A pipeline to OzWheat joy

Emmett Leyne

www.github.com/EPLeyne/DS_SynthesisProject

AGRICULTURE AND FOOD
www.csiro.au

I am a Research Technician from Agriculture and Food in the Cereal Genomics Program. Before Data School I was primarily in the world of 'Data Production' in the molecular biology field working with DNA and RNA. I manage the High Throughput Genomics Facility and am Data Custodian for the OzWheat project. I was self taught in R and Python coding but could only do a small amount.

My Synthesis Project

The OzWheat Diversity Panel is a collection of 289 lines of wheat that are historically important in Australia. The OzWheat project aims to harness the big data of crop genomics and machine learning to predict yields. As part of the OzWheat project RNA is extracted from all samples across the panel from multiple sites and growing conditions, and sequenced. My Data School synthesis project was to automate the process of sequence analysis to knit the tools together in one easy pipeline. The goal was to be able to run the entire pipeline with three inputs: the raw data file location, the project name and a CSIRO ident. The raw data were the RNA sequence data of a small pilot trial conducted in 2016. The approach is to build a master batch file that will hold all the variables and call the separate scripts for the individual tools.

My Digital Toolbox

- Python
- HPC
- Shell/Bash

Favourite tool

Even though I haven't used it in the project my favourite tool is R and the Tidyverse. The syntax of R is much easier for me, although I find Python better for non-statistical purposes.

My time went …

Too fast, there were a few speed bumps along the way that really slowed me down. I learned that I like the parts that other's don't like data munging and error solving, but I can get fixated on those if I let it. The most difficult step was to create lists of files in certain formats for each tool that was used in the pipeline. I ended up having to re-write major parts of the working script to do it. A lot of mental energy went into deciding how to manage the data, especially once I recognised the size of the files that were outputted by the scripts. I would have also liked to have completed some data analysis.

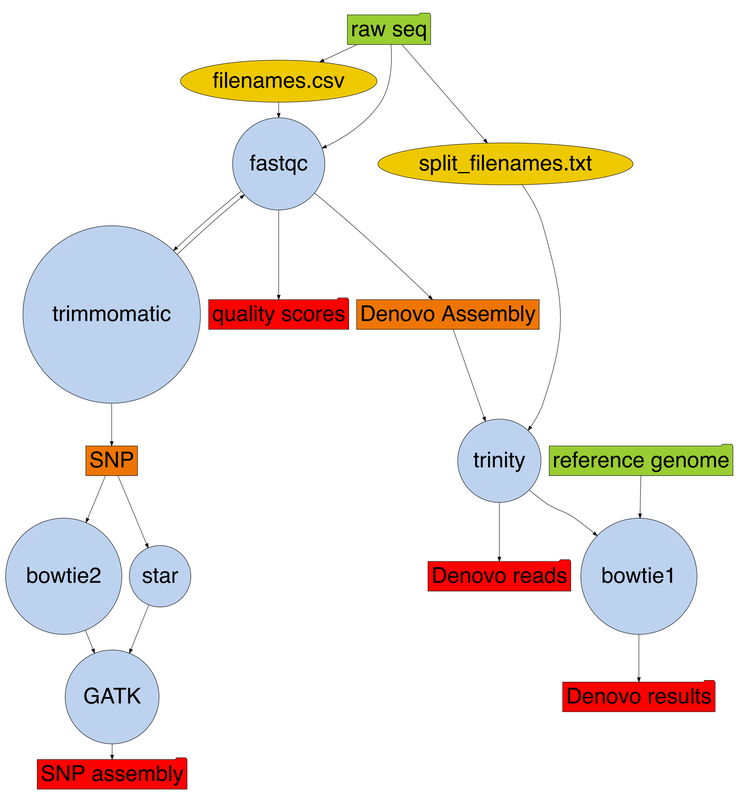
Next steps

- To complete the pipeline and integrate existing pipeline methods such as SnakeMate.
- Make the pipeline more flexible by giving researchers options of which tool they would like to use.
- Complete the Data Charter for the OzWheat project using the Tidy data lessons from Data School.

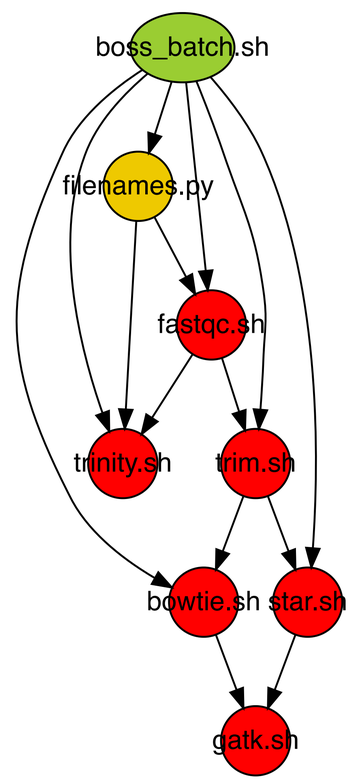
Pipeline

The pipeline of the sequence analysis with the green folders representing the inputs, the blue circles are the programs on the HPC used and the red folders are the outputs.

Workflow of OzWheat Pipeline



Connections from the master file to the tools



MY DATA SCHOOL EXPERIENCE

Before the Data School my data skills were based on Excel, Excel, Excel, and in all my projects all data management was based on the discretion of individual staff members. Data loccations, file names and variable names were an eccentric mix across all the people in a project.

Since Data School and in an external collaboration with NPI I managed the field trials in R instead of Excel as I would have previously. This decision has paid off many times as the external collaborators repeatedly asked for changes to be made or sent me new data for field trials that I was able to make the changes very quickly.

I have also been made the data custodian of the OzWheat project and have created a Data Charter from the lessons in Data School that defines all aspects of the data collection including file names and locations, standardised variable names, date formats, etc. This Charter has since been used by the AF Digital Coordinator as an example for other projects and can be directly attributable to the Data School.