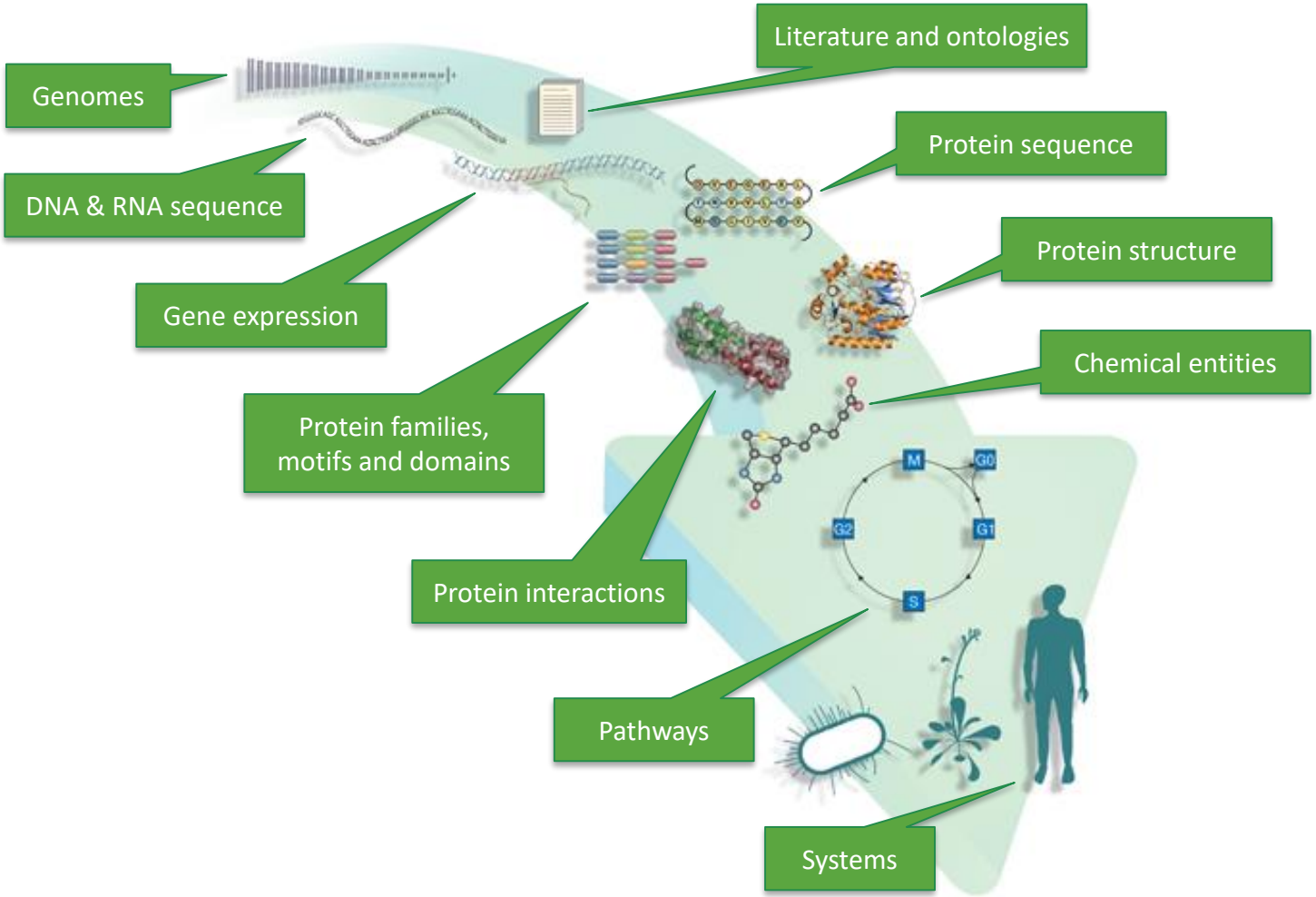
Enabling Australian Genomics research through enhancements to the Genomics Virtual Lab

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

Development partners

Supported by

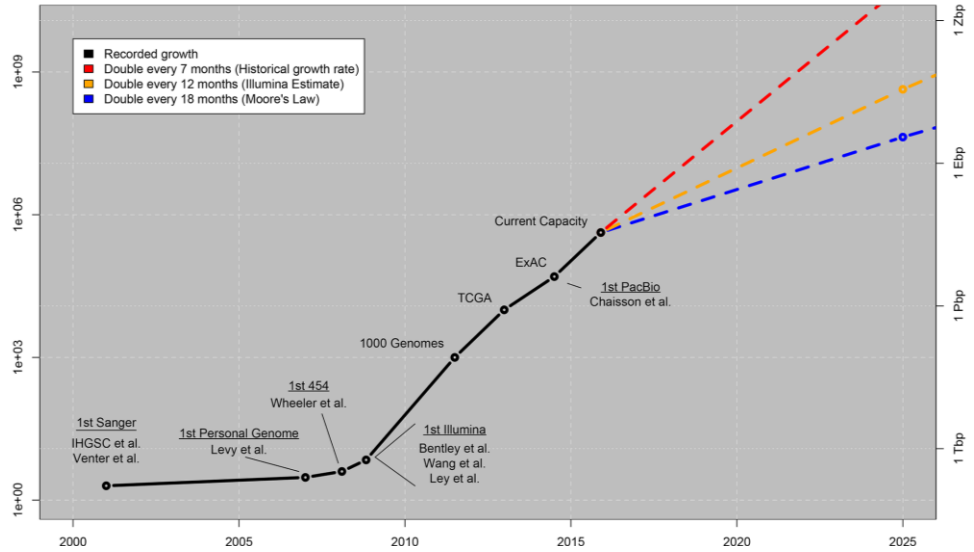
From Genomes to Systems

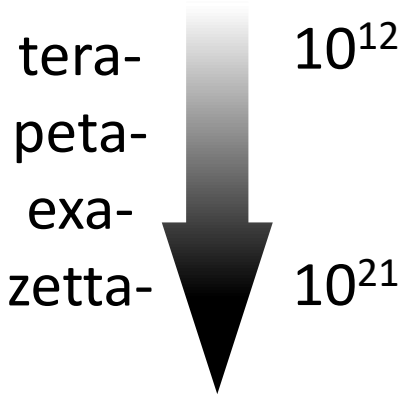


EMBL Australia Bioinformatics Resource

Big Data: Acquisition, Storage, Distribution, and Analysis

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement





Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C et al., PLOS Biology (2015)

What is Next Gen Sequencing?

Next Generation Sequencing:

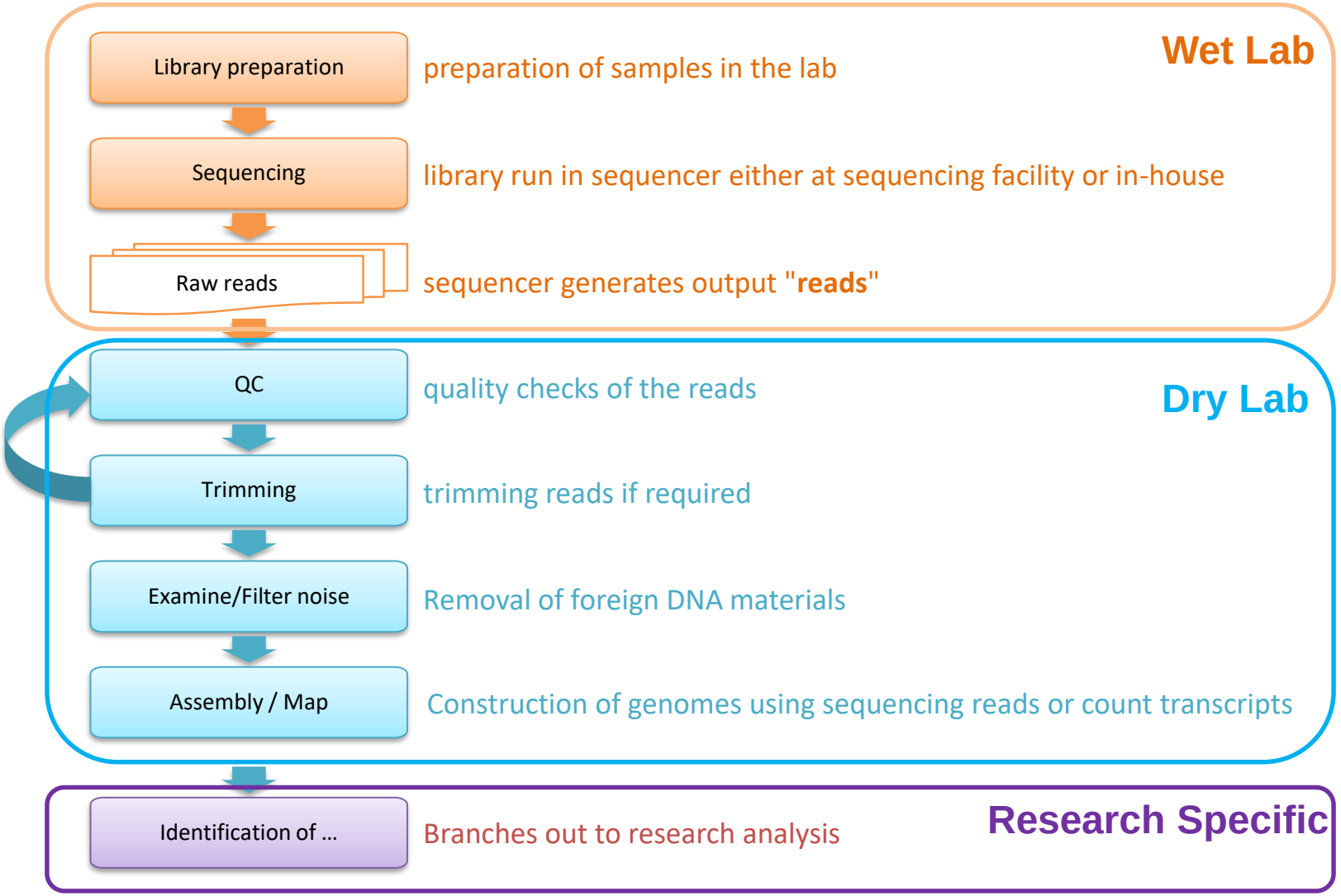
- high-throughput sequencing
- massive parallel short (and now long) read sequencing
- deep sequencing
- Sequence gathered by rounds* of extension or degradation on single* molecules, not specific base termination leading to size fractionated pools of DNA fragments

In reality NGS really just refers to the scale of sequencing. For Example:

ABI Sequencers, Venter Institute – 2007
1 Human Genome in total (years!)

Illumina HiSeq 2000 and Xs, Sanger.co.uk
~24 Human Genomes a day at 30x

Overview of NGS data flow

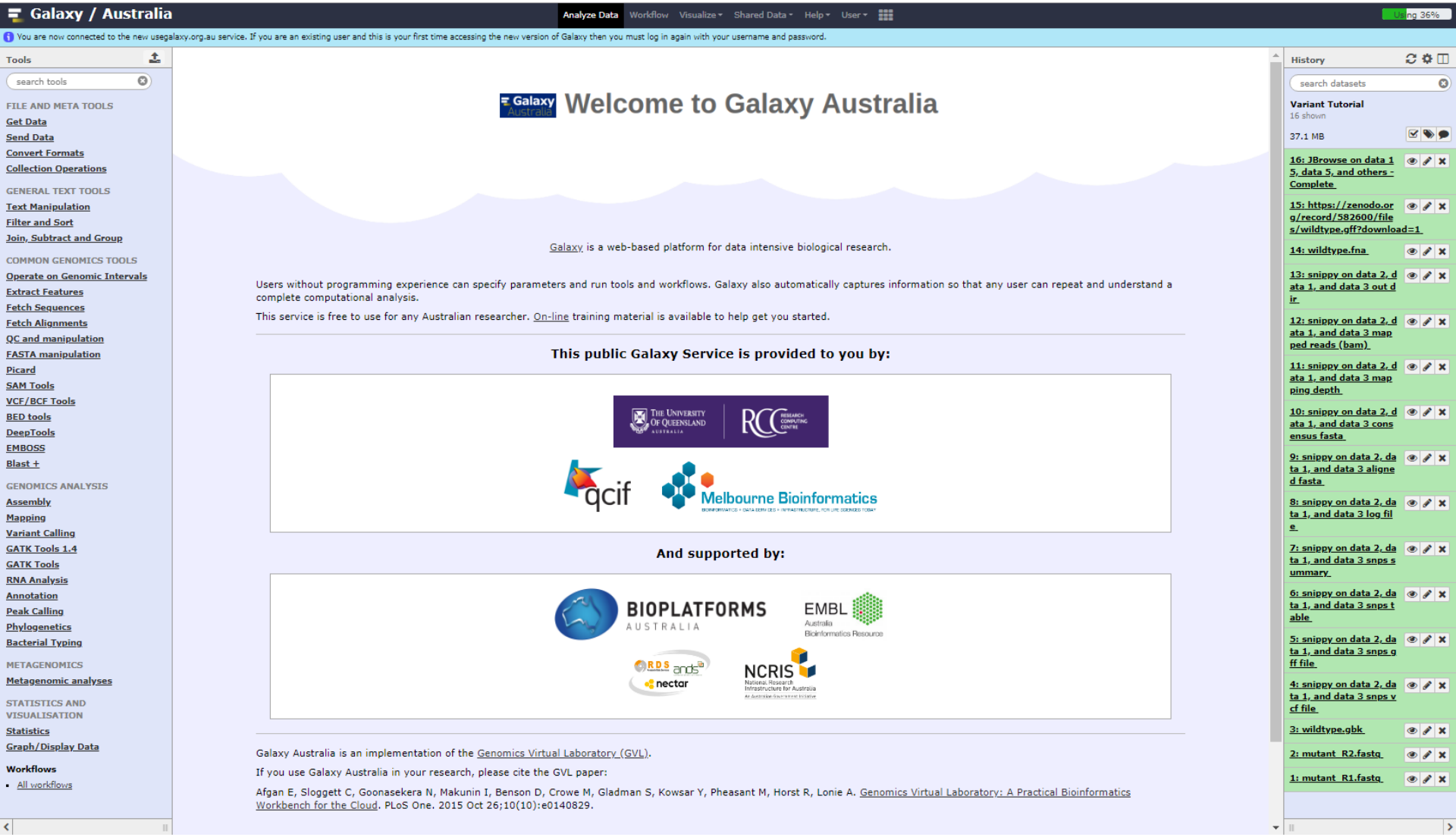


Tools – Academic and Clinical

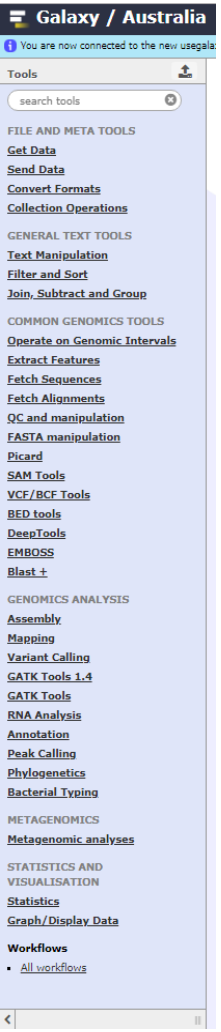
- **Freeware**
  - **Genome Analysis Toolkit (GATK)**
  - **Virtual Labs / Machines**
    - Galaxy
    - R Studio (Bioconductor)
    - Command Line
- **Commercial**
  - **Agilent**
    - Cartagenia Bench Lab for Molecular Pathology
  - **Illumina**
    - BaseSpace
  - **Qiagen**
    - CLC-Bio Suite of Analysis Products
    - Ingenuity Pathway Analysis
    - Ingenuity Variant Analysis
    - ANNOVAR
  - **ThermoFisher**
    - Ion Reporter
  - *Google Genomics*
  - *Microsoft Genomics*
  - *Oracle Healthcare Precision Medicine*

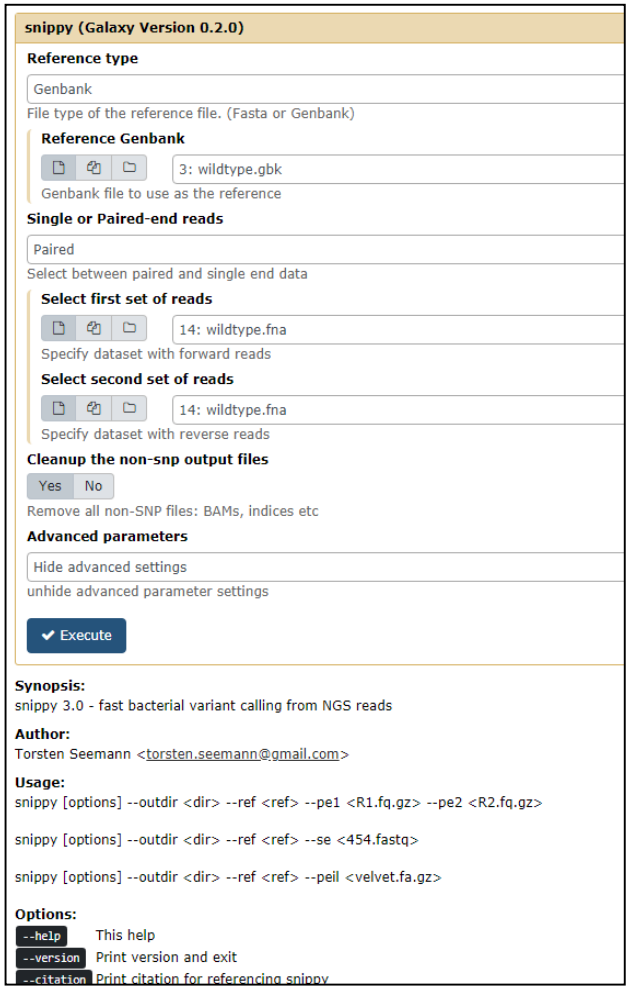
  http://grouthbio.com/Genome_Software_Service.php

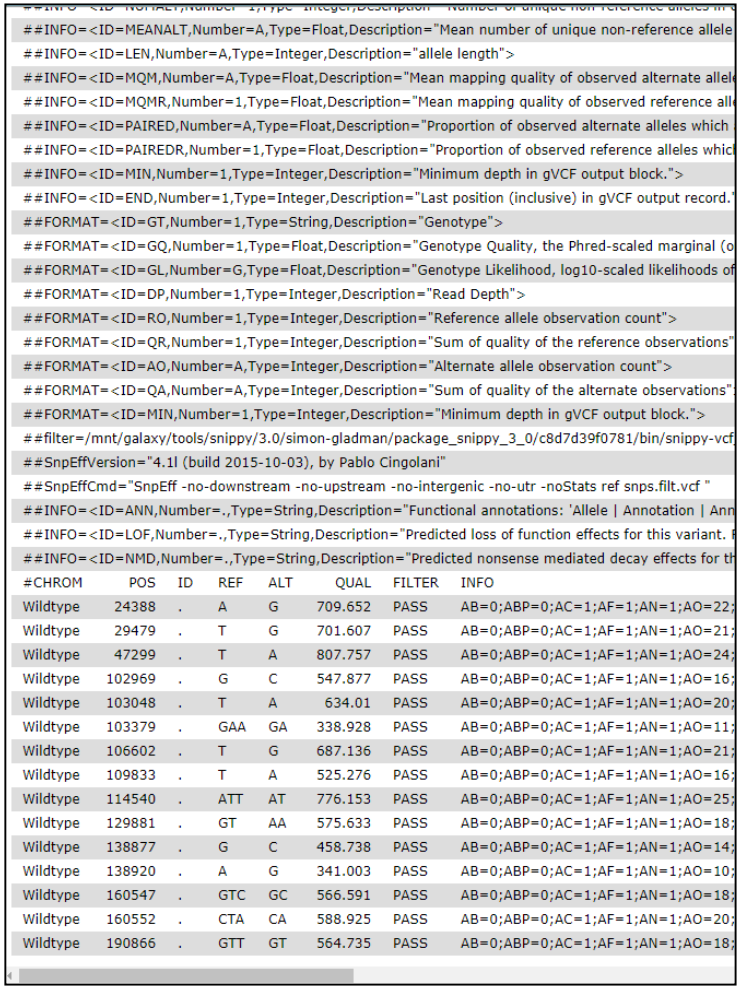
What is Galaxy

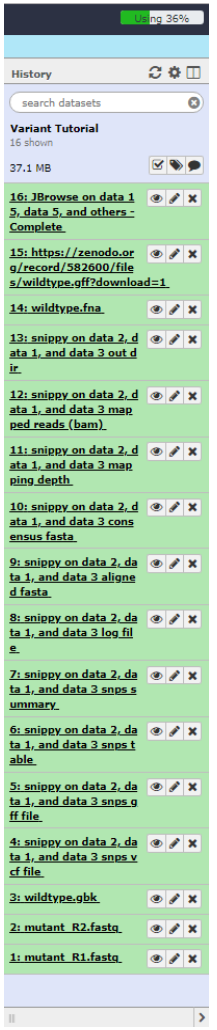


Interactive Window



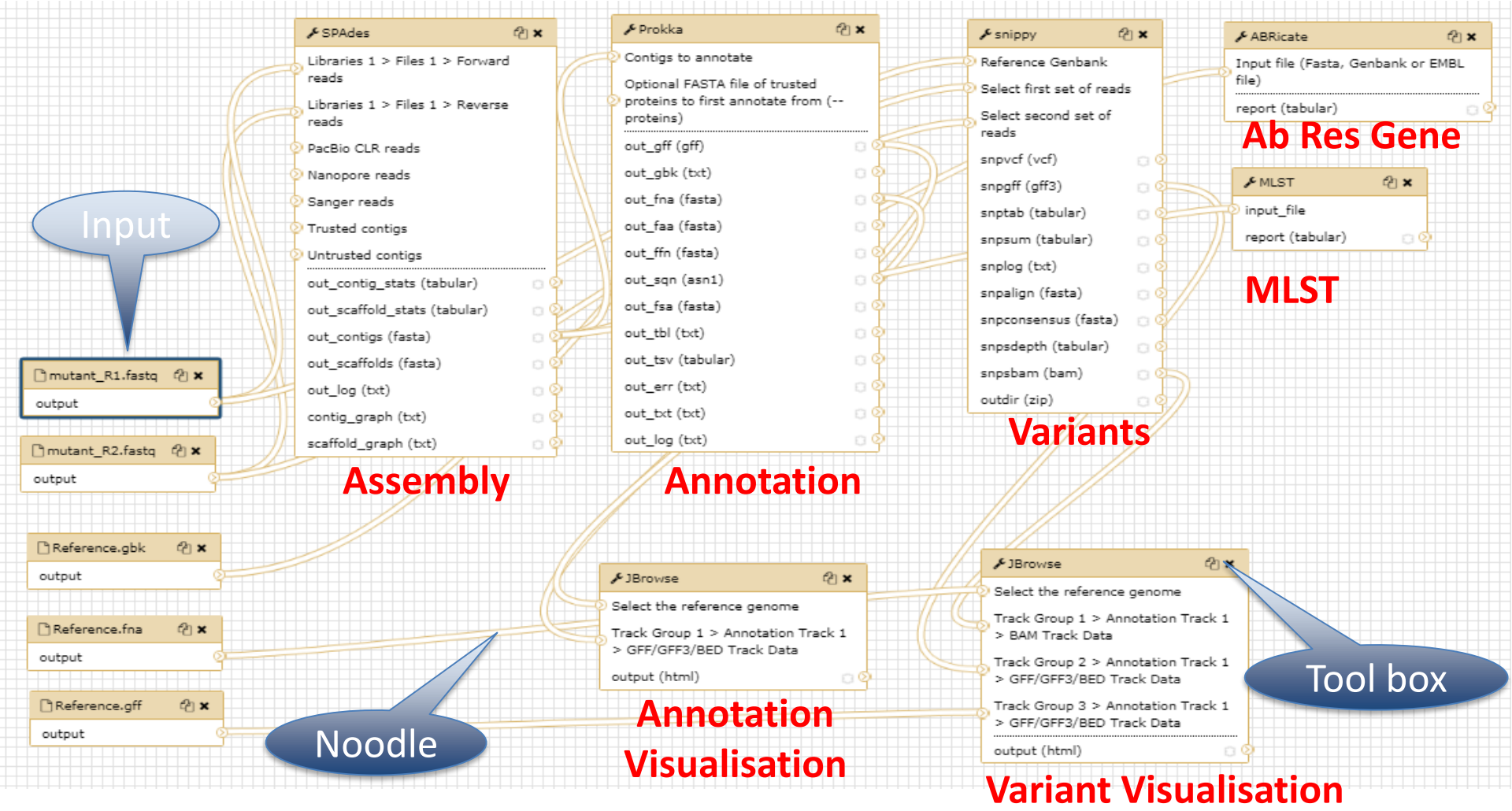






Workflow Screenshot

A Galaxy workflow is a series of tools and dataset actions that run in sequence as a batch operation



Galaxy Australia

- Galaxy Australia has recently been upgraded to:
  - incorporate new features
  - extended compute and quotas – *linked to HPC resources*
  - more training resources
    - Establish a national network of Galaxy Trainers (via EMBL-ABR Nodes and others)
    - Undertaking 4 virtual/physical national training Galaxy events
    - Rationalise Australian developed Training Material: https://galaxy-au-training.github.io/tutorials/
  - Over 600 tools and tool versions (for legacy analyses)
  - Over 200 reference genomes, indexed for rapid analyses

Galaxy Australia - Usage

An active and engaged user community

2268

registered users.

608 active users (last 90 days)

User growth 2016 - 2018

Registered users in Australia from:

30 Australian Universities

21 Medical Research Institutes or Organisations

9 Other Research Organisations

Users per domain

Users represented across

338

organisations

Users represented across

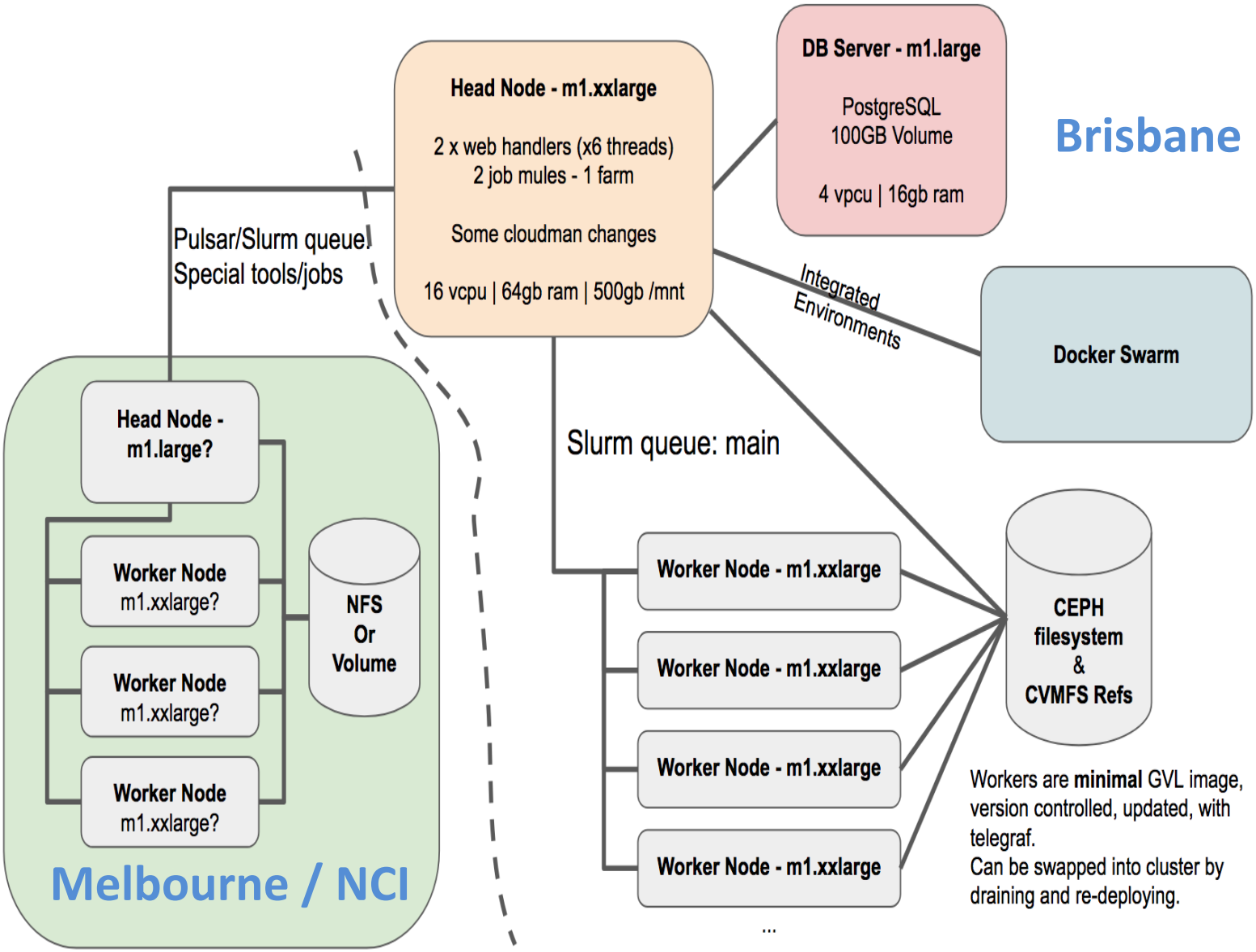
50

countries

Galaxy Australia – project activity

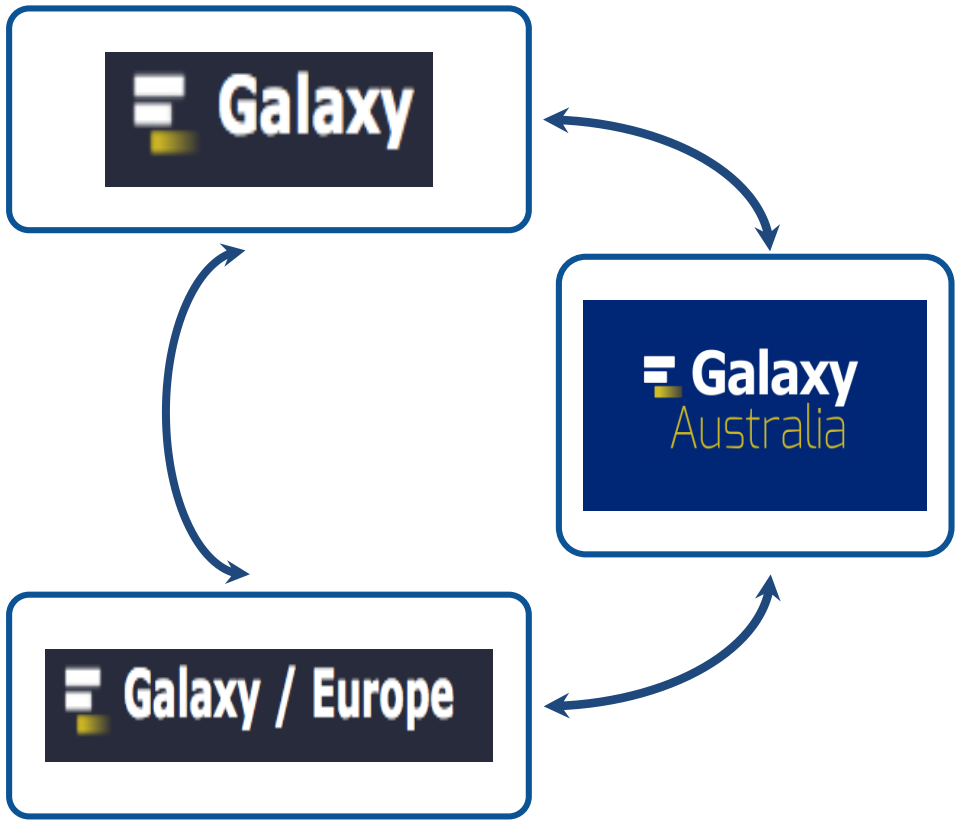
- Adding more tools and references
  - User request form for both types, with justification
  - Route to support sanction and novel tools
- Adding distributed and “bespoke” compute
  - Pulsar servers at University of Melb. / NCI
  - High memory at UQ
- Adding responsiveness feedback to the user
  - Number of running jobs
  - Average job run time

Galaxy Australia - Architecture

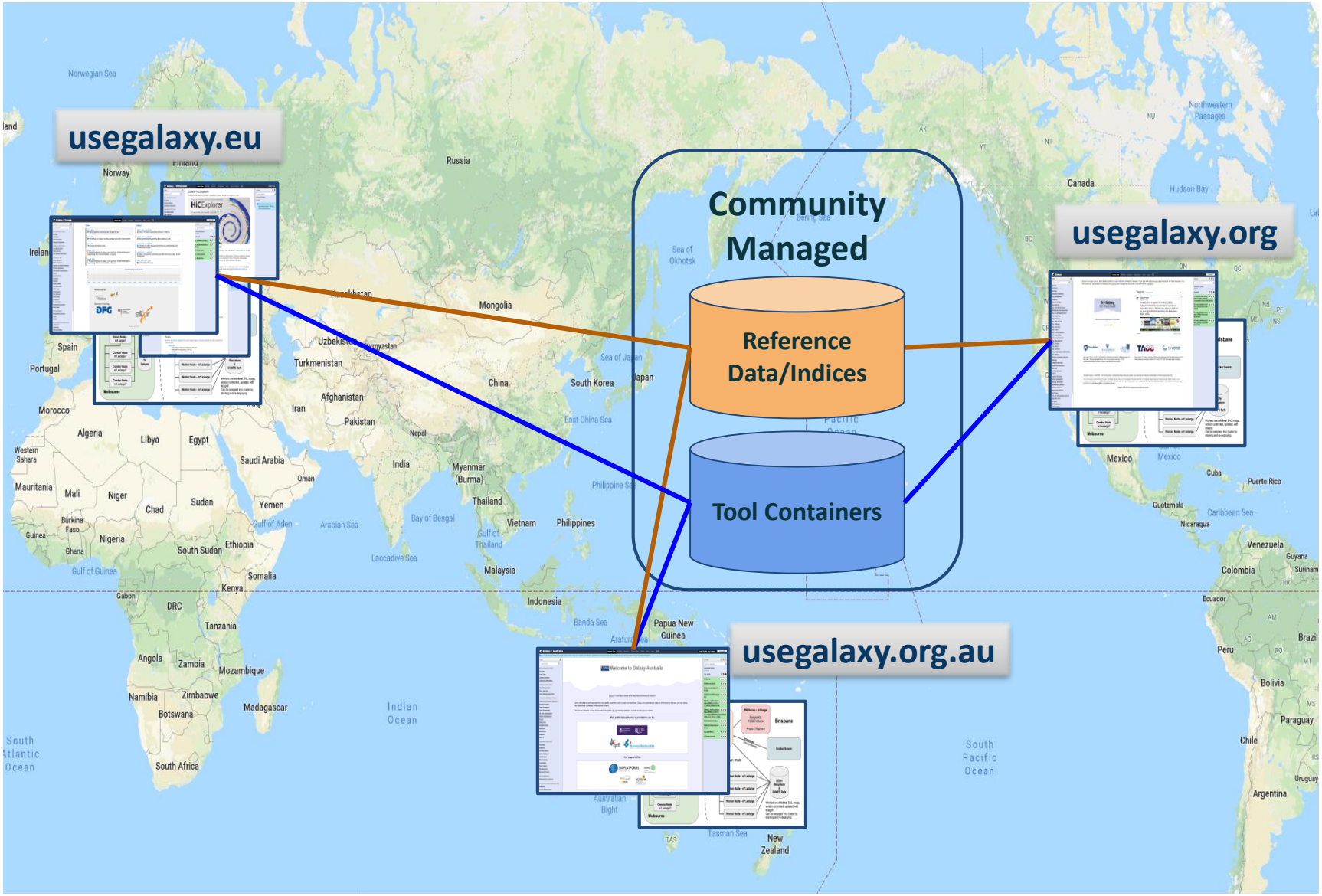


UseGalaxy.*
Harmonising the look and feel with Other Global Galaxy Services

- Run the latest stable Galaxy release
- Present the user with similar tool list layout
- Alignment of Galaxy Training Network's core tutorials, including the same training datasets available



usegalaxy.* - Proposed Global Architecture



Galaxy Australia – future activity

- What will Galaxy Australia be in a few years
  - Grow in scope (tools and compute) to support long read and hybrid analysis
    - Pacbio RSII and Sequel
    - ONT MinION and PromethION
  - Grow in Pulsar machines to support generic and bespoke analyses
  - Increased alignment with usegalaxy.*
- *Financial packages and cost/benefit*
  - *Galaxy Australia users sourcing ONLY commercial software is **10x greater** cost than using our service*

Stay in Touch

- Galaxy Australia: Twitter https://twitter.com/galaxyaustralia
- Galaxy Australia Community https://www.embl-abr.org.au/galaxyaustralia

- And of course a final reminder:
  - Galaxy Australia **https://usegalaxy.org.au**

Galaxy Australia Team Members

Gareth Price (PM) - QFAB

Simon Gladman - Melb Bionf

Derek Benson - UQ-RCC

Anna Syme - Melb Bioinf

Igor Makunin - UQ-RCC

Nuwan Goonasekera - Melb Bionf

Christina Hall - Melb Bionf

Helen van der Pol - Melb Bioinf