

## 08 | Bioinformatics: DNA identity



**Aim** View and interpret the chromatogram from DNA sequencing. Use the DNA barcode sequence file to identify the invertebrates sampled using bioinformatics.

**Activity outline** Explain the data displayed in a chromatogram and how it provides information on the accuracy of the DNA sequence generated from Sanger sequencing. Interpret how accurate the DNA barcode sequences are from the chromatograms, using Chromas (free to download software).

Open the sequence file and select the region of quality DNA sequence from the DNA barcode. Using this DNA barcode sequence, perform a BLAST (Basic Local Alignment Search Tool) to identify the invertebrate sampled.

Students should share information about the invertebrates that they have identified. Can they all identify their invertebrate from the DNA sequence? Has anyone got a quality DNA sequence that is not similar to known invertebrates in the database?

**Age range** Key stage 4 and above (14 years and older)

**Timing** 20 min - interpretation of the chromatogram  
 20 min - using bioinformatics to BLAST the sequence file against a database  
 20 min - sharing information found about the invertebrates sampled from their DNA barcode sequence

**Venue** Classroom with laptops or a computer room

**Resources**

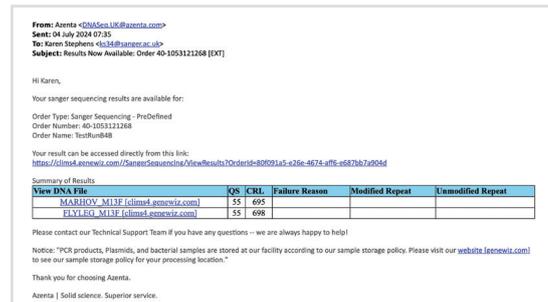
- Access to chromatograms and sequence files
- Ability to view chromatograms using Chromas software
- Instructions for the bioinformatics and BLAST activity (with [additional online guide](#))
- **Presentation:** 08\_P\_Bioinformatics-DNA-identity

## Preparation

### Accessing results

For each sample sent for DNA sequencing there should be a **chromatogram** and a **sequence file** returned. These will be accessible from an email.

- The email you get will give a link, through which your results can be accessed. Click on the link and it will take you to the Azenta Life Sciences webpage.



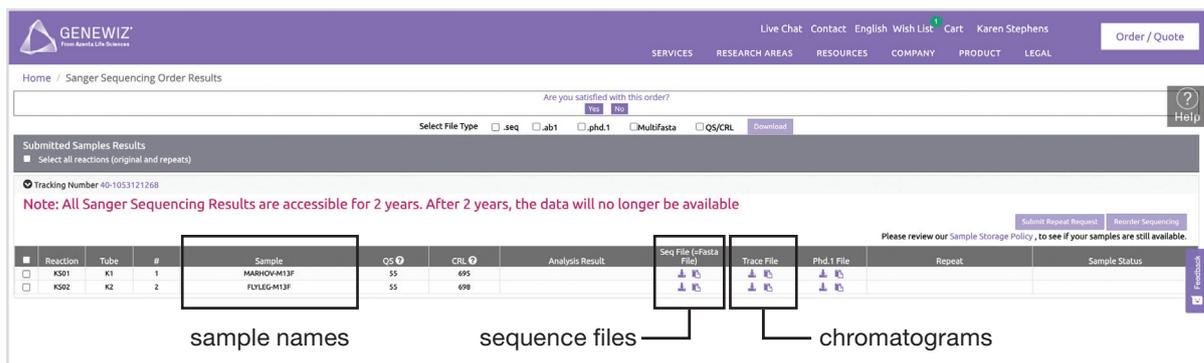
- In order to view your results using this link, you will need to log in with the following username and password.

- Username: **ks34@sanger.ac.uk**
- Password: **Barcoding4Beginners.**

This will allow you access not only to your DNA sequence, but to data generated by other schools and colleges participating in Barcoding for beginners, embracing the Wellcome Sanger Institute commitment to Open Access data.



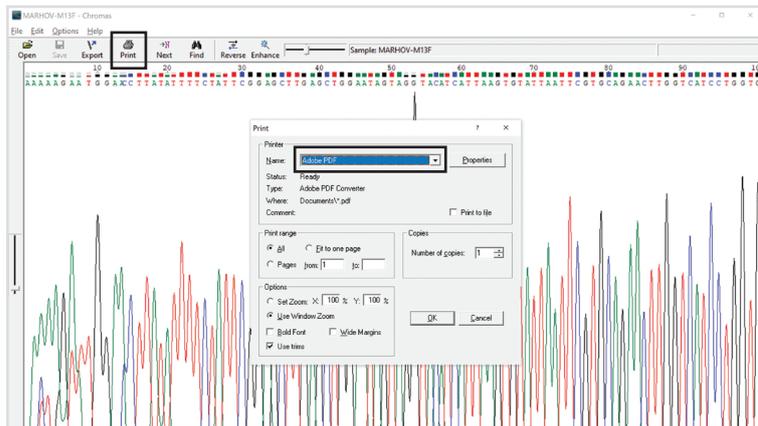
- Once you have logged in, you can download the chromatogram and sequence files.
  - In the column labelled **Sample** you will see the 7 letter sample names. This allows you to identify which sample the results are for.
  - The column showing **Trace file** has the **chromatograms**, which can be used to judge the accuracy of DNA sequencing. These are **.ab1** files, which can be opened using the free Chromas software. This can be downloaded from: [technelysium.com.au/wp/chromas](http://technelysium.com.au/wp/chromas)
  - The column labelled **Seq file (=FASTA file)** has the **DNA sequence** for the sample as a **.seq** file, which can be opened in Microsoft Word.



## Chromatogram (.ab1 file)

You can explore the chromatogram using the **Chromas software**, but the easiest way to obtain the chromatogram for viewing and analysis is to save it as a pdf file.

- Select **print** from the top bar, then set printer name to **Adobe pdf**, and save it in a place and use a name that you will remember.



- When you open the pdf file you will be able to see more than 710 bp on an A4 sheet.

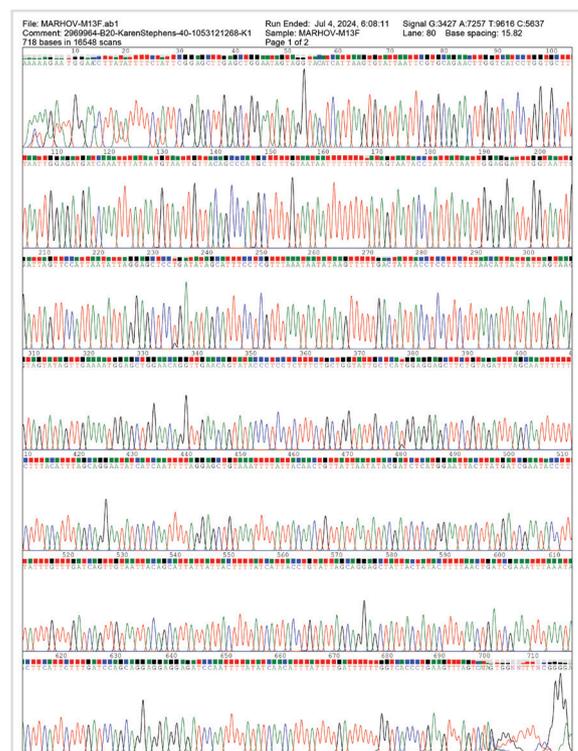
A chromatogram plots the strength of the fluorescence detected against time (synonymous with DNA length). The colour of the fluorescence is used to assign the sequence of the bases, with the strength of fluorescence giving a measure of the certainty / quality of the sequence. An algorithm is used to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

From the view of the chromatogram in a pdf file, you will be able to see at a glance:

- The fluorescent tags used for the different bases (although the yellow tag for guanine is replaced with black on the trace).
- The quality of the sequence. Sharp, evenly-spaced peaks indicate excellent sequence quality.
  - ▲ *It is usual to have about 40 bases of poor sequence quality at the start of the trace, as very short fragments do not migrate predictably by gel electrophoresis.*

From this students should be able to comment on the accuracy of their sequence.

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



## Sequence (.seq file)

When you open the .seq file in Microsoft Word, it will give you the 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.

- Save the sequence file in a place, and use a name that you will remember.

```
>MARHOV-M13F_B20
AAAAAGAATGGAACCTTATATTTTCTATTCGGAGCTTGAGCTGGAATAGTAGGTACATCATT
AAGTGTATTAATTCGTGCAGAACTTGGTCATCCTGGTGCTTAAATGGAGATGATCAAATTT
ATAATGTAATTGTTACAGCCCATGCTTTTGTAAATAATTTTTTTTATAGTAATACCTATTATA
ATTGGAGGATTTGGTAATTGATTAGTTCATTAATATTAGGAGCTCCTGATATAGCATTTC
TCGTTTAAATAATAAAGTTTTGACTATTACCTCCTCTTAAACATTATTATTAGTAAGTA
GTATAGTTGAAAATGGAGCTGGAACAGGTTGAACAGTATACCTCCTCTTCTGCTGGTATT
GCTCATGGAGGAGCTTCTGTAGATTTAGCAATTTTTCTTTACATTTAGCAGGAATATCATC
AATTTTAGGAGCTGTAAATTTTATTACAACGTGTTAATATATACGATCTCATGGAATTACTT
ATGATCGAATACCTTTATTTGTTGATCAGTTGTAATTACAGCATTATTATTACTTTTATCA
TTACCTGTATTAGCAGGAGCTATTACTATACCTTTAACTGATCGAAAATTTAAATACTTCATT
CTTTGATCCAGCAGGAGGAGGAGATCCAATTTTATATCAACATTTATTTGATTTTTTGGTC
ACCTGAAGTTTAGTCATAGTGNNTTTTNCGGGGAA
```

- Identify any unknown nucleotides (N) and select a length of sequence that does not include any Ns for comparison to databases of known sequence.

```
>MARHOV-M13F_B20
AAAAAGAATGGAACCTTATATTTTCTATTCGGAGCTTGAGCTGGAATAGTAGGTACATCATT
AAGTGTATTAATTCGTGCAGAACTTGGTCATCCTGGTGCTTAAATGGAGATGATCAAATTT
ATAATGTAATTGTTACAGCCCATGCTTTTGTAAATAATTTTTTTTATAGTAATACCTATTATA
ATTGGAGGATTTGGTAATTGATTAGTTCATTAATATTAGGAGCTCCTGATATAGCATTTC
TCGTTTAAATAATAAAGTTTTGACTATTACCTCCTCTTAAACATTATTATTAGTAAGTA
GTATAGTTGAAAATGGAGCTGGAACAGGTTGAACAGTATACCTCCTCTTCTGCTGGTATT
GCTCATGGAGGAGCTTCTGTAGATTTAGCAATTTTTCTTTACATTTAGCAGGAATATCATC
AATTTTAGGAGCTGTAAATTTTATTACAACGTGTTAATATATACGATCTCATGGAATTACTT
ATGATCGAATACCTTTATTTGTTGATCAGTTGTAATTACAGCATTATTATTACTTTTATCA
TTACCTGTATTAGCAGGAGCTATTACTATACCTTTAACTGATCGAAAATTTAAATACTTCATT
CTTTGATCCAGCAGGAGGAGGAGATCCAATTTTATATCAACATTTATTTGATTTTTTGGTC
ACCTGAAGTTTAGTCATAGTGGNNTTTTNCGGGGAA
```

- ▲ *Although the region of the cytochrome c oxidase subunit I (barcode) gene that we sequence is 710 bp, additional DNA sequence is added by the primers and DNA sequencing technique used*

## Before the session

Make sure that **chromatogram files** (as .ab1 or a pdf) and **DNA sequence files** (as .seq or Word documents) are accessible to students from the start of this session.