Galaxy
AUSTRALIA

Galaxy Australia – the Bring Your Own Data platform enabling multi-omics analysis for biology researchers

Dr Gareth Price, Service Manager of Galaxy Australia (Queensland Facility for Advanced Bioinformatics)

Melbourne Bioinformatics
BIOINFORMATICS + DATA SERVICES + INFRASTRUCTURE, FOR LIFE SCIENCES TODAY

QFAB qcif
BIOINFORMATICS | BIOSTATISTICS | BIODATA | BIO IT

ARDC
Australian Research Data Commons

BIOPLATFORMS
AUSTRALIA

2019
iAWARDS
Queensland iAward Winner

Biosciences: the nature of the Australian research community

30,000 health/biosciences researchers

18,000 health/biosciences HDR students

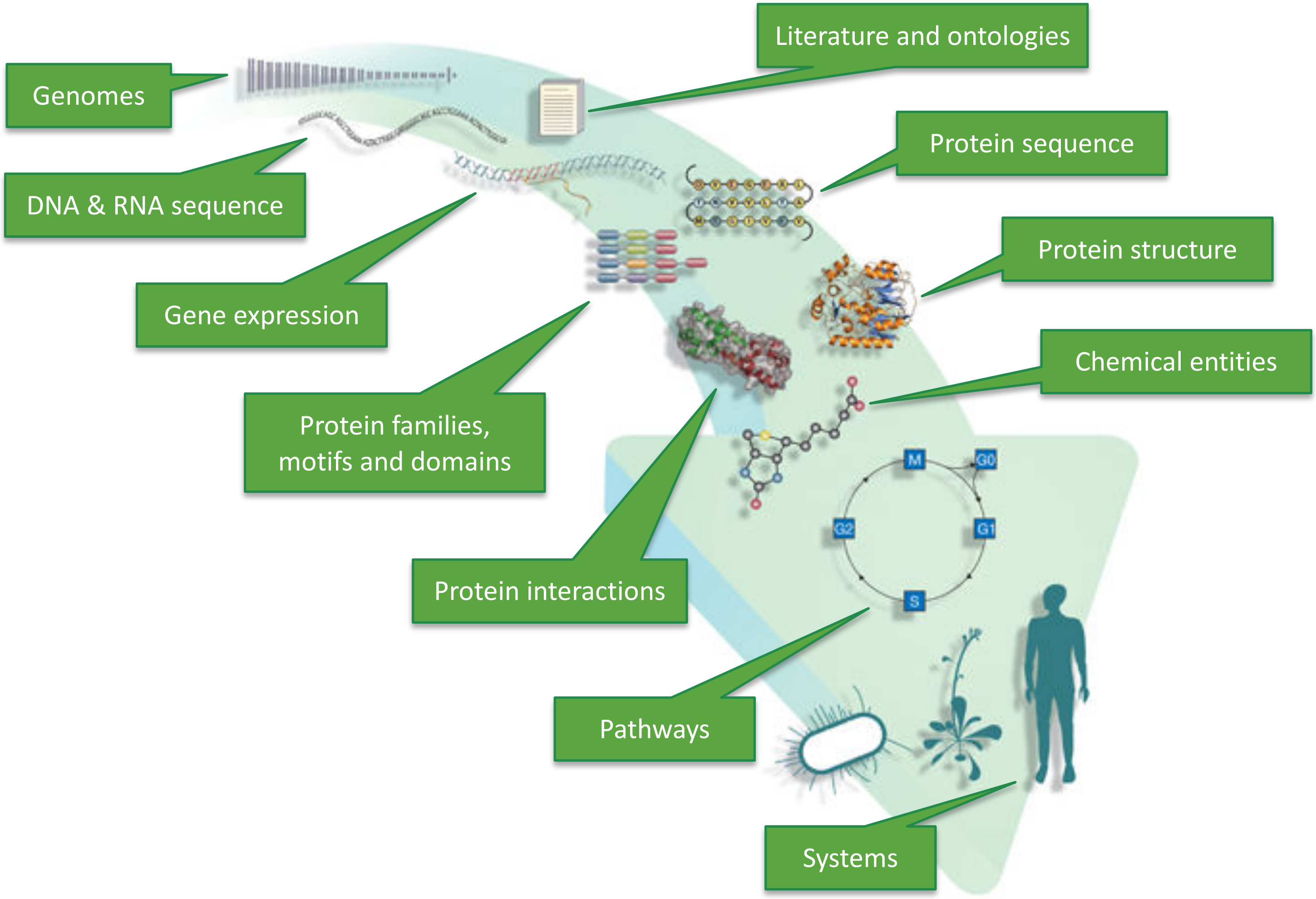
48,000 health/biosciences PG course work students

(163,000 + 40,000 =) 200,000 health/biosciences UG students

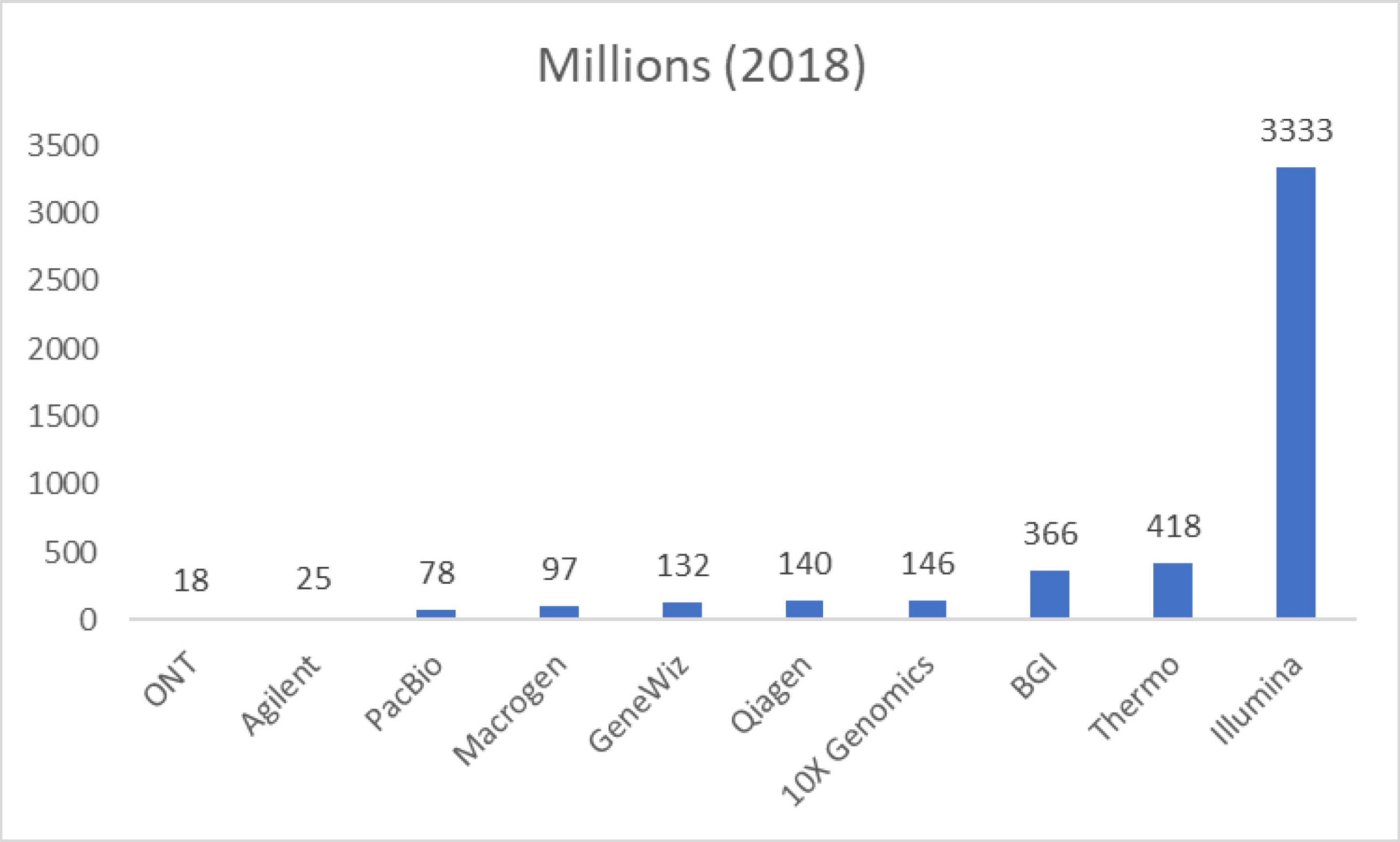
1,000 to 1,500 bioinformatician/computational biologists

=> the IT smarts need to be packaged up and delivered cleanly

From Genomes to Systems

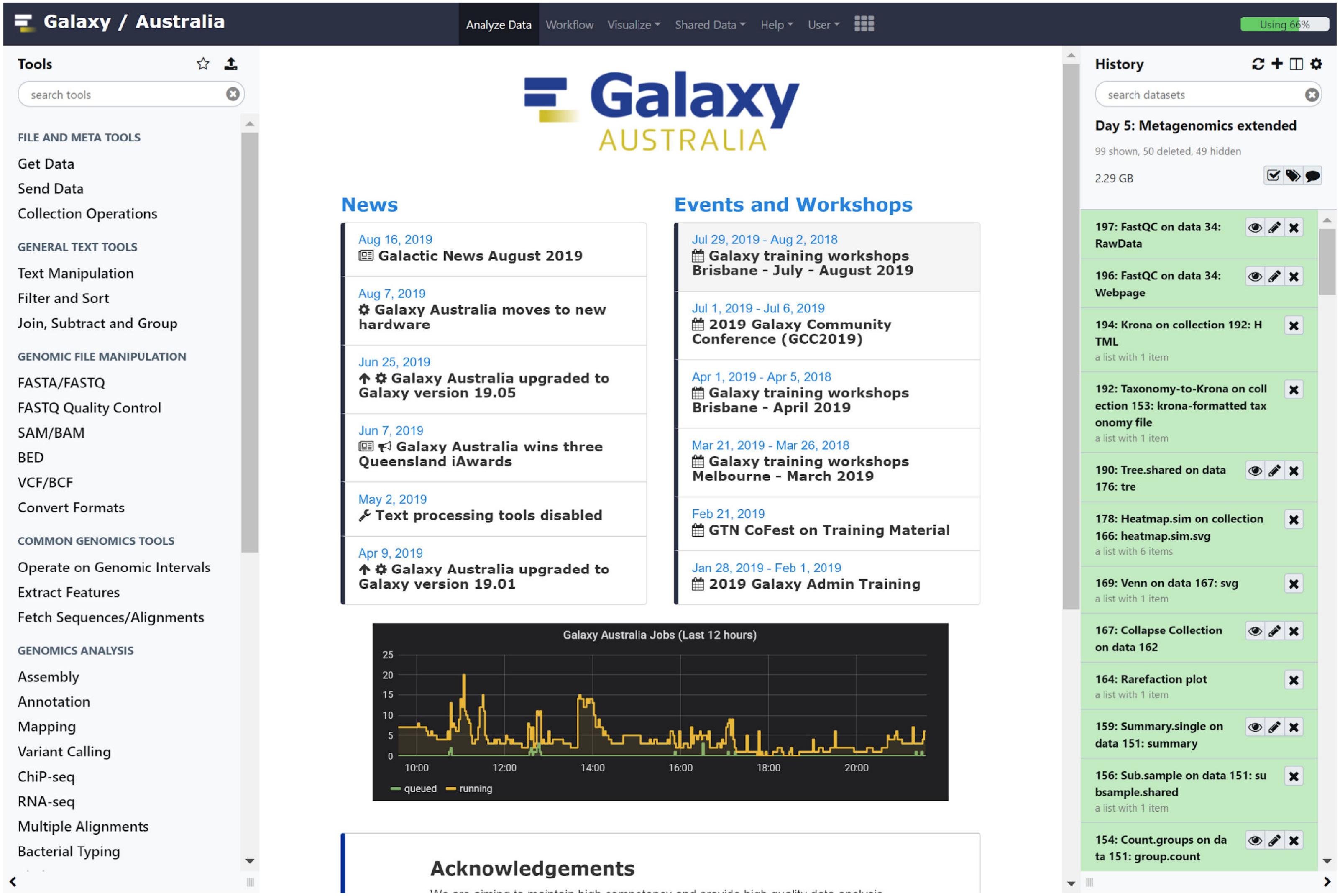


The Production landscape of genomics



https://www.genengnews.com/a-lists/top-10-sequencing-companies/

usegalaxy.org.au



FILE AND META TOOLS
- Get Data
- Send Data
- Collection Operations

GENERAL TEXT TOOLS
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group

GENOMIC FILE MANIPULATION
- FASTA/FASTQ
- FASTQ Quality Control
- SAM/BAM
- BED
- VCF/BCF
- Convert Formats

COMMON GENOMICS TOOLS
- Operate on Genomic Intervals
- Extract Features
- Fetch Sequences/Alignments

GENOMICS ANALYSIS
- Assembly
- Annotation
- Mapping
- Variant Calling
- ChiP-seq
- RNA-seq
- Multiple Alignments
- Bacterial Typing

News

Aug 16, 2019
Galactic News August 2019

Aug 7, 2019
Galaxy Australia moves to new hardware

Jun 25, 2019
Galaxy Australia upgraded to Galaxy version 19.05

Jun 7, 2019
Galaxy Australia wins three Queensland iAwards

May 2, 2019
Text processing tools disabled

Apr 9, 2019
Galaxy Australia upgraded to Galaxy version 19.01

Events and Workshops

Jul 29, 2019 - Aug 2, 2018
Galaxy training workshops Brisbane - July - August 2019

Jul 1, 2019 - Jul 6, 2019
2019 Galaxy Community Conference (GCC2019)

Apr 1, 2019 - Apr 5, 2018
Galaxy training workshops Brisbane - April 2019

Mar 21, 2019 - Mar 26, 2018
Galaxy training workshops Melbourne - March 2019

Feb 21, 2019
GTN CoFest on Training Material

Jan 28, 2019 - Feb 1, 2019
2019 Galaxy Admin Training

Acknowledgements

Day 5: Metagenomics extended

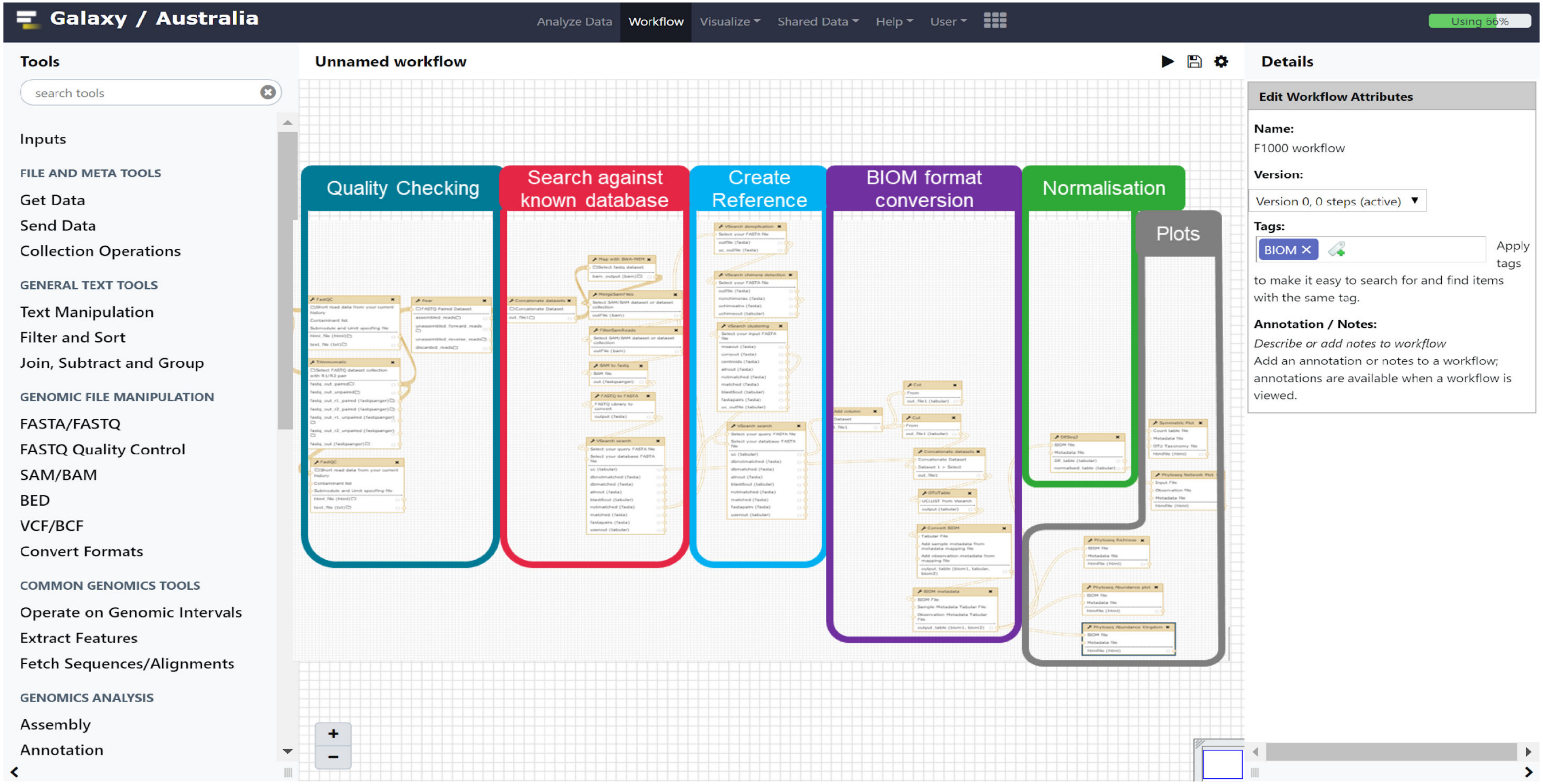
99 shown, 50 deleted, 49 hidden

2.29 GB

- 197: FastQC on data 34: RawData
- 196: FastQC on data 34: Webpage
- 194: Krona on collection 192: HTML
  a list with 1 item
- 192: Taxonomy-to-Krona on collection 153: krona-formatted taxonomy file
  a list with 1 item
- 190: Tree.shared on data 176: tre
- 178: Heatmap.sim on collection 166: heatmap.sim.svg
  a list with 6 items
- 169: Venn on data 167: svg
  a list with 1 item
- 167: Collapse Collection on data 162
- 164: Rarefaction plot
  a list with 1 item
- 159: Summary.single on data 151: summary
- 156: Sub.sample on data 151: subsample.shared
  a list with 1 item
- 154: Count.groups on data 151: group.count

Galaxy AUSTRALIA

Efficiency through Galaxy controlled scheduling



Name: F1000 workflow

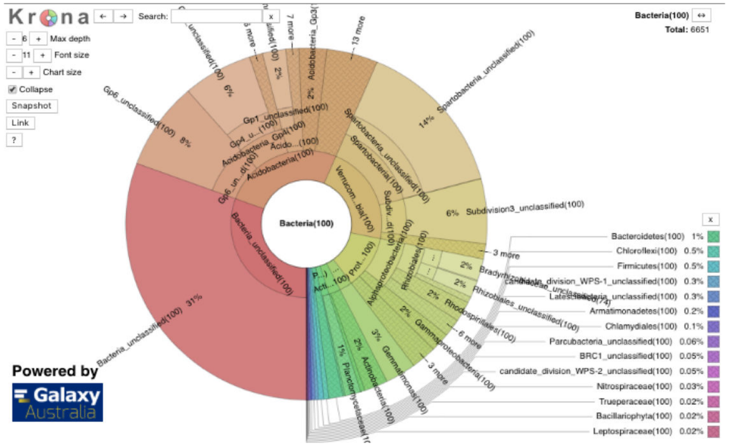
Version: Version 0, 0 steps (active)

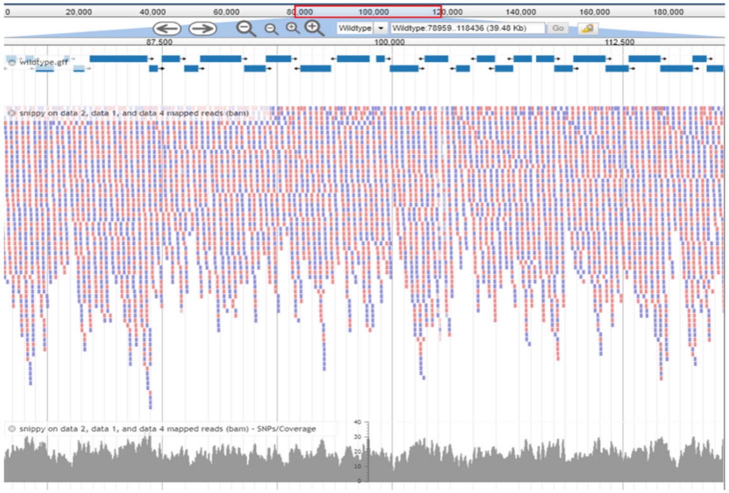
Tags: BIOM ✕ Apply tags to make it easy to search for and find items with the same tag.

Annotation / Notes: *Describe or add notes to workflow* Add an annotation or notes to a workflow; annotations are available when a workflow is viewed.

Taking the IT out of bioinformatics

- 946 tools supporting analysis applying more than 200 reference datasets
- simple data upload with optional rule based data management
- retains “histories” of analyses
- a large number of cutting edge tools
- an app store (over 7,000 tools)
- All aspect peer reviewable and transparent
  - Workflows
  - Citations / origins
  - Histories
  - Data

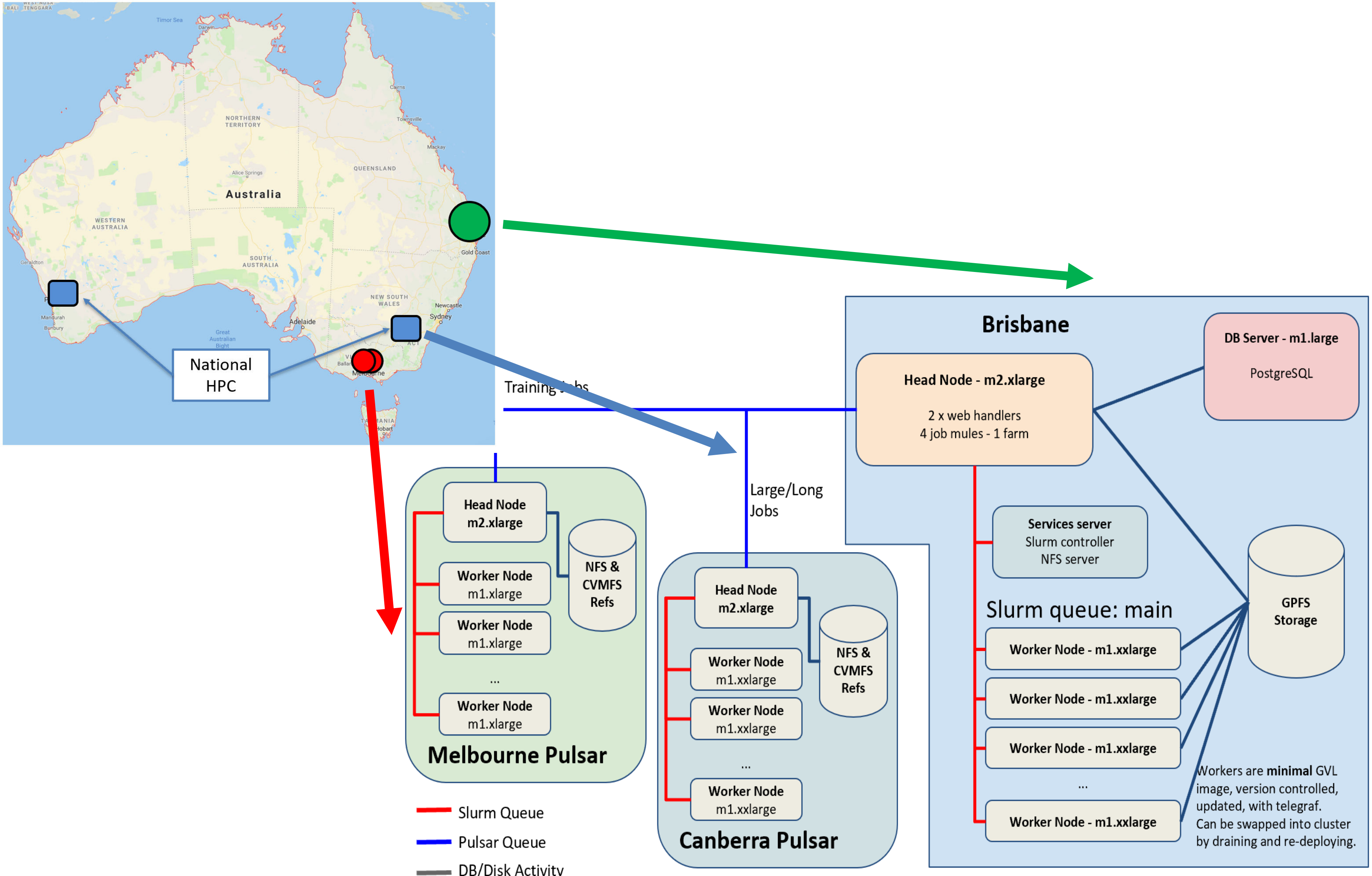



0
20,000
40,000
60,000
80,000
100,000
120,000
140,000
160,000
180,000
Wildtype
Wildtype:78959..118436 (39.48 Kb)
Go
87,500
100,000
112,500
wildtype.gff
snippy on data 2, data 1, and data 4 mapped reads (bam)
snippy on data 2, data 1, and data 4 mapped reads (bam) - SNPs/Coverage

A Brief History in Time

- Genomics Virtual Lab for public and private VLs, with Galaxy as a managed public service
  - Galaxy-Melb, Galaxy-Qld and Galaxy-Tut (Melbourne)
    - Tool version alignment / tutorial consequences
    - Duplicate staff roles
    - Tut was re-imaged over and over, no longevity to user accounts
- Rebrand Gal-Qld to Galaxy Australia and move to org.au
  - Staged migration of Melb to Aus
- Bring online 2 x Pulsar (repurpose Galaxy-Melb allocation and new allocation at NCI Canberra)
- Hybrid training events: 4 run in Galaxy with local facilitators and a lead trainer.
  - 11 of sites, 1064 of participants
- Policy, policy and policy – have a lot to thank Galaxy Europe for on this front

Galaxy Australia – distributed architecture



National HPC

Training Jobs

Large/Long Jobs

Brisbane

Head Node - m2.xlarge

2 x web handlers

4 job mules - 1 farm

DB Server - m1.large

PostgreSQL

Services server

Slurm controller

NFS server

Slurm queue: main

Worker Node - m1.xxlarge

Worker Node - m1.xxlarge

Worker Node - m1.xxlarge

...

Worker Node - m1.xxlarge

GPFS Storage

Workers are **minimal** GVL image, version controlled, updated, with telegraf.
Can be swapped into cluster by draining and re-deploying.

Melbourne Pulsar

Head Node m2.xlarge

Worker Node m1.xlarge

Worker Node m1.xlarge

...

Worker Node m1.xlarge

NFS & CVMFS Refs

Canberra Pulsar

Head Node m2.xlarge

Worker Node m1.xxlarge

Worker Node m1.xxlarge

...

Worker Node m1.xxlarge

NFS & CVMFS Refs

🟥 Slurm Queue

🟦 Pulsar Queue

⬛ DB/Disk Activity

Galaxy AUSTRALIA

Community Impact Summary

A complete bioinformatics analysis platform with:

>621k bioinformatics jobs run

946 bioinformatics tools

213 reference genomes

... and on average 2 new tools or references added weekly on user request

Supporting published research across many areas:

- Cardiac
- Toxicology and Drug Discovery
- Cancer
- Vision
- Agriculture and land management
- Obesity
- Ecology
- Infection and immunity
- Evolution

54 publications mentioned the GVL

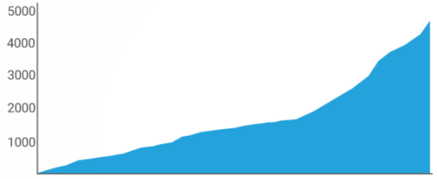
31 studies used the GVL in their publication

An active and engaged user community

4545 registered users.

724 active users (last 90 days)

User growth 2016 - 2019 Q2



Registered users in Australia from:

30 Australian Universities

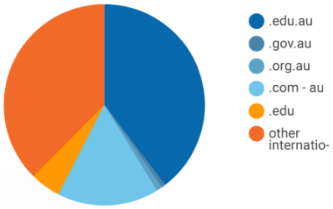
22 Medical Research Institutes or Organisations

19 Other Research Organisations

4 teaching courses utilising GVL annually

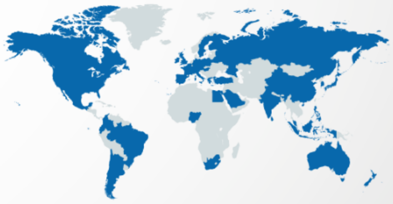
1064 people trained in hands-on workshops

Users per domain



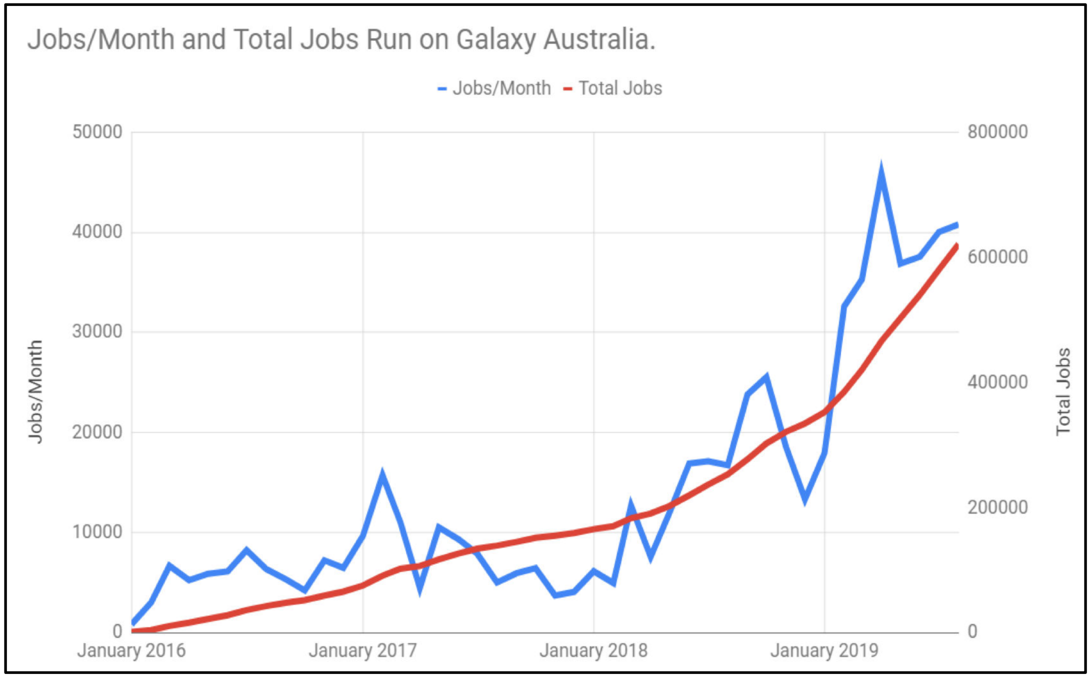
Users represented across 420 organisations

Users represented across 49 countries



Galaxy AUSTRALIA

Community Impact



"We have sequenced well over 500 transcriptomes and genomes, and routinely use Galaxy Australia for many bioinformatics processes.

It is easy to use, has high computational power, a sophisticated support structure and enables global collaboration through straightforward data sharing.

We greatly appreciate the service."

Dr Fabio Cortesi & Prof Justin Marshall, Queensland Brain Institute

Galaxy
AUSTRALIA

Growing our Australian Community

- Ongoing funding (for 4years) with a view to establish an enduring national research infrastructure
- adding to service functionality by adding metabolomics and phylogenetics
- greater security through institutional authentication, linking this to higher resourcing for authenticated users
- providing data sharing and data movement options through AARNET Cloudstor
- anticipating needs of other additional communities – esp. single cell and genome assembly and ultra long read technologies
- adding new resources: Galaxy slave servers (aka Pulsars) around the country

METABOLOMICS
AUSTRALIA

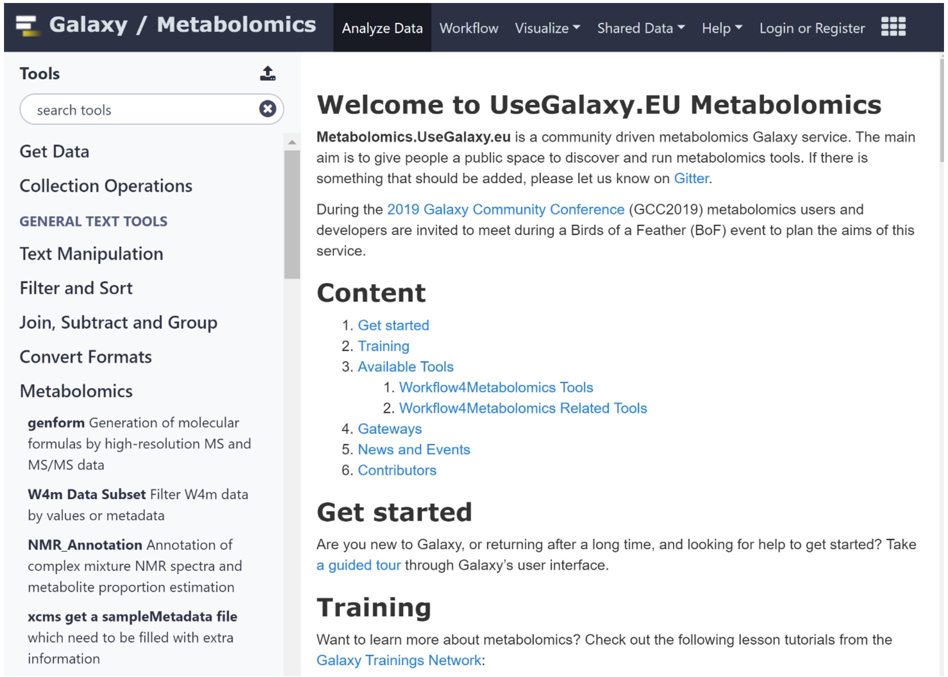
OMG
OZ MAMMALS GENOMICS

AUSTRALIAN
ACCESS FEDERATION

aarnet
Australia's Academic and Research Network

Growing our Australian Community – Metabolomics Australia

- Provide vendor agnostic tools and visualisations
- Leverage an increasingly mature global open source community



genform Generation of molecular formulas by high-resolution MS and MS/MS data

W4m Data Subset Filter W4m data by values or metadata

NMR_Annotation Annotation of complex mixture NMR spectra and metabolite proportion estimation

xcms get a sampleMetadata file which need to be filled with extra information

Welcome to UseGalaxy.EU Metabolomics

Metabolomics.UseGalaxy.eu is a community driven metabolomics Galaxy service. The main aim is to give people a public space to discover and run metabolomics tools. If there is something that should be added, please let us know on Gitter.

During the 2019 Galaxy Community Conference (GCC2019) metabolomics users and developers are invited to meet during a Birds of a Feather (BoF) event to plan the aims of this service.

Content

1. Get started
2. Training
3. Available Tools
    1. Workflow4Metabolomics Tools
    2. Workflow4Metabolomics Related Tools
4. Gateways
5. News and Events
6. Contributors

Get started

Are you new to Galaxy, or returning after a long time, and looking for help to get started? Take a guided tour through Galaxy's user interface.

Training

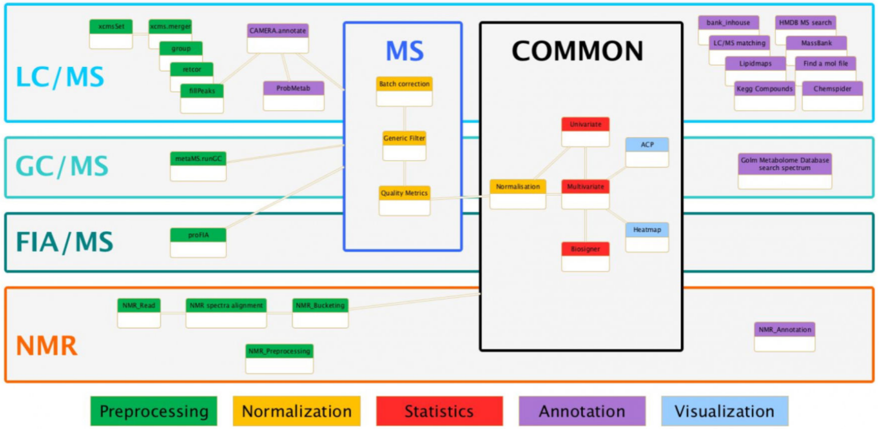
Want to learn more about metabolomics? Check out the following lesson tutorials from the Galaxy Trainings Network:

STEP 1
Request an account

STEP 2
Connect to Galaxy

STEP 3
Upload your data

STEP 4
Make reproducible science

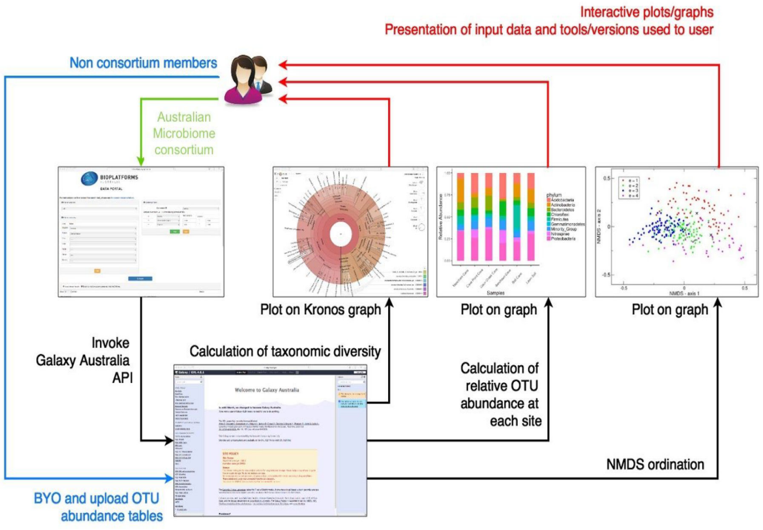


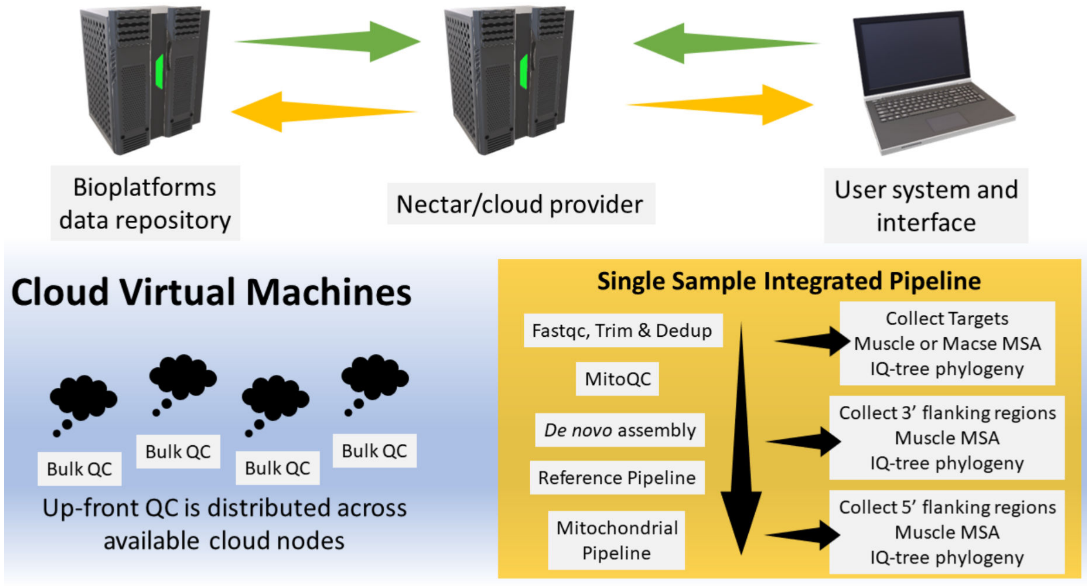
Galaxy AUSTRALIA

OMG OZ MAMMALS GENOMICS

Growing our Australian Community – Oz Mammal Genomes and Target Capture Panel pipelines

- Build on 2018 experience with BPA Data Portal Australian Microbiome
- User interact at BPA Data Portal (and potentially never leave!) to trigger pipeline on Galaxy Australia
- Analysed data returned to Data Portal and made publicly visible


Interactive plots/graphs
Presentation of input data and tools/versions used to user
Non consortium members
Australian Microbiome consortium
BIOPLATFORMS
DATA PORTAL
Plot on Kronos graph
Plot on graph
Plot on graph
Invoke Galaxy Australia API
Calculation of taxonomic diversity
Calculation of relative OTU abundance at each site
NMDS ordination
Welcome to Galaxy Australia
BYO and upload OTU abundance tables



Growing our Australian Community – authentication and storage links

- Galaxy Project is an open collaborative global project
- Users on public Galaxy services rarely are required to authenticate
- But with authentication comes opportunity for tailored or seamless servicing

- Link Galaxy to AAF
  - Link to institutional HPC destinations

- Link Galaxy to AARNET Cloudstor
  - Data Input
  - Data Export
  - Service provider mediated Input

AUSTRALIAN ACCESS FEDERATION

aarnet
Australia's Academic and Research Network

usegalaxy.* - a global platform and support network

Distributed reference data between .* servers

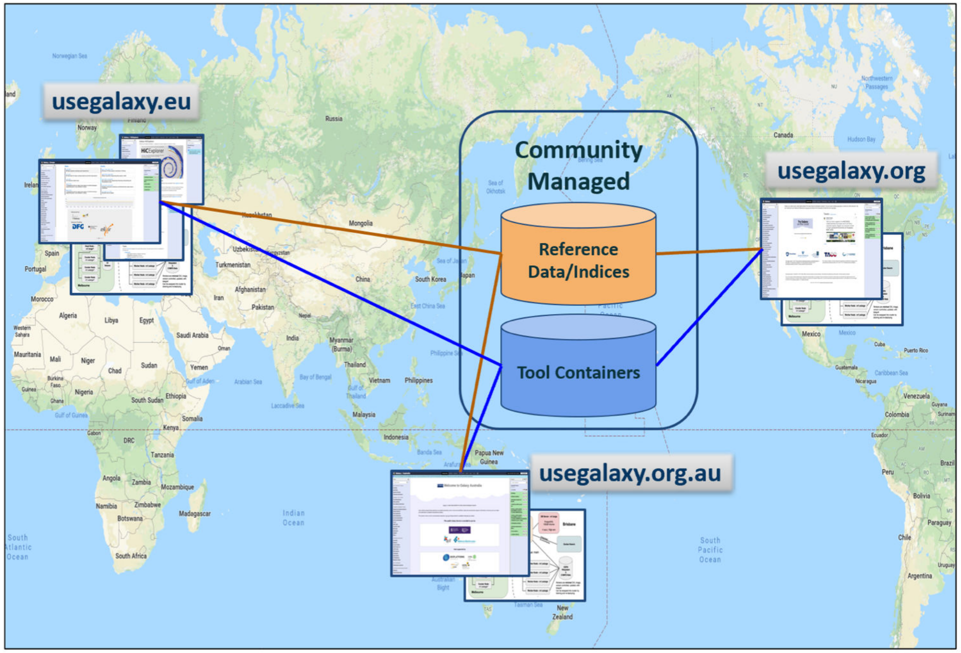
- reduced System Administration per locale
- Australian contribution to global efforts
- users are not restricted to “local” content

Intergalactic Data Commission

- formed in 2018
- regulation, automation and documentation of the CVMFS reference collections
- Australian representation on the IDC

Galaxy Project Executive Steering Committee

- formed in 2019
- Australian representation on the Committee



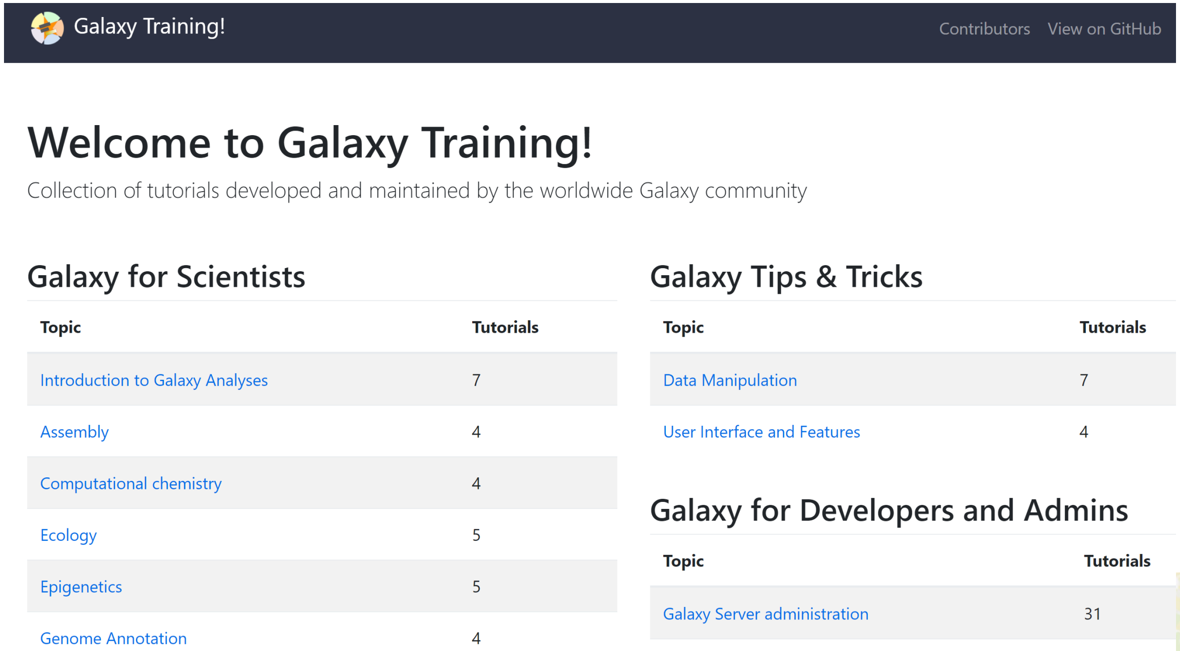
usegalaxy.* - a global platform and support network

Galaxy Training Network

- Australian content contributions
- Multi-language options
- Simple and comprehensive options
- Many peer reviewed best practices
- Synergised tool set

More usegalaxy.* services

- Growing number of countries and regions are forming their own usegalaxy.* service
- Increasing content development
- Distributing SysAdmin activity
- Thursday's keynote address: **The Development of ASEAN Federated Identity and Login Management, and Galaxy ASEAN Community**



Galaxy Training! Contributors View on GitHub

Welcome to Galaxy Training!

Collection of tutorials developed and maintained by the worldwide Galaxy community

Galaxy for Scientists

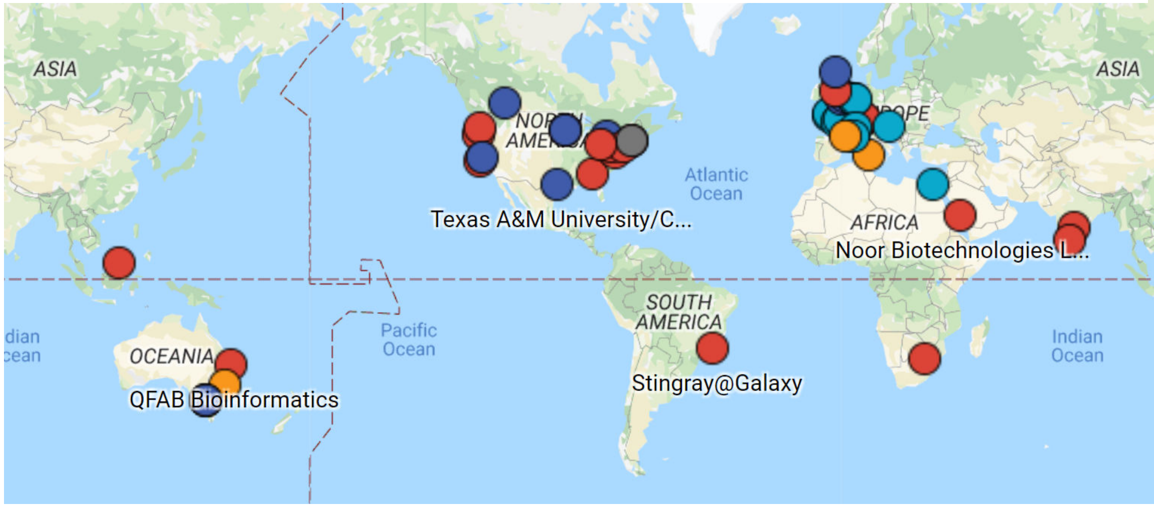
Topic	Tutorials
Introduction to Galaxy Analyses	7
Assembly	4
Computational chemistry	4
Ecology	5
Epigenetics	5
Genome Annotation	4

Galaxy Tips & Tricks

Topic	Tutorials
Data Manipulation	7
User Interface and Features	4

Galaxy for Developers and Admins

Topic	Tutorials
Galaxy Server administration	31



Computational Power

- Per patient sequencing data size is not changing, but the scope of analysis is/will:
  - WES and WGS
  - Short read for SNV
  - Long read of CNV and SV
  - Multi-omics (epigenome and transcriptome)
- Number of requests growing
- Choices are:
  - More computational power
  - Better algorithms
  - Appropriate use of computational architecture

Solutions for Data Analysis

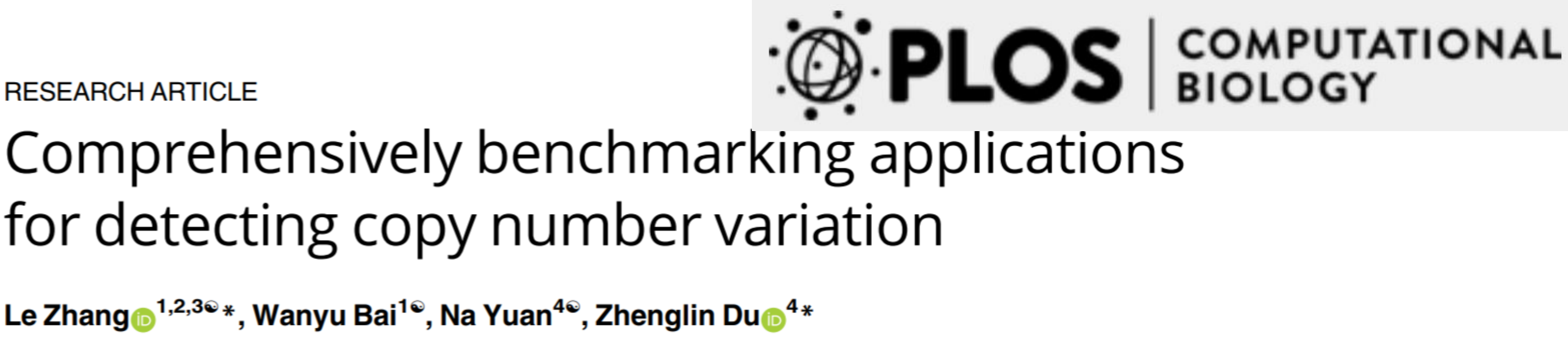
- **Freeware**
  - Galaxy
  - R Studio
  - Command Line / HPC

Most equivocal solution

Galaxy Australia user numbers (as of Sept 2018 - 2268) as CLC-Bio users is a difference of 750k funds (+750K in kind) vs approx. $11million annual licence fees

This does not include the cost of computers to support CLC-Bio installations

- **Commercial**
  - **Office**
    - Excel
  - **Agilent**
    - Cartagenia Bench Lab for Molecular Pathology
  - **Illumina**
    - BaseSpace
  - **Qiagen**
    - CLC-Bio Suite of Analysis Products
  - **ThermoFisher**
    - Ion Reporter

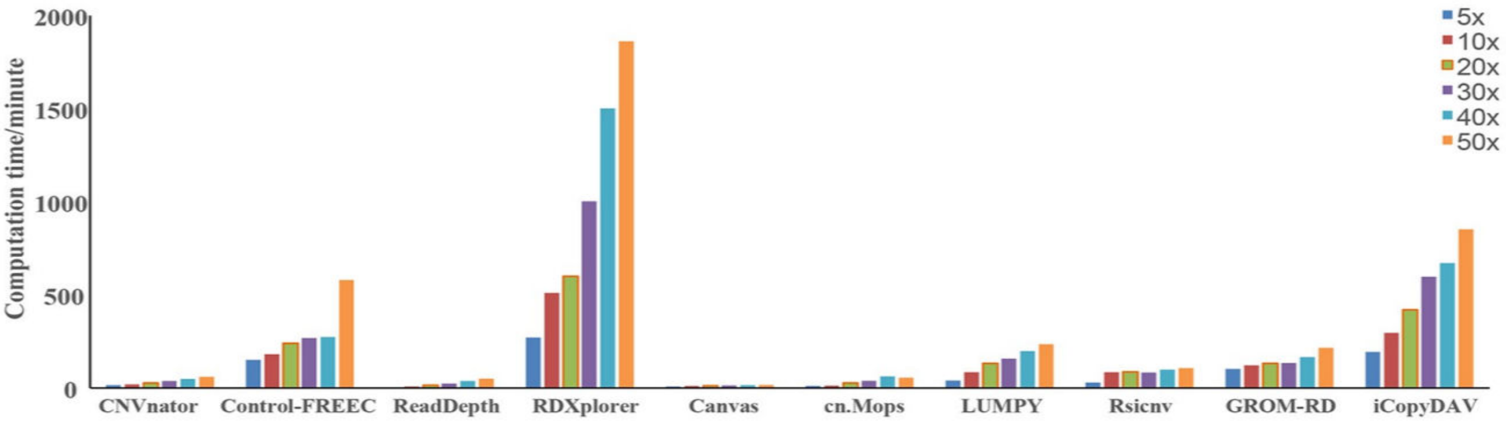


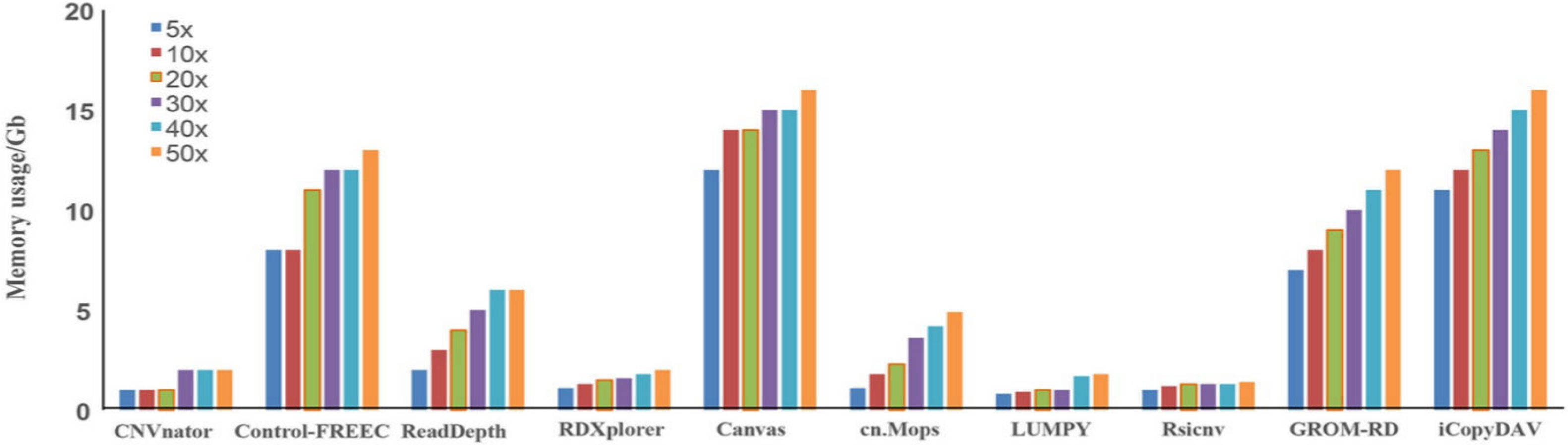
PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Comprehensively benchmarking applications for detecting copy number variation

Le Zhang[1,2,3]*, Wanyu Bai[1], Na Yuan[4], Zhenglin Du[4]*



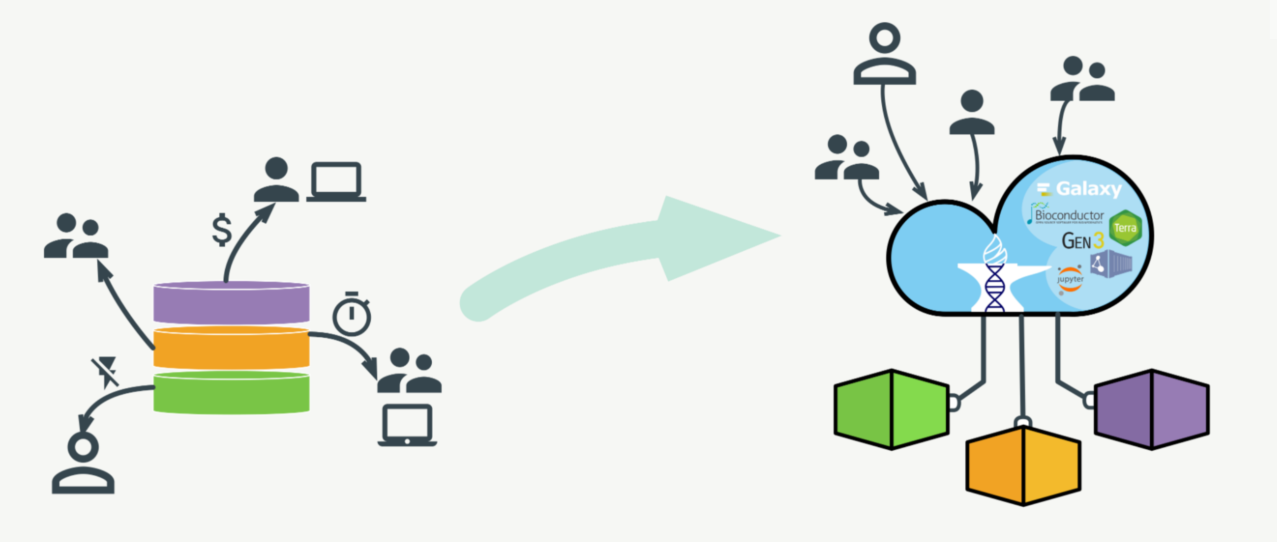


2019: https://doi.org/10.1371/journal.pcbi.1007069

Computational Power - more

- Cloud Life Sciences (formerly Google Genomics)
  - *Broad Institute replaced its in-house genome sequence analysis computers and storage with Google Cloud Platform, which delivers greater speed, scalability, and data security.*
- AWS Genomics
- Galaxy Australia

AnVIL



https://anvilproject.org/

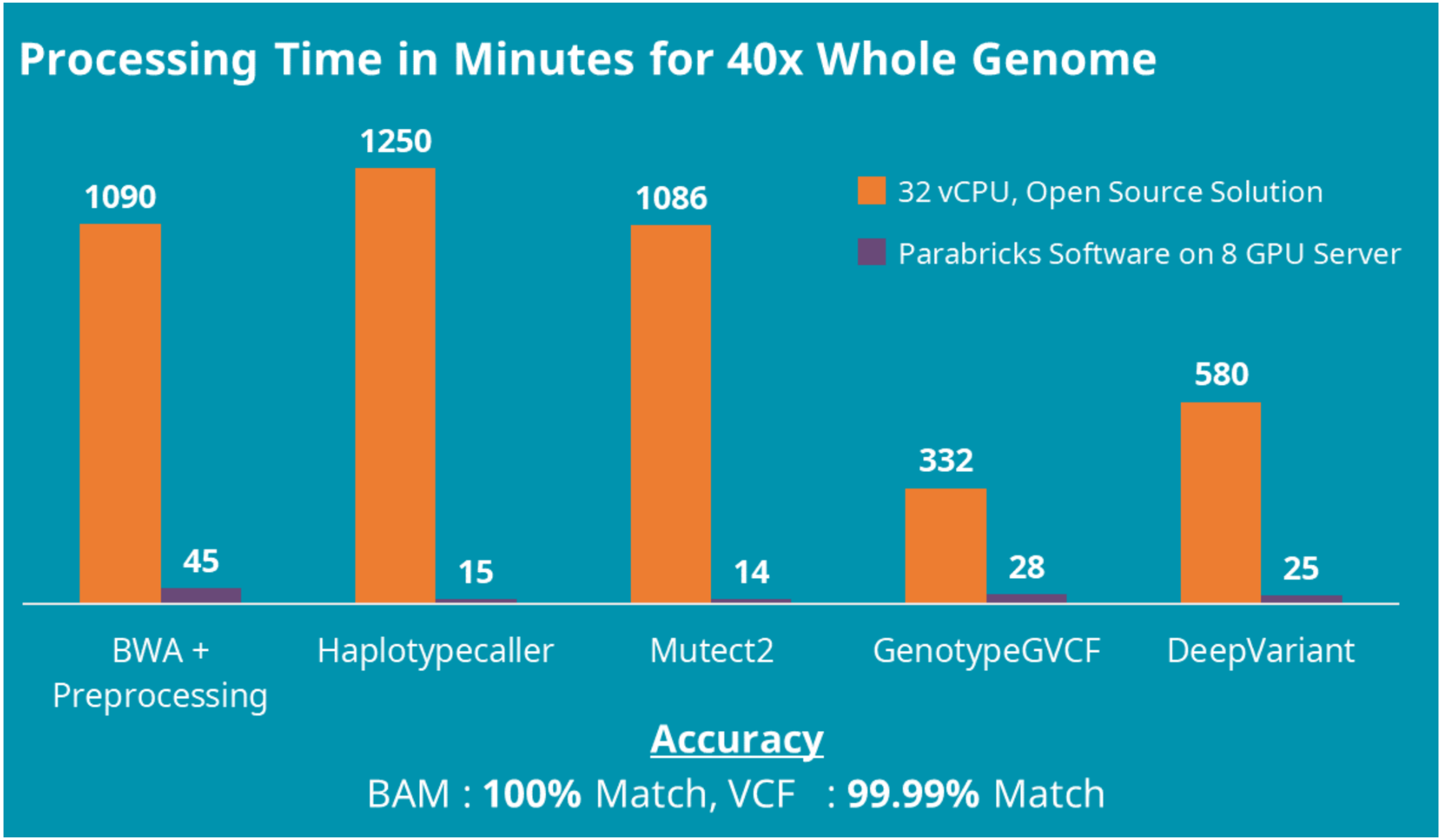
- Terra is an analysis platform that allows users to access data, run analysis tools, and collaborate, powered by Google Cloud Platform.
- [Gen3](Gen3) is a cloud-based software platform for managing, analyzing, harmonizing, and sharing large datasets.
- Dockstore is an open platform used by the GA4GH for sharing Docker-based tools described with the Common Workflow Language (CWL), the Workflow Description Language (WDL), or Nextflow (NFL).

Computational Power – better algorithms / architecture

- Analytical code can be deployed in a systematic way (Ansible, Docker, Singularity, Windows Updates!) but this makes assumption about the computer architecture

- Solution: Graphics processing unit (GPU) card or Field Programmable Gate Array (FPGA) card
  - Parabricks
  - Illumina / Dragen
  - ONT – NVIDIA

Parabricks



Processing Time in Minutes for 40x Whole Genome

Accuracy

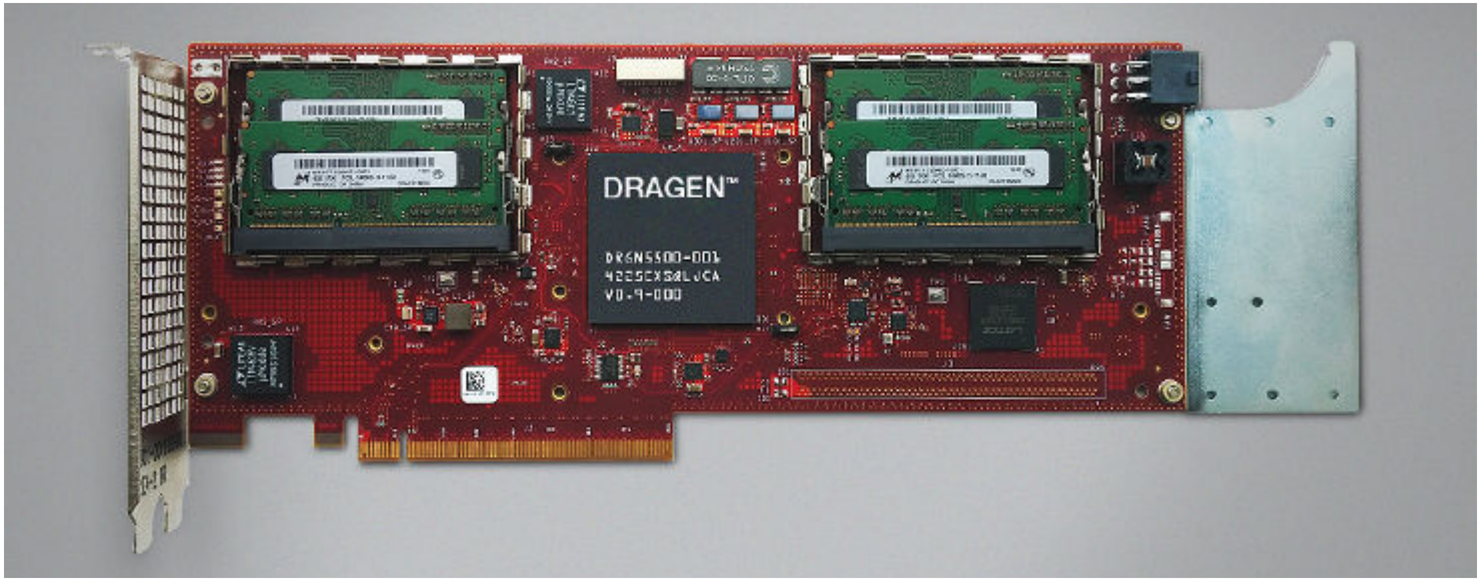
BAM : **100%** Match, VCF : **99.99%** Match

Illumina - Dragen

- The DRAGEN Platform can process NGS data for an entire human genome at 30× coverage in about 25 minutes on premise vs. > 15 hours with a traditional CPU-based system. It set two world speed records for genomic data analysis.

- The DRAGEN Platform can reduce on-premise investments in server clusters and utilization of cloud computing resources.

- Available on premises and on the cloud (Basespace, at AWS Sydney)

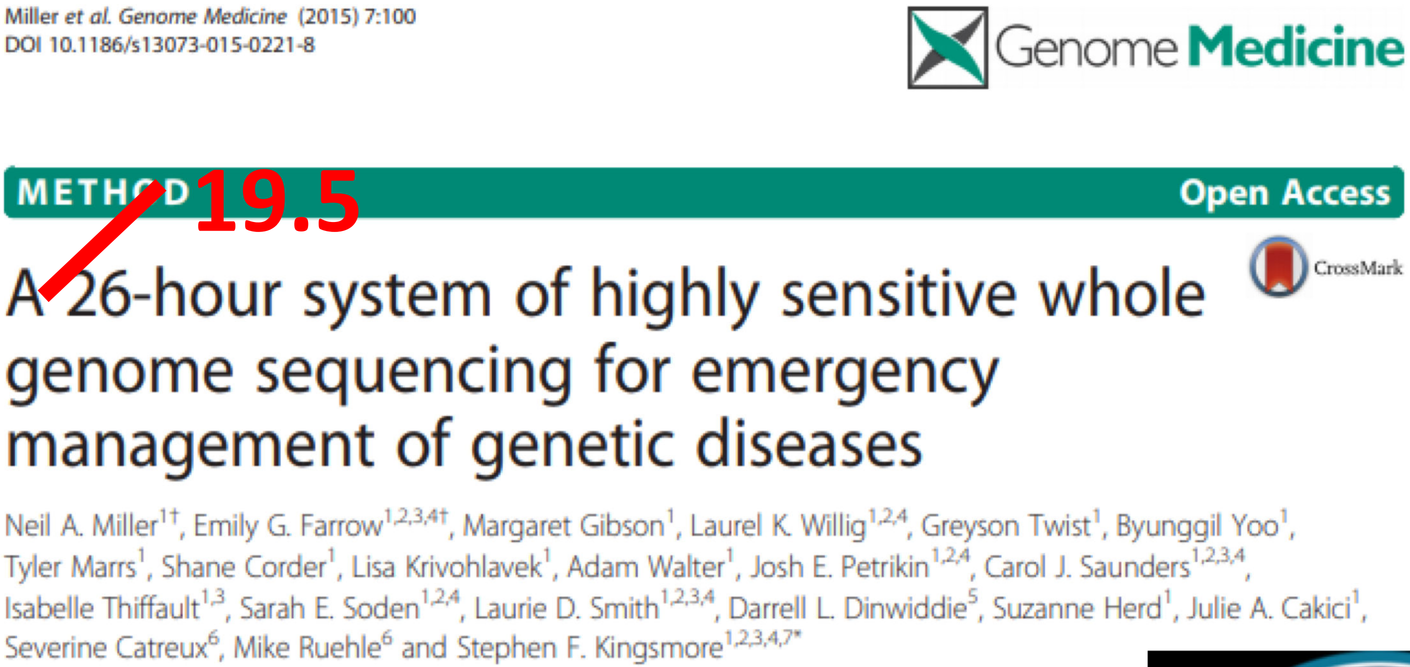


Oxford Nanopore Technologies - NVIDIA

- With MinIT on NVIDIA AGX, they’re approaching a 10x performance improvement over previous versions to help unlock real-time human and plant genomics. Its benchtop PromethION product is powered by NVIDIA Volta GPUs and can crank out a human genome for under $800.

- *Stay tuned for Illumina to do the same with Dragen (GATK)*

You know you've made it when..



Miller *et al. Genome Medicine* (2015) 7:100
DOI 10.1186/s13073-015-0221-8

Genome Medicine

METHOD 19.5 Open Access

CrossMark

A ~~26~~-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases

Neil A. Miller[1†], Emily G. Farrow[1,2,3,4†], Margaret Gibson[1], Laurel K. Willig[1,2,4], Greyson Twist[1], Byunggil Yoo[1], Tyler Marrs[1], Shane Corder[1], Lisa Krivohlavek[1], Adam Walter[1], Josh E. Petrikin[1,2,4], Carol J. Saunders[1,2,3,4], Isabelle Thiffault[1,3], Sarah E. Soden[1,2,4], Laurie D. Smith[1,2,3,4], Darrell L. Dinwiddie[5], Suzanne Herd[1], Julie A. Cakici[1], Severine Catreux[6], Mike Ruehle[6] and Stephen F. Kingsmore[1,2,3,4,7*]

- *Dr. Kingsmore receives the GUINNESS WORLD RECORDS™ certificate for the fastest genetic diagnosis.*
- **San Diego—Feb. 12, 2018**
- https://www.rchsd.org/about-us/newsroom/press-releases/new-guinness-world-records-title-set-for-fastest-genetic-diagnosis/

Galaxy Australia Team Members

Gareth Price

Igor Makunin

Thom Cuddihy

Nick Rhodes

Simon Gladman

Nuwan Goonasekera

Special Thanks:

Sarah Richmond, Ecoscience Research Cloud (ecocloud)

Derek Benson, CSIRO

Anna Syme, Royal Botanical Gardens, AU

Grahame Bowland, QCIF