

# An Australian open-source tool and workflow platform





BIOINFORMATICS | BIOSTATISTICS | BIODATA | BIOIT







Australian Research Data Commons



### 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

### A Brief History in Time, 1

- Galaxy-Melb (Melbourne, Victoria), Galaxy-Qld (Brisbane, Queensland) and Galaxy-Tut*orial* (Melbourne, Victoria)
  - Tool version alignment / tutorial consequences
  - Duplicate staff roles
  - Tut, re-imaging, no longevity to user accounts
- Rebrand Gal-Qld to Galaxy Australia and "agreed" 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. An "all-hands" event with SysAdmin constantly monitoring
  - 11 of sites, 1064 of participants

### A Brief History in Time, 2

- Enabled third party API triggered data send, user ID generation and workflow
- "Victim of our own success", more users, more requests, more jobs
- Reconfiguration of Pulsar to better handle small jobs (best attempt at aligning to training datasets)
- Policy, policy and policy have a lot to thank Galaxy Europe for on this front
  - Quota, retention time, API guarantees
  - Still working on a "who we are"
  - Still working on "when we will contact you"



The Australian Microbiome Initiative

#### Connecting Australian Data Repositories with Galaxy Australia



#### What we know from our users

- From training
- From user support
- From tool installation requests
- Workflows on "standard NGS"
- We need to service a lot of standard analysis





## Galaxy Solutions for Data

Analysis

#### Freeware

- Galaxy
- R Studio
- Command Line / HPC

#### Most equivalent 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

- Office
  - Excel

mercial

- Agilent
  - Cartagenia Bench Lab for Molecular Pathology
- Illumina
  - BaseSpace
- Qiagen
  - CLC-Bio Suite of Analysis Products
  - Ingenuity Pathway Analysis
  - Ingenuity Variant Analysis

#### ThermoFisher

Ion Reporter

#### State of Galaxy in Australia



Galaxy Australia

Galaxy Melbourne / Tutorial Now Pulsar



### **Community Impact**





### **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 and Prof Justin Marshall | Queensland Brain Institute





- adding to service functionality with planned (and funded) work into 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
- adding new resources: Galaxy slave servers (aka Pulsars) around the country
- leading the regional Galaxy community: launching at International Conference on Bioinformatics 2019, Jakarta, September 10-12









### Galaxy Australia Data Retention Policy – or our Fair Go policy

	Storage quota	Data retention period
Registered Australian researchers	600GB	90 days
Other registered users	100GB	90 days
Unregistered users	5GB	NA

User access defined by registration Email @domain.edu.au OR @domain.org.au OR @domain.edu - only in the case of known Australian Universities not on the .au domain OR all other @domains

Please contact us (help@genome.edu.au) if your institution does not conform to this rule but you understand it should be defined as performing publicly funded Australian research



"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 – future activity

- Take our well-valued GVL paper (Afgan *et al,* 2015) and value add locally with Galaxy Australia paper – currently under review at BMC Bioinformatics
- Balance between Operations and Innovation
- Grow in scope (tools and compute) to support long read and hybrid analysis
  - Pacbio RSII and Sequel
  - ONT MinION and PromethION
- Grow number of Pulsar services, to support generic and bespoke analyses (GPU Pulsar enabled with DRAGEN and / or Parabricks)

### Galaxy Australia – Pathfinder activity

- Enable new communities (Metabolomics, *single cell*), new functionality (phylo -genetics -genomics)
- Cost/benefit of providing Pulsar to GA or enable the local community
- Linking for all Australian researchers national data storage (Cloudstor) with Galaxy Australia
- Explore practically and policy level access to HPC resources in Australia
- Internationally: can a Galaxy Australasia or Galaxy Oceania exist?
  - INCOB2019 conference and associated Goblet meeting
  - Genomics Aoteroa (New Zealand)
  - Asia-Pacific Advanced Bioinformatics Network





#### usegalaxy.\* - a global platform and support network

#### Distributed reference data between .\* servers

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

#### Intergalactic Data Commission

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

#### **Galaxy Project Executive Steering Committee**

- O formed in 2019
- O Australian representation on the Committee



#### **Galaxy Australia Team Members**

Gareth Price

Simon Gladman\*

Igor Makunin

Nuwan Goonasekera \*

Thom Cuddihy

Nick Rhodes

Christina Hall

Helen van der Pol

Special Thanks:

Sarah Richmond, Ecoscience Research Cloud (ecocloud)

Derek Benson, CSIRO\*

Anna Syme, Royal Botanical Gardens, AU\*

Grahame Bowland, Centre for Comparative Genomics

Kylie McKenzie Carmel Akhavari

