DNA Barcoding

Method	Page number
DNA Barcoding	
Introduction	2
The process	7
Part 1 - Sample preparation	8
PCR amplification information	9
PCR protocol	11
Gel electrophoresis	13
PCR product purification	14
PCR product sequencing	15
Part 2 - Identification	17
Sequence quality control	17
Clustering	19
Database searching - BOLD	24
Blast searching	26
Problems using databases	29

Introduction

DNA barcoding in an academic setting has two aims: firstly, to assign unknown individuals to species, and secondly to enhance the discovery of new species (Hebert et al., 2003a; Stoeckle, 2003; Blaxter, 2003; Blaxter, 2004). The term "DNA barcoding" originates from the idea of Universal Product Codes for manufactured goods being applied to DNA sequences for different species. Creation of the 'barcode' involves the PCR amplification and sequencing of a conserved gene sequence (typically around 600bp of the mitochondrial Cytochrome oxidase I gene (COI)). In simple terms, these sequences or barcodes are then aligned and a tree is produced; if a suitable gene has been used, the clusters should identify meaningful groups of individuals as distinct taxa.

Mitochondrial genes are particularly attractive for this application because of their lack of introns. Hebert et al. (2003b) have shown that in arthropods, most species have more than 50 substitutions in each 500bp of their COI gene – more than enough for species identification. However, a common misconception is that each species has it's own unique sequence for the entire 600bp region of COI. Most species have a degree of intraspecific variation between individuals or populations, therefore typically between five and ten individuals from each species are sequenced, in order to define the variation within the taxa studied.

The concept of sequencing a region of DNA and using it to identify species is not new; however, in recent years the scope and utilisation of DNA barcoding has expanded rapidly. The creation of the International Barcode of Life (iBOL: http://ibol.org/phase1/) has led to the standardisation of protocols for DNA barcoding and many projects are ongoing. Recent projects have targeted birds (Hebert et al, 2004; Kerr et al, 2007), fish (Ward et al, 2005), bats (Clare et al 2007), fungi (Seifert et al 2007), and invertebrates (Ball et al., 2006).

DNA barcoding and plant pathogens/pests

Identifying invertebrate pests and fungal/bacterial pathogens to species-level using morphology alone can be time consuming and requires specialist skills and knowledge. Many invertebrates can be morphologically cryptic in their juvenile stages and may require culturing to gain a positive identification, a

process that can take many weeks. Furthermore, an identification cannot be made if these samples are dead on arrival at the laboratory or die during culturing. In addition, there is a well-documented decline (e.g. Coomans, 2002; Hopkins and Freckleton 2002) in the availability of experts in the field of morphological taxonomy, due at least in part to changing trends in teaching at Universities. Thus, maintaining a critical mass of expertise in these fields in order to provide a service is often difficult.

DNA barcoding could become a valuable tool in this arena. In addition to assigning unknown individuals to species and enhancing the discovery of new species, the technique can be used to identify unknown specimens. Given a validated dataset of sequences obtained from morphologically identified species, an unknown individual or juvenile may be identified by placing its sequence in the tree and seeing which species it clusters with. In contrast to the traditional test (e.g. ELISA test) that produces a positive or negative result for the presence of one organism, DNA barcoding can be thought of as a molecular identification tool. The technique also has a number of technological advantages since it is relatively simple, requiring only PCR (and access to the relevant primer sequences) and sequencing (a readily available service which is both rapid and inexpensive). The future of DNA barcoding as an application in the plant health arena will ultimately be determined by the availability of validated databases of sequences.

Many projects are currently underway to produce validated datasets of DNA barcodes. Their applications include forensics (Nelson et al., 2007) and elucidating cryptic species (Hulcr et al., 2007), in addition to identifying economically important species for biosecurity (Armstrong & Ball, 2005, 2006; Brunner et al., 2002). The International Barcode of Life is also coordinating many barcoding projects for a range of taxa and Genbank and other repositories (e.g. BOLD: http://www.boldsystems.org) contain many DNA barcode sequences useful for identifying plant pests.

A further application of DNA barcoding data that is currently emerging is the use of next generation sequencing and DNA metabarcoding in which mixture or populations of organisms can be identified (e.g. Toju and Baba 2018).

One of the most important aspects of a DNA barcoding effort is the initial identification of material to be sequenced. The downstream identification of any

unknown specimen is only as good as the data used for comparison. Web-based software produced by Ratnasingham and Hebert (2007) provides an identification engine that contains the DNA barcodes held to date, all of which were positively identified by morphology prior to sequencing. As the number of species, and the number of barcodes per species, in the public domain becomes larger, the power of the technique increases. However, beyond the academic aspects, it is important to consider the likely applications and potential practitioners of the method, and to engage with this community at an early stage. This community is the source of identified and validated material, without which the barcoding effort will be worthless. It is important that all the required information is captured for the material to be barcoded and that this information is both relevant to the person using the approach to achieve an identification, and links to voucher specimens. The concept of DNA vouchers and also digital vouchers from which the DNA was extracted also need to be addressed.

Commonly used DNA barcoding primers

Targets / gene	**Primer name**	**Primer sequence (5' – 3')**	**Size**	**Tm**	**Reference**
Nematodes / 18S	SSU18A	AAAGATTAAGCCATGCATG	~1000bp	52°C	Floyd et al (2002)
	SSU26R	CATTCTTGGCAAATGCTTTCG			
Invertebrates/ COI	LCO1490	GGTCAACAAATCATAAAGATATTGG	~700bp	52°C	Folmer et al (1994)
	HCO2198	TAAACTTCAGGGTGACCAAAAAATCA			
Fungi / ITS	ITS1	TCCGTAGGTGAACCTGCGG	~650bp	55°C	White et al (1990)
	ITS4	TCCTCCGCTTATTGATATGC			
Whitefly / COI	2195Bt	TGRTTTTTTGGTCATCCRGAAGT	~867bp	52°C	Mugerwa et al 2018
	C012/Bt-sh2	TTTACTGCACTTTCTGCC			

References

Armstrong, K. F. & Ball, S. L. (2005). DNA barcodes for biosecurity: invasive species identification. *Philosophical Transactions of the Royal Society B, 360,* 1813-1823

Ball, S. L. & Armstrong, K. F. (2006). DNA barcodes for insect pest identification: a test case with tussock moths (Lepidoptera: Lymantriidae). *Canadian Journal of Forest Research, 36,* 337-350

Blaxter, M. (2003). Molecular systematics: Counting angels with DNA. *Nature, 421,* 122–12

Blaxter, M. (2004). The promise of a DNA taxonomy. *Philosophical Transactions of the Royal Society B, 359,* 669–679

Brunner, P. C., Fleming, C. & Frey, J. E. (2002). A molecular identification key for economically important thrips species (Thysanoptera: Thripidae) using direct sequencing and a PCR-RFLP-based approach. *Agricultural and Forest Entomology, 4,* 127-136

Clare, E. L., Lim, B. K., Engstron, M. D., Eger, J. L. & Hebert, P. D. N. (2007). DNA barcoding of neotropical bats: species identification and discovery within Guyana. *Molecular Ecology Notes, 7,* 184-190

Coomans, A. (2002). Present status and future of nematode systematics. *Nematology, 4,* 573-582

Floyd, R., Abebe, E., Papert, A., and Blaxter, M (2002) Molecular barcodes for soil nematode identification. *Molecular Ecology* 11:839-850.

Folmer, O., Black, M., Hoeh, W., Lutz, R., Vrijenhoek, R. (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, 3: 294-299.

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

Hebert, P. D. N., Ratnasingham, S. & deWaard, J. R. (2003b). Barcoding animal life: Cytochrome C oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society London B, 270,* S96-S99

Hebert, P. D. N., Stoeckle, M. Y., Zemlak, T. S. & Francis, C. M. (2004). Identification of birds through DNA barcodes. *PloS Biology, 2,* e312

Hopkins, G. W. & Freckleton, R. P. (2002). Declines in the numbers of amateur and professional taxonomists: implications for conservation. *Animal Conservation, 5,* 245-249

Hulcr, J., Miller, S. E., Setliff, G. P., Darrow, K., Mueller, N. D., Hebert, P. D. N. & Weiblen, G. D. (2007). DNA barcoding confirms polyphagy in a generalist moth, *Homona mermerodes* (Lepidoptera: Tortricidae). *Molecular Ecology Notes, 7,* 549-557

Kerr, K. C. R., Stoeckle, M. Y., Dove, C. J., Weight, L. A., Francis, C. M. & Hebert, P. D. N. (2007). Comprehensive DNA Barcode coverage of North American Birds. *Molecular Ecology Notes, 7,* 535-543

Mugerwa, H., Seal, S.E., Wang, H., Patel, M.V., Kabaalu, R., Omongo, C., Alicai, T., Tairo, F., Ndunguru, J., Sseruwagi, P., & Colvin, J. (2018). African ancestry of New World, Bemisia tabaci-whitefly species. *Scientific Reports*. http://doi.org/10.1038/s41598-018-20956-3

Nelson, L. A., Wallman, J. F. & Dowton, M. (2007). Using COI barcodes to identify forensically and medically important blowflies. *Medical and Veterinary Entomology, 21,* 44-52

Ratnasingham, S. & Hebert, P. D. N. (2007). BOLD: the Barcode of Life Data System (www.barcodinglife.org).

Seifert, K. A., Samson, R. A., deWaard, J. R., Houbraken, J., Levesque, C. A., Moncalvo, J. M., Louis-Seize, G. & Hebert, P. D. N. (2007). Prospects for fungus identification using CO1 DNA barcodes, with Penicillium as a test case. *Proceedings of the National Academy of Sciences USA, 104,* 3901-3906

Stoeckle, M. (2003). Taxonomy, DNA and the barcode of life. *BioScience, 53,* 2-3

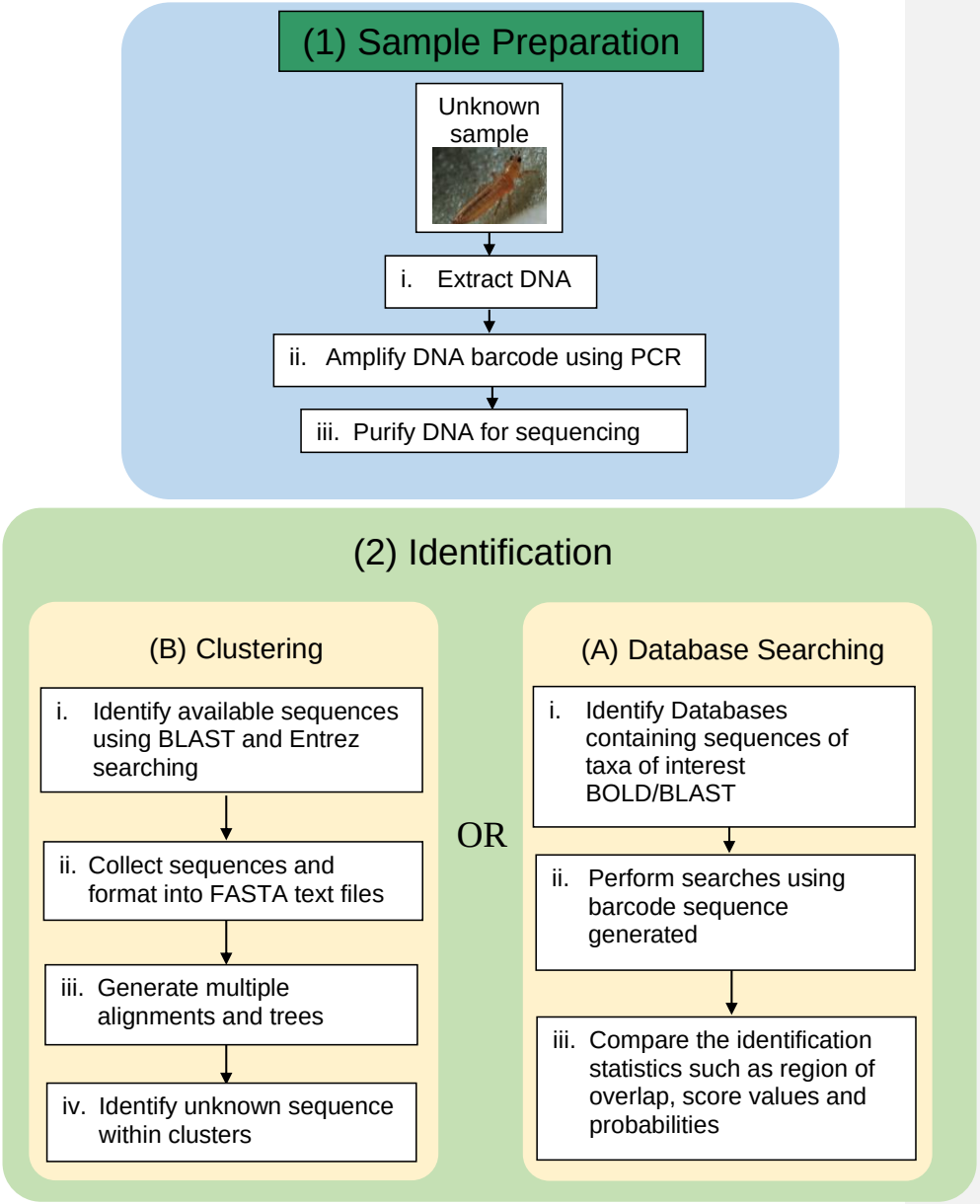
Toju, H., & Baba, Y. G. (2018). DNA metabarcoding of spiders, insects, and springtails for exploring potential linkage between above- and below-ground food webs. *Zoological letters*, *4*, 4. doi:10.1186/s40851-018-0088-9

Ward, R. D., Zemlak, T. S., Innes, B. H., Last, P. R. & Hebert, P. D. N. (2005). DNA barcoding Australia's fish species. *Philosophical Transactions of The Royal Society of London B, 360,* 1847-1857

White, T.J., Bruns, T.D., Lee, S.B., and Taylor, J.W. (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics, p. 315-322. In M.A. Innis, D. H. Gelfand, J.J. Sninsky, and T.J. White (ed.), PCR protocols. A guide to methods and applications. Academic Press, San Diego, Calif.

DNA barcoding: the process

DNA barcoding is a simple method based around two main processes: (1) Generation of the sequence from the unknown (2) comparing the sequence of the unknown to a database of sequences with known identity, which can be done in two ways (a) clustering or (b) database searching.



DNA Barcoding Practical Part 1

(1) Sample preparation

(i) Insect nucleic acid extraction protocol

For invertebrate identification carry out an extraction from known species as controls. The method (Boonham N *et al., (*2002). *Journal of Virological Methods* **101,** 37-48) is suitable for small invertebrates, thrips, whitefly, aphids and nematodes. All centrifugation steps are carried out at 14,000 rpm, in a microfuge, unless stated otherwise

Protocol

1. Individual insects are ground in a 1.5ml microcentrifuge tube (using pellet grinders and matching tubes) with 50µl nuclease-free water and stored on ice.
2. Chelex resin (Chelex100, Biorad) (50µl of a 50% w:v slurry) is added to each sample.
3. The samples are heating at 94°C for 5 minutes on a thermocycler.
4. The tubes are centrifuged for 5 minutes, the pellet discarded and the supernatant stored at -20°C.

Commented [RC1]: Changed to correct volume for grinder

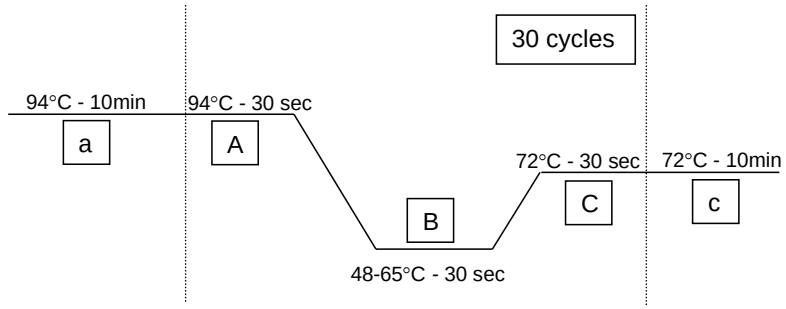
(ii) PCR amplification information

Background

There are a huge range of PCR and RT-PCR reagents available for performing PCR amplification. These can be in the form of individual reagents mixed together by the user, or master-mixes which contain everything except for the primers, some even contain loading buffer (e.g. Reddymix, AB-Gene). In many cases master-mixes are less expensive and more convenient for diagnostic use.

PCR-Cycling

Typically, PCR cycling is performed in 3 steps, (A) denaturation, (B) annealing and (C) extension. This is frequently preceded by a melting step (a) and followed by a longer extension step (c). Typically, the temperature of the annealing step is the most critical part and related to the Tm of the primers. The length of time for each step is related to the length of the product and finally the number of cycles is related to the target copy number and the efficiency of the PCR. A typical PCR cycle for diagnostic use (products in the 200-500bp range) is as follows:



During optimisation of a new PCR it is typical to start with an annealing temperature several degrees lower than the Tm of the primers, and to improve specificity and reduce miss-priming gradually increase the annealing temperature until single sharp DNA bands of the correct size are observed.

Work flows, controls and contamination

Ideally different areas of the lab (or different labs) will be dedicated to the different parts of the PCR process, to ensure freedom from contamination. The parts to separate are (i) extraction, (ii) PCR set up (iii) DNA spiking and (iv) Post PCR. Steps (ii) and (iii) can be combined for conventional PCR but for real-time PCR it is essential that all four processes should be separated ideally with dedicated equipment such as pipettes.

Controls for contamination should be included in each stage of the process, thus for each set of extractions a known healthy control should be included (ideally of the same species or a closely related species) this will be tested alongside the diagnostic samples and will identify any contamination during the extraction process.

Water controls of several kinds should be included in the process as follows:

1. Tubes capped following the addition of master mix: this indicates how clean the reagents being used are.
2. Tubes left open during DNA spiking but closed afterwards: this highlights and cross contamination during set up (especially when using plates).
3. Finally tubes where water is added at the end of the process to indicate cross contamination from sample to sample or associated with the pipette during set up.

PCR Protocol

DreamTaq Green PCR Master Mix (Thermo Fisher Scientific) is a ready-to-use master-mix containing Taq DNA Polymerase, buffer, $MgCl_2$, and dNTPs. It also contains a loading buffer and dye to enable direct loading of PCR products onto the gel prior to electrophoresis.

PCR reaction should be set up on ice, with the stock reagents sitting on ice. Typical reaction conditions are 25µl; although this can be scaled up or down depending on need.

The mastermix has been prepared in advance according to following table:

Reagent	**Starting [conc]**	**Final [conc]**	**Volume /µl (1 reaction)**	**Volume /µl (x reactions)**
Primer 1: PF - 2195Bt	10µM	0.5 pmol/µl	2	
Primer 2: PR C012/Bt-sh2	10µM	0.5 pmol/µl	2	
Mastermix	2x	1x	25	
SDW	-	-	19	
Reaction volume	**-**	**-**	**48**	

To assemble the reaction mix the mastermix and primers were thawed on ice. Each tube was mixed using a vortex mixer briefly and centrifuged briefly (or flicked) to collect the contents into the base of the tubes without air bubbles. To calculate the total mix required, the volume of each component was multiplied by the numbers of samples to be tested + 1 (e.g. 4 samples = x 5).

To set up the PCR reaction:

1. The mastermix is stored on ice. Vortex briefly and centrifuge briefly (or flick) to collect the contents into the base of the tube without air bubbles.
2. Pipette 48µl of master into PCR tubes (either 0.5µl or 0.2µl tubes, strips or plates, depending on the thermal-cycler being used) labelled with the sample numbers.

Commented [RC2]: Rearranged to show set up in advance

3. Using a different pipette, pipette 2µl of DNA template to the appropriate tube. Changing tips after each pipetting step.
4. Transfer the tubes to the thermal-cycler and subject to the following cycling conditions:

94°C	2 min	
94°C	30 s	40 cycles
52°C	30 s	
72°C	1 min	
72°C	10 min	

Gel electrophoresis

To enable the assessment of the PCR amplification (size and amount of product produced) gel electrophoresis is carried out followed by visualisation of the product using gel staining and UV illumination.

A 1 % agarose gel solution has been pre-prepared. 0.25 g agarose and 25 ml 1 X TBE buffer was heated in a microwave to boiling at 100°C. The solution was allowed to cool for approximately 5 minutes before adding 2.5 µl of the stain GelGreen.

1. Prepare a 1% agarose gel as follows:
   A. Wearing nitrile gloves[1] and using heat protection (gloves or paper towel) collect a prepared vial of 1 % agarose (containing GelGreen[2]) from the 65 °C water bath. Swirl the agarose mixture.
   B. Pour into the gel case with comb in place and leave to set for 20-30 minutes.
   C. Once set, carefully remove the rubber dams from both ends of the gel.
2. Place set gel in an electrophoresis tank so that it is submerged under 1x TBE buffer and carefully remove the comb. Pipette 5µl Mass Marker ladder into the first lane well.
3. Pipette between 5-10µl of each PCR sample into subsequent wells. Run at 90V for 45 minutes.
4. Remove gel from the tank and wearing nitrile gloves, carefully place gel on a UV transilluminator and view the PCR products.

Commented [RC3]: Rearranged to show how set up in advance

Commented [RC4]: Reduced voltage from 130 to 90 as our tanks are small – could go higher voltage and shorter run time if need be

Commented [RC5]: Cut staining steps that aren't applicable

[1] The DNA can be stained using ethidium bromide or products such as Gel Red. Note all DNA intercalating dyes should be treated with caution and gloves should always be worn when handling anything that has come into contact with the solutions.

[2] The stain can be added to the gel or the gel can be stained after running by soaking the gel in a solution containing the stain. If gel containing stain is used, the running buffer in the gel tank will contain the stain after electrophoresis is complete.

PCR Product purification

Following the observation of a clear distinct amplified product of the correct size following electrophoresis the remaining PCR product should be purified using a commercially available kit (e.g. QIAquick PCR Purification Kit (Qiagen) or GeneJET (ThermoFisher)).

GeneJET PCR Purification Kit

Before starting

1. Ethanol (45ml of 96–100% ethanol) was added to the wash buffer (9ml) concentrate before use.
2. All centrifugation steps are carried out at 17,900 x g (13,000 rpm)
3. Examine the Binding Buffer for precipitates before use. Re-dissolve any precipitate by warming the solution to 37°C and cooling to room temperature. Wear gloves when handling binding buffer.

Commented [RC6]: Changed from students adding to added for them

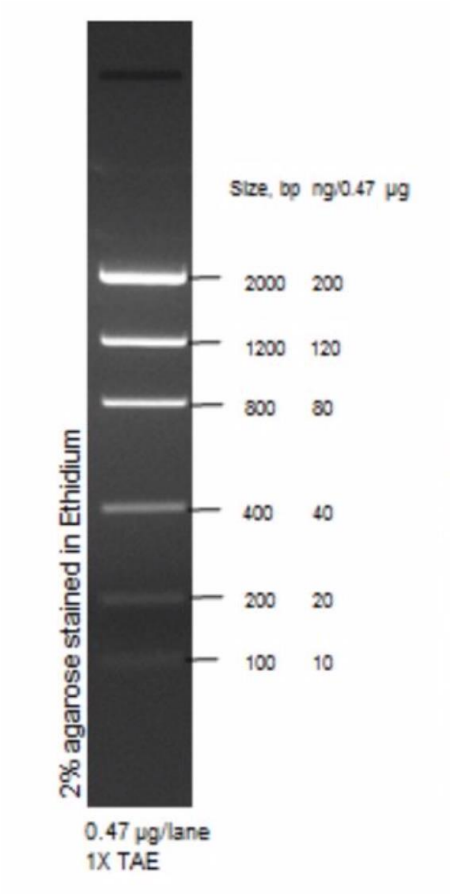
Protocol

1. Add 1 volume of binding buffer to 1 volume of the PCR sample and mix. (e.g. add 40µl of binding buffer to 40µl PCR sample).
2. Place a GeneJET spin column in a collection tube and transfer a maximum of 800µl of sample from step 1 into the column. Centrifuge for 30–60 s and discard flow-through.
3. Place the GeneJET column back into the same tube.
4. To wash, add 700µl wash buffer to the GeneJET column and centrifuge for 30–60 s.
5. Discard flow-through and place the GeneJET column back in the same tube. Centrifuge the column for an additional 1 min to remove residual wash buffer.
6. Place the GeneJET column in a clean 1.5ml microcentrifuge tube.
7. To elute DNA, add 50µl Elution buffer (or water pH 7.0–8.5) to the centre of the GeneJET membrane and centrifuge the column for 1 min. For increased DNA concentration, add 20-30µl of elution buffer to the centre of the GeneJET membrane, let the column stand for 1 min, and then centrifuge.
8. Discard the column and label the tube – store at -20°C until use.

PCR Product sequencing

The resulting DNA should then be quantified, either by using a spectrophotometer (e.g. nano drop) or by running a small amount alongside a DNA mass ladder and making a visual comparison (see example below).

Mass ladder contains individual chromatography-purified DNA fragments which are designed for sizing and estimating the concentration of double-stranded DNA in the range of 100 bp to 2,000 bp.



DNA Sequencing

Protocol

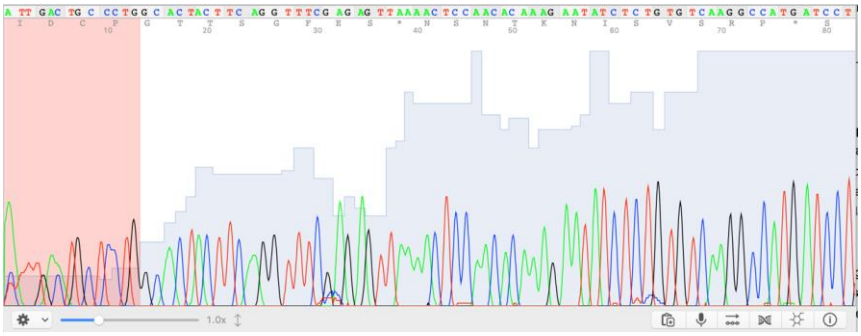
The PCR product should be adjusted to a final volume of 15μl and sent to a sequencing service along with one of the PCR primers adjusted to a final volume of 1. The amounts of DNA that are required are as follows:

PCR product size	Amount of DNA	Primer
100-200 bp	1-3 ng per reaction	3.2 picomoles per reaction
200-500 bp	3-10 ng per reaction	
500-1000 bp	5-20 ng per reaction	
1000-2000 bp	10-40 ng per reaction	
>2000 bp	40-100 ng per reaction	

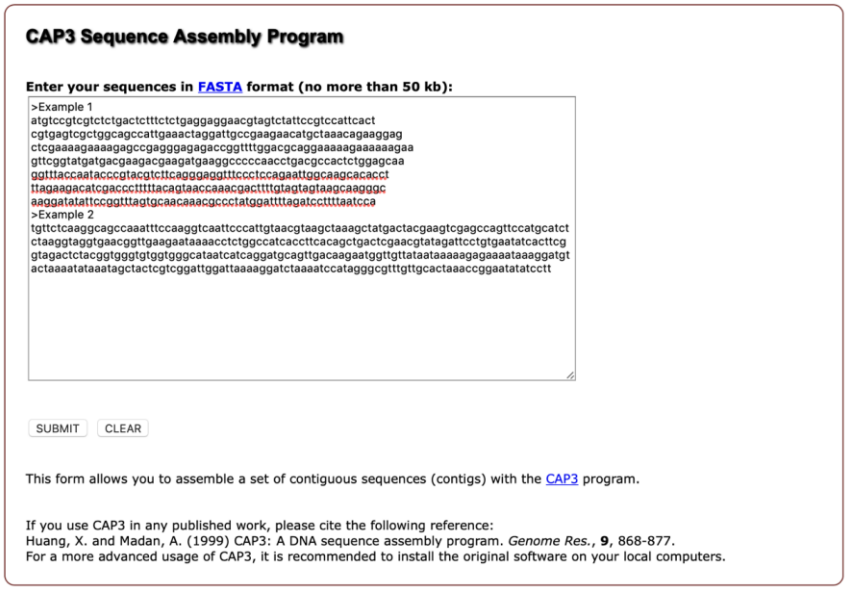
DNA Barcoding Practical Part 2 - Identification

Sequence quality control

Prior to analysis the quality of the sequence should be assessed, removing any low-quality sequence. Free sequence editing software is available, Chromas (https://technelysium.com.au/wp/chromas/) available for Windows and 4Peaks (https://nucleobytes.com/4peaks/index.html) for MacOS. The example below shows a sequence chromatogram overlaid with quality scores. The area of low-quality sequence is highlighted in red should be deleted.



After editing, if the product has been sequenced from both ends, a single contiguous sequence (contig) should be generated from the two separate sequence files. This can be done on-line using the CAP3 sequence assembly tool (http://doua.prabi.fr/software/cap3). Simply paste the two sequences into the on-line tool in FASTA format and press submit, select the contig from the results page.



CAP3 Sequence Assembly Program

Enter your sequences in FASTA format (no more than 50 kb):

```
>Example 1
atgtccgtcgtctctgactctttctctgaggaggaacgtagtctattccgtccattcact
cgtgagtcgctggcagccattgaaactaggattgccgaagaacatgctaaacagaaggag
ctcgaaaagaaaagagccgagggagagaccggttttggacgcaggaaaaagaaaaaagaa
gttcggtatgatgacgaagacgaagatgaaggcccccaacctgacgccactctggagcaa
ggtttaccaatacccgtacgtcttcagggaggtttccctccagaattggcaagcacacct
ttagaagacatcgacccttttttacagtaaccaaacgacttttgtagtagtaagcaagggc
aaggatatattccggtttagtgcaacaaacgccctatggattttagatccttttaatcca
>Example 2
tgttctcaaggcagccaaatttccaaggtcaattcccattgtaacgtaagctaaagctatgactacgaagtcgagccagttccatgcatct
ctaaggtaggtgaacggttgaagaataaaacctctggccatcaccttcacagctgactcgaacgtatagattcctgtgaatatcacttcg
gtagactctacggtgggtgtggtgggcataatcatcaggatgcagttgacaagaatggttgttataataaaaagagaaaataaggatgt
actaaaatataaatagctactcgtcggattggattaaaaggatctaaaatccatagggcgtttgttgcactaaaccggaatatatcctt
```

SUBMIT CLEAR

This form allows you to assemble a set of contiguous sequences (contigs) with the CAP3 program.

If you use CAP3 in any published work, please cite the following reference:
Huang, X. and Madan, A. (1999) CAP3: A DNA sequence assembly program. *Genome Res.*, **9**, 868-877.
For a more advanced usage of CAP3, it is recommended to install the original software on your local computers.

(2) Identification

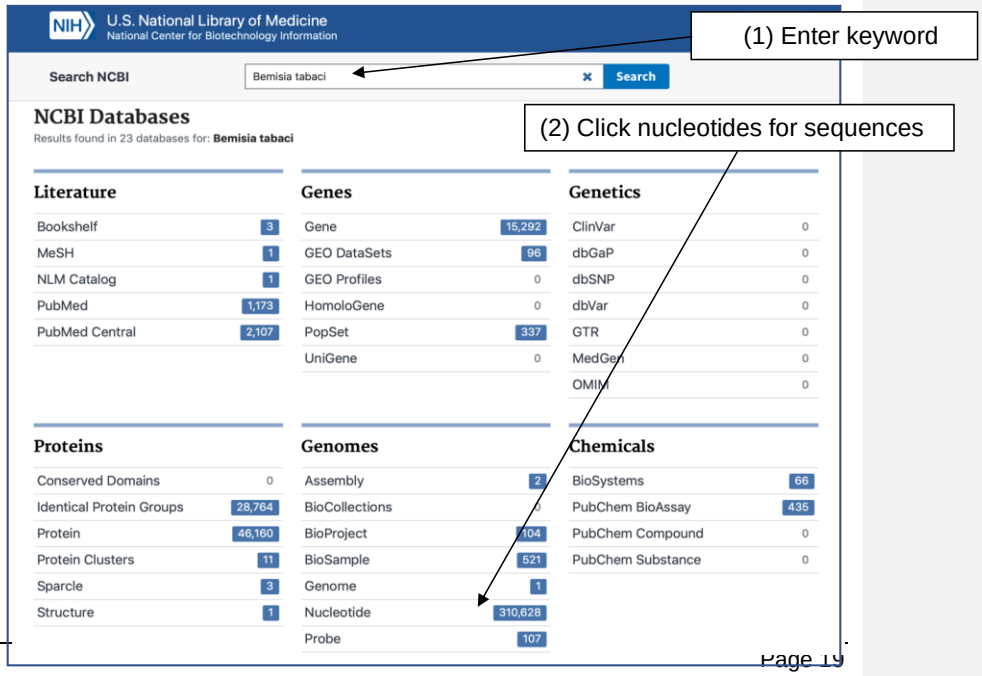
(A) Clustering

(i) Acquiring sequences for clustering

The National Centres maintain a single database called the non-redundant database. A submission to any of the centres results in the permeation of the data into all the databases. In Europe you submit to EMBL; in Japan to the DNA databank of Japan; and in the US to the NCBI. The data is then shared among all these systems. There are a large, and ever growing, number of databases that you can search against. The databases can be searched in a number of ways, the most useful methods are searching for an individual sequences of interest i.e. using a keyword, followed by a range of similarity based searches, i.e. searching a database for sequences similar to a query sequence.

ENTREZ search

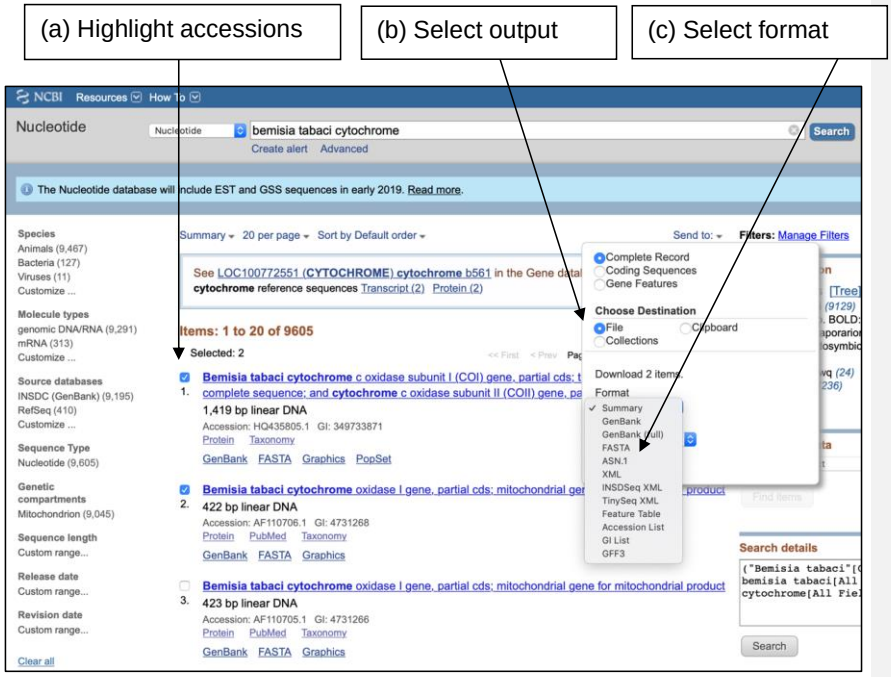
A keyword search that allows you to search a range of databases including the nucleotide databases. This can be accessed in a number of ways, a simple ENTREZ search (https://www.ncbi.nlm.nih.gov/search/) searches all possible databases, reporting back all hits. For example a search for *Bemisa tabaci.*



Hyperlinks lead you through into the sequence accessions, papers, taxonomy etc. where the information can be either retrieved by cut and pasting or saving to file in a range of formats.

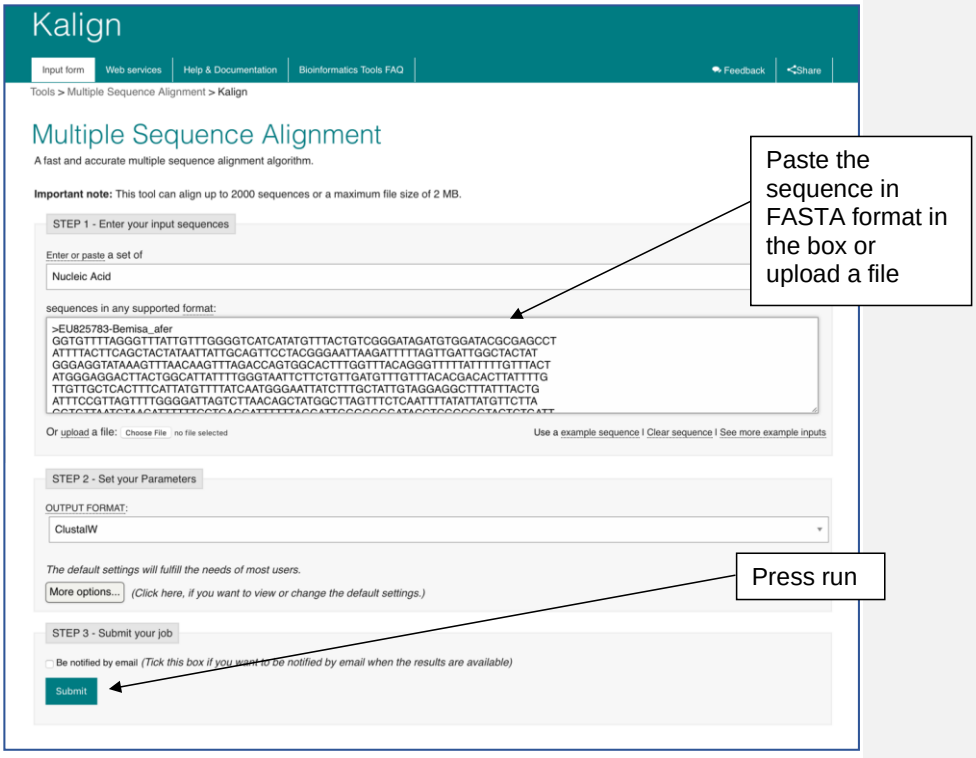
(ii) Retrieving sequence information

Information can be removed individually following the hyperlinks, or in groups by (a) highlighting the individual items, then (b) selecting the output (e.g. clipboard or to file) followed by (c) the format in which you want the information (e.g. FASTA) as follows. For sequence information the most useful formats are the summary, the Genbank file (the sequence accession) and FASTA format (for further analysis software).



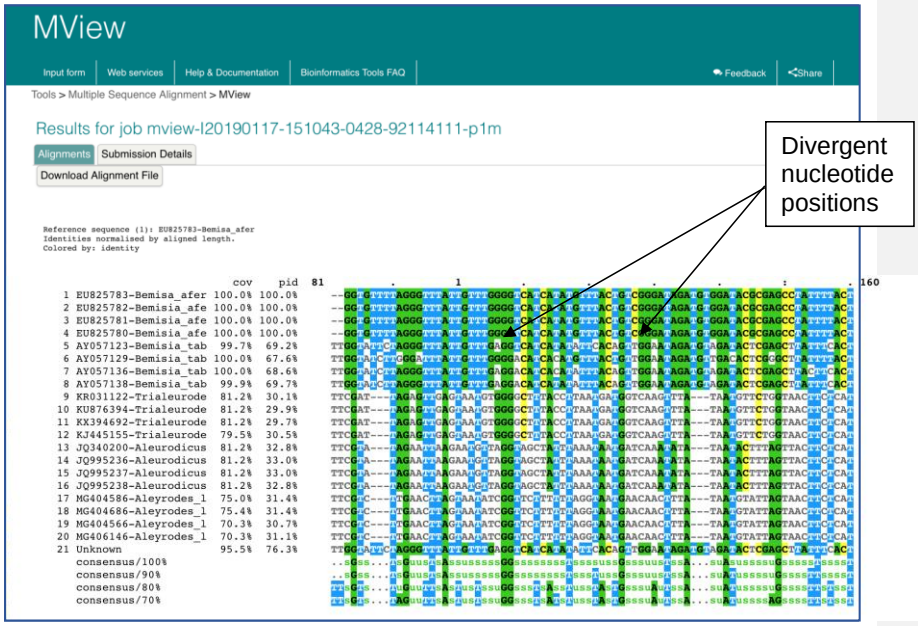
(iii) Making a multiple sequence alignment (MSA)

Multiple sequence alignments allow you to identify nucleotides with identity to other sequences. It is also the first part of the process for generating phylogenies for deriving evolutionary relationships and identifying similarity between sequence. It is also a useful tool to use prior to designing diagnostic primers (PCR, LAMP or RPA), since it allows you to identify regions of sequence that have conservation (between members of the same species) and divergence (between members of closely related species). There are a range of algorithms that can be used, one of the most common is CLUSTAL and can be performed at several web sites (e.g. https://www.ebi.ac.uk/Tools/msa/) an example is shown below.



(iv) Sequence clustering

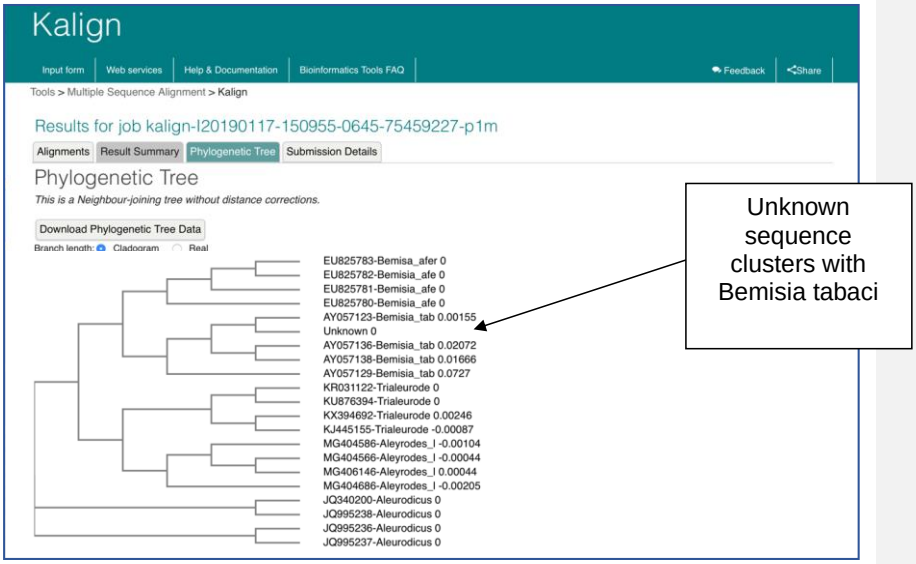
The alignment can be annotated to identify conserved and divergent nucleotides by sending the alignment to MView.



Drawing phylogenetic trees

The alignments can also be used for drawing clustering trees that highlight similarity between sequences. These can be used to identify unknown sequences within a group of sequences with a known identity.

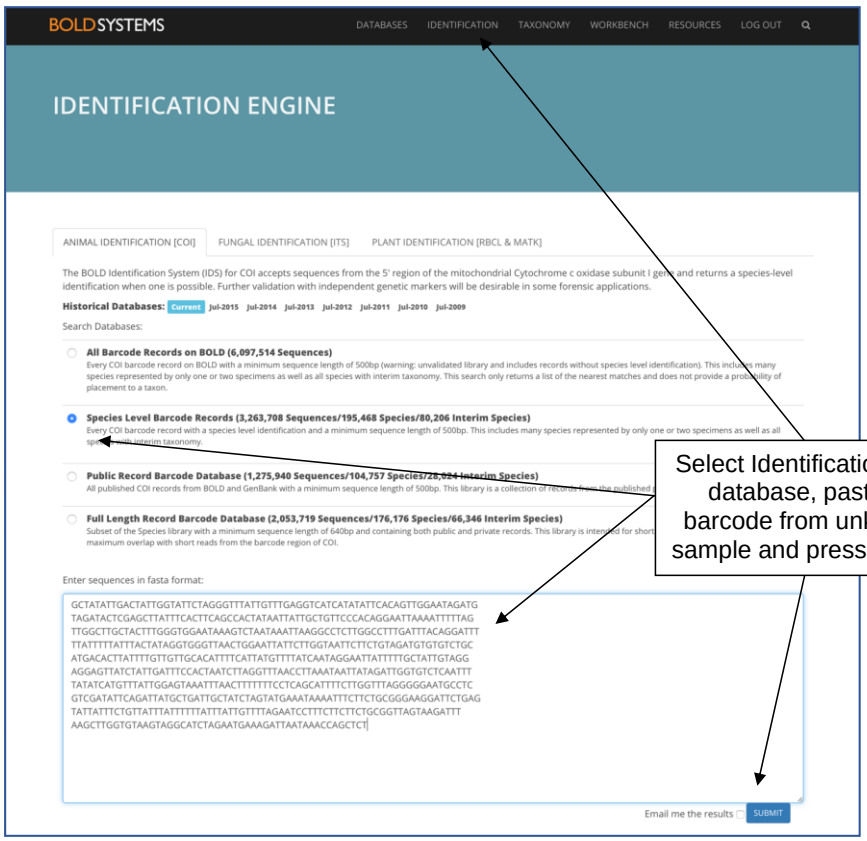
Clustering where multiple sequences from each taxa of interest are included in the analysis is the safest approach to achieving identification. Ideally, sequences from different populations, geographical regions and different analysis labs should be included for each taxa of interest. The analysis enables assessment of inter and intra species sequence variation (within each cluster) and a sequence fitting within a cluster can then be indicative of a reliable identification. Clearly, careful selection of sequences for the most relevant taxa with which to build the trees is important to achieve accurate identification of unknown sequences. BLAST can be a useful tool for identifying potential sequences and taxa to include into the multiple-sequence alignment prior to clustering with the unknown sequences.



(B) Database searching

(i) BOLD

Some specific software is available for barcode identification which is helpful since it automatically defines the database for you. The best software currently is the BOLD IDS (The Barcode of Life Data Systems Identification System) present at (http://www.boldsystems.org/views/idrequest.php) this software however is not specific to plant pathogens and pests.



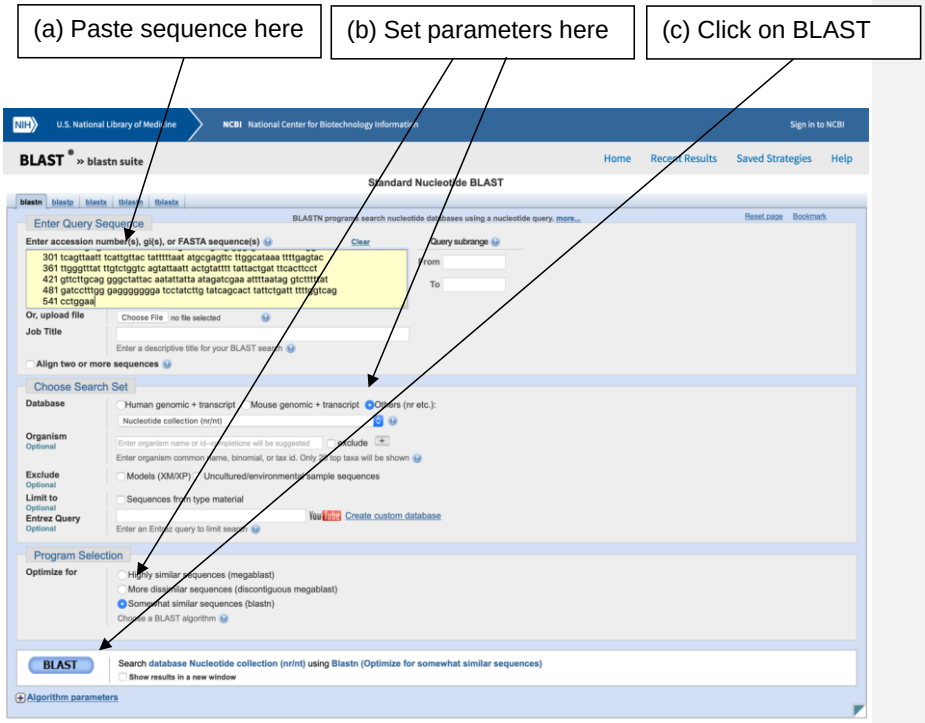
The database selected will vary depending on the presence of validated sequence in the database. It is best to begin using the reference database then try the species level database and finally the 'all barcode records'. It should be noted however that if your species does not appear in the database identification will not be made.

Identify the most closely related species in the database

BLAST searching

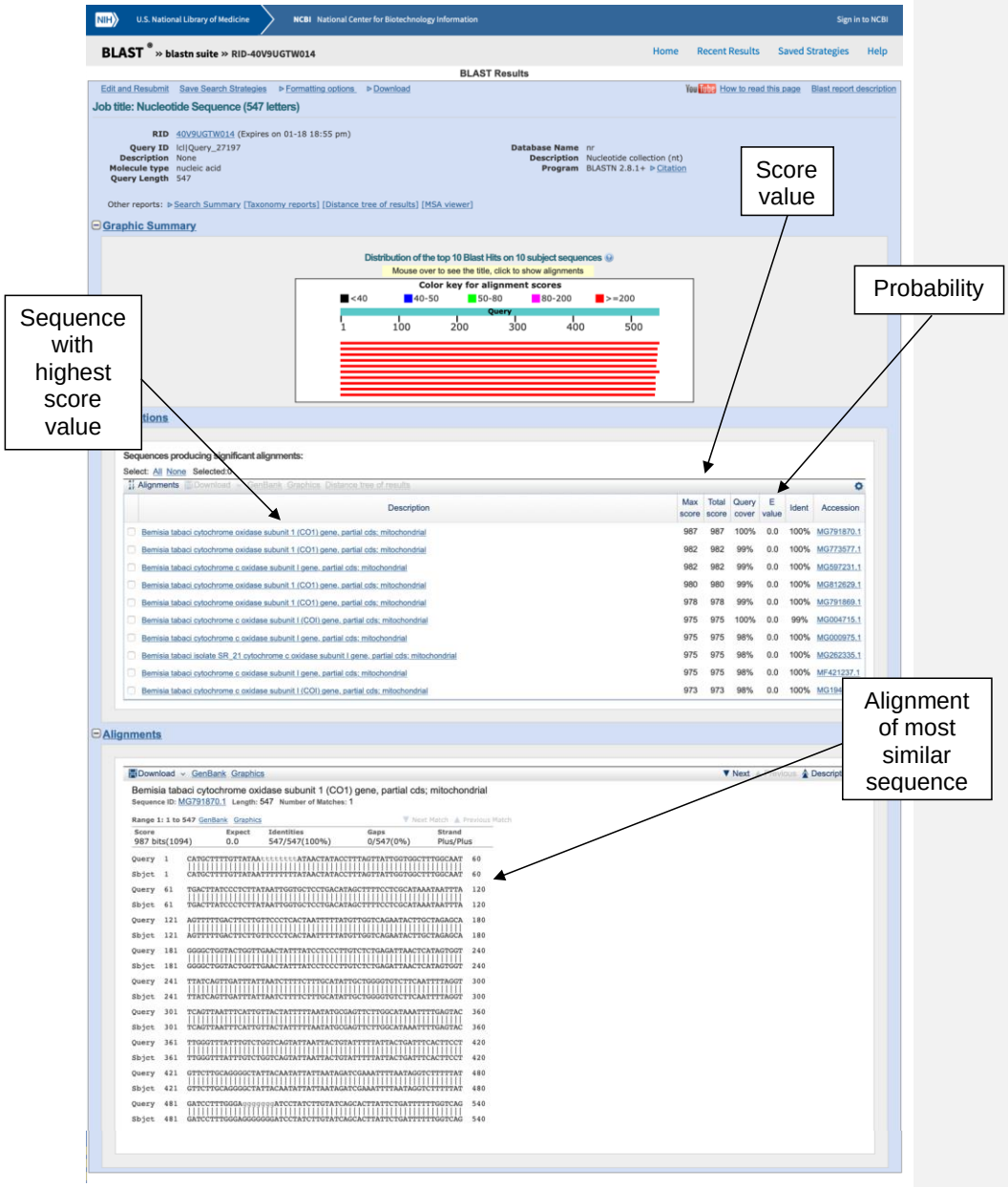
The Basic Local Alignment Search Tool (BLAST) (https://blast.ncbi.nlm.nih.gov/Blast.cgi) for comparing gene and protein sequences against others in public databases.

The most useful database search for DNA barcode use is nucleotide BLAST (or BLASTn). This returns information from the database about a query sequence, essentially allowing you to identify 'unknown' sequences or confirm the identity of a known sequence. Remember, like BOLD the results only reflect what is present on the database, thus negative results (no significant matches may be difficult to interpret). Paste in your query sequence into the link, set the database to nr (non-redundant) and optimise for 'somewhat similar' sequences.



Interpreting the results

After submitting a BLAST query the results will be displayed graphically (colour coding score values), followed by a list of highest scoring matches and finally alignments of the Highest Scoring Sequence Pairs (HSP's).



Interpreting the score values

Complete explanation of the interpretation of the score values is beyond the scope of this introductory protocol booklet, and can be found at (http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html). However, it is important to understand the basics.

The score is the basis of the BLAST search. In an alignment of a pair of sequences each nucleotide is given a score depending on whether it matches or not. If there is no match the score is negative, and if there is a match the score is positive. For each region of alignment all the scores are added up, though the score can never go below zero. A BLAST search is looking for regions of sequence that align together with a probability that is greater than random chance thus giving a high score value - these are called High-scoring Segment Pairs (HSP).

When you get a BLAST result you also get a probability value. This value is the probability that the HSP occurs by random. For example, if you take a completely random sequence the same length as your query sequence there is a chance that you could get exactly the same sequence by random. This probability is based on the length of the query sequence and the total length of the database. If you are comparing a short sequence it is more likely that a random sequence could give you the same result. If you are comparing a sequence to a large database of sequences, there is more chance that you have a random sequence in there that matches your query sequence than if you are comparing your sequence to a small database.

A probability of 0 means that there is essential no chance your match was random. A probability of 10^{-50} (reported as 1 E -50) means that there is 1 in 10^{50} chance a random sequence of the same length would generate this score value. This is not very likely, and so most hits returned with a score of 10^{-50} are probably real. In contrast a result of 0.1 means that there is a 1 in 10 chance of a random sequence of the same length generating this score value thus it is probably not significant.

Problems using database searching

Whilst BLAST searching is the simplest and as a result most widely used approach to identification it is not the most reliable. If sequences for the taxa being searched are not within the database being searched the most similar sequence present will be returned as the best match – the score values, % identity and coverage should be examined to investigate if this close match is a reliable identification. Coverage of sequences within the database (e.g. multiple sequences of a single taxa) can also cause problems with using BLAST algorithms for the identification of species. Problems can also be encountered if samples are incorrectly identified prior to sequencing and loading onto the database, here even an identical match would be misleading, looking for matches to multiple sequences from different studies is a good approach to identifying 'rogue' sequences in the database. Some databases (e.g. BOLD and qBANK) use voucher specimens and curated sequences to provide more confidence that the material was identified accurately before sequencing.