# Installing R, RStudio and dartR

We will be using **R**, **RStudio** and **dartR** (Mijangos, Gruber*, et al.*, 2022) for data analyses. Follow the below tutorial to install them:

<https://github.com/green-striped-gecko/dartR/wiki/Installation-tutorial>

# Accessing and downloading platypus data

The platypus data that we are going to analyse is stored in ***Gadi***, Southern Hemisphere's fastest supercomputer. Gadi is run by the National Computational Infrastructure (***NCI***) Australia, the nation’s most highly integrated high-performance research computing environment.

## NCI sign up

To get access to ***Gadi***, you need first to open an ***NCI*** account using the link below.

<https://my.nci.org.au/mancini/signup/0>

There are 5 steps that you have to go through. In step 3, you will be asked for a project name, for this, type the word “platypus” in the search box and then select the two platypus projects that will be shown. Then finish the remaining steps.

After you submit this information, wait for Jaime to grant you permission to access the project.

## Connecting to Gadi

The easiest way to connect to ***Gadi*** is by using a file transfer client application. One of this kind of applications that is simple to use is ***Filezilla***. Use the below link to download it and then install it.

<https://filezilla-project.org>

To connect to ***Gadi***, open ***Filezilla*** and enter the following information:

1. Click on Menu > File > Site Manager
2. In the window that will appear click on the “New site” button. In the General tab, to the left of the window, choose from the drop menu the option: **SFTP - SSH File Transfer Protocol**.
3. In the ‘host’ field, enter: **gadi-dm.nci.org.au**
4. In the ‘User’ field, enter your NCI user name
5. In the ‘Password’ field, enter your NCI password
6. You can leave port blank (default = 22)
7. Click ‘Quick connect’

A small introduction about how to use ***Filezilla*** can be found in the attached pdf file “**Introduction\_to\_filezilla.pdf**”.

To access the platypus data type: ***/scratch/mo73/final\_platypus*** in the **Remote Site** box located in the right panel. Then drag this folder to your local computer (left panel).

## Description of the data

We have four different datasets.

The first dataset is whole genome sequencing data consisting of 26 platypus samples that were collected in the unregulated Tenterfield Creek (n=11) and below the dam (n=8), and above the dam (n=7) in the regulated Severn River (Mijangos, Bino*, et al.*, 2022).

The second dataset is SNP data genotyped using DArTseq (Mijangos, Bino*, et al.*, 2022). The dataset consist of 214 platypus samples (108 females, 106 males) collected from nine different rivers (five regulated by major dams and four unregulated) across four regions in south-east Australia. The name of the files from this dataset have the word “dart”.

The third dataset is whole genome sequencing data consisting of 57 platypus samples from across the whole species range in eastern mainland Australia and Tasmania (Martin et al., 2018). The name of the files from this dataset have the word “oxford”.

The fourth dataset is the same dataset of the 57 platypus as above, but using different methods for read mapping (NextGenMap; Sedlazeck et al., 2013), genotype calling (Octopus; Cooke et al., 2021) and mapped to the latest reference genome (Zhou *et al.*, 2021). The name of the files from this dataset have the word “Martin”.

Datasets and genomic resources have been split by chromosome and stored in subfolders within the “***final\_platypus***” folder. Each chromosome folder contains all the data files required for the analyses. There are six more folders within the ***final\_platypus*** folder:

***scripts*** folder - contains scripts used for read mapping, SNP calling, splitting files, remapping SNPs, loading data, validating data and first analyses. Pipeline steps are described in the document “**steps\_pipeline.docx**”.

***info*** folder - contains metadata information of the datasets, files for remapping and general information about the project.

***literature*** folder- contains reference articles, including the last two articles on platypus genomics (Martin *et al.*, 2018; Zhou *et al.*, 2021) and their supplementary information.

***original\_files*** folder - contains the original files of Oxford, DArT and GFF files.

***programs*** folder - contains some of the programs (for macOS) used in the scripts.

***reference\_genomes*** folder - contains the reference genomes (FASTA) used for the DArT, Oxford and 26 genomes datasets.

# Data analyses

## Checking interferon genes

First have a quick look at the reports of some interferon proteins in the platypus using the link below.

<https://www.uniprot.org/uniprotkb?query=%28taxonomy_id%3A9258%29%20ifn>

Can you identify in which chromosomes are located the interferon genes?

## Loading data into RStudio

First create a project in RStudio. Here is a small tutorial:

<https://bookdown.org/daniel_dauber_io/r4np_book/starting-your-r-projects.html#creating-an-r-project>

Then move the “final\_platypus” folder to your working directory in RStudio. Then open the script “analyses\_platypus.R” within the folder “scripts”. This script is a pipeline that loads a specific chromosome using the function from the script “load\_platy.R”, subsets loci based on information on the GFF file and does a couple of basic analyses.

For example to load the data for chromosome X1, you should type “X1” in line 22 of the script.

I used the MHC genes as an example for this script.

If you want have a go and run all the following lines.

dartR has a couple of functions that will help us to analyse genetic diversity in our datasets (gl.report.heterozygosity and gl.report.diversity). Have a look at their help files. One way to access these help files is by typing in the R console “?” and then the name of the function as shown below.

> ?gl.report.heterozygosity

> ?gl.report.diversity

# Learning resources

This link is an [R-refresher tutorial](http://georges.biomatix.org/storage/app/media/uploaded-files/Tutorial_1_dartR_RStudio_Refresher_22-Dec-21.pdf), [this one](http://georges.biomatix.org/storage/app/media/uploaded-files/tutorial3bdartrdatastructuresandinputfromsourcesotherthandartlmagv2-2.pdf) is a tutorial for reading genetic data into dartR and this is a [data manipulation tutorial](http://georges.biomatix.org/storage/app/media/uploaded-files/tutorial4dartrdatamanipulation22-dec-21-3.pdf).

The below link will take you to a folder with resources to learn R.

<https://unisyd-my.sharepoint.com/:f:/g/personal/jose_mijangosaraujo_sydney_edu_au/EjJB3KYKd75FrslR_htLuK8BXrM_vVGLR9rJ3uDXqNO2cw?email=agar8199%40uni.sydney.edu.au&e=EuLu8q>

The below link will take you to a folder with resources for scientific writing.

<https://unisyd-my.sharepoint.com/:f:/g/personal/jose_mijangosaraujo_sydney_edu_au/Ej9DLFvwVAhIvWc9n-MJJmwBDPIqr3pCkBuGYqhwpBWUHA?email=agar8199%40uni.sydney.edu.au&e=0OYTsK>

Bioinformatics glossary:

<http://www.discoveryandinnovation.com/bioinformatics/glossary.html>

# References

Cooke, D. P., Wedge, D. C.*, et al.* (2021). A unified haplotype-based method for accurate and comprehensive variant calling. *Nature Biotechnology, 39*(7), 885-892.

Martin, H. C., Batty, E. M.*, et al.* (2018). Insights into platypus population structure and history from whole-genome sequencing. *Molecular Biology and Evolution, 35*(5), 1238-1252.

Mijangos, J. L., Bino, G.*, et al.* (2022). Fragmentation by major dams and implications for the future viability of platypus populations. *Communications Biology, 5*(1), 1127.

Mijangos, J. L., Gruber, B.*, et al.* (2022). dartR v2: an accessible genetic analysis platform for conservation, ecology, and agriculture. *Methods in Ecology and Evolution, 13*, 2150–2158.

Sedlazeck, F. J., Rescheneder, P.*, et al.* (2013). NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics, 29*(21), 2790-2791.

Zhou, Y., Shearwin-Whyatt, L.*, et al.* (2021). Platypus and echidna genomes reveal mammalian biology and evolution. *Nature, 592*(7856), 756-762.