Costs of Running Production Genomics Workflows On-Premise as well as on Commercial Clouds

Garvan Institute of Medical Research

Summary

While the field of genome analysis is rapidly evolving, a number of tools have matured to become the de facto standard. The Garvan Institute has rapidly grown to become Australia's leading genome-powered medical research institute, and this has meant that we have moved production genome analysis workflows out of the hands of individual researchers. Instead, these workflows are built by engineers to be run at scale on-premise, on commercial clouds (Amazon Web Services and Microsoft Azure) and at the National Computational Infrastructure (NCI). Here we pose the following questions:

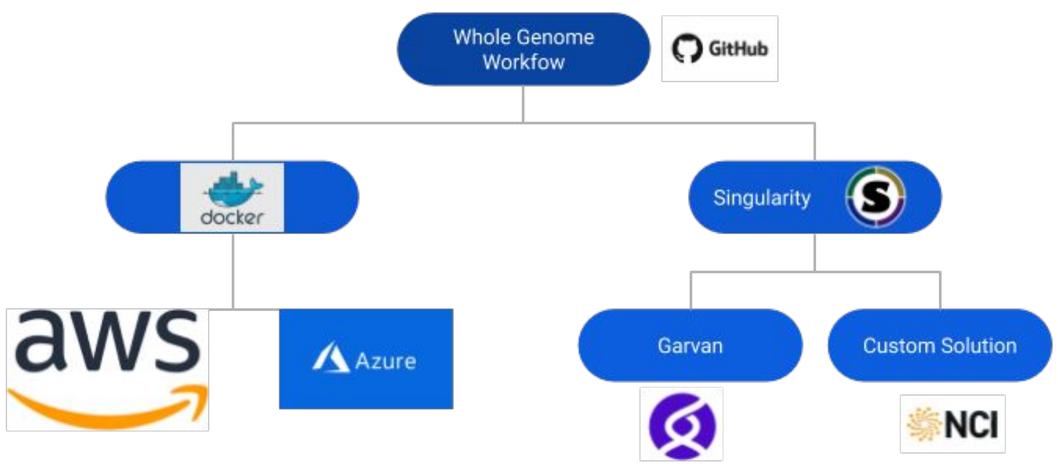
1. Since we built our 5000 core on-premise computational infrastructure using grants - what does it cost us to run these workflows at Garvan, and how do these costs compare when running the identical workflows on AWS, Azure and the NCI?
2. What are the underlying costs of running these analyses that include, power, maintenance, staffing etc?

Approach

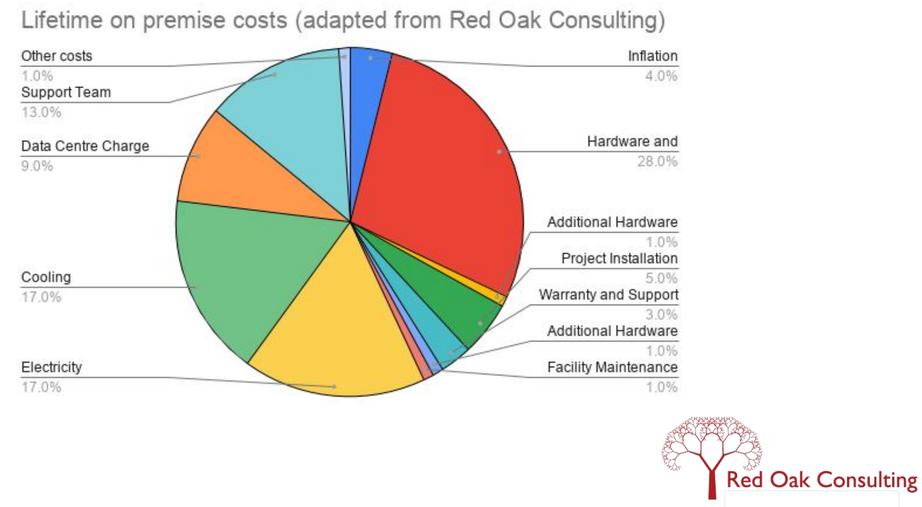
We looked to use a single source code to deploy our analysis on AWS, Azure and the NCI. The specific analysis that we used is our best-practice whole genome and exome workflows that essentially comprise the running of the tools BWA-MEM together with the Genome Analysis Toolkit (GATK). As we looked to understand the underlying costs of running these analyses we came up with a number of key questions:

- Capital versus Operational expenditure
- Data egress
- Central grant/funding agencies
- Budgeting - At Garvan budgets are allocated on an annual basis. If research computing costs (HPC and others) is to move into Opex, how do Garvan researchers secure long-term funding commitments to insulate themselves from the vagaries of the annual budgeting cycle?
- How can researchers more accurately predict their cost exposure?
- Cloud is a new provisioning and usage model for the researchers. There is now an increased (overwhelming) level of choice. What tools should be developed to help users decide on what services to use?

Garvan also had conversations with Red Oak Consulting, experts in HPC procurement in the UK in order to refine it's thinking in this area.



Findings, Analysis and Emerging Themes



Typical on-premise data centre costs. Costs identified by Red Oak Consulting (top) are contrasted with the Garvan Institute's on-premise solution (bottom).

Whole Genome Workflow	AWS Cost	Azure	NCI	Garvan
Wall time (hours)	21	22	12	11
CPU cost ($AUD)	$10	$14	$17	3 (estimated)
Egress	$30	$28	zero	zero
Total	$40-$45	$43-$46	$17	$3 (estimated)

Superficially, Garvan's low costs of computing genomes on premise sounds very encouraging. However when we contrasted the team's make-up (two engineers) with mostly non-redundant tasks it became increasingly clear to us that our services are particularly vulnerable. We also have become increasingly mindful that in the recruitment of engineers Garvan is up against the cloud providers, and other tech giants and financial services. In terms of remuneration, Garvan also cannot compete with these companies, and this fact alone may result in us (like colleagues around the world) not being able to support our own infrastructure. This will drive us towards running increasing numbers of workloads on the cloud.

Issues, Barriers and Project Expenditure

No issues or barriers were identified.

Project Expenditure

Engineer (1 FTE equivalent for 4 months)

Acknowledgement:

This research/project is supported by the Australian Research Data Commons (ARDC). The ARDC is enabled by NCRIS.

Australian Research Data Commons

Contact Details

Warren Kaplan (PhD), Head Data Sciences Platform

w.kaplan@garvan.org.au