Analysis of the transcriptional landscape in a knock-out parasite

Background

Plasmodium falciparum is the causative agent of the most dangerous form of malaria in humans. The reference genome for *P. falciparum strain 3D7* was determined and published about 15 years ago (Gardener et al., 2002). Since then, the genomes of several other species of *Plasmodium* that infect humans or animals have been elucidated.

Malaria is widespread in tropical and subtropical regions, including parts of Asia, Africa, and the Americas. Each year, there are approximately 350–500 million cases of malaria, killing more than one million people, the majority of whom are young children in sub-Saharan Africa.

As working within human is not feasible, mouse models are used like *Plasmodium berghei* or *Plasmodium chabaudi*. They can easily be maintained in the lab and mirror specific phenotypes as found in human malaria.

A colleague of yours knocked out a gene (**PBANKA_KO**) in the rodent malaria parasite ***Plasmodium berghei***. Over the last three months, she generated different biological replicates of the wild type (**WT**) and the knock-out (**KO**) to understand differences between both lines.

She needs to find out what the function of the gene is to finalise her grant application and she has a meeting with her boss to go through the **FINAL** results this afternoon!!!

Unfortunately, on the way to work, she broke both her hands cycling and cannot do the analysis!!!! Can you help her out?

Your job is to understand the implication of the knock-out of the gene.

- What influence does it have?
- What is the name of the knock-out gene?
- How did you determine those?

Use your knowledge from the RNA-seq module. It might be worth giving the tasks of genome mapping/visualisation and the transcriptome mapping/analysis to different team members so that you can run them at the same time. *Careful, some of the needed files might not be present...*

Useful sites:

- www.plasmodb.org - Gene information and GO enrichment (ask your instructor if/when you get to this point)
- www.genedb.org
- ftp://ftp.sanger.ac.uk/pub/project/pathogens/gff3/CURRENT/ - Reference genome and annotation (look for *Pberghei*)

Can you save your poor colleagues and save the grant...?

Required materials

To complete this project, we have created a `github` repository with some of the files you will need to get the data and complete the tasks. Thus, you first need to clone the repo in your VM. To do this, go to `~/course_data/` and type the following command:

```
git clone https://github.com/xibarrasoria/WTAC_NGS_Chile2021_projectRNAseq_1b.git
```

This will create a copy of the repository. Change into the directory that was created and have a look at the different files available.

Getting the data

The data generated by your colleague has been deposited in public repositories: NCBI's Gene Expression Omnibus and EBI's European Nucleotide Archive. Luckily, you have access to the information about which samples are needed to complete the work for the grant.

Sample name	Experimental condition	Replicate	GEO accession	ENA run
WT1	wild type	1	GSM2131951	SRR3437887
WT2	wild type	2	GSM2131957	SRR3437899
WT3	wild type	3	GSM2131963	SRR3437911
KO1	knockout	1	GSM2131953	SRR3437891
KO2	knockout	2	GSM2131959	SRR3437903
KO3	knockout	3	GSM2131965	SRR3437915

Inside the `data` directory, you can find:

- `samples.txt` has the URLs pointing to each of the FASTQ files.
- `download_raw_data.sh` is a bash script that reads the `samples.txt` file and uses `curl` to download each of the files.

To download the data, we would need to execute `download_raw_data.sh`.

Depending on your internet connection, this might take around an hour, since the files are fairly large. Instead, we have created *downsampled* files that contain only ~1.5M fragments per sample, which are a lot smaller. However, you have the scripts to download the original (complete) files, if you want to try analysing the complete data later.

For now, download the data from here: https://drive.google.com/file/d/1v95T3broSVxMbxuNQz5dQrrwgyaFxRiO/view?usp=sharing. Save the file in the `data` directory. Then use the `tar` command to extract all the files.

We also have access to the reference genome for *Plasmodium berghei*, a *gff3* file that contains the gene annotations, and the corresponding transcript sequences (note that the files are compressed).

Your colleague changed the gene_id of the knocked-out gene to **PBANKA_KO**, to make it easier to track it throughout the analyses (instead of having to remember the gene_id).

Now you are ready to start your analyses. Think about what steps are required to answer the questions set out above; how will you make sure that the data is of good quality and that the experiment has worked; and remember some files might be missing.

Happy data analysis!!

Written by: Victoria Offord, with modifications by Ximena Ibarra