Enabling Genomic Analysis to Improve Risk Characterisation in Australia's Red Meat Industry

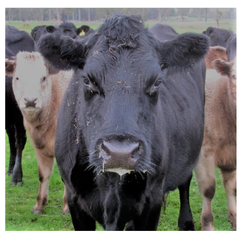
A reproducible analysis workflow using the CSIRO Galaxy and HPC services

Derek Benson, Tim Ho, Scott Chandry, Glen Mellor

Galaxy[1] is a biomedical platform that enables scientists to connect powerful computational analysis tools into pipelines which can be offloaded to high performance computing (HPC) systems. We applied Galaxy to a scientific problem important to Australia's red meat industry analysing hundreds of genomic samples simultaneously on the CSIRO's Pearcey HPC system.

A Significant Health Problem

While most strains of *Escherichia coli* (*E. coli*) are relatively harmless, Shiga toxin-producing *E. coli* (STEC) are a significant public health issue worldwide[2]. The development of bloody diarrhea as well as the severe and life-threatening disease in humans commonly referred to as haemolytic-uremic syndrome (HUS) depends on the carriage, expression and production of virulence genes that encode Shiga toxins (Stx). Classifying *E. coli* isolates by their virulence genes is proving to be an effective method of risk characterisation and a useful tool for estimating the human disease potential of STEC serotypes derived from red meat systems.



But how do we process and categorise a large number of sequenced isolates?

Like most prokaryotes, *E. coli* contains a single circular DNA molecule and in *E. coli* this is about 5.4 million bases long. We analysed 384 samples produced by the Illumina platform with each sample containing millions of reads, 150 bases in length. These short reads needed to be assembled into contigs prior to searching for virulence genes, a process that is computationally intensive and time consuming using a high-performance workstation.

To significantly increase the computational power available for sequence analysis, the CSIRO Galaxy service submits analyses to the CSIRO HPC systems and the Galaxy workflow shown in Figure 1 automates what is an enormous task and enables it to be completed with very little user interaction.

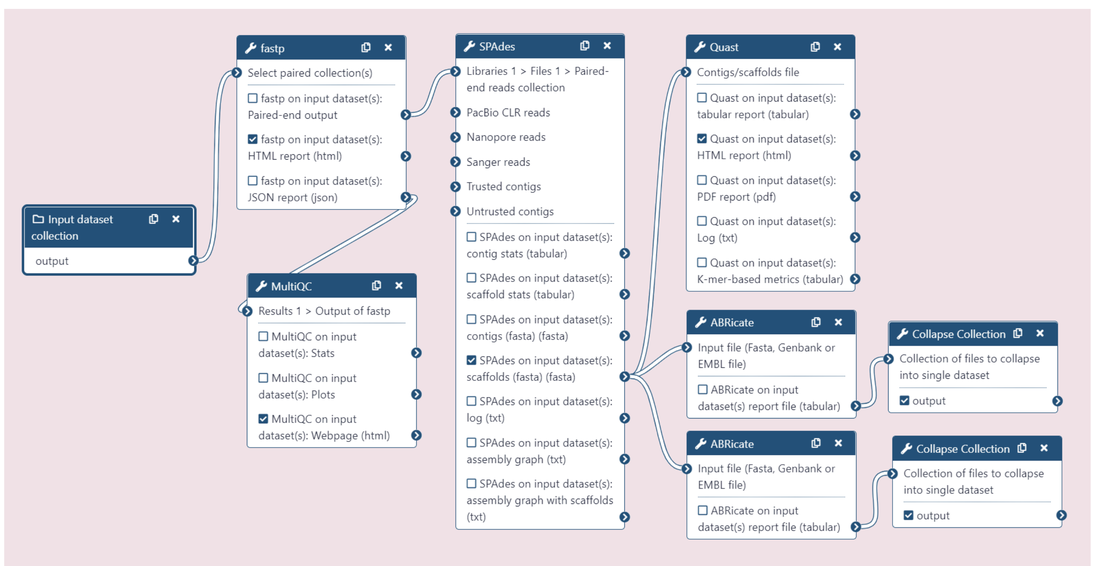


Figure 1: Galaxy workflow showing quality control and reporting using fastp and MultiQC, assembly of the bacterial genome and reporting with SPAdes and Quast, and searching for genes with Abricate[3].

High Performance Computing Resources and Storage Underpinning the analysis.

The Galaxy service makes extensive use of the HPC facility at CSIRO (Figure 2) to improve the speed of processing with HPC resources dynamically matched to the size of Files that needed input data and the tools being used.

- The HPC system has fast access to input datasets on CSIRO Bowen storage via an InfiniBand network.
- A lot of IO cycles were transferred to a high performing BeeGFS filesystem mounted across the cluster.
- Embarrassingly parallel steps were scaled horizontally across multiple CPUs.
- Capacity was such that all 384 paired end input samples were able to be processed simultaneously.
- The pipeline accurately detected virulence genes that are important for risk profiling Australian STEC into international defined risk groups.

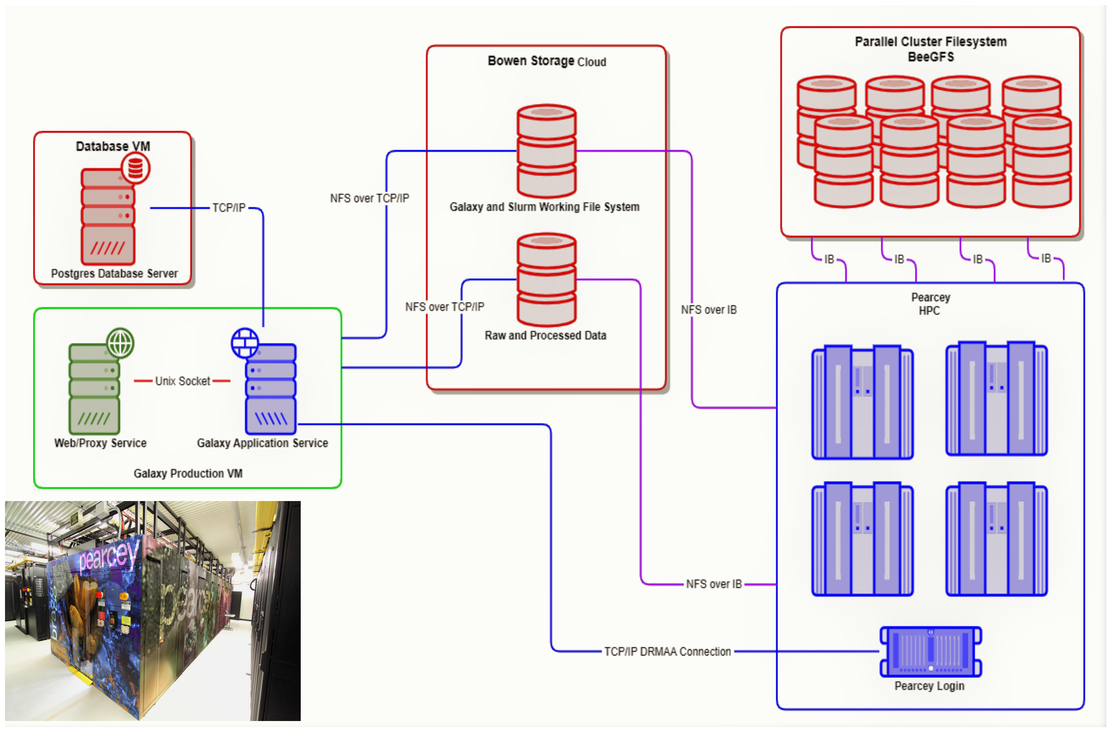


Figure 2: CSIRO Galaxy service and supporting HPC infrastructure including a shot of Pearcey in the bottom left..

Future Work

Parsing of reports to produce risk groups based on international risk typing schemes could be automated in Galaxy to provide a rapid and high throughput risk profile output for the Australian Red Meat industry. Likewise, the incorporation of relationship analysis tools into the pipeline would enable industry to rapidly assess the genetic relationships between cattle and human derived isolates thus providing an extra layer of information from which to risk profile STEC.

As Australia's national science agency and innovation catalyst, CSIRO is solving the greatest challenges through innovative science and technology.

CSIRO. Unlocking a better future for everyone.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge funding from Meat & Livestock Australia and the Commonwealth Scientific and Industrial Research Organization (CSIRO).

FOR FURTHER INFORMATION

Derek Benson
Information Management Technology
derek.benson@csiro.au
https://www.csiro.au/en/Research/Technology/Scientific-computing

REFERENCES

1, The Galaxy Platform - https://doi.org/10.1093/nar/gky379
2, Microbiological Risk Assessment Series 32, Report. FAO/WHO
3, Seemann T, Abricate, Github https://github.com/tseemann/abricate;VFDB - doi:10.1093/nar/gkv1239;Ecoli_VF - https://github.com/phac-nml/ecoli_vf;EcOH - doi:10.1099/mgen.0.000064