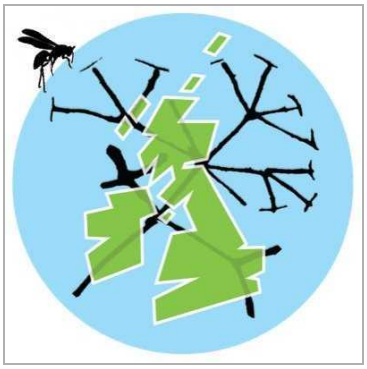

JAN 20, 2023

OPEN ACCESS

Protocol Citation: Laura L Forrest, David Bell, Michelle Hart 2023. DNA Barcoding Standard Operating Protocol, Plants and Lichens at RBGE, Sanger Sequence Data. **protocols.io** https://protocols.io/view/dna-barcoding-standard-operating-protocol-plants-a-cmwcu7aw

License: This is an open access protocol distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working
We use this protocol and it's working

Created: Jan 17, 2023

Last Modified: Jan 20, 2023

PROTOCOL integer ID: 75428

Keywords: DToL, RBGE, Plant Barcoding, Darwin Tree of Life, Sanger sequencing, contig, electropherogram, chromatogram, BOLD, BLASTn, rbcL, psbA-trnH, ITS2, ITS

DNA Barcoding Standard Operating Protocol, Plants and Lichens at RBGE, Sanger Sequence Data

Forked from DNA Barcoding Standard Operating Protocol, Plants and Lichens at RBGE, Sample Data

In 1 collection

Laura L Forrest[1], David Bell[1], Michelle Hart[1]

[1]Royal Botanic Garden Edinburgh

Laura L Forrest
Royal Botanic Garden Edinburgh

ABSTRACT

This is part of the collection DToL Taxon-specific Standard Operating Procedures for the Plant Working Group (protocols.io). The SOP collection contains guidance on how to process the various land plant taxa within the scope of the Darwin Tree of Life project. The guidance specifically refers to the tissue samples needed for DNA barcoding (which takes place at the Royal Botanic Garden (RBGE)). Every specimen is submitted for DNA barcoding first before potentially being sent to the Wellcome Sanger institute.

This DNA barcoding SOP outlines the processing of plant and lichen data for the Darwin Tree of Life project (DToL) at the Royal Botanic Garden Edinburgh (RBGE).

DNA barcoding is used as part of the species identification process AND sample tracking (to check that the genome sequence corresponds to the material that was sent and that there have been no sample mix-ups).

Definition: Land plants (Embryophyta) and lichens
Including: Bryophyta, Marchantiophyta, Anthocerotophyta, Lycopodiophyta, Polypodiophyta, Pinophyta, Cycadophyta, Ginkgophyta, Gnetophyta, Magnoliophyta, lichenized fungi

Excluding: all non-lichenized fungi

GUIDELINES

Including: Bryophyta, Marchantiophyta, Anthocerotophyta, Lycopodiophyta, Polypodiophyta, Pinophyta, Cycadophyta, Ginkgophyta, Gnetophyta, Magnoliophyta, lichenized fungi

Note

Previous versions of the Plant Working Group SOPs can be found here:
RBGE DToL Sample collection Standard Operating Procedure Vascular Plants
RBGE DToL Sample collection Standard Operating Procedure Bryophytes
SOP RBGE Plant DNA Barcoding sample submission

Naming trace files

- At RBGE we use an excel sheet to generate trace file names.
- Our **DToL Sequencing form** concatenates the morphological identification (genus and specific epithet), the DNA number (our EDNA number, which is also the BOLD sample ID), the BOLD primer name and the submitting Genome Acquisition Lab in a standardized format, to be incorporated into the sequence read file name (simplifying the process of deposition data onto BOLD).

Note

The ABI file names must be less than 100 characters to upload successfully onto BOLD; we have included a column on our DToL Sequencing form that gives the total predicted character length of the ABI file name, to allow you to identify and adjust any file names that will be over 99 characters.

Sequence editing

1. One sequencing has been completed, we upload the reads from each marker and each major lineage (liverwort, moss, hornwort, fern+lycophyte, seed plant) into a separate file for assembly and editing. We use GeneCodes Sequencher software for this.
2. Bidirectional reads are assembled by name into contigs, retaining the morphological identification and the BOLD sample ID number as the contig name. Primer sequences are deleted and base calls manually checked.
3. Once sequences have been manually edited, the contigs are converted to text, aligned (if coding), the alignment checked for indels or stop codons (if coding), and exported from Sequencher in either aligned (rbcL) or unaligned (ITS, ITS2, psbA-trnH, trnL-trnF) FASTA format.

Note

While automating this process would be significantly faster, manual editing of trace data increases the amount of useable data obtained, particularly in cases where sequence quality is variable across a trace, as is particularly common in traces with one or more short sequence repeat regions.

Verifying sequence data

1. The FASTA files are checked for laboratory mistakes, contamination issues and major editing errors, 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. This is also useful for spotting any sequences that should be reverse complemented. Potential editing mistakes can be resolved by rechecking the original trace files.
2. Where there are problems (e.g. fungal sequences generated for ITS2), any trace files that have already been uploaded to BOLD should be flagged as contamination.
3. The Genome Acquisition Lab and/or plant collector should be notified of any identification issues that arise with their specimens.

Note

We enter any identification queries into a DToL DNA barcoding ID verifications Google sheet as soon as we become aware of them. This sheet also allows us to keep a record of any identification changes, and where updates have been made.

Preparing sequencing data for BOLD

Sequence reads:

1. Our FASTA text files require some modification before they can be uploaded to BOLD, to remove the taxon name, and either delete any text in the read name following the BOLD sample ID, or add a pipe character (“|”) directly after the BOLD sample ID.
2. FASTA files named by the BOLD Sample ID (or the BOLD Process ID) can be uploaded to the relevant BOLD project (for DTOL samples this is the EDTOL project), under the appropriate barcoding marker and sequencing centre (e.g. Royal Botanic Garden, Edinburgh), by copying and pasting the FASTA text.

Trace Files:

1. 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 so the forwards and reverses don't match...) and primer sequences and low quality regions have been deleted, the SCF 3.0 files for the project can be exported from Sequencher into a new folder, ready to upload to BOLD.
2. Using the Terminal or Windows PowerShell and the ls command, get a list of all the SCF file names in the folder. Pasting these into an excel file and using the "convert to text" command is an easy way to extract the BOLD sample ID, primer and locus information from the name of each trace file.
3. We use a spreadsheet to match BOLD sample IDs with BOLD Process IDs, allowing the BOLD Version 3.0 trace upload form to be completed.
4. The trace files are then uploaded to the BOLD **EDTOL project** in a zipped folder that contains the reads and the completed Version 3.0 form.

> Note
>
> The BOLD Version 3.0 form 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 and direction.
>
> The form and the accepted locus and primer names are available from the BOLD website; the BOLD primer names relevant for this project are also tabulated in DNA Barcoding Standard Operating Protocol, Plants and Lichens at RBGE, Lab methods: PCR and Sequencing (protocols.io).

Passing / failing samples

> Note
>
> 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 considered to have passed barcoding.

If the taxon is missing from the reference database:

- 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? Then the sample can still be considered to have passed barcoding.
- If the sample is 100% or a very close match to something that is not a close relative of its expected taxon, it can fail barcoding despite the gap in the reference database.

Note

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 pipeline 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 can be accepted based on the morphological identification.

If there is a 100% match to a different species and a lower match to the expected taxon:

- Are the species sometimes synonomised? Taxonomies can vary (e.g. species has been split recently so that older GenBank sequences represent a different species concept, e.g. *Conocephalum conicum/Concephalum salebrosum*).
- 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?

- Specimens where the morphology-based identification conflicts with the DNA-barcoding-based identification for any marker should be flagged, lab protocols reviewed for potential contamination issues, and the submitting collector/Genome Acquisition Lab contacted for verification.

Note

- We enter all identification queries into a shared DToL DNA barcoding ID verifications Google sheet as soon as we find them, so that they can be acted on as quickly as possible.
- Until any issue is resolved, we edit the name of the specimen in BOLD to reflect any uncertainty, with a taxonomic note added to explain that there is currently a conflict between the morphological and DNA-based identification of the specimen.

Following morphological reverification of flagged samples:

- Specimens which still have different morphological and DNA barcoding identifications after the collector/Genome Acquisition Lab has reverified the voucher can 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 appears problematic, DNA extractions may be attempted from the herbarium vouchers instead.
- In cases where the morphological identification of the specimen is revised in line with the DNA barcode data, the specimen in BOLD is edited to reflect the correct taxon name.

Note

If the taxonomic identification of a collection changes, it 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.

Reference databases

- **UK seed plants:** these have been DNA barcoded for two plastid markers (rbcL, matK) 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.
- **UK lycophytes:** these have been DNA barcoded for three plastid markers (rbcL, matK, psbA-trnH) and one nuclear marker (ITS2). This data is in an open BOLD project.
- **UK ferns:** For several fern lineages good quality data is available on GenBank (i.e. projects involving well respected taxonomists; projects involving multiple

samples per species).

- **UK liverworts and hornworts:** These have been DNA barcoded for three plastid markers (rbcL, matK, psbA-trnH) and one nuclear marker (ITS2). This data is on private RBGE servers or private projects in BOLD (liverworts and hornworts). For several lineages, good quality data is available on GenBank (i.e. projects involving well respected taxonomists; projects involving multiple samples per species).
- **UK mosses:** A limited number of UK mosses (e.g. Bryaceae) have been DNA barcoded for three plastid markers (rbcL, matK, psbA-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, is being 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).
- **UK lichens:** UNITE is commonly used for fungal barcoding.

Note

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.

	A	B	C	D
		Date	***Changes***	***Contributors***
	1.0	August 2020	First draft	Laura L Forrest, Michelle L Hart
	1.1	January 2021	Revisions	Laura L Forrest

Previous Version History, RBGE DToL DNA Barcoding Standard Operating Procedure

Working SOP, checked by experts

Trace file names

1 Before submitting samples for sequencing, generate trace file names that are in a standardized format as this will facilitate downstream processing.

Note

Including the BOLD sample ID or process ID, the locus and primer name in the trace file is strongly advised. It is also helpful to include the provisional morphological identification, as this lets you easily see if BLAST results are as expected.

Sequence editing

2 Assemble the forward and reverse traces, trim off low quality areas and primer sequences, and resolve any disagreements between strands.

3 Export the sequence reads in FASTA concatenated format.

Verifying sequence read and taxonomic identity

4 BLASTn against NCBI to check the top matches (by identity), and look for significant differences in the alignments, e.g. indels. Recheck these against the trace files in your assembly.

5 Export the revised sequence reads in FASTA concatenated format.

Putting sequence data onto BOLD

6 FASTA files: rename the sequence identifiers in the definition line to fit BOLD format. Paste the sequences into the BOLD project.

7 Trace files: Export the trace files in SCF format, fill in the BOLD spreadsheet, and upload the files and spreadsheet to BOLD in a zipped folder.

DNA barcode check

8 Check ID using a mixture of BLASTn against NCBI, BOLD taxon ID trees and Full DB searches, and comparisons with in-house reference databases.

> **Note**
>
> For plants, basic searches can be run within BOLD using rbcL or matK, but it is not currently possible to run full DB searches for any other plant barcoding marker.

9 Inform the collector/Genome Acquisition Lab of any samples that are not a good match to the expected taxonomy so that they can check the voucher specimen and photographs.

10 Update taxon names on BOLD to reflect the most accurate current identification of the sample.

> **Note**
>
> This can mean removing the species epithet, genus name, etc. from samples where the molecular and morphological identifications are in conflict, up to the taxonomic level where both datasets agree, while the issue is resolved. The DToL plant data on BOLD is open access, so any errors should be resolved as soon as possible.