



Enabling Australian Genomics research through enhancements to the Genomics Virtual Lab

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

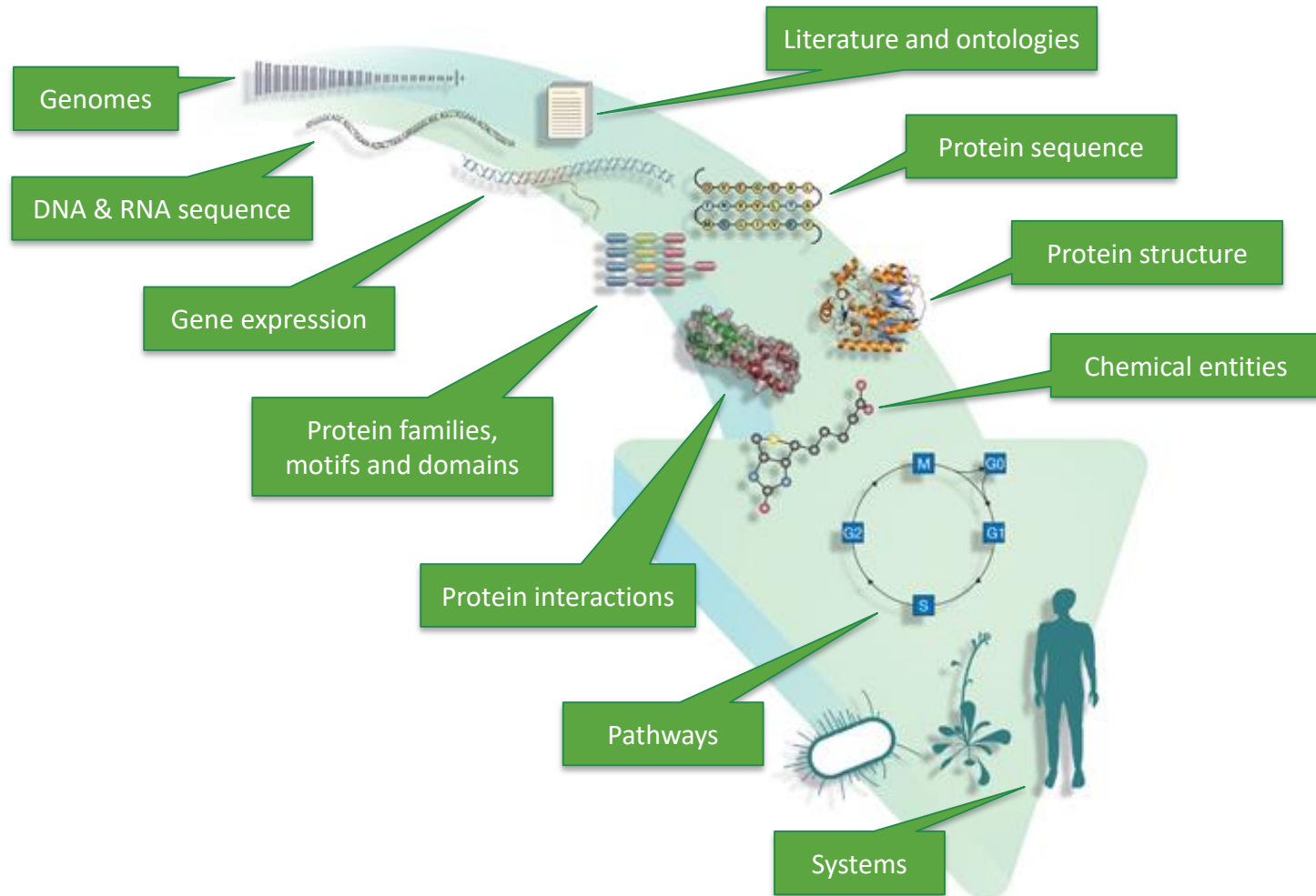
Development partners



Supported by

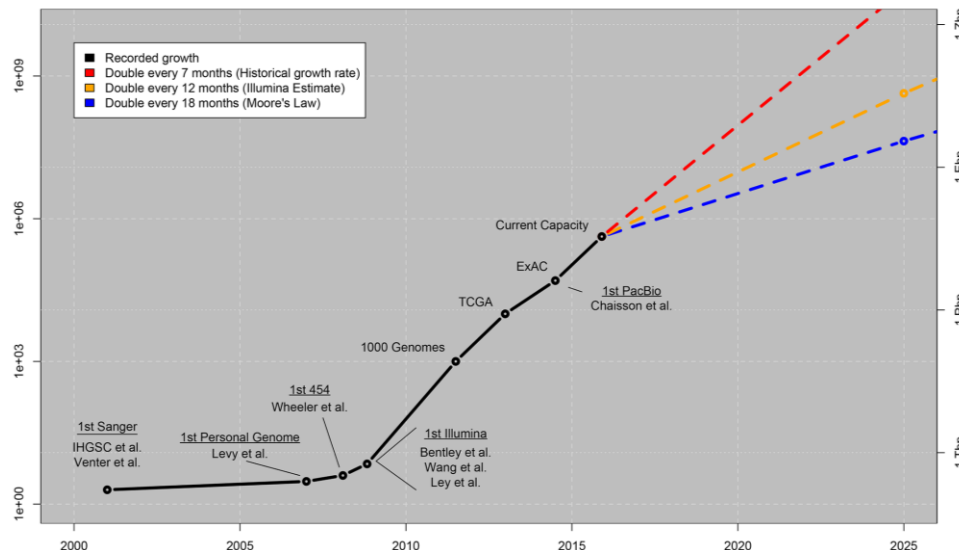


From Genomes to Systems



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



tera-
peta-
exa-
zetta-

10¹²

10²¹

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:

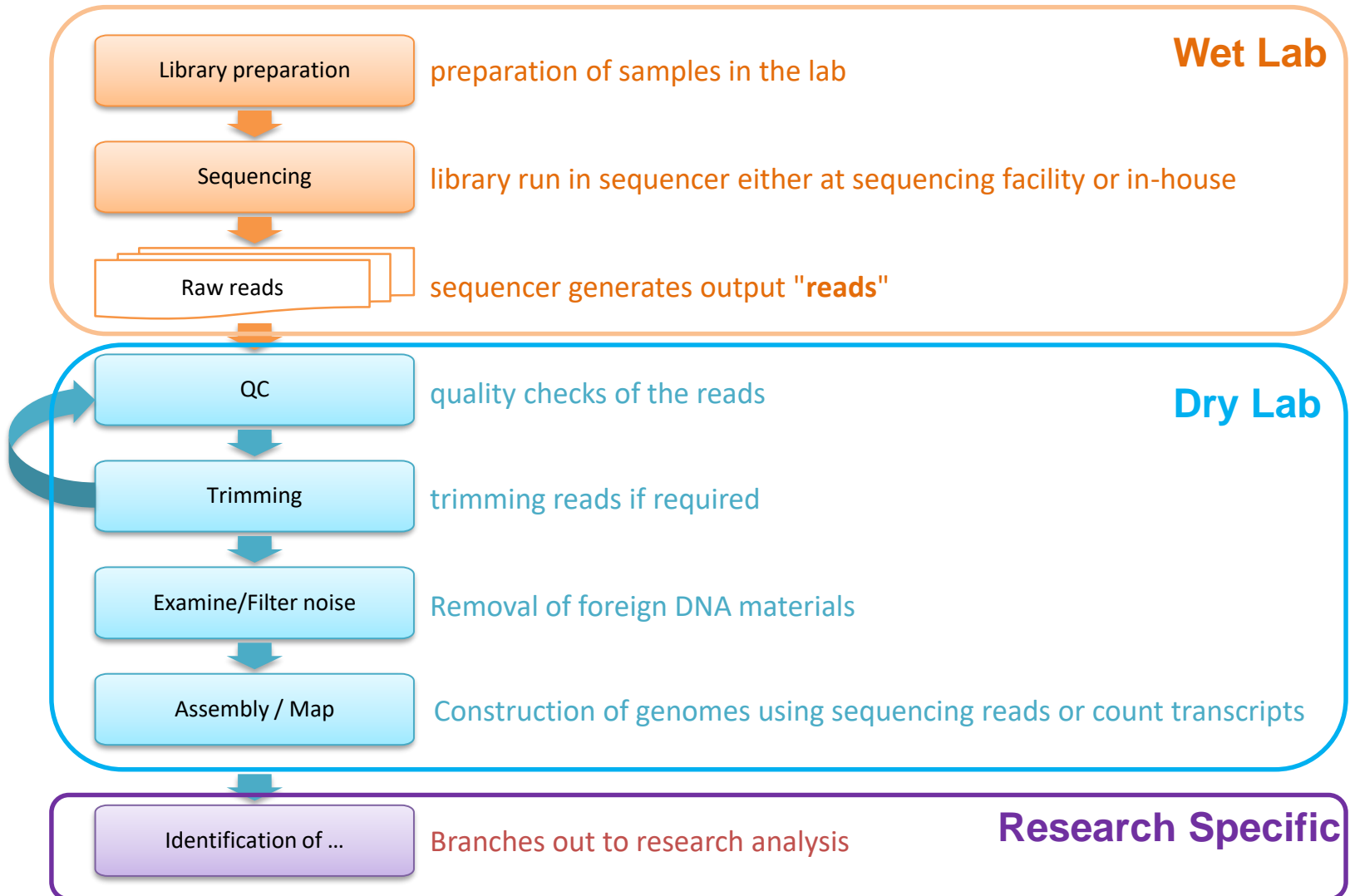


ABI Sequencers, Venter Institute – 2007
1 Human Genome in total (years!)



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



Galaxy / Australia

Analyze Data Workflow Visualize Shared Data Help User

36%

You are now connected to the new usegalaxy.org.au service. If you are an existing user and this is your first time accessing the new version of Galaxy then you must log in again with your username and password.

Tools

search tools

FILE AND META TOOLS

Get Data

Send Data

Convert Formats

Collection Operations

GENERAL TEXT TOOLS

Text Manipulation

Filter and Sort

Join, Subtract and Group

COMMON GENOMICS TOOLS

Operate on Genomic Intervals

Extract Features

Fetch Sequences

Fetch Alignments

QC and manipulation

FASTA manipulation

Picard

SAM Tools

VCF/BCF Tools

BED tools

DeepTools

EMBOSS

Blast

GENOMICS ANALYSIS

Assembly

Mapping

Variant Calling

GATK Tools 1-4

GATK Tools

RNA Analysis

Annotation

Peak Calling

Phylogenetics

Bacterial Typing

METAGENOMICS

Metagenomic analyses

STATISTICS AND VISUALISATION

Statistics

Graph/Display Data

Workflows

All workflows

Welcome to Galaxy Australia

Galaxy is a web-based platform for data intensive biological research.

Users without programming experience can specify parameters and run tools and workflows. Galaxy also automatically captures information so that any user can repeat and understand a complete computational analysis.

This service is free to use for any Australian researcher. [On-line](#) training material is available to help get you started.

This public Galaxy Service is provided to you by:



And supported by:



Galaxy Australia is an implementation of the [Genomics Virtual Laboratory \(GVL\)](#).

If you use Galaxy Australia in your research, please cite the GVL paper:

Afgan E, Sloggett C, Goonasekera N, Makunin I, Benson D, Crowe M, Gladman S, Kowsar Y, Pheasant M, Horst R, Lonie A. [Genomics Virtual Laboratory: A Practical Bioinformatics Workbench for the Cloud](#). PLoS One. 2015 Oct 26;10(10):e0140829.

History

search datasets

Variant Tutorial

16 shown

37.1 MB

16: JBrowse on data 1

5, data 5, and others -

Complete

15: <https://zenodo.org/record/582600/file/s/wildtype.off?download=1>

14: wildtype.fna

13: snippy on data 2, d

ata 1, and data 3 out d

ir

12: snippy on data 2, d

ata 1, and data 3 map

ped reads (bam)

11: snippy on data 2, d

ata 1, and data 3 map

ping depth

10: snippy on data 2, d

ata 1, and data 3 cons

ensus fasta

9: snippy on data 2, d

ata 1, and data 3 aligne

d fasta

8: snippy on data 2, d

ata 1, and data 3 log fil

e

7: snippy on data 2, d

ata 1, and data 3 snps s

ummary

6: snippy on data 2, d

ata 1, and data 3 snps t

able

5: snippy on data 2, d

ata 1, and data 3 snps o

ff file

4: snippy on data 2, d

ata 1, and data 3 snps v

cf file

3: wildtype.gbk

2: mutant_R2.fastq

1: mutant_R1.fastq

Interactive Window

Galaxy / Australia

You are now connected to the new usgalaxy

Tools

search tools

FILE AND META TOOLS

Get Data

Send Data

Convert Formats

Collection Operations

GENERAL TEXT TOOLS

Text Manipulation

Filter and Sort

Join, Subtract and Group

COMMON GENOMICS TOOLS

Operate on Genomic Intervals

Extract Features

Fetch Sequences

Fetch Alignments

QC and manipulation

FASTA manipulation

Picard

SAM Tools

VCF/BCF Tools

BED tools

DeepTools

EMBOSS

Blast+

GENOMICS ANALYSIS

Assembly

Mapping

Variant Calling

GATK Tools 1-4

GATK Tools

RNA Analysis

Annotation

Peak Calling

Phylogenetics

Bacterial Typing

METAGENOMICS

Metagenomic analyses

STATISTICS AND VISUALISATION

Statistics

Graph/Display Data

Workflows

All workflows

snippy (Galaxy Version 0.2.0)

Reference type

Genbank

File type of the reference file. (Fasta or Genbank)

Reference Genbank

3: wildtype.gbk

Genbank file to use as the reference

Single or Paired-end reads

Paired

Select between paired and single end data

Select first set of reads

14: wildtype.fna

Specify dataset with forward reads

Select second set of reads

14: wildtype.fna

Specify dataset with reverse reads

Cleanup the non-snp output files

Yes No

Remove all non-SNP files: BAMs, indices etc

Advanced parameters

Hide advanced settings

unhide advanced parameter settings

Execute

Synopsis:

snippy 3.0 - fast bacterial variant calling from NGS reads

Author:

Torsten Seemann <torsten.seemann@gmail.com>

Usage:

snippy [options] --outdir <dir> --ref <ref> --pe1 <R1.fq.gz> --pe2 <R2.fq.gz>

snippy [options] --outdir <dir> --ref <ref> --se <454.fastq>

snippy [options] --outdir <dir> --ref <ref> --pe1 <velvet.fa.gz>

Options:

--help This help

--version Print version and exit

--citation Print citation for referencing snippy

```
##INFO=<ID=NONREF,Number=1,Type=Integer,Description="Number of unique non-reference alleles in"
##INFO=<ID=MEANALT,Number=A,Type=Float,Description="Mean number of unique non-reference allele
##INFO=<ID=LEN,Number=A,Type=Integer,Description="allele length">
##INFO=<ID=MQM,Number=A,Type=Float,Description="Mean mapping quality of observed alternate alle
##INFO=<ID=MQMR,Number=1,Type=Float,Description="Mean mapping quality of observed reference all
##INFO=<ID=PAIRED,Number=A,Type=Float,Description="Proportion of observed alternate alleles which
##INFO=<ID=PAIREDR,Number=1,Type=Float,Description="Proportion of observed reference alleles which
##INFO=<ID=MIN,Number=1,Type=Integer,Description="Minimum depth in gVCF output block.">
##INFO=<ID=END,Number=1,Type=Integer,Description="Last position (inclusive) in gVCF output record."
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality, the Phred-scaled marginal (o
##FORMAT=<ID=GL,Number=G,Type=Float,Description="Genotype Likelihood, log10-scaled likelihoods of
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=RO,Number=1,Type=Integer,Description="Reference allele observation count">
##FORMAT=<ID=QR,Number=1,Type=Integer,Description="Sum of quality of the reference observations"
##FORMAT=<ID=AO,Number=A,Type=Integer,Description="Alternate allele observation count">
##FORMAT=<ID=QA,Number=A,Type=Integer,Description="Sum of quality of the alternate observations"
##FORMAT=<ID=MIN,Number=1,Type=Integer,Description="Minimum depth in gVCF output block.">
##filter=/mnt/galaxy/tools/snippy/3.0/simon-gladman/package_snippy_3_0/c8d7d39f0781/bin/snippy-vcf
##SnpEffVersion="4.1l (build 2015-10-03), by Pablo Cingolani"
##SnpEffCmd="SnpEff -no-downstream -no-upstream -no-intergenic -no-utr -noStats ref.snps.filter.vcf "
##INFO=<ID=ANN,Number=.,Type=String,Description="Functional annotations: 'Allele | Annotation | Ann
##INFO=<ID=LOF,Number=.,Type=String,Description="Predicted loss of function effects for this variant.
##INFO=<ID=NMD,Number=.,Type=String,Description="Predicted nonsense mediated decay effects for th
#CHROM POS ID REF ALT QUAL FILTER INFO
Wildtype 24388 . A G 709.652 PASS AB=0;ABP=0;AC=1;AF=1;AN=1;AO=22;
Wildtype 29479 . T G 701.607 PASS AB=0;ABP=0;AC=1;AF=1;AN=1;AO=21;
Wildtype 47299 . T A 807.757 PASS AB=0;ABP=0;AC=1;AF=1;AN=1;AO=24;
Wildtype 102969 . G C 547.877 PASS AB=0;ABP=0;AC=1;AF=1;AN=1;AO=16;
Wildtype 103048 . T A 634.01 PASS AB=0;ABP=0;AC=1;AF=1;AN=1;AO=20;
Wildtype 103379 . GAA GA 338.928 PASS AB=0;ABP=0;AC=1;AF=1;AN=1;AO=11;
Wildtype 106602 . T G 687.136 PASS AB=0;ABP=0;AC=1;AF=1;AN=1;AO=21;
Wildtype 109833 . T A 525.276 PASS AB=0;ABP=0;AC=1;AF=1;AN=1;AO=16;
Wildtype 114540 . ATT AT 776.153 PASS AB=0;ABP=0;AC=1;AF=1;AN=1;AO=25;
Wildtype 129881 . GT AA 575.633 PASS AB=0;ABP=0;AC=1;AF=1;AN=1;AO=18;
Wildtype 138877 . G C 458.738 PASS AB=0;ABP=0;AC=1;AF=1;AN=1;AO=14;
Wildtype 138920 . A G 341.003 PASS AB=0;ABP=0;AC=1;AF=1;AN=1;AO=10;
Wildtype 160547 . GTC GC 566.591 PASS AB=0;ABP=0;AC=1;AF=1;AN=1;AO=18;
Wildtype 160552 . CTA CA 588.925 PASS AB=0;ABP=0;AC=1;AF=1;AN=1;AO=20;
Wildtype 190866 . GTT GT 564.735 PASS AB=0;ABP=0;AC=1;AF=1;AN=1;AO=18;
```

History

search datasets

Variant Tutorial

16 shown

37.1 MB

16: JBrowse on data 1 5, data 5, and others - Complete

15: https://zenodo.org/record/382600/files/wildtype.gff?download=1

14: wildtype.fna

13: snippy on data 2, data 1, and data 3 out dir

12: snippy on data 2, data 1, and data 3 mapped reads (bam)

11: snippy on data 2, data 1, and data 3 mapping depth

10: snippy on data 2, data 1, and data 3 consensus fasta

9: snippy on data 2, data 1, and data 3 aligned fasta

8: snippy on data 2, data 1, and data 3 log file

7: snippy on data 2, data 1, and data 3 snps summary

6: snippy on data 2, data 1, and data 3 snps table

5: snippy on data 2, data 1, and data 3 snps gff file

4: snippy on data 2, data 1, and data 3 snps vcf file

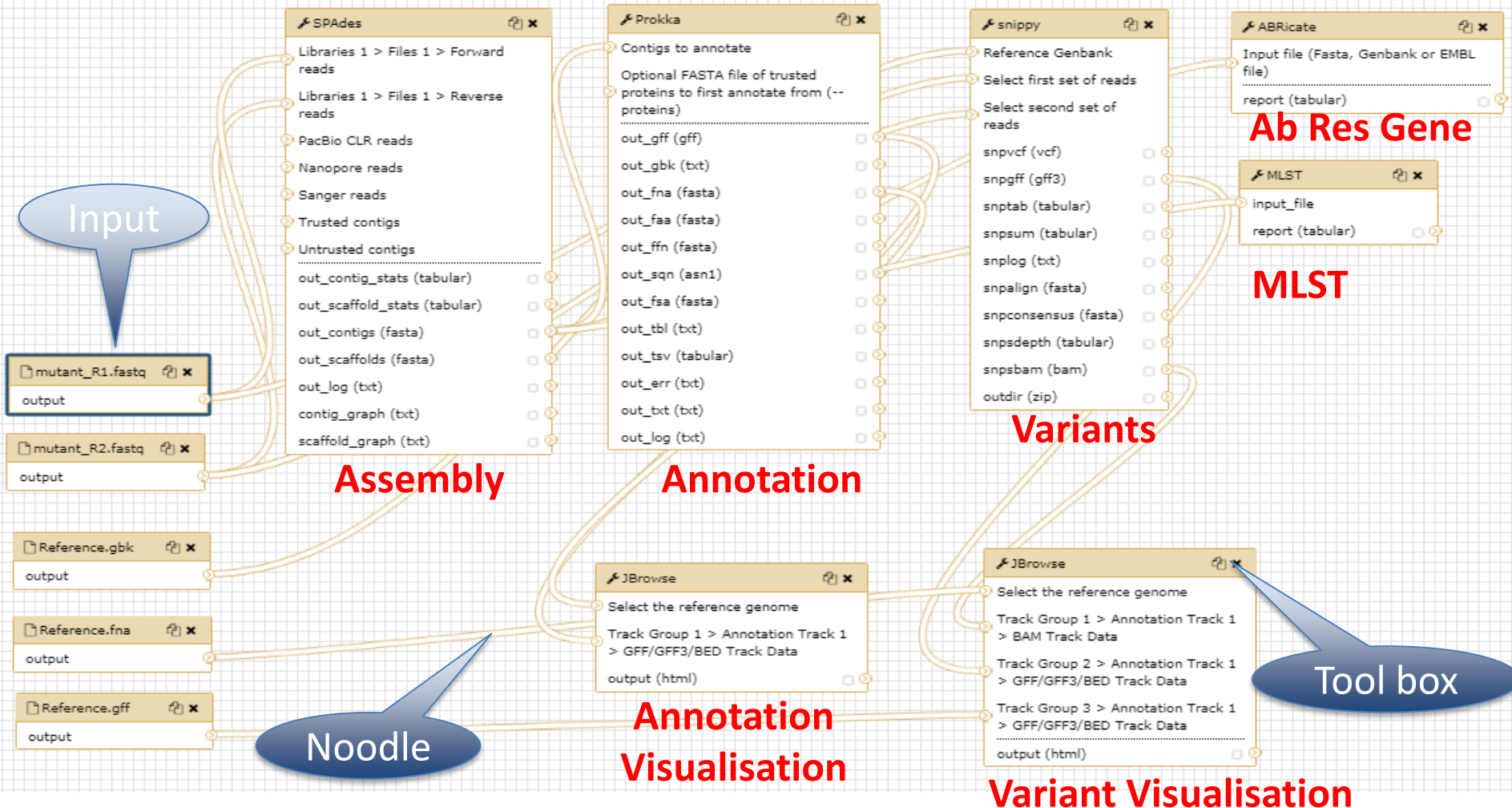
3: wildtype.gbk

2: mutant_R2.fastq

1: mutant_R1.fastq

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



- .edu.au
- .gov.au
- .org.au
- .com - au
- .edu
- other international

Users represented across

338

organisations

Users represented across

50

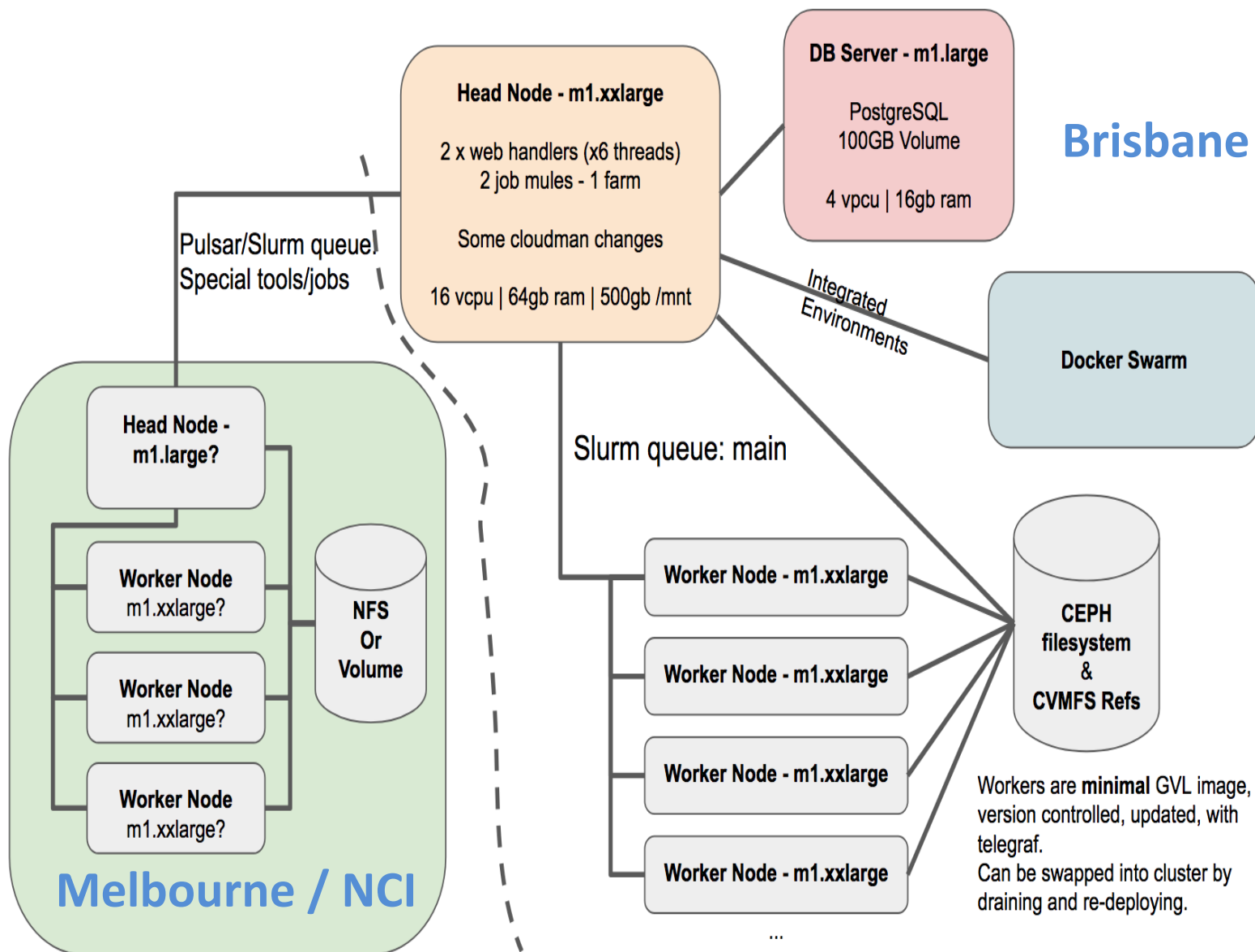
countries



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

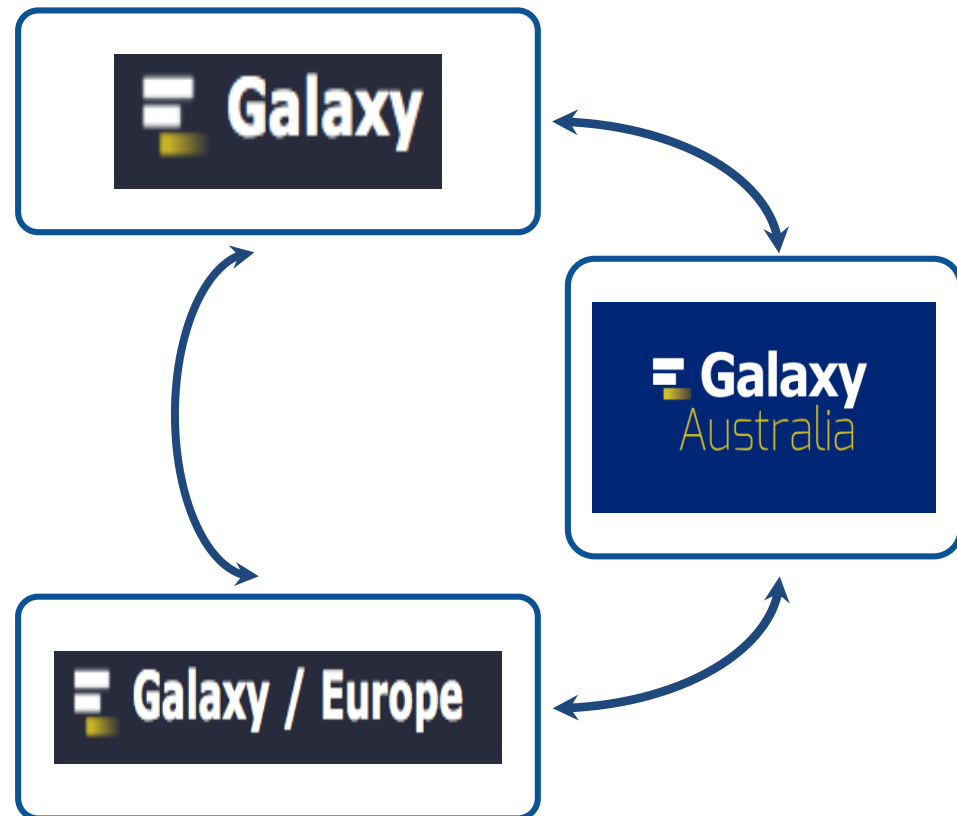
Galaxy Australia - Architecture



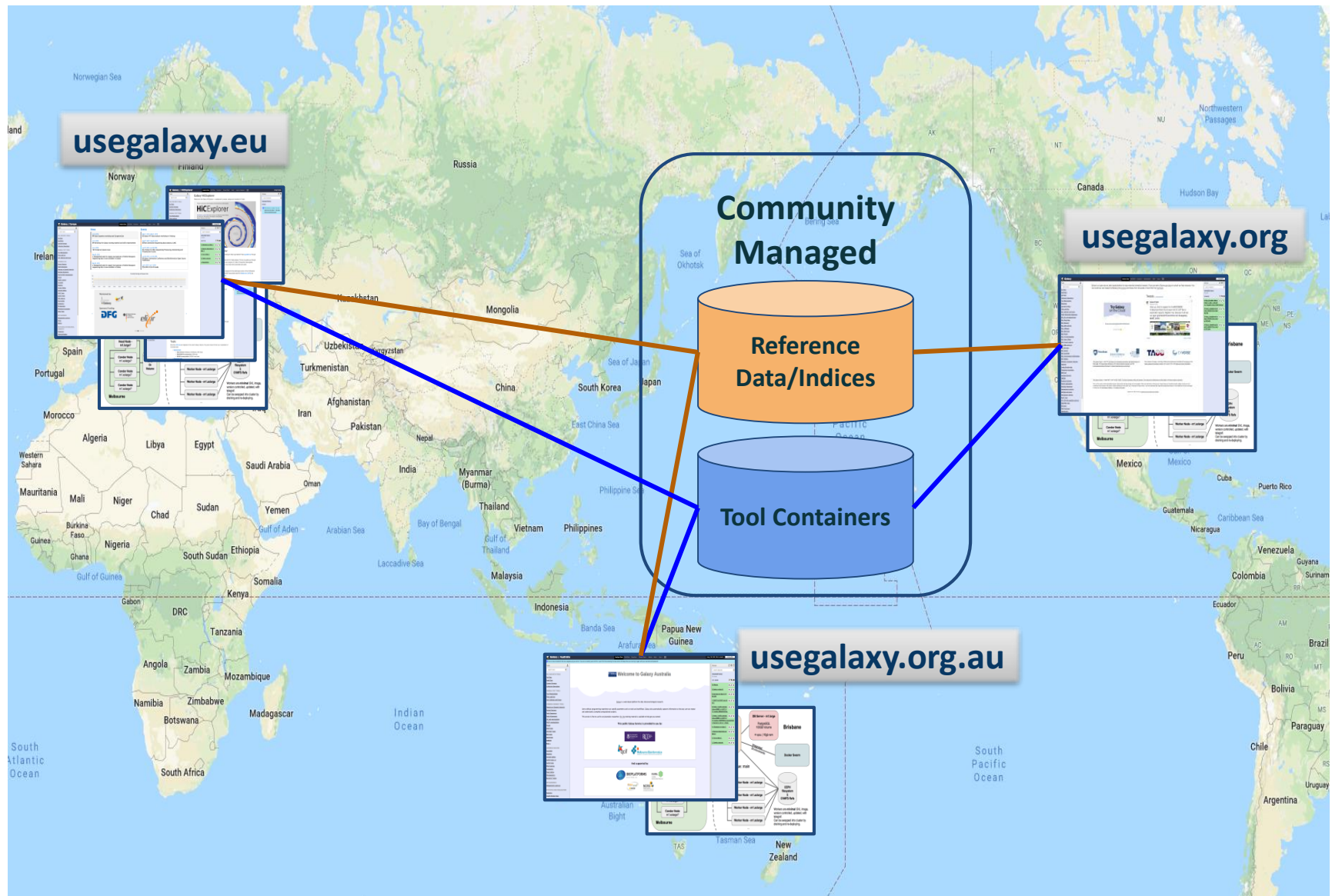
UseGalaxy.*

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

