QIMR Berghofer
Medical Research Institute

# Running Genomics Analyses in the Cloud

*John Pearson*
*HISA HIC, Melbourne*
*August 2019*

# Background

*Queensland Genomics:*

- $25 million initiative created by the Queensland Government in 2016
- Aim: to translate research genomics into healthcare practice
- The Genomic Information Management (GIM) capability: provide informatics infrastructure including genome analytics and a genome data repository

*Genomic analysis track record (2014-2019):*

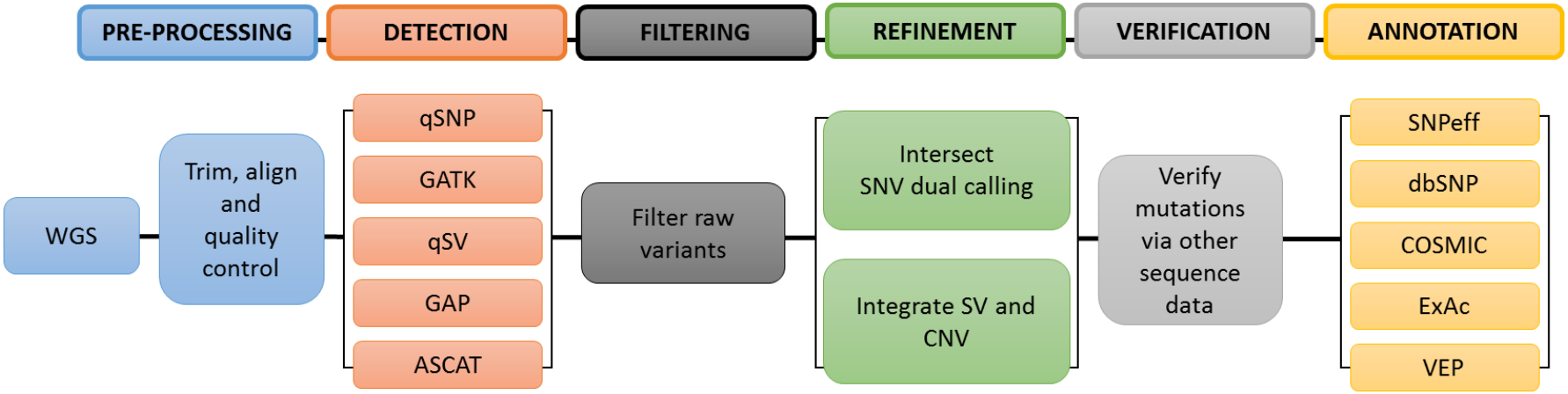
- 195 Studies
- 11403 Donors (patients)
- 31640 Samples
- 6109 Microarrays
- 28591 analysis-ready BAM files
- 100116 Analyses

Today we will outline some of the lessons learned by the GIM team as they deployed a cloud-based genomic analysis platform for Queensland Genomics.

# Our cancer somatic/germline NGS analysis pipeline:

- Approximately 30 pieces of software
- 200-500 jobs per tumour/normal pair
- Jobs have to be run in a particular order and files and data must be passed between jobs
- Workflows defined in WDL (Workflow Description Language)
- Cromwell WDL execution engine





# Genomics Data Sets are BIG:

A volume of the Encyclopedia Britannica:
500 double-sided pages, 8 million characters


x 1

8
megabytes

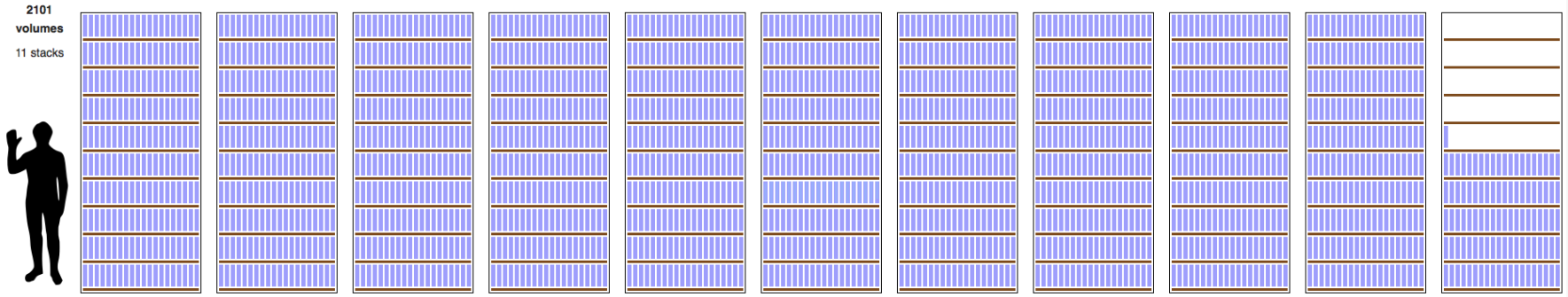
# Genomics Data Sets are BIG:

A printed copy of the English version of Wikipedia: 4.7 million articles, 2.7 billion words.

x 2,100

17

gigabytes





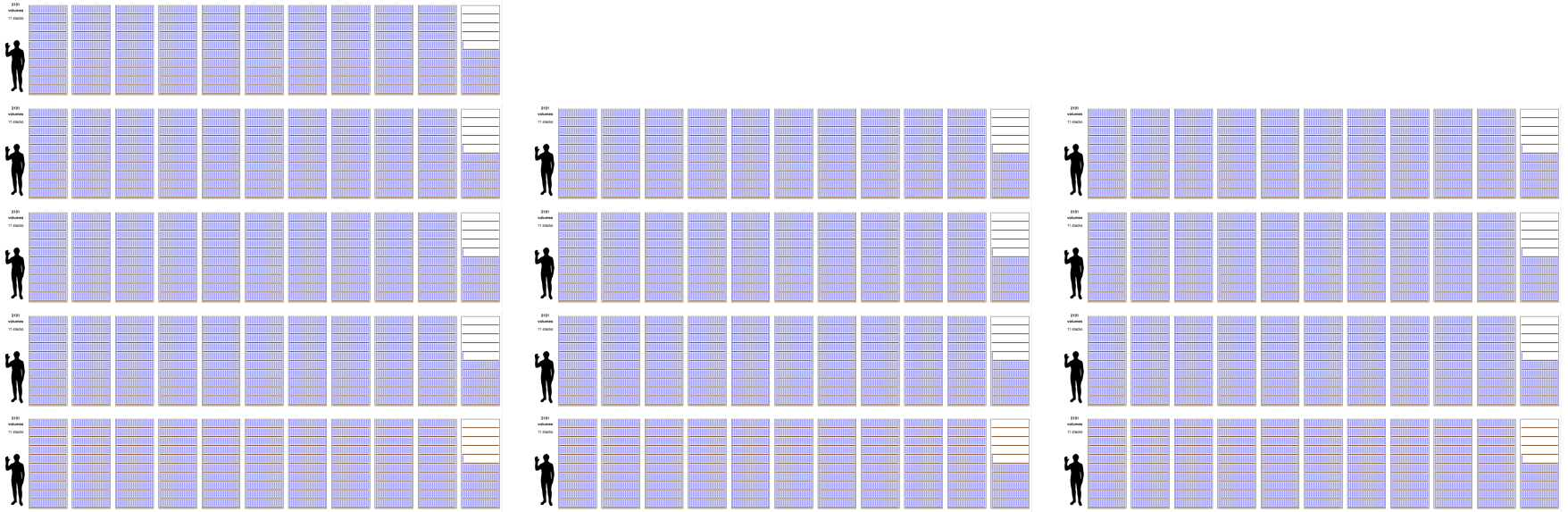
# Genomics Data Sets are BIG – one patient:

A single patient's cancer genome data: tumour sample @60x, normal sample @30x

x 37,500

300

gigabytes



# Lessons Learned - Topics

- Adapting to the new environment
- Interacting with cloud systems
- Storage
- Compute
- Data Transfer
- Security

# Lessons Learned - Adapting to the new environment

- Use of on-premise computing often appears to be “free” whereas you pay for every single service that you use on cloud
- Failed runs still have to be paid for so testing is critical
- Uploading data is usually free but download or transfer of data usually costs
- Purchasing on-premise computing is a capital expenditure whereas cloud computing is an operational expense.
- Cloud must be paid for monthly
- Bills can vary widely from month-to-month
- Planning can be difficult because billing occurs after the work is complete
- Cloud computing may not easily map onto existing tender and procurement processes

# Lessons Learned - Interacting with cloud systems

- Cloud vendors often provide multiple ways to interact with their system including
  - a web-based console
  - a set of command-line interface (CLI) utilities
  - one or more application programming interfaces (API's) that can be used to create custom software that can interact with and manage services on the cloud platform
- It is usually easiest to start using the web console
- It is usually smartest to move quickly to use of the CLI and API interface

# Lessons Learned - Storage

- Cloud and on-premise storage systems are often different
  - Long-term cloud storage systems are usually object storage
  - On-premise systems are usually block storage
- To adapt on-premise workloads for cloud:
  - Software will need to be re-written to work with cloud storage *OR*
  - Data files need to be copied from object cloud storage so that software can see it
- Output files that are to be kept need to be copied back to object storage
- Paying for cloud storage every single month often “clears the mind” and leads to workflows that only keep data that is truly required
- Cloud storage systems often have security, resilience, redundancy and encryption features built-in BUT these benefits only apply if cloud users understand, incorporate and implement them
- Cloud is no different than on-premise in that poor architecture design leads to poor outcomes

# Lessons Learned - Compute

- Cloud compute nodes are virtual and come in many configurations in terms of compute power, memory and input/output throughput – highly flexible
- You may be unable to find a node type that matches your on-premise compute
- More powerful nodes may be relatively more expensive than less powerful nodes so doing a job twice as fast may cost three times as much – you will have to do your own cost/benefit analysis
- Design workflows to fail early – it is better for a long process to do as much checking as possible and fail early (and cheaply) rather than fail while writing the final output file because you still pay for the failed workflow
- Consider containerised workflows – containers can let you run the same workflow on cloud and on-premise

# Lessons Learned – Data Transfer

- Data transfer is a critical piece of adopting cloud as it is the moment when the security context of the data changes - secure cloud **before** transferring data
- Data transfer can be a significant cost in time and money
  - uploads are usually free but downloads or transfers are usually expensive
  - download costs could be similar to the cost for storing data for 20 months
- For large data transfers in or out of cloud, consider using data transfer appliances if offered by your cloud vendor – often cheap and quick
- Cloud storage can make data transfer disappear entirely – one organisation can selectively open up their cloud storage to another organisation so data can be shared/analysed without any actual data transfer – no lost, delayed, misdirected, stolen data drives

# Lessons Learned – Security

- Challenges and principles for managing security of genomic data are shared between on-premise and cloud environments but the implementations differ
- We still need to control:
  - Access — traffic control; can you drive right up to my house from the road or only once I open my driveway gates?
  - Authentication — once you are at my house, are you who you say you are?
  - Authorisation — given you are who you say you are, what actions are you allowed to perform?
- Basic principles still apply:
  - Ask for, accept and store only the data you need - don't take on unnecessary privacy liabilities
  - Always use the principle of least privilege - give each class of operations staff the permissions they need to accomplish their work, and no more
  - Create an audit trail — use named accounts and log access to sensitive information – cloud usually has big capabilities in this area
  - Perform security audits and reviews regularly and preferably use an unrelated external organisation.

# Conclusions

- Cloud is not necessarily more or less complicated than on-premise compute but it is different
- Organisations will need to make some internal changes to support cloud use
- With cloud you pay for what you use so you really need to understand what you are doing because it is really necessary and what you are doing because you have always done it that way
- Cloud has a steep learning curve so start early, don’t try to rush an adoption at the last second

# Acknowledgements:

QIMR Berghofer Medical Research Institute

NHMRC

Queensland Genomics

The Ian Potter Foundation

Australian Genomics Health Alliance

Perpetual

