How To Get A DNA Barcode And What To Do With It?

Monday 18 August 2014 08:30-16:30

The Library (Room GD14), University of Nottingham Malaysia Campus

ORGANISERS
John James Wilson, Museum of Zoology, Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia
johnwilson@um.edu.my
Siti Azizah M Nor, School of Biological Sciences, Universiti Sains Malaysia, Penang, Malaysia
sazizah@usm.my
Noor Adelyna Mohammed Akib, Division of Research & Innovation, Universiti Sains Malaysia, Penang, Malaysia
adelyna_akib@yahoo.com

SPECIAL PRESENTATIONS
Megan Milton, Biodiversity Institute of Ontario, University of Guelph, Guelph, Ontario, Canada
mmilton@uoguelph.ca
Foo Yoke Wai, MyTACG Bioscience Enterprise, Selangor, Malaysia
ywfoo.mytacg@gmail.com

WORKSHOP ASSISTANTS
Kong Wah Sing (Gary)
singkongwah@um.edu.my
Lee Ping Shin (Ping)
pingshin@um.edu.my
Narong Jaturas (Narong)
golf_mdsc@hotmail.com
Brandon Mong (Brandon)
mong432@gmail.com
SUMMARY

- The first part of the workshop will cover the basic molecular methods required to generate a DNA barcode: sample preparation > DNA extraction > PCR > DNA sequencing. This will be covered through lecture, videos, and demonstration.

- The second part will cover the steps needed to be taken after sequences are returned from the sequencer: sequence editing > sequence alignment > upload of sequence to BOLD > submission of sequence to GenBank > identification of the sequence > biodiversity assessment using a group of sequences. This will be hands-on session.

SCHEDULE

08:30  Registration and welcome
09.00  Introduction to BOLD (Megan)
10.00  Q&A session on BOLD
10:15  Tea break
10.30  What is a DNA barcode, examples of DNA barcoding studies (Siti)
11:00  Wet laboratory techniques, DNA extraction, PCR
11:30  DNA sequencing (Foo)
11:45  Q&A session on wet lab techniques
12:00  Lunch break
13:00  Download and installation of programs needed for next sections
13:15  Sequence editing, CodonCode Aligner
14:00  Sequence alignment, BioEdit
14:15  Sequence analysis – checking for contaminants and identity, BLAST
14:30  Sequence analysis – grouping into MOTUs, ABGD
14:45  Tea break
15:00  Sequence analysis – tree building, distance analysis, MEGA (Adelyna)
16:00  Introduction to next-generation sequencing DNA barcoding
16:30  End
A. Specimen collection and storage
Care should be taken to prevent DNA degradation in your collected samples. Animal samples should be collected into an individual tube of 99.9% EtOH which is later deposited into a freezer (-20°C). If this is not possible, longer-term storage at room temperature should suffice. The EtOH should be frequently changed to ensure it remains at high concentration (as water diffuses out of the specimen, the ethanol decreases) [1]. Plant samples (leaf tissue) should be stored in archival quality coin envelopes placed in a sealed container or plastic zip bag with sufficient silica gel to dry the tissue. Do not put any material destined for DNA extraction in drying ovens. Silica gel is the best way to dry and preserve the tissue. Keep envelopes in the dark away from light [2].

B. DNA extraction
Most commercial kits use silica-based filters but a wide variety of options are available [3]. For fresh animal tissues an economical option is alkali-based extraction which is available in a kit [4] or can be home-made [5]. If using a commercial kit, simply follow the manufacturer’s instructions.

C. PCR
The reagents needed for PCR can be purchased as a kit or a pre-mix. Lots of choices are available and the crucial component is the taq polymerase [6]. Depending on your sample (old, degraded DNA, or fresh) and budget, choose the best taq you can. Another critical factor can be the choice of primers. Many primers for the standard DNA barcode regions can be found in the BOLD primer database [7]. Standard PCR conditions for animal barcoding are:

<table>
<thead>
<tr>
<th>PCR RECIPE</th>
<th>PCR THERMOCYCLING CONDITIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ddH2O 16.5µl</td>
<td>Initial denaturation 94°C (1min)</td>
</tr>
<tr>
<td>10x buffer 2.5µl</td>
<td>5 cycles: Denaturation 94°C (30s), Annealing 45°C (40s), Extension 72°C (1min)</td>
</tr>
<tr>
<td>MgCl2 1.25µl</td>
<td>35 cycles: Denaturation 94°C (30s), Annealing 45°C (40s), Extension 72°C (1min)</td>
</tr>
<tr>
<td>dNTPs 0.125µl</td>
<td>Final extension 72°C (10min)</td>
</tr>
<tr>
<td>Forward primer 0.25µl</td>
<td>Reverse primer 0.25µl</td>
</tr>
<tr>
<td>Taq polymerase 0.12µl</td>
<td></td>
</tr>
</tbody>
</table>

D. DNA sequencing
Most researchers opt to send their PCR products for sequencing by a commercial company (e.g. Macrogen, 1st Base). As a rough guide you should expect to pay between US$5-10 for each sequence.

UNDERSTANDING WHAT HAPPENS DURING PCR IS ESSENTIAL FOR SEQUENCE EDITING

### TEMPLATE DNA

<table>
<thead>
<tr>
<th>Forward Strand</th>
<th>Reverse Strand</th>
</tr>
</thead>
<tbody>
<tr>
<td>5' AAAATTTTCCCGGGG</td>
<td>3' LLLLLYYA9G99999999</td>
</tr>
<tr>
<td>3' LLLLLYYA9G99999999</td>
<td>5'</td>
</tr>
</tbody>
</table>

#### FIRST PCR CYCLE

- **Forward Primer**
- **Copy Forward (1)**
- **Copy Reverse (1)**

#### SECOND PCR CYCLE

- **Copy Forward (2)**
- **Copy Reverse (1)**
- **Copy Forward (1)**
- **Copy Reverse (2)**

#### THIRD PCR CYCLE

- **Copy Forward (3)**
- **Copy Reverse (2)**
- **Copy Forward (2)**
- **Copy Reverse (3)**

#### FOURTH PCR CYCLE

- **Copy Forward (4)**
- **Copy Reverse (3)**
- **Copy Forward (3)**
- **Copy Reverse (4)**
IMPORTANT VOCABULARY

Chromatogram = Electropherogram = Trace file

Reverse complement
Example: >AAATTTCCGGG< becomes <CCCGGAAATT<

Contig: Two or more sequences grouped into same folder (usually one Forward and one Reverse) to produce a consensus

Consensus sequence
>ATCGATCG> FORWARD
<ATCGATCG< REVERSE
>ATCGATCG> CONSENSUS

N = A or T or C or G
Messy sequence has a lot of Ns
1. Your sequences will probably come back from the sequencing company in a zip folder by email.
2. Unpack (extract) the zip file to your desktop. The files you are interested in have an extension .ab1, e.g. NUMBEROFSAMPLE_F.ab1 and NUMBEROFSAMPLE_R.ab1. If you sequence company is not providing you with files like this ask them why not. Delete the other files in the folder.
3. Open CodonCode Aligner and choose Create a new project and press OK.
4. Go to File>Import>Add Folder navigate to the desktop and select the folder of traces. Click Open>Import.
5. To see the files you just imported press ► besides the Unassembled Samples folder
6. The .ab1 files should be of the form NUMBEROFSAMPLE_F.ab1 where the second part “F” refers to the direction, i.e. Forward.
7. Sort the files by quality by double-clicking on Quality. Any sequences that are of very poor quality (look for a big difference between the sequence length and the quality score) can be deleted by highlighting the sequence and clicking Edit>Move to Trash.
8. Next we will group our sequences by direction for easy editing.
9. Select the Config menu and move the cursor over Advanced Assembly. From the options that appear select Assemble in Groups.
10. A new window will appear. Click the button Define name parts...
11. There are two name parts in our file names (see above). The first part of our file names refers to the number of the sample and for our purposes the option in the Meaning menu (first row) can be left as Clone. Since the sample number is followed by an underscore, choose _ (underscore) in the Delimiter menu next to Clone (if it isn’t already selected).
12. For the second row choose Direction in the Meaning menu. We can ignore the Delimiter menu for the Direction part because there is nothing following the direction in our file names.
13. Delete all additional name parts that may appear in the window (if any), and next click Preview... to check how CodonCode Aligner is interpreting the sample names.
14. Click Close to exit the preview. Click OK to return to the Assemble in Groups window.
15. We first want to assemble our samples according to direction. Choose Direction in the Name part: dropdown menu. Then click Assemble. You should now have two folders, one called F with the forward sequences and one called R with the reverse sequences. [Note: if you only sent your PCR products for sequencing in one direction (with one primer) then you will only have one folder.]
16. We will deal with the reverse sequences first. The first step is to reverse complement the sequences (see page 5). Highlight the R folder, select Edit>Reverse complement.
17. Next we need to cut the primer from the sequence (why? see page 4). Double click the R folder to open it. For the reverse sequences, you need to find the forward primer motif and delete it from the beginning of the consensus sequence at the bottom of the window. You will find the primer around 30 nucleotides from the end of the raw sequence. For example, you would need to delete the section of the sequence marked below in bold and everything to the left of it. Highlight it on the consensus sequence and press the Backspace key on the keyboard. ➤ ATTCACCACATCATAAGATATTGGAACATTATATTATTTTTGGTATTGATCAGG.....
18. Next go to the opposite end of the consensus, the far right. Delete the consensus sequence from the point where the sequence gets messy. This will be show in green highlight, and probably with the presence of gaps ("--"). For example, delete the section marked in bold and everything to the right of it. Highlight it on the consensus sequence and press the Delete key on the keyboard. Close the window.

...TCTTTTTTTGACCCTGCTGGTGGAGG-G-TTT-GGT-AA-T-TTT-G-C→

19. Double click the F folder to open it. Go to the far right of the consensus sequence and find the reverse complements reverse primer motif at the very end (why? see page 5). This should be around 680bp on the raw sequence. For example, you would delete the section marked in bold and everything to the right of it. Highlight it on the consensus sequence at the bottom of the window and press the Delete key.

...GGGGAGACCCTATTCTTTATCAACATTTATTT TGAT TTTTTGGACATC CAGAACTTTA→

20. Next go to the opposite end of the consensus, the far left, and delete the consensus sequence from the point where the sequences get messy. This will show in green highlight, and probably with the presence of gaps ("--"). For example, you would delete the sequence in bold and everything to the left of it. Highlight the region on the consensus sequence and press the Backspace key on the keyboard. Close the window.

←AT - GC - T - TTT - TTT - G - A - TGTIT - ATCAGGACTAATTGGAACTTC

21. Dissolve both the F and R folders by highlighting them and clicking the button marked with a red X.

22. Now we are going to combine the forward and reverse sequence from each specimen into a contig. Select all the sequences and open the Contig menu. Move the cursor over Advanced Assembly. From the options that appear select Assemble in Groups. This time choose Clone in the Name part: menu, then click Assemble. [Note: if you only sent your PCR products for sequencing in one direction (with one primer) then you will need to check each sequence individually rather than checking a consensus (contig).][Note: specimens which only sequenced successfully in one direction will have files which remain in the Unassembled Samples folder.]

23. Open each folder (contig) in turn by double-clicking. Correct ambiguous positions (shown in red and/or as N) and gaps ("--") in the consensus sequence by checking the original traces. This is done by double-clicking on the consensus sequence. Always check both trace files (forward and reverse) and compare them. Note, the corrected consensus sequence should have NO gaps.

24. Generally if traces conflict (i.e. different colored peaks appear in the same location on the forward and reverse chromatograms) you can decide which is more reliable based on sequence quality (e.g. less background noise, taller peaks).

25. Check the contigs first, then check the individual single sequences in the Unassembled Samples folder.

26. To export the consensus sequences, highlight all the folders, go File>Export>Consensus Sequences..., choose Current selection. Open the Options and select Include gaps in FASTA but deselect all other options. Press Export. Save the file to the desktop as sequences.fas.

27. To export single direction sequences, go File>Export>Samples..., choose Current selection. Press Export. Save the file to the desktop as sequences_single.fas.
STEP-BY-STEP SEQUENCE ALIGNMENT USING BIOEDIT

BioEdit is free for use by any and all interested parties, but is no longer being regularly maintained. The program can be downloaded from http://www.mbio.ncsu.edu/bioedit/bioedit.html

1. Open the file sequences.fas in BioEdit
2. You can then File>Import>Sequencing alignment file to add sequences_single.fas to the alignment.
3. Make sure Mode is set to Edit using the dropdown menu.
4. Another dropdown menu will become visible to the right of the Edit dropdown. Make sure this is set to Insert.
5. Sequences that have ended up in the FASTA file in the wrong orientation may be corrected by highlighting the sequence name by clicking the cursor on it, clicking the Sequences menu at the top of the screen. Moving the cursor down the dropdown to Nucleic Acid and clicking Reverse complement.
6. Sequences all need to be 658 bp and aligned to each other (before proceeding for further analysis or before uploading to BOLD and GenBank). This can be done by typing additional Ns at the beginning and end of the sequences the BioEdit Edit mode. Be sure to check across the whole of the alignment of the sequences that you have added the correct number of Ns.
7. In the figure below featuring a 50-bp barcode for simplification, LOP001-11 is of full length. LOP002-11 needs 6 Ns adding to the left side of the sequence to become aligned, while LOP003-11 needs 4 Ns adding to the right side of the sequence to be 50 bps long. LOP005-11_F needs to be reverse complemented to be in the same orientation as the other sequences.
8. Sequences which were not part of a consensus (i.e., when one direction failed but the single sequence is of sufficient length and quality for submission to BOLD) may appear in the FASTA still tagged with the direction. This needs to be deleted, e.g., the sequence named LOP005-11_F should be renamed as LOP005-11.
9. If you are having trouble with the alignment, a good quality (i.e., 658[0n]) sequence can be downloaded from BOLD and imported into your BioEdit file as a guide, e.g., MHAHC824-05 (Fig. 3). Be sure to delete this sequence before saving the file.
10. Save the file (File > Save).

**Unaligned FASTA**

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOP001-11</td>
<td>AACTTTATATTTATTTTGGGATTTTGACAGGAAATAGTAGGAACCTCTT</td>
</tr>
<tr>
<td>LOP002-11</td>
<td>ATATTTATATTTTTGAGCTGGAAATAGTAGGAACCTCTT</td>
</tr>
<tr>
<td>LOP003-11</td>
<td>AACTCTATATTTTATTTTGGGACGAGGAAATAGTAGGAACCTCTT</td>
</tr>
<tr>
<td>LOP004-11</td>
<td>TCTATATATTTTATTTTGGGACGAGGAAATAGTAGGAACCTCTT</td>
</tr>
<tr>
<td>LOP005-11_F</td>
<td>ATATTTATATTTTATTTTGGGACGAGGAAATAGTAGGAACCTCTT</td>
</tr>
</tbody>
</table>

**Aligned FASTA**

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOP001-11</td>
<td>AACTTTATATTTATTTTGGGATTTTGACAGGAAATAGTAGGAACCTCTT</td>
</tr>
<tr>
<td>LOP002-11</td>
<td>NNNNNNNATATTTATTTTATTTTGGGACGAGGAAATAGTAGGAACCTCTT</td>
</tr>
<tr>
<td>LOP003-11</td>
<td>AACTCTATATTTTATTTTGGGACGAGGAAATAGTAGGAACCTCTT</td>
</tr>
<tr>
<td>LOP004-11</td>
<td>NNNNTCTATATTTTATTTTGGGACGAGGAAATAGTAGGAACCTCTT</td>
</tr>
<tr>
<td>LOP005-11</td>
<td>AACTTTATATTTTATTTTGGGACGAGGAAATAGTAGGAACCTCTT</td>
</tr>
<tr>
<td>MHAHC824-05</td>
<td>AACTTTATATTTTATTTTGGGACGAGGAAATAGTAGGAACCTCTT</td>
</tr>
</tbody>
</table>
SEQUENCE ANALYSIS TOOLS

For sequence analysis we find it is best to use online tools to ensure you are using the latest versions of the programs and datasets.

You can open the FASTA file in MS word, then copy and paste the sequences into these websites.

For checking sequences for contamination and/or establishing a taxonomic identity, we would commonly “blast” our sequences against the BOLD and GenBank databases.


**Sequence Similarity Searching** = a method of searching sequence databases (libraries) by using alignment to a query sequence. By statistically assessing how well database and query sequences match, we can infer homology and transfer information (such as putative species membership) to the query sequence.

**Max score** = highest alignment score (bit-score) between the query sequence and the database sequence segment. *Higher the better. 1000 is very good.

**Total score** = sum of alignment scores of all segments from the same database sequence that match the query sequence (calculated over all segments). *Not so applicable for protein-coding genes.

**Query coverage** = percent of the query length that is included in the aligned segments. *Not meaningful for our purpose.

**E-value** = number of alignments expected by chance with a particular score or better. The expect value is the default sorting metric and normally gives the same sorting order as Max score. *The closer to 0 the better. Most important metric for barcoding.
To group sequences into MOTU (molecular operational taxonomic units), a popular tool is the ABDG (automatic barcode gap discover).


ABGD uses an automatic recursive procedure to converge on the best patterns for the dataset and arranges DNA barcodes into clusters (partitions) accordingly. The median number of ABDG partitions can be used as the basis for MOTU (species) as this has produced good correspondence with traditional species in empirical studies.

Another extremely popular program for sequence analysis is MEGA. MEGA is free to download from here: [http://www.megasoftware.net/](http://www.megasoftware.net/) and has an extensive user manual.