Practical 3

Re-Sequencing Genomes

You've learnt so far how to assemble and annotate a whole genome from scratch. That's a very useful skill but, why stop there? Sequencing is getting cheaper and cheaper and now it is possible to compare the genome sequence of multiple individuals at a time. After all, biology is about populations and diversity, and that's the last thing we will explore in this series of practicals: genome diversity.

3.1 Comparing genome sequences

It all started with a few contaminated apricots, now you know that a pathogenic strain of *Campylobacter jejuni* was behind the outbreak. You know the genome sequence, its gene content, and you even analyzed its expression profile. But how does this genome compare to other evolutionary related genomes? How similar is to *Campylobacter coli* which you used to assemble the contigs into scaffolds in the first practical?

The best way to answer this question is by whole-genome alignment. You will remember from the first part of this module that short DNA sequences can be aligned using standard algorithms. Now, you're gonna use similar strategies to align whole genome sequences. You already know the software, `Mauve`, but you will use it to compare multiple fully assembly genomes. To do so, you will align the genomes of your query *Campylobacter jejuni* genome, another *C. jejuni* strain, and other related species: *C. coli* and *C. fetus*.

Task 1: Align genomes with MAUVE

1. Uncompress the `campylo.zip` file provided.

2. Open the `Mauve` program and go to 'File->Align with progressive-Mauve'.

3. With 'Add Sequence' open the files for the four *Campylobacter* genomes provided.

4. Click on 'Align'.

5. The program will ask for an output file name. Pick any name you like (such as 'campylo_align'), and press Enter.

6. The program will take a few minutes to align the four genomes.[1]

7. When the process is done, you will see a window showing a four way alignment (Figure 3.1). Each color box is a conserved block, connected with lines accross the genomes.

8. You can hide selected genomes by cliking on the '-' on top of the red 'R'.

9. Play around with the alignments and discuss with your mates.

Questions to be addressed in the final report

- Which genome is more similar to your query genome and which one is more distant?

- Discuss in your report what you see in the alignment. You can include a figure of the alignment ('Tools->Expore->Export Image').

[1] Mauve uses a 'greedy approach' for sequence alignment, as we briefly discussed in the main lecture.

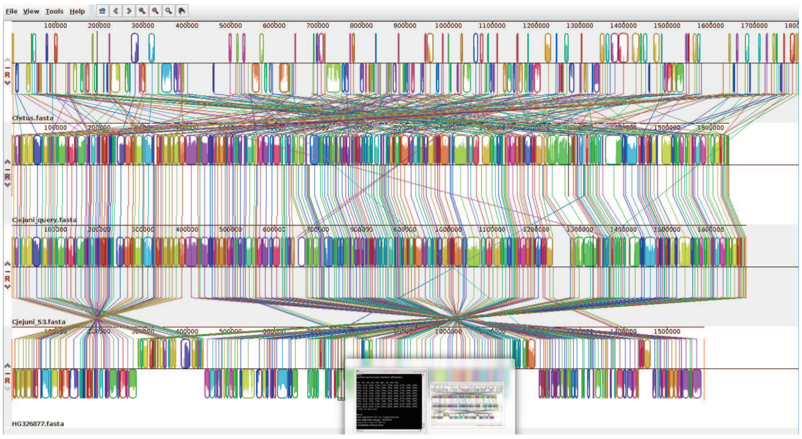

Figure 3.1: Mauve genome alignment program

3.2 Re-sequencing genomes

Epidemiological research in the outbreak due to contaminated apricots determined that a new strain of *C. jejuni* is particularly dangerous: it is lethal in most infected patiens! Most likely, this strain is likely the result of one or a few recent mutations. You just got a sample from this particularly dangerous bacteria and got it sequenced. Now, you're about to scan it for point mutations, that is, single nucleotide polymorphisms (SNP)[2] between the two samples. This is a common technique used in the analysis of population, and it is known as SNP-calling.

As in the previous practical, you will use our Galaxy server. Remember you can use any of these two addresses:

- `http://bsproteomics/galaxy/`
- `http://bsbioinfo/galaxy/`

Task 2a: SNP calling with Freebayes

- Go to Galaxy (see previous practical handbook for details) and log-in with your account.

[2] pronounced *snips*

- Upload the `fastq` or `fastqsanger` file provided. That contains sequence reads from the mutant lethal strain.
- Upload the *C. jejuni* query genome as a `fasta` file.
- Open the `Bowtie2` program.
- Select the `fastq` file you just uploaded.
- Change 'Use a built-in index' to 'Use one from history' and select the genome you uploaded.
- 'Execute'. Now your reads are going to be mapped to the reference genome you provided. The output file will be, as in the first of these practicals, a 'bam' file.
- Open the `Freebayes` program. Actually, in our Galaxy server it is called 'Call SNPs with Freebayes'.
- 'Reference file' should be the uploaded genome, 'Bam alignment file' is the output of `Bowtie2`. 'Execute'![3]
- Open the output file ('eye' icon) and you will see some comment lines to be ignored (beginning with '##') followed by the SNPs detected between the reference genome and our mutant strain.
- To sort out the SNPs detected by quality (QUAL) we open the program 'Sort' in Galaxy and select the output file from Freebayes. Then we sort numerically in descending order using column 6.
- The first two entries in the sorted table have a high quality. Choose the one with the lowest 'AB' value.[4]
- Write down the position of the detected SNP as you will be using this information in the next task.

Up to this point you have characterize the most likely point mutation that occurred in the mutant bacteria strain. Actually, many more mutations,

[3] Every time you see the word 'bayes' or 'bayesian' in a program means that the program use sophisticated probabilistic models. It other words, it's going to be slow. Be patience, the output of Freebayes can take a few minutes.

[4] 'AB' value is in the INFO column. 'AB' is a number between 0 and 1, which is the ration of the reference allele with respect to the alternative allele. The lowest the value, the more likely there's a SNP in this position.

including insertions and deletions have happened, but for simplicity we will focus solely in the most significant of the changes. Now, we should know were this mutation is in the genome. Let's have a look at a genome browser and find out.

Task 2b: Mapping a mutation with a genome browser

- In a web browser go to `http://archaea.ucsc.edu/cgi-bin/hgGateway?db=campJeju`
- This is a Genome Browser of *C. jejuni*, that is, a graphical display of a genome sequence we can navigate. In the 'position or search term' field type 'chr:' followed by the position of the SNP you detected. For instance, 'chr:123456'.[5]
- The browser will display only one nucleotide, which is where the mutation happened. Zoom out until you see the name of the gene where the SNP has been found. At this stage you may need to play around the browser and familiarize yourself with the information it contains.
- Click on the gene name (something like 'Cj' followed by a number).
- In this page you will find some useful about the affected gene.

Questions to be addressed in the final report

- What are the values of the top two detected SNPs?
- What is the 'AB' value of the selected SNP, and what it means.
- For this SNP, what is the reference genome allele and what is the allele in the mutant strain?
- Which gene is affected by this mutation?
- Why do you think this mutation may have produced a more dangerous bacteria?[6]

[5] In a Genome Browser you have to write the chromosome number, for instance, chr1 or chr 21. Since *C. jejuni* has only one chromosome, you write only 'chr'.

[6] There's no correct answer here, but this is a great opportunity for you to speculate based on the predicted function of the gene.

Well done! You've successfully analyzed the genome of the new pathogenic strain. Thanks to your efforts (and those of your colleagues) applied pharmacologists are now working in a new medication and epidemiologists know how to quickly detect the bacteria. The World needs scientists like you to fight against disease. Have you consider a career in the exciting field of genomics?

3.3 Genome-Wide Association Studies

This part of the practical is here only for completeness and should not be covered in the SPF. The goal is that you become familiar with GWAS[7] datasets and how to navigate them.

In the lecture we discussed how Genome-Wide Association Studies (GWAS) can help us to identify genes associated to diseases. Here we will explore a Manhattan plot to identify loci associated to Rheumatoid Arthritis. We will use the Integrated Genome Viewer (IGV), a Java application developed by the Broad Institute.

Task ∞: Browsing GWAS experiments with IGV

- Download the gwas file provided (via Moodle) and unzip it.
- Go to `http://www.broadinstitute.org/igv/` and register as a user to download the IGV
- Launch the application with 750 MB, do not update the Java version of your computer and ignore all warning messages.
- After a short while you will have a window open. Go to the top left corner click on the 'Human hg18'. This is not the version our data was mapped to. Select 'More...' and find version hg19. Upload it.
- Click on File->Load from file, and select your GWAS dataset.
- You will see a Manhattan plot which covers all human chromosomes. Without further instructions, explore it and try to identify which gene/s were associated to Rheumatoid Arthritis in this particular dataset.

[7] pronounced *gee-was*

3.4 Further reading

Details on SNP calling and software can be found in:

- Nielsen R et al (2011) Genotype and SNP calling from next-generation sequencing data *Nat Rev Genet* 12:443-451.

An affordable review on GWAS is:

- Bush WS and Moore JH (2012) Genome-Wide Association Studies . *PLoS Comput Biol* 8:e1002822.