A useful tool for genome investigation and analysis, with a focus on *Rubus idaeus*

## Introduction/Motivation

*Rubus idaeus*, commonly known as red raspberries, are popular fruits often consumed for their health benefits relating to the high fibre and phytochemical contents[1].Yet, despite their popularity, there is a deficiency in knowledge surrounding the raspberry genome[2]. Many studies have produced a genome map from various raspberry varieties, yet none are complete, and thus valuable information regarding genomic mechanisms and their role in raspberry development and yield may be unknown[3]. Consequently, it is important to develop the knowledge surrounding the raspberry genome. The production of a tool to facilitate genome and DNA sequence investigation will enable this, by allowing for easy transcription, translation and GC content identification of desired DNA sequences[4]. Furthermore, the development of a tool to facilitate sequence alignment for the multiple genome sequences from the raspberry varieties will allow for identification of conserved regions between varieties of raspberries, and will support the development of knowledge surrounding potential gene targets for the improvement of valuable traits in the fruit[5]. This study will aim to produce a resource to enable the simplification of these processes, to develop knowledge surrounding the raspberry genome. This could have important future implications, by improving the ease by which simple genetic investigative processes can be completed. This tool will be functional with all other compatible DNA sequences, allowing application to many other research areas away from crop sciences, such as ecology and conservation research, among others.

## Methods

The genome investigation tool was produced using Python, version 3.13.1, and utilised the Biopython package[6], version 1.84. The code to run the project can be accessed from the following link: https://github.com/lottiewilson02/SWBio_Short_Project. MAFFT (Multiple Alignment using Fast Fourier Transform)[7] was also installed, version 7.511, using the Command Line. The subprocess module, the io module for StringIO and AlignIO, and the urllib module were also installed.

For the DNA sequence investigation, the DNA sequence of the “Glen Ample” raspberry variety was used. The following code then allows for complementation of the DNA sequence, alongside reverse complementation, transcription from DNA to RNA, back-transcription, and translation from RNA to the corresponding proteins. The code then was produced to calculate the percentage of the Guanine (G) and Cytosine (C) content of the DNA sequence. The output is printed for each tool, both to provide the required result and to allow for identification of any errors.

Furthermore, the code was then created to allow for the alignment of multiple DNA sequences, to enable further investigation into the identity of the chosen sequences. This was tested used DNA sequences from different varieties of *R. idaeus*, along with a sequence from *Rubus argustus* as a comparison. The code was produced to allow for the importation of the URL link to the file containing the sequences required for analysis, with the description of each sequence then printed to check the full sequence list was present. To create the multiple sequence alignment, each sequence object was converted into a string, containing the description of each sequence, followed by the relevant sequence. A subprocess, using MAFFT, was then ran to perform the alignment of the sequence, with the alignment output printed. Following the alignment, a check was written to ensure that the length of all sequences matched to ensure the alignment had been performed correctly.

## Results

The code for the sequence investigation tools was tested using a section of the DNA sequence from the ‘Glen Ample’ variety of *R. idaeus*, Table 1, to allow for the identification of any errors. For the complementation and reverse complementation, transcription, translation

to proteins, and the calculation of the GC content, no errors were identified. Accurate sequences were produced in the output for all sections, as seen in Table 1, with the correct value provided for the percentage GC content for the sequence, as in Table 1.

**Table 1: The results of each sequence investigation tool, with the output produced.** The input sequence was shown, a section from the genome of the "Glen Ample" variety of *Rubus idaeus*. The tools produced allowed for complementation, reverse complementation, transcription, back transcription, translation and the calculation of the percentage content of Guanine and Cytosine bases in the sequence.

| Tool for sequence investigation | Output |
|---|---|
| Input (chosen sequence, Glen Ample) | CAACAATCTCGACAGGTGCCTCAGGGACAGCTTCTTCAGCAG |
| Complementation | GTTGTTAGAGCTGTCCACGGAGTCCCTGTCGAAGAAGTCGTC |
| Reverse complementation | CTGCTGAAGAAGCTGTCCCTGAGGCACCTGTCGAGATTGTTG |
| Transcription (input DNA to RNA) | CAACAAUCUCGACAGGUGCCUCAGGGACAGCUUCUUCAGCAG |
| Back transcription | CAACAATCTCGACAGGTGCCTCAGGGACAGCTTCTTCAGCAG |
| Translation | QQSRQVPQGQLLQQ |
| GC% | 40.95 |

The alignment of multiple sequences used DNA sequences from different varieties of the *R. idaeus* species, alongside a sequence from *R. argustus* to act as a comparison. The sequences were imported and parsed successfully, with all sequences present with the correct corresponding ID upon checking. The alignment was also performed successfully, with the results showing correct identification of conserved regions between each genome. This is supported by the length of each sequence post-alignment, with all sequences showing a length of 327, indicating a successful alignment.

## Conclusion

Overall, we can conclude that this tool to assist with simple genomic processes and calculations, along with the alignment tool, successfully allows for the desired processes to be carried out. This will enable the development of knowledge surrounding the genomics of raspberries, which can the assist with breeding targets to ensure that the fruit can be grown without compromise in fruit yield or quality. This will have many applications in future research, allowing for quick and easy transcription, translation and alignment for a wide range of DNA sequences, allowing application to many commercial crops. However, further work may be required for the refinement of this tool for it to be in competition with similar resources, such as ExPASy[8]. Such tools have developed a more professional interface, alongside enabling a greater range of functions to be carried out. An example of this could be the addition of an open reading frame identification tool, to identify coding regions of the DNA sequences, along with an alignment analysis tool, to provide a score to assess the accuracy of the multiple sequence alignment produced by this tool. This refinement would improve the effectiveness and user experience of this tool[9].

## References


**[1]** Franck, M. *et al.* (2020). Nutrients, 12(12), 3858.
**[2]** Price, R. J. *et al.* (2023). PLoS ONE, 18(5), 285756–285756.
**[3]** Ishka, M.R. *et al.* (2023). F1000Research, 12, 1257–1257.
**[4]** Benjamini, Y. *et al.* (2012). Nucleic Acids Research, 40(10), 72.
**[5]** Hall, A.E. *et al.* (2002). Plant Physiology, 129(4), 1439–1447.
**[6]** Cock, P.J. *et al*. (2009). Bioinformatics, *25*(11), 1422–1423.
**[7]** Katoh K. *et al*. (2002). Nucleic Acids Res. 2002(30), 3059–3066.
**[8]** Gasteiger E. *et al*. (2003). Nucleic Acids Res. 31, 3784-3788.
**[9]** Clausen, M. *et al.* (2024). *Genetics in Medicine Open*. 2, 101814–101814.