
    - Do these include the expected taxon? If so the barcode data does not contradict the morphological identification and the sample passes.

- **If there is a 100% match to a different species and a lower match to the expected taxon:**
    - Are the species sometimes synonomised? Taxonomy may be known to vary (e.g. species has been split recently so that older GenBank sequences represent a different species concept, e.g. *Conocephalum conicum*.
    - Is the expected taxon sampled from a very different geographic location (e.g. from another continent)? Species concepts are not always applied uniformly geographically, e.g. *Metzgeria conjugata* in UK and North America. Morphologically similar things on different continents are often given the same name until genetic research shows them to be distinct.
    - Is the mismatch due to wobbles in either the sample sequence or the reference sequence? (Particularly common for heterozygous or concatenated loci like ITS).
    - Is there length variation in the reference database, so that a higher similarity can be caused by less variable stretches of the marker being included in the database for a less closely related taxon?

Samples where the morphology-based identification conflicts with the DNA-barcoding-based identification for any marker are flagged, lab protocols reviewed for potential contamination issues, and the submitting GAL contacted for verification. I**dentification queries are entered into the DToL DNA barcoding ID verifications Google sheet** https://docs.google.com/spreadsheets/d/1QUaxQzYWBDbeFOesrM7GsrUvctHw-XVqgfFYaQinjdE/edit?usp=sharing. Until any issue is resolved, the name of the specimen in BOLD is edited to reflect uncertainty (e.g. by removal of the epithet), with a taxonomic note added to explain that there is currently a conflict between the morphological and DNA-based identification of the specimen.

Accessions with unexplained conflict between the GAL's reverification and the molecular data may be re-extracted from the silica gel dried tissue, after in-house taxonomic review of the silica gel dried tissue for obvious issues (e.g. mixed collections). In cases where the silica gel dried tissue is clearly problematic, DNA extractions may be attempted from the herbarium vouchers instead.

On the other hand, in cases where the taxonomic identification of the sample changes in line with the DNA barcode data, the specimen in BOLD will be edited to reflect the correct taxon name. The sample may no longer represent a DToL priority species for genome sequencing, and could even have already been sampled for the project. However, it will still have all the associated "extended specimen" metadata and physical herbarium and DNA samples, so remains a DToL collection even if it is not required for whole genome sequencing.

All DToL plant barcoding data on BOLD is open access.

Once the reads have been manually edited in Sequencher and passed basic quality checks (e.g. that none of the plates have accidentally been reversed) and primer sequences and low quality regions have been have been deleted, the **SCF2.0 files** for the project are exported from Sequencher. There is a **BOLD conversion worksheet** that matches **BOLD sample IDs** with **BOLD Process IDs**, allowing the **BOLD Version 3.0 trace upload form** to be completed. The trace files are then uploaded to the **BOLD EDTOL project** in a zipped folder containing the reads and the completed **Version 3.0 form** (this includes the **read file names** (which must be less than 100 characters), the **BOLD Process_IDs**, the **BOLD locus name**, the **BOLD PCR primer names** and the **BOLD sequencing primer name**) (the form and the accepted locus and primer names are available from the BOLD website).

Notes

DNA extraction plates, and sequence batches, can contain a mixture of DToL genome samples and barcode reference samples. The reference samples will not necessarily have the associated material to get through the DToL manifest validation – e.g. for mosses a lot of reference samples will be historic collections from the Herbarium and lack the same level of sampling permissions. The pipeline needs to include a way of processing both DToL samples and reference samples.

Non-chlorophylous plants – there are frequently stop codons and indels in plastid barcode markers due to decay / pseudogenization (which leads BOLD to flag the data as unreliable due to stop codons), as well as potential problems with contamination from the host plant.