Practical 2

Functional annotation of genomes

From our previous exercise we came up with a fully assembled genome sequence of our unknown pathogenic bacteria. Now it's time to know what all these letters (A,C,G,T) mean. In this practical you will use different techniques to detect and annotate genes in your genome. Although the programs here discussed are used in Bacteria annotation, the same principles apply to the annotation of eukaryotic genomes.

2.1 Detecting and annotating genes

Well done! You and your team successfully sequenced and assembled the genome of the bacteria that has been intimidating the world. But we still don't know much about this bacteria. For the moment you know is a *Campylobacter*-type species with a genome of about 2Mb in size. You, as the expert genome scientist in your team, are in charge of making sense of the whole genome sequence.

First, you need to predict genes in the genome sequence. As we discussed in the lectures, in bacterial genomes it is relatively easy to detect protein-coding genes. A number of web servers allow you to do this online, without running any program in your computer. You will submit your sequence to BASys, which combine different tools to fully annotate a bacterial genome sequence. These servers may be slow, and for that reason you will submit your genome sequence but then you will explore the results already generated a few days ago, so you don't have to wait.

Task 1: Automatic annotation of bacterial genomes

1. Open a web browser and go to `https://www.basys.ca/`
2. Choose one of the servers.
3. Fill in the form including your email address.
4. The Genome identifier is a name you give for your own records (it can be anything).
5. The bacteria is Gram Negative (as all other *Campylobacter*).
6. The contig is circular.
7. Click on the 'browse' tab and select the genome file provided by your instructor (available from Moodle).
8. Click on 'submit'

Now, the annotation process may take a few days. You will receive an email when the annotation is ready. However, instead of waiting for the results, we have previously run an annotation process a week ago, and you can find the results in the address provided by your instructor.

1. Open the annotation web page from Moodle.
2. Click on View Map. You may use this figure in your final report (save it first).
3. Right-click on the 'Gene Fasta File' and save it in your computer. This file contains the genes predicted for your genome, and you will need it in the next task.
4. If you click on 'Expand' you will be able to look into the details of the genome annotation.

5. In 'View Table' from the main menu you will find a list of predicted genes with their annotation.

6. Click on the first gene: BASYS00001

7. You will find useful information on nucleotide and amino acid sequences, whether it is in an operon or not, and functional annotation based on Gene Ontology.

8. Copy the nucleotide sequence and do a BLAST search against the NCBI 'non-redundant' database. The first hit will tell you which species is this.

Questions to be addressed in the final report

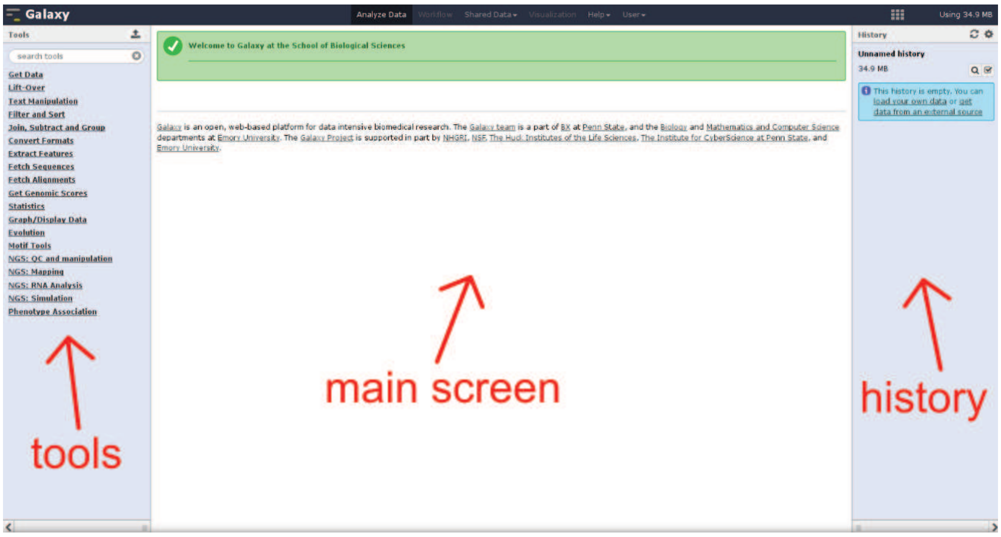
- How many genes did you annotate with BASys?
- Briefly describe the annotation of one of the genes of your choice.
- From which species is the bacterial sample?

2.2 Gene expression analysis

Now you have a series of genes annotated to your genome. How could you make sure that they are expressed? Well, the only way is to measure gene expression. While you were sequencing and annotating this bacteria, colleagues at another institution have extracted and sequenced the RNA content of some bacterial samples. They have deposited this information in a public repository, and you only have to get this data and compare them with your genome annotation. You will be using the full power of Galaxy, a web-based set of programs for bioinformatic analysis.

You will be using the Galaxy server at the School of Biological Sciences, which is only accessible within campus. There are two addresses to access to it:

- `http://bsproteomics/galaxy/`



Welcome to Galaxy at the School of Biological Sciences

Galaxy is an open, web-based platform for data intensive biomedical research. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

This history is empty. You can load your own data or get data from an external source

Figure 2.1: The School of Biological Sciences Galaxy server

- `http://bsbioinfo/galaxy/`

Create an account in the Galaxy server by going to User/Register in the top menu. Use your U. Essex student email account, and set a password of your choice and a public name. Please note that if you use different Galaxy server you will need different accounts for each server.

Before to start, familiarize yourself with Galaxy. Open our local Galaxy server in a web browser. Galaxy has three main windows (Figure 2.1). On the left you will see the tools window from which you will be able to upload files and run analyses. On the right you will find a history tab. In it, all your analyses will be recorded, and all files generated during the process will be listed. It is important that you keep record of these files so you know what are they at any time. The window in the center is the main window, and it will show the parameters and options of the programs you want to run, as well as the output files.

Task 2a: Measuring transcript levels with RNA-seq

- Go to Galaxy and log-in with your account details.
- First you need to pre-load the genes for which you want to measure the

expression level. Go to Get data->Upload File in the left panel.

- In the File Format field select 'fasta', and select the fasta file with gene sequences[1] with the 'Browse...' button. Press 'Execute'.
- On the left panel click on 'Get Data' and find the 'EBI SRA'. That will lead you to ENA, the European Nucleotide Archive.
- In the search field type 'ERX014110'. This is the accession number to the dataset we are interested in.
- Click on 'Search'.
- You will find some useful information about how the expression information was obtained. Write down the number of reads produced in this sequencing reaction.
- In a table at the bottom you will see the 'Submitted files (Galaxy)' column, click on the 'Fastq file 1' link below.
- After a while you will have a whole RNA-seq experiment loaded into your Galaxy account.[2]
- Click on the 'pencil' icon next to the uploaded file, and in the 'Datatype' tab change the data type to 'fastqsanger'.[3]
- We want to map RNA read into our annotated genes. To do so we will use the program `Bowtie` installed in Galaxy. In the 'Search tools' field in the left panel type 'bowtie2' and click on the program.
- Select the `fastq` file you extracted from the database.
- In the 'Will you select a reference...' tab select 'Use one from history' and select the *gene* one.
- Click on 'Execute'. Your RNA reads will be mapped to the annotated genes.

[1] You downloaded this file from the BASys annotation page

[2] The `fastq` format is like the one you saw in the previous practical, but this time each sequence is a fragment of an RNA expressed in the cell.

[3] Some programs in Galaxy do not recognize `fastq` files unless they are defined as `fastqsanger` formatted files.

At the end of this process a `bam` file will be generated. A `bam` file stores the information of where the expressed sequence reads are mapped. In other words, how much transcript is each gene producing. This program takes a while, so for your convenience you will find the result file as well as other `bam` files stored in the Galaxy server. In the next step you will be comparing the expression profiles of the two samples of *Campylobacter*.

Task 2b: Comparing multiple samples

- In Galaxy go to the 'Shared Data->Data libraries' from the top menu. Select the 'campylobacter expression' folder.
- Select all four datasets (samples) and click on 'Go' to import to your current Galaxy history.
- Click in the 'Galaxy' icon (top-left) to return to the main window.
- In the 'search tool' field type 'BAM'. Click on the 'SAM/BAM To Counts' link. This program allows you to convert the (not particularly informative) BAM format into a table of read counts.
- In the 'Short names for samples' box type "sample_1,sample_2,sample_3,sample_4"
- Select 'BAM' format in the box below.
- Now, select 'sample_1' as a BAM file, click on 'Add new Additional BAM file'. Select 'sample_2', and so for until you have all four samples.
- 'Execute'
- When the process is done you will have a table with counts in your history tab. Click on the 'eye' icon and you will see the read counts per each gene in each of the four samples.[4]

Well done, you can now visualize how the genes you annotated are expressed from four different samples. Actually, samples 1 and 2 come from one laboratory, and samples 3 and 4 come from another. You suspect that expression levels may be different between these two laboratories as they may have analyzed different bacteria strains. To determine whether there are

[4] Remember for the lectures, the read count is the number of RNA molecules sequenced from the sample, and it's a measure of the expression level of the gene.

differences in gene expression, you know will do the (probably) most common Genomics/Bioinformatics type of analysis: differential gene expression. More specifically, you're about to detect which genes are higher expressed in samples from laboratory A with respect to samples from laboratory B.

Task 2c: Differential gene expression

- Find the 'De Seq' tool in Galaxy and open it.
- Select the counts file you just generated and in the 'Column Types in counts file' field type "labA,labA,labB,labB". That will indicate that the first two samples came from laboratory A and the other two from laboratory B.
- In 'Comparison type' write "labA,labB", which tells the program that you want to compare samples from labA with samples from labB.
- 'Execute'
- After a while two files will be generated. One of them, the 'Top Table' will list those genes for which the expression was different between the two datatsets.
- Click on the 'eye' icon next to the 'Top Table' file. 'Fold change' column is the expression level of sample A divided by the expression level of sample B. 'Padj' is the p-value adjusted for multiple testing.
- Genes overexpressed in labA samples will have a fold change greater than 2 (this is an arbitrary threshold). Also, we are interested in those with a significant change, which we now define as having a adjusted p-value below 0.01. Galaxy can help us to parse this list of genes.
- Find the 'Filter' tool and run it over the 'Top Table' output file. The condition to filter is "c5>2 and c8<0.01".[5] 'Execute'.
- Since you're only interested in the list of genes, find the 'cut' tool and in the 'cut columns' field type "c1".
- 'Execute' and open ('eye' icon) the resulting output file. This is the list of genes overexpressed in lab A samples.

[5] That means, genes with a fold change (column 5) greater than 2, and a adjusted p-value (column 8) lower than 0.01.

- Keep this window open as you will be analyzing some of this information in the next task.

Questions to be addressed in the final report

- How many reads did the RNAseq experiment produce?
- Find examples of highly expressed genes and lowly or no expressed genes.
- How many genes are upregulated in laboratory A samples?

2.3 Functional annotation of gene lists

To know the function of a gene is relatively easy. To understand the function of hundreds of genes at the same time, is not that easy. As we discuss in a lecture, there is a way of finding Gene Ontology annotation terms who are enriched for a particular list of genes. Next, you're about to find which functions are upregulated (genes are overexpressed) in the samples comming from laboratory A.

Task 3: Gene Ontology enrichment analysis

- Open a new web-browser window and go to `http://geneontology.org/page/go-enrichment-analysis`
- This tool allows you to quickly evaluate GO term enrichements. Paste in the 'Enrichment analysis' window the list of genes from the previous Task.
- Select 'Biological Process' and the species 'E. coli'.[6]
- 'Submit'
- Explore the results.

[6] We select *E. coli* because the gene names given during the annotation process were taken from this species.

Questions to be addressed in the final report

- Which GO term is the most significantly enriched?
- By looking at the top enriched GO term, can you conclude which metabolic process is different between the two samples?

Other databases and resources

In this tutorial you explored some databases such as ENA and Gene Ontology. However, there are many other that you may find useful in your future genomic investigations. Here are some of them:

- **GeneBank:** Genes and genome sequences. (`http://www.ncbi.nlm.nih.gov/genbank`)
- **GEO:** Gene expression data. (`http://www.ncbi.nlm.nih.gov/geo`)
- **OMIM:** Genes involved in Mendelian (monogenic) diseases. (`http://www.ncbi.nlm.nih.gov/omim`)
- **GeneCards:** Extensive information on human genes (`http://www.genecards.org/`)
- **Xfam:** This site directs you to Pfam, Rfam, Dfam... which are databases specific for proteins (Pfam), RNA (Rfam)...(`http://xfam.org/`)
- **BioGRID:** Interaction networks (`http://en.wikipedia.org/wiki/BioGRID`)
- **RepBase:** Database of transposable elements (`http://www.girinst.org/repbase/`)

2.4 Further reading

The tools you used are standard in bacterial genome annotation. For eukaryotic genomes, the same principles apply, but the programs are more sofisticated as they have to account for introns, alternative splicing, and more complex gene structures in general. You will find more information on Bacterial genome annotation in the following review paper:

- Richardson EJ and Watson M (2012) The automatic annotation of bacterial genomes. *Brief Bioinformatics* 14:1-12.

A superb introduction to Eukaryotic genome annotation can be found in:

- Yandell M and Ence D (2012) A beginner's guide to eukaryotic genome annotation. *Nature Review Genetics* 13:329.