

08 | Bioinformatics: DNA identity



Equipment list

- Computer for:
 - viewing viewing chromatograms
 - using a BLAST to identify invertebrate samples from the DNA barcode sequence
- Instructions for viewing and interpreting chromatograms
- Instructions for using a BLAST to identify the organism that a DNA barcode came from

Instructions | Interpreting the chromatogram from DNA sequencing

What does a chromatogram show?

During Sanger sequencing the DNA bases are labelled with fluorescent tags:

A = green

C = blue

G = yellow

T = red

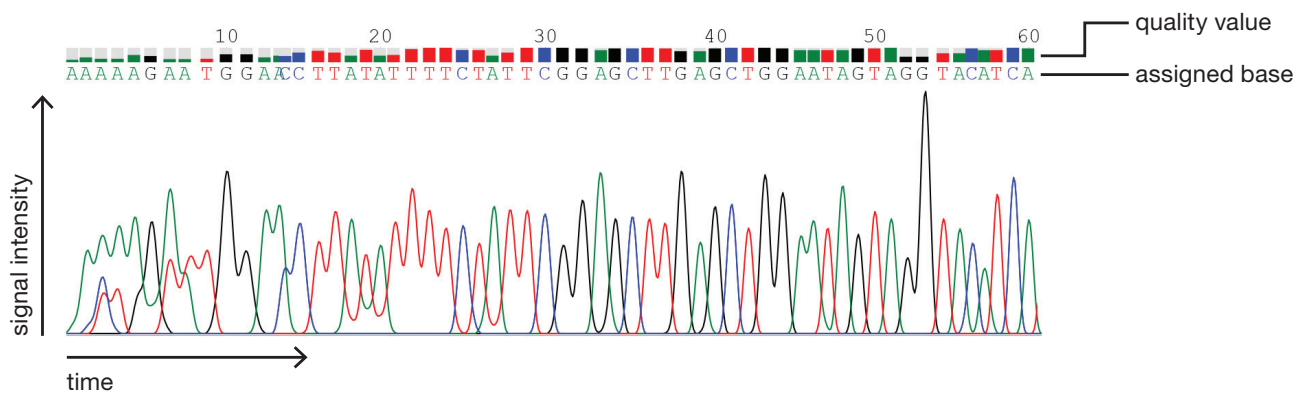
Then the bases are separated by size using capillary gel electrophoresis (separating DNA by size through gel in a very narrow capillary using an electrical current).

A laser causes excitation of the fluorescent tags, which a detector records.

This gives a **chromatogram**, which plots the strength of the fluorescence detected against time (synonymous with DNA length). The colour of the fluorescence is used to assign the sequence of the bases, with the strength of fluorescence giving a measure of the certainty / quality of the sequence. An algorithm is used to assign bases.

- The first part of the sequence is often poor (for about 40 bases) as very short fragments do not migrate predictably during gel electrophoresis.
- The best data is usually between 100 and 500 bp into the sequence – you can see this as sharp, evenly-spaced peaks.
- Towards the end of the trace the peaks are not as strong and become less evenly-spaced.

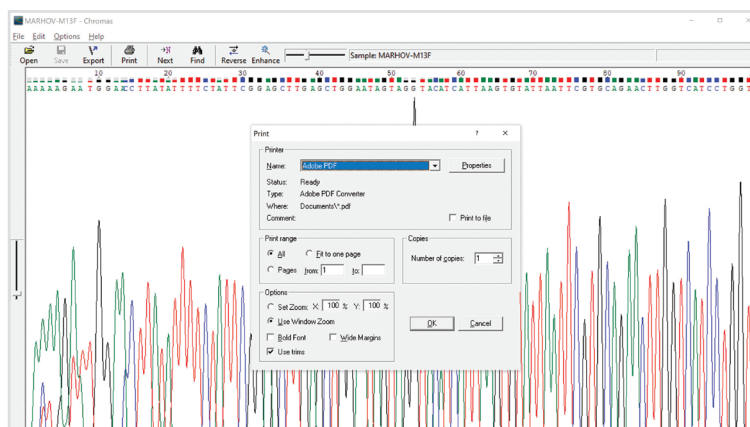
View of a chromatogram



Time, which correlates with base position, is plotted on the x axis. Signal intensity, in relative fluorescence units, is plotted on the y axis. Analysis software processes the raw data and assigns bases to the peaks to generate the final DNA sequence. Quality values are calculated for each assigned base, providing a measure of confidence in the assigned base.

1. Open your chromatogram file

- You may be given access to your chromatogram as an **.ab1** file, to be opened using Chromas software, or as a **pdf**.
- If you are given an **.ab1** file, you can explore the chromatogram using the free Chromas software. This can be downloaded from: technelysium.com.au/wp/chromas
- To then obtain the chromatogram for viewing and analysis, select **print** from the top bar, then set printer name to **Adobe pdf**. Save it in a place, and use a name that you will remember.



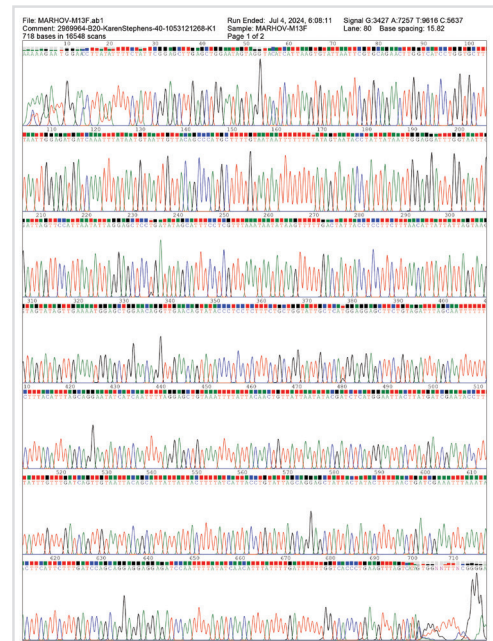
- If you are given a **pdf**, you will not need to do the 2 previous steps.

□ 2. View your chromatogram and assess DNA sequence accuracy

- Zoom the chromatogram to fit onto one page. You will be able to see more than 710 bp on an A4 sheet.
- Discuss the accuracy of your DNA barcode sequence (using information from the chromatogram) with your partner, teacher and class.

For example, my sequence was accurate as the fluorescent peaks detected were sharp and evenly-spaced and only the first 21 bp were inaccurate.

- Make sure that you save the chromatogram in a file format, location and using a file name that you will be able to find and use again.



□ 3. Identify the DNA sequence to compare to databases of known sequence

- You may be given access to your chromatogram as an **.seq** file, which can be opened in Microsoft Word, or already saved as a **Microsoft Word** file.
- When you open the file it will show the DNA sequence from the chromatogram.

It is in FASTA format, with > indicating the name of the DNA sequence, and the sequence then starting on the next line. FASTA (Fast-All) format allows the sequence to be easily used in sequence alignment with existing databases.

```
>MARHOV-M13F_B20
AAAAAGAATGGAACCTTATATTTTCTATTCGGAGCTTGAGCTGGAATAGTAGGTACATCATT
AAGTGTATTAATTCGTGCAGAACTTGGTCATCCTGGTCTTAAATTGGAGATGATCAAATTT
ATAATGTAATTGTTACAGCCCATGCTTTTGTAAATAATTTTTTTTATAGTAATACCTATTATA
ATTGGAGGATTGGTAATTGATTAGTTCCATTAATATTAGGAGCTCCTGATATAGCATTCC
TCGTTTAAATAATATAAGTTTGGACTATTACCTCCTTAAACATTATATTAGTAAGTA
GTATAGTTGAAATGGAGCTGGAACAGGTTGAACAGTATACCTCCTCTTCTGCTGGTATT
GCTCATGGAGAGCTTCTGTAGATTAGCAATTTTCTTTACATTAGCAGGAATATCATC
AATTTTAGGAGCTGTAAATTTTATACAACGTATTATTAATATACGATCTCATGGAATTACTT
ATGATCGAATACCTTTATTTGTTTGATCAGTTGTAATTACAGCATTTATTACTTTTATCA
TTACCTGTATTAGCAGGAGCTATTACTATACTTTAACTGATCGAAATTTAAATACTTCATT
CTTTGATCCAGCAGGAGGAGGATCCAATTTTATATCAACATTTATTTTGATTTTGGTC
ACCTGAAAGTTTAGTCATAGTGGNNTTTNCGGGGAA
```

- If there are a large number of **N** (unknown base) recorded early and late in your sequence then this part of the DNA sequence is not very accurate and will not yield good results when trying to identify your invertebrate.
- Find any unknown nucleotides (N) and select a length of sequence that does not include any Ns for comparison to databases of known sequence. Indicate in your file which parts of the sequence are of high quality and will be used in analysis by highlighting the text.
- Make sure that you save the DNA sequence in a file format, location and using a file name that you will be able to find and use again.

Instructions | Using a BLAST to identify your invertebrate sample

A **BLAST** (**B**asic **L**ocal **A**lignment **S**earch **T**ool) is used to compare DNA sequences. Now that you have your DNA barcode sequence files you can use a BLAST against a database of known sequences to identify the invertebrate that you sampled at the start of the project.

DNA barcode is the name given to the DNA sequence of a gene found in the mitochondrial DNA of all animals. The mitochondrial cytochrome oxidase subunit 1 gene is used as the DNA barcode. It is a useful tool for identifying organisms as the gene sequence is constant within a species, but varies between species.

From this point onwards if you would prefer to use online instructions follow this link:

app.tango.us/app/workflow/08-Using-BLAST-to-identify-your-sample-787fce2ee4fb418ab27423321d8fc046

1. Find the National Centre for Biological Information (NCBI) website with free software for DNA sequence comparison

- Type **NCBI BLAST** into an internet search engine.
- Click on the link for **BLAST: Basic Local Alignment Search Tool**.
- Select **Nucleotide BLAST**. This should take you onto the blastn tab.

2. Match the DNA barcode sequence from your sampled invertebrate against a database of DNA sequences, to find which organism the DNA barcode came from

- Use the accurate portion of the sequence file returned after DNA sequencing, saved in a new text file, in FASTA format.

FASTA format contains a > indicating the name of the DNA sequence, and the sequence then starting on the next line. FASTA (Fast-All) format allows the sequence to be easily used in sequence alignment with existing databases.

- Copy the >, name and sequence for your DNA barcode.
- On the blastn tab, in the white box in the **Enter Query Sequence** section, copy and paste the >, name and sequence for your DNA barcode.
- In the **Choose Search Set** section, the **Standard databases (nr etc.)** should be checked.
- Select **Nucleotide collection (nr/nt)** from the drop down menu.
- In the **Program Selection** section, optimise for **Highly similar sequences (megablast)**.
- Click the blue **BLAST** button. Algorithms will try to find the best match for your DNA barcode by comparing it to all of the DNA sequences stored in its database. Depending on how many searches are submitted at the same time as yours this may take a few minutes.

3. Understanding the results

- Scroll down until you see 4 tabs.
- On the **Descriptions** tab, you can see the scientific name (binomial classification) of the organism and the name of the DNA sequence that matches your query.

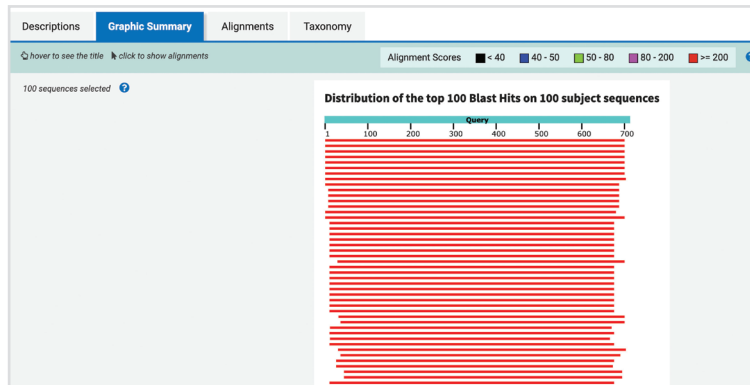
The screenshot shows the 'Descriptions' tab of a BLAST search results page. It displays a table of sequences that produced significant alignments. The table has columns for Description, Scientific Name, Max Score, Total Score, Query Cover, E value, Per. Ident, Acc. Len, and Accession. The results list various E. coli strains and their corresponding cytochrome c oxidase subunit I (COX1) gene sequences. The E values are consistently low, indicating high similarity to the query sequence.

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
E. coli O157:H7 complete genome	E. coli O157:H7	1254	1254	97%	0.0	99.42%	16175	NC_036481.1
E. coli O157:H7 complete genome	E. coli O157:H7	1254	1254	97%	0.0	99.42%	1509	K0262632.1
PREDICTED: E. coli O157:H7 complete genome	E. coli O157:H7	1254	1254	97%	0.0	99.42%	2375	XM_055895641.1
E. coli O157:H7 complete genome	E. coli O157:H7	1254	1254	97%	0.0	99.42%	16014	MT872709.1
E. coli O157:H7 complete genome	E. coli O157:H7	1254	1254	97%	0.0	99.42%	1529	KC192981.1
E. coli O157:H7 complete genome	E. coli O157:H7	1254	1254	97%	0.0	99.42%	4081	MT410784.1
E. coli O157:H7 complete genome	E. coli O157:H7	1249	1249	97%	0.0	99.28%	16547	OK299507.1
E. coli O157:H7 complete genome	E. coli O157:H7	1249	1249	98%	0.0	99.14%	16832	OK244022.1
E. coli O157:H7 complete genome	E. coli O157:H7	1238	1238	96%	0.0	99.56%	685	PP741878.1
E. coli O157:H7 complete genome	E. coli O157:H7	1225	1225	95%	0.0	99.40%	680	OK345240.1
E. coli O157:H7 complete genome	E. coli O157:H7	1225	1225	95%	0.0	99.40%	682	OK63271.1
E. coli O157:H7 complete genome	E. coli O157:H7	1225	1225	95%	0.0	99.40%	682	OK63270.1
E. coli O157:H7 complete genome	E. coli O157:H7	1225	1225	95%	0.0	99.40%	682	OK63269.1
E. coli O157:H7 complete genome	E. coli O157:H7	1223	1223	95%	0.0	99.55%	675	MT516428.1
E. coli O157:H7 complete genome	E. coli O157:H7	1216	1216	97%	0.0	98.41%	16141	NC_008754.1
E. coli O157:H7 complete genome	E. coli O157:H7	1210	1210	93%	0.0	99.85%	658	OK065259.1

- In the **Description** column, each line shows the species name, a sample reference made up of numbers and letters, then what the DNA sequence is. You should see the term **cytochrome c oxidase subunit I**, which is the gene used as a barcode for animal species, or **mitochondrion**, as this is where the DNA barcode gene is located.
- In the **Scientific name** column it gives the binomial classification of the organism. Click on this to see more information on classification and the common name, or type the binomial classification into an internet search engine to find out what the common name for the invertebrate is.
- In a column to the right, the **E value** or Expectation value, is the number of alignments with the query sequence that would be expected to occur by chance in the database.

Lower E values mean that the chance of the alignment occurring randomly is very low, so the probability that the sequence retrieved is related to the query sequence is high.

- On the **Graphic Summary** tab, you can see whether the sequence alignments are for the whole of the query sequence or just part of it.



- Remember that the longer alignments provide more precise results.
Precision = the closeness of agreement between independent measurements obtained under the same conditions.
- On the **Alignments** tab there is a detailed view of each sequence from the database aligned to the query sequence.

Query	Subject	Score	Expect	Identities	Gaps	Strand
3	1535	1254 bits(679)	0.0	688/692(99%)	2/692(0%)	Plus/Plus
AAAGA-A-TGGAACCTTATATTTCTATTCGGAGCTTGAGCTGGAATAGTAGTACATCA	AAAGATATTGGAACTTATATTTCTATTCGGAGCTTGAGCTGGAATAGTAGTACATCA	60				
61	1594					
TTAAGTGATTAAATTCGTGAGAACTTGGTCATCTGGTCTTTAATTGGAGATGATCAA	TTAAGTGATTAAATTCGTGAGAACTTGGTCATCTGGTCTTTAATTGGAGATGATCAA	120				
121	1654					
ATTTATAATGTAATTTTACAGCCCATGCTTTGTAAATAA*****ATAGTAATACCT	ATTTATAATGTAATTTTACAGCCCATGCTTTGTAAATAA*****ATAGTAATACCT	180				
181	1714					
ATTATAATTGGAGGATTTGGTAATTGATTAGTCCATTAAATATTAGGAGCTCCTGATATA	ATTATAATTGGAGGATTTGGTAATTGATTAGTCCATTAAATATTAGGAGCTCCTGATATA	240				
1715	1774					
ATTATAATTGGAGGATTTGGTAATTGATTAGTCCATTAAATATTAGGAGCTCCTGATATA	ATTATAATTGGAGGATTTGGTAATTGATTAGTCCATTAAATATTAGGAGCTCCTGATATA					

- To present the results of your scientific identification of your invertebrate using a DNA barcode, you should include this alignment. Click the **Download** button in the top left hand corner and select **text (aligned sequences)**, then press continue.
- Save the alignment in a location you can find again, and with a name that you will remember.

4. Share your results

- An important part of science is the ability to communicate results to others.
- Comment on the accuracy of your DNA barcode sequence (using information from the chromatogram).
- Are you able to identify your invertebrate species by comparing your DNA barcode sequence to a database using a BLAST search?
- If unable to identify your invertebrate species is it because of lack of accurate sequence or is the sequence accurate and the species appears not to be present in the database?