A cloud-based system to enable streamlined access to, and analysis of, continental-scale environmental metagenomics data by non-genomics researchers

Jeff Christiansen, Derek Benson, Grahame Bowland, Samuel Chang, Simon Gladman, Gareth Price, Anna Syme, Tamas Szabo, Mike Thang, Andrew Bissett
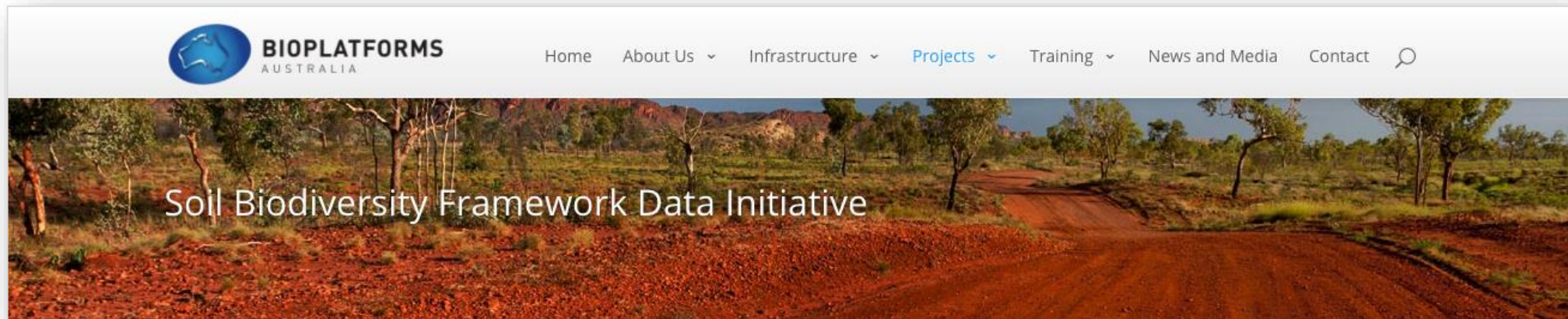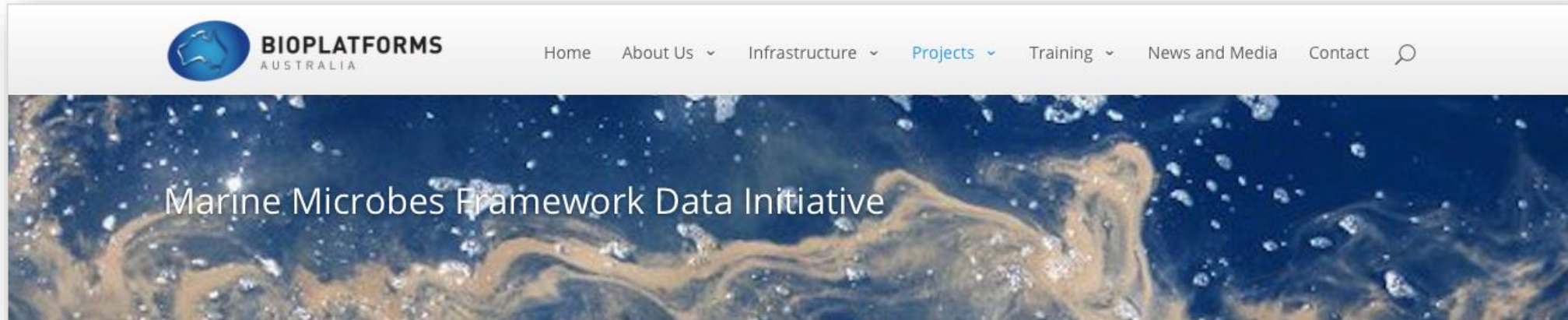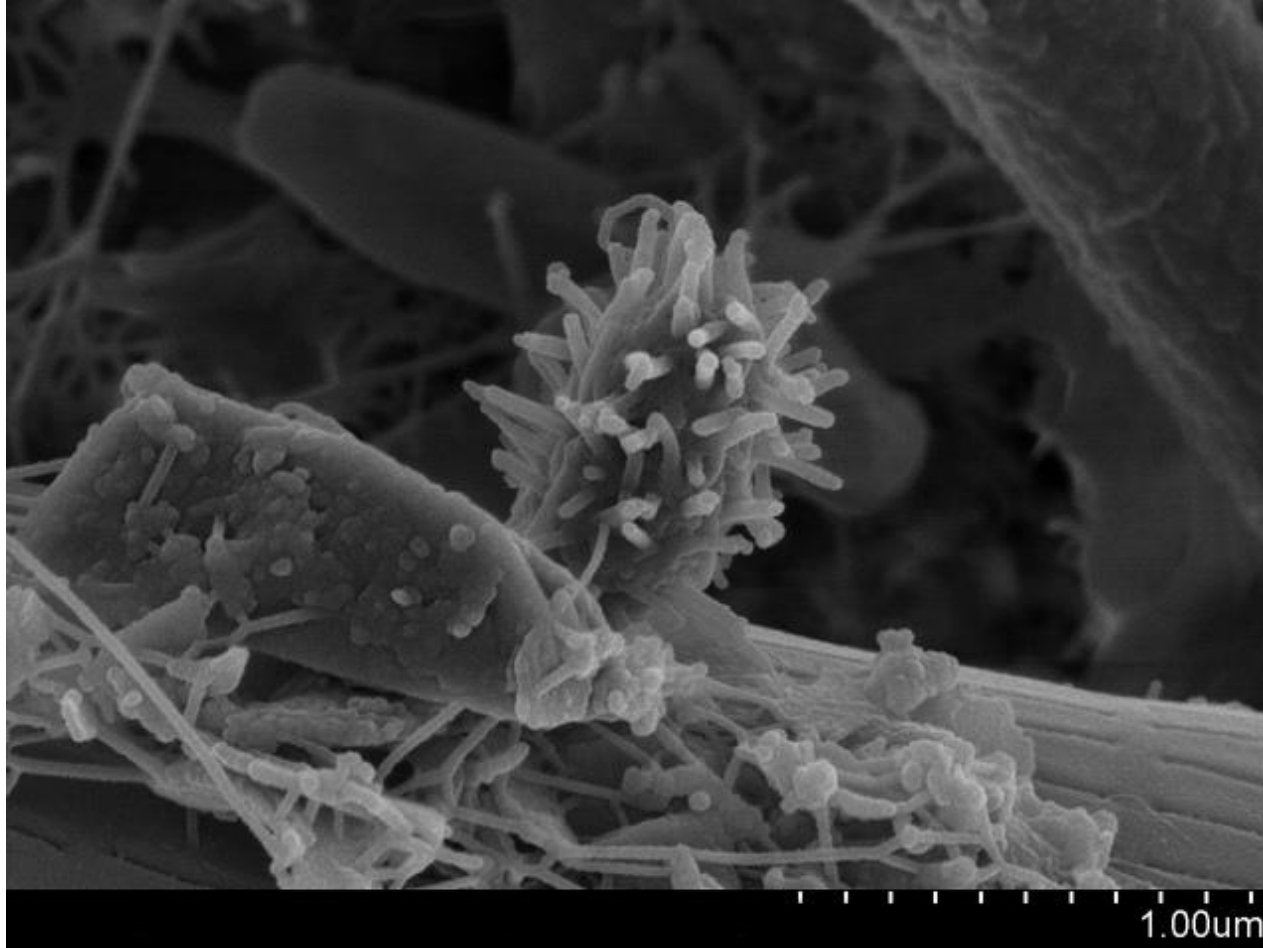
ARDC funded Bio–RDC project

- Supporting Researchers using Environmental Metagenomics Data

ARDC funded Bio–RDC project

- The environment is teeming with microorganisms

A community of bacteria growing on a single grain of sand collected from intertidal sediment on a beach near Boston, USA

Image courtesy of the Lewis Lab at Northeastern University. Image created by Anthony D'Onofrio, William H. Fowle, Eric J. Stewart and Kim Lewis. Source: https://www.flickr.com/photos/adonofrio/4482210139/in/gallery-137201199@N07-72157661521584722/

ARDC funded Bio–RDC project

- The metagenomics approach

Data format:

1000s of columns (**samples**)

1000000s of rows (**sequences**)

Corresponding biological taxa
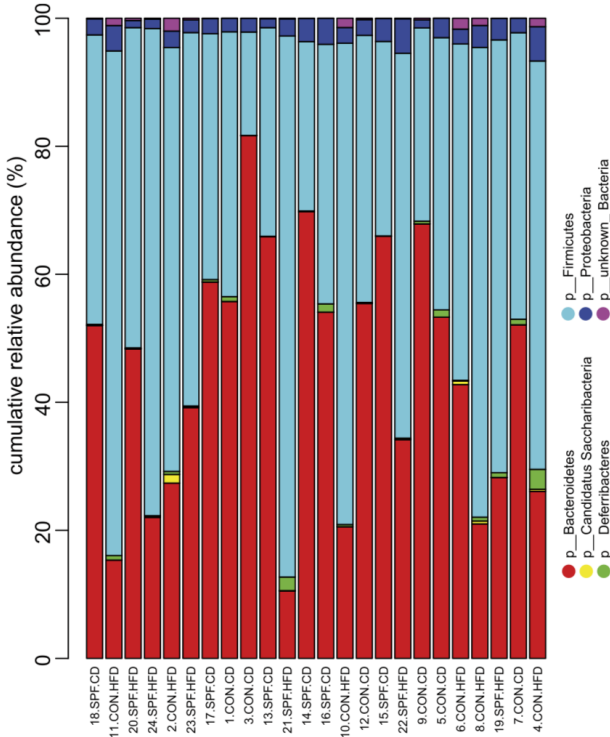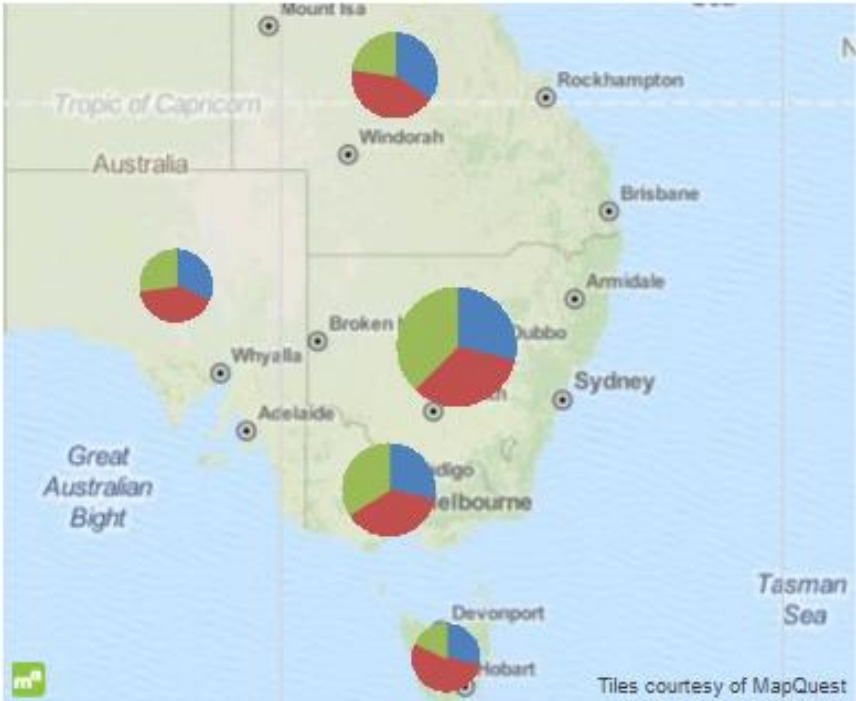(Kingdom > Phylum > Class > Order > Family > Genus > Species)

1000000s of rows (**sequences**)

Contextual information
(e.g. soil, environment features)
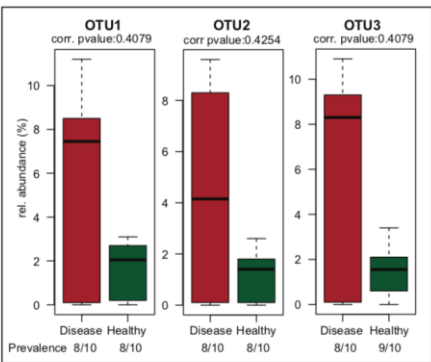
1000s of **samples**

ARDC funded Bio–RDC project

- Supporting Researchers using Environmental Metagenomics Data

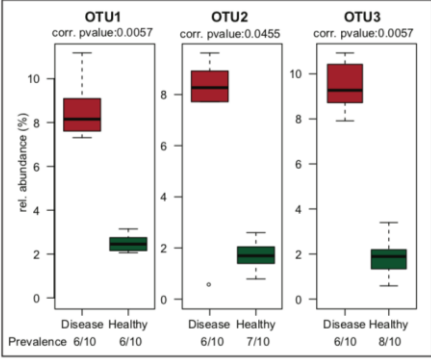- Streamlining access to data and tools for analysis of that data

ARDC funded Bio–RDC project

- Supporting Researchers using Environmental Metagenomics Data

- Streamlining access to data and tools for analysis of that data

- Extending and linking existing data infrastructures:
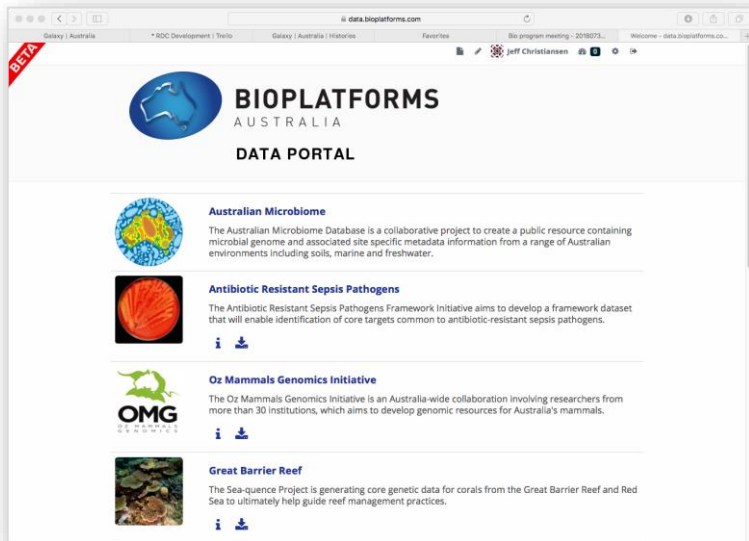
Data Storage   Data Analysis

ARDC funded Bio–RDC project

- Supporting Researchers using Environmental Metagenomics Data

- Streamlining access to data and tools for analysis of that data

- Extending and linking existing data infrastructures:

- Development Team: CCG, QCIF/Melb Bionf, Andrew Bissett CSIRO, ALA (obs)
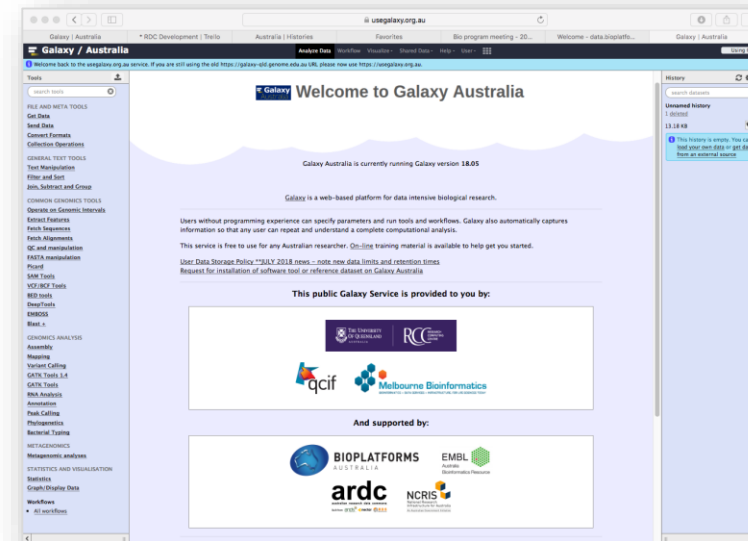
ARDC funded Bio–RDC project

- Supporting Researchers using Environmental Metagenomics Data

- Streamlining access to data and tools for analysis of that data

- Extending and linking existing data infrastructures:

- Development Team: CCG, QCIF/Melb Bionf, Andrew Bissett CSIRO, ALA (obs)

- Timeline: Feb 2018 – Dec 2018 (10 months)
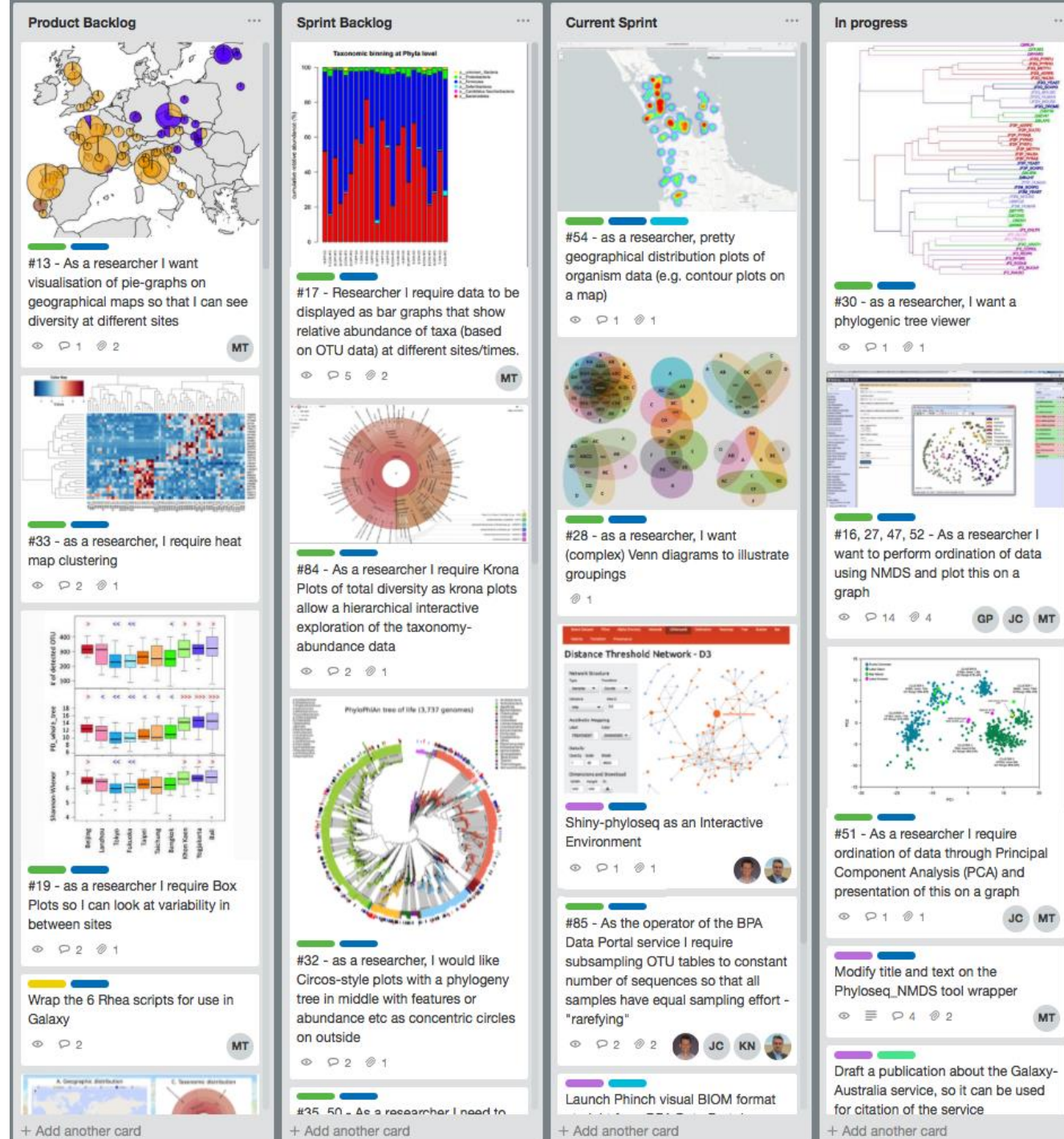
ARDC funded Bio–RDC project

- Researcher's Requirements

ARDC funded Bio–RDC project

- Researcher's Requirements

- Gathered at BPA coordinated
  Marine Microbes Meeting
 March 2018

ARDC funded Bio–RDC project

- Researcher's Requirements

- Gathered at BPA coordinated Marine Microbes Meeting March 2018

- Functionality to analyse OTU data across sites/populations:

  - Ordination
  - Clustering
  - Rarefaction
  - Search by Sequence Similarity
  - Plots (Box, Bar, Venn etc)
  - Maps (bubbleplots, heatmaps)
 - Browsing Taxonomic abundance data
  - BYO OTU data

ARDC funded Bio–RDC project

- Researcher's Requirements

- Gathered at BPA coordinated Marine Microbes Meeting March 2018

- Functionality to analyse OTU data across sites/populations:
  - Ordination
  - Clustering
  - Rarefaction
  - Search by Sequence Similarity
  - Plots (Box, Bar, Venn etc)
  - Maps (bubbleplots, heatmaps)
  - Browsing Taxonomic abundance data
  - BYO OTU data

- Prioritised by Andrew Bissett
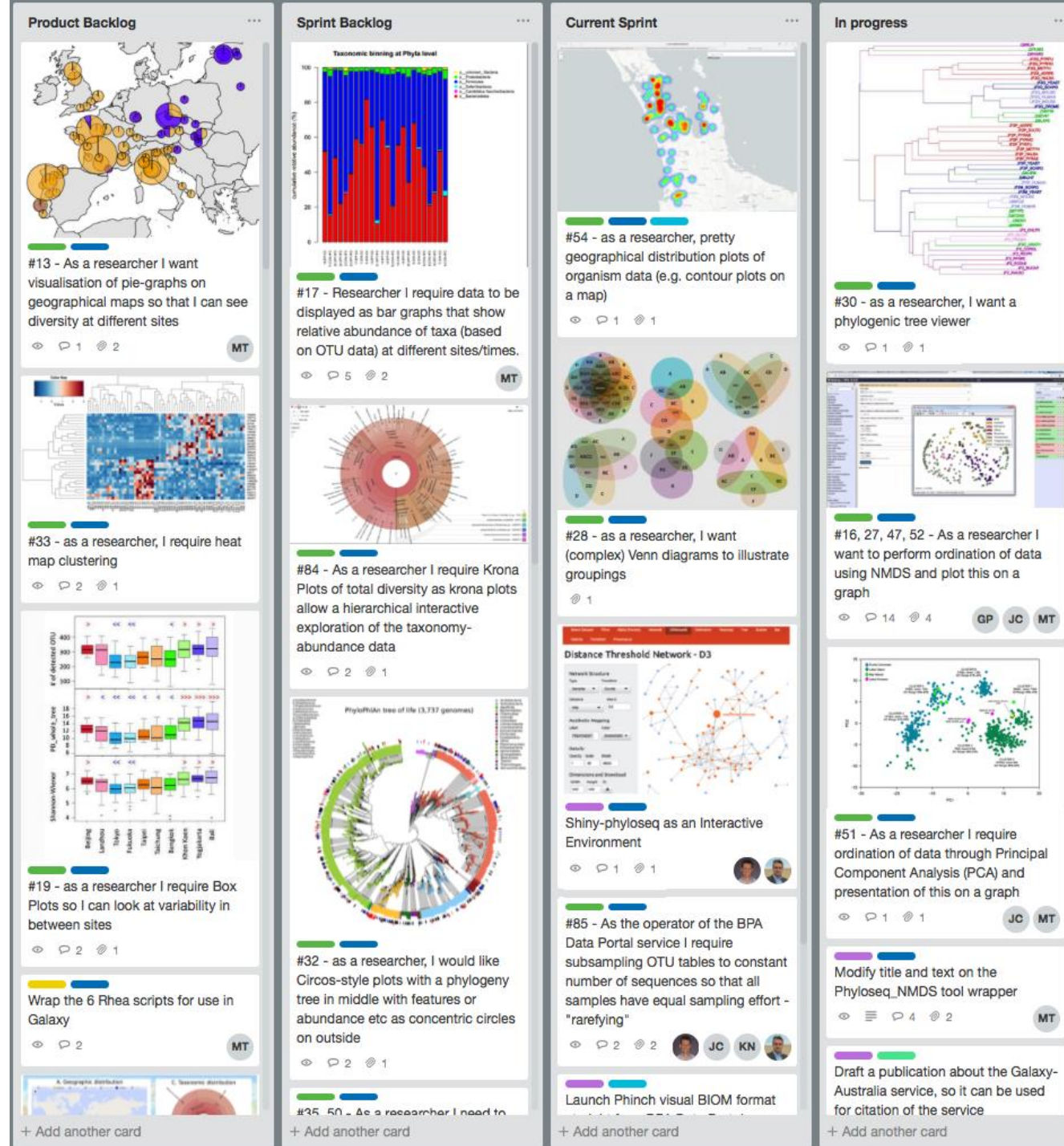
Workflow:
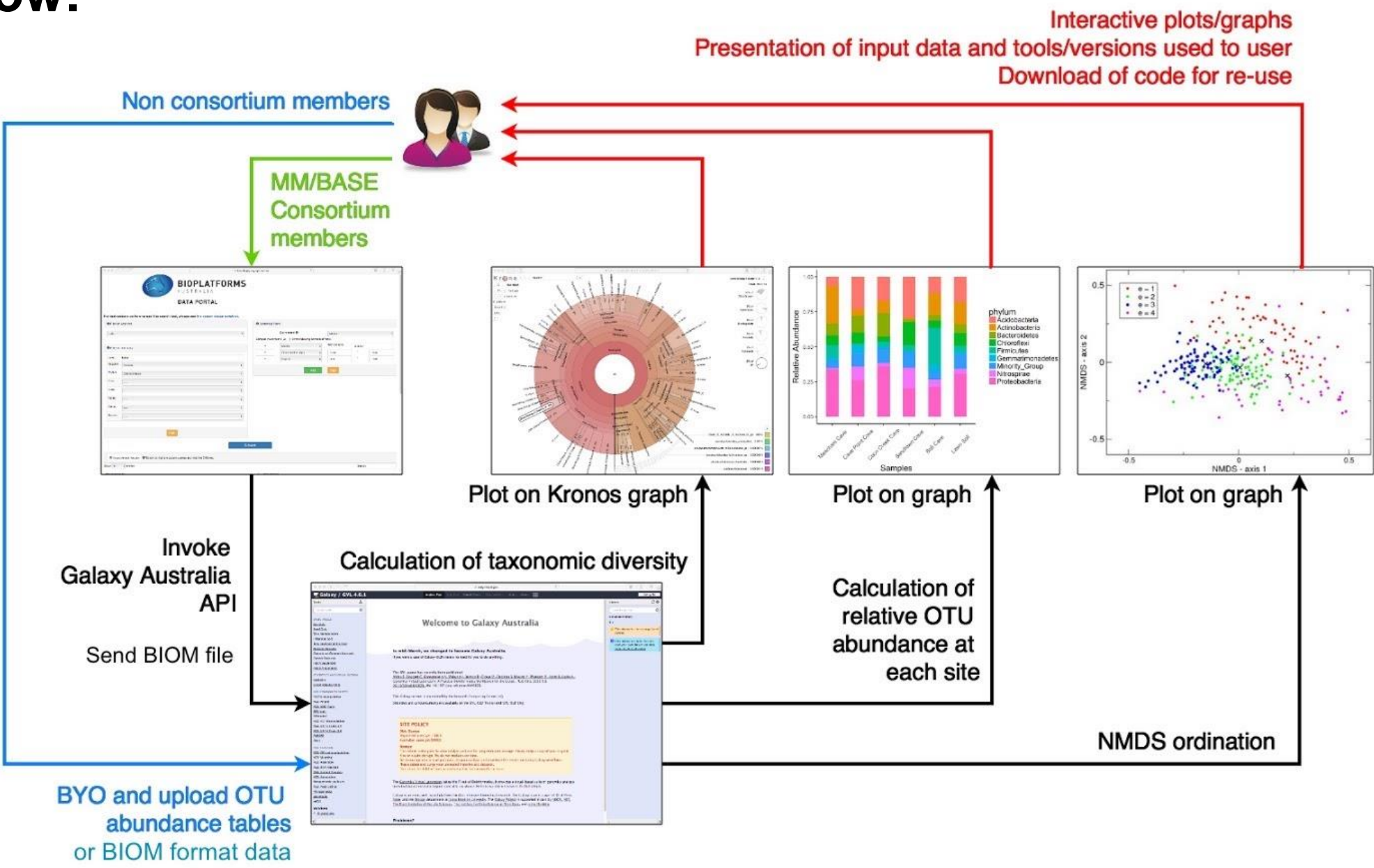
What have we developed so far?

What have we developed so far?

NOTE: ALL DEMOS ARE DONE ON STAGING SERVERS (BPA DATA PORTAL AND GALAXY AUSTRALIA)

FUNCTIONALITY WILL BE RELEASED TO PRODUCTION SERVERS ONLY AFTER THOROUGH QUALITY ASSURANCE, THROUGHOUT Q4 2018

BIOPLATFORMS
AUSTRALIA

Australian Microbiome Home   Search   Map   Contextual   👤 jeff.christiansen@qcif.edu.au

DATA PORTAL

Filter by Amplicon ⓘ and Taxonomy ⓘ

| Amplicon | is ⇕ | --- ⇕ |

Taxonomy

Kingdom	is ⇕	--- ⇕
Phylum	is ⇕	--- ⇕
Class	is ⇕	--- ⇕
Order	is ⇕	--- ⇕
Family	is ⇕	--- ⇕
Genus	is ⇕	--- ⇕
Species	is ⇕	--- ⇕

Clear

ⓘ Contextual Filters

| Environment ⓘ | is ⇕ | --- ⇕ |

Contextual Filters

Add   Clear

🔍 Search

🌐 Show results on Map   📥 Submit to Galaxy   📥 Export Search Results (CSV)   📥 Export Search Results (BIOM)

Account made for BPA Data Portal users on Galaxy Australia

Data sent via Galaxy API

Currently this takes a little while

~3mins later…

Species | is ⇅ | --- ⇅

Clear

🔍 Search

🌐 Show results on Map | 🖥 Submit to Galaxy | ⬇ Export Search Results (CSV) | ⬇ Export Search Results (BIOM)

Submission to Galaxy in Progress ...   ✕

If you are new to Galaxy Australia, please see this Getting started guide   ✕

BPA Sample ID	BPA Project
7035	Soil
7036	Soil
7049	Soil
7050	Soil
7051	Soil
7052	Soil
7061	Soil
7062	Soil
7067	Soil
7068	Soil

Previous | Page 1 ⇅ of 31 | 10 rows ⇅ | Next

Currently this also takes a little while

~4mins later…

Plenty of options for increasing speed
e.g. farming out jobs from specific tools
to various backends (cloud, HPC etc)

Bioplatforms OTU Search Facility

Galaxy | Australia

Galaxy / Australia

Analyze Data   Workflow   Visualize ▾   Shared Data ▾   Help ▾   User ▾

Using 2%

❌ This is the galaxy.au staging server. Please email d@uq.edu.au if the are problems with the service.

Tools

phyloseq

Metagenomic analyses

Phyloseq Biom Filtering biom file filter

Phyloseq Ordination Plot ordination plotting

FROGSSTAT Phyloseq Multivariate Analysis Of Variance

FROGSSTAT Phyloseq Sample Clustering of samples using different linkage methods

FROGSSTAT Phyloseq Import Data from 3 files: biomfile, samplefile, treefile

FROGSSTAT Phyloseq Composition Visualisation with bar plot and composition plot

FROGSSTAT Phyloseq Alpha Diversity with richness plot

FROGSSTAT Phyloseq Beta Diversity distance matrix

FROGSSTAT Phyloseq Structure Visualisation with heatmap plot and ordination plot

Workflows

- All workflows

✅ 1 job has been successfully added to the queue – resulting in the following datasets:

2: Phyloseq Ordination Plot – PCoA (Principal Coordinate Analysis)_bray (Bray-Curtis distance).html

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

History

search datasets

Australian Microbiome: amplicon is '27f519r_bacteria'; phylum is 'Bacteroidetes'; environment is 'Soil'; Vegetation Type is 'Grassland'; Vegetation Type is 'Forest'

2 shown

20.87 MB

2: Phyloseq Ordination Plot – PCoA (Principal Coordinate Analysis)_bray (Bray-Curtis distance).html

1: AustralianMicrobiome -2018-10-09T081822.biom.zip

Ordination Plots

Utilises..

- R script from the Phyloseq package…

phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data

Paul J. McMurdie, Susan Holmes*

Department of Statistics, Stanford University, Stanford, California, United States of America

Abstract

Background: The analysis of microbial communities through DNA sequencing brings many challenges: the integration of different types of data with methods from ecology, genetics, phylogenetics, multivariate statistics, visualization and testing. With the increased breadth of experimental designs now being pursued, project-specific statistical analyses are often needed, and these analyses are often difficult (or impossible) for peer researchers to independently reproduce. The vast majority of the requisite tools for performing these analyses reproducibly are already implemented in R and its extensions (packages), but with limited support for high throughput microbiome census data.

Results: Here we describe a software project, phyloseq, dedicated to the object-oriented representation and analysis of microbiome census data in R. It supports importing data from a variety of common formats, as well as many analysis techniques. These include calibration, filtering, subsetting, agglomeration, multi-table comparisons, diversity analysis, parallelized Fast UniFrac, ordination methods, and production of publication-quality graphics; all in a manner that is easy to document, share, and modify. We show how to apply functions from other R packages to phyloseq-represented data, illustrating the availability of a large number of open source analysis techniques. We discuss the use of phyloseq with tools for reproducible research, a practice common in other fields but still rare in the analysis of highly parallel microbiome census data. We have made available all of the materials necessary to completely reproduce the analysis and figures included in this article, an example of best practices for reproducible research.

Conclusions: The phyloseq project for R is a new open-source software package, freely available on the web from both GitHub and Bioconductor.

Ordination Plots

Utilises..
- R script from the Phyloseq package   'wrapped' for use in the Galaxy GUI

Other tools to include:

Other tools to include:

Many are available in the Galaxy Toolshed "App Store", so trivial to install locally:

Other tools to include:

Many are available in the Galaxy Toolshed "App Store", so trivial to install locally:

e.g. Mothur, QIIME, Krona have been installed

Other tools to include:

We can also 'wrap' any R, python, or other script/command:

Other tools to include:

We can also 'wrap' any R, python, or other script/command:

e.g.   - More Phyloseq functions
         - Rhea pipeline (for normalization, alpha diversity, beta diversity,
                    Taxonomic binning, Serial Group Comparisons, Correlations):

Three user types to support

#35, 50 - As a researcher I need to export or be able to re-access the code I used in an analysis in this system, so I can re-deploy/re-use the code in future analyses - on this or other systems

👁 💬 1

The system to be deployed must support users who require a Graphical User Interface to conduct complex analyses

📎 1

#93 - As a researcher, I require a 'one click biogeography' service where I can upload a sequence, and retrieve information about samples where species correspond to that sequence are found, as well as information on the taxonomic diversoty at those sites, so I can quickly assess environmental metagenomics survey data that is relevant to my own species of interest

👁 💬 2 📎 2

R/python code users   GUI users   One click

Three user types to support

#35, 50 - As a researcher I need to export or be able to re-access the code I used in an analysis in this system, so I can re-deploy/re-use the code in future analyses - on this or other systems

R/python code users

The system to be deployed must support users who require a Graphical User Interface to conduct complex analyses

GUI users

#93 - As a researcher, I require a 'one click biogeography' service where I can upload a sequence, and retrieve information about samples where species correspond to that sequence are found, as well as information on the taxonomic diversoty at those sites, so I can quickly assess environmental metagenomics survey data that is relevant to my own species of interest

One click

Three user types to support

#35, 50 - As a researcher I need to export or be able to re-access the code I used in an analysis in this system, so I can re-deploy/re-use the code in future analyses - on this or other systems

👁 💬 1

The system to be deployed must support users who require a Graphical User Interface to conduct complex analyses

📎 1

#93 - As a researcher, I require a 'one click biogeography' service where I can upload a sequence, and retrieve information about samples where species correspond to that sequence are found, as well as information on the taxonomic diversoty at those sites, so I can quickly assess environmental metagenomics survey data that is relevant to my own species of interest

👁 💬 2 📎 2

R/python code users   GUI users   One click

Direct launch of R-studio in the hosted Galaxy environment

Three user types to support

#35, 50 - As a researcher I need to export or be able to re-access the code I used in an analysis in this system, so I can re-deploy/re-use the code in future analyses - on this or other systems

The system to be deployed must support users who require a Graphical User Interface to conduct complex analyses

A. Geographic distribution   C. Taxonomic distribution

B. Environmental context   D. Downloadable flat files

Alignment results file
Hits sequences
Hits abundances
Environmental data file

#93 - As a researcher, I require a 'one click biogeography' service where I can upload a sequence, and retrieve information about samples where species correspond to that sequence are found, as well as information on the taxonomic diversoty at those sites, so I can quickly assess environmental metagenomics survey data that is relevant to my own species of interest

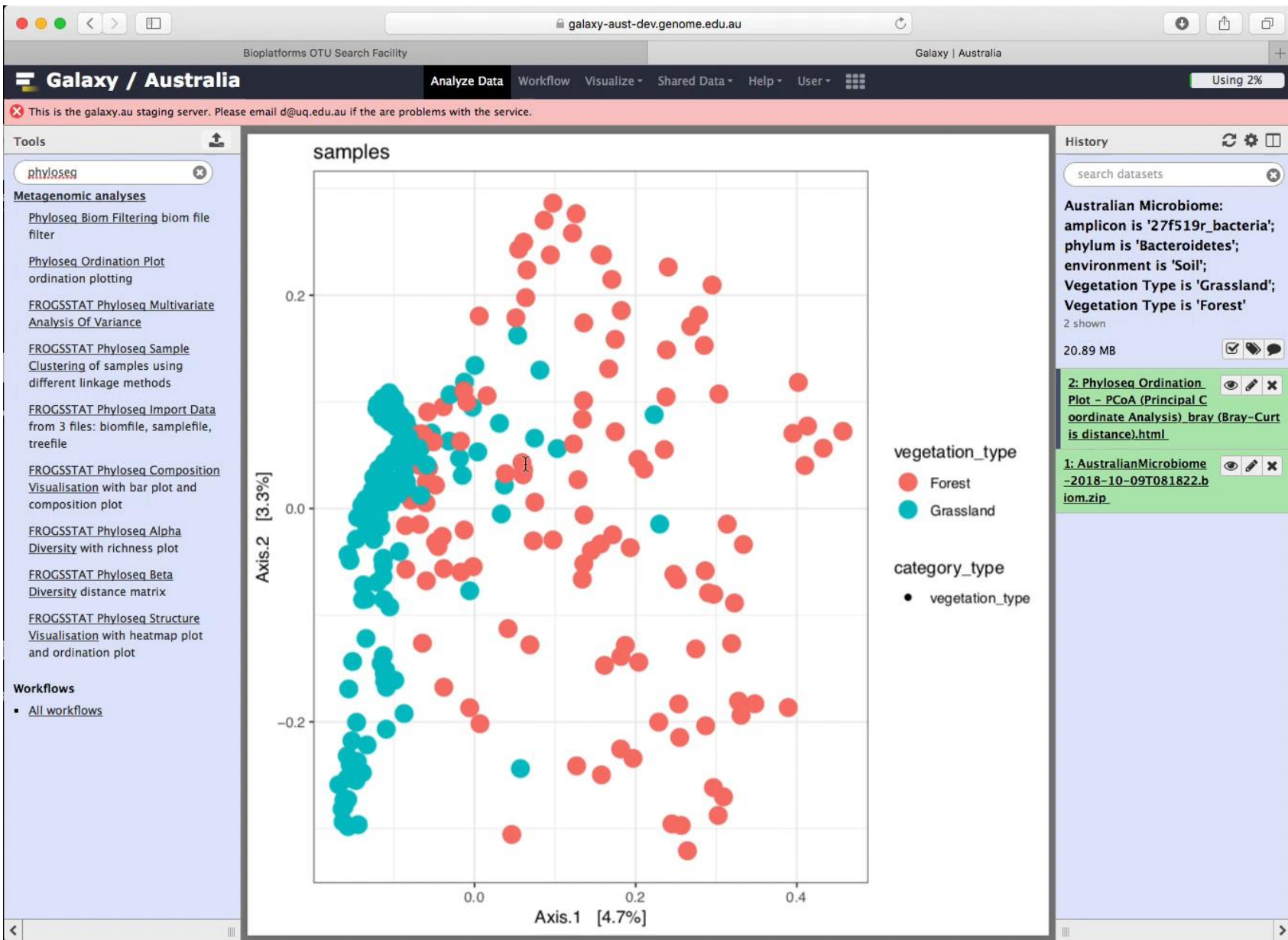R/python code users   GUI users   One click

Generating Reports

- Can simply invoke multiple Galaxy Tools at once as a 'workflow'

- Then present the outputs to the user in a webpage e.g.

- Could send a link via email

Status (I)

- Eight months into ten month project

- Demo system available for testing

- Recent/upcoming usability sessions (Brisbane, Melbourne, Hobart)

- As tools/functions pass QC, they will be deployed to production servers
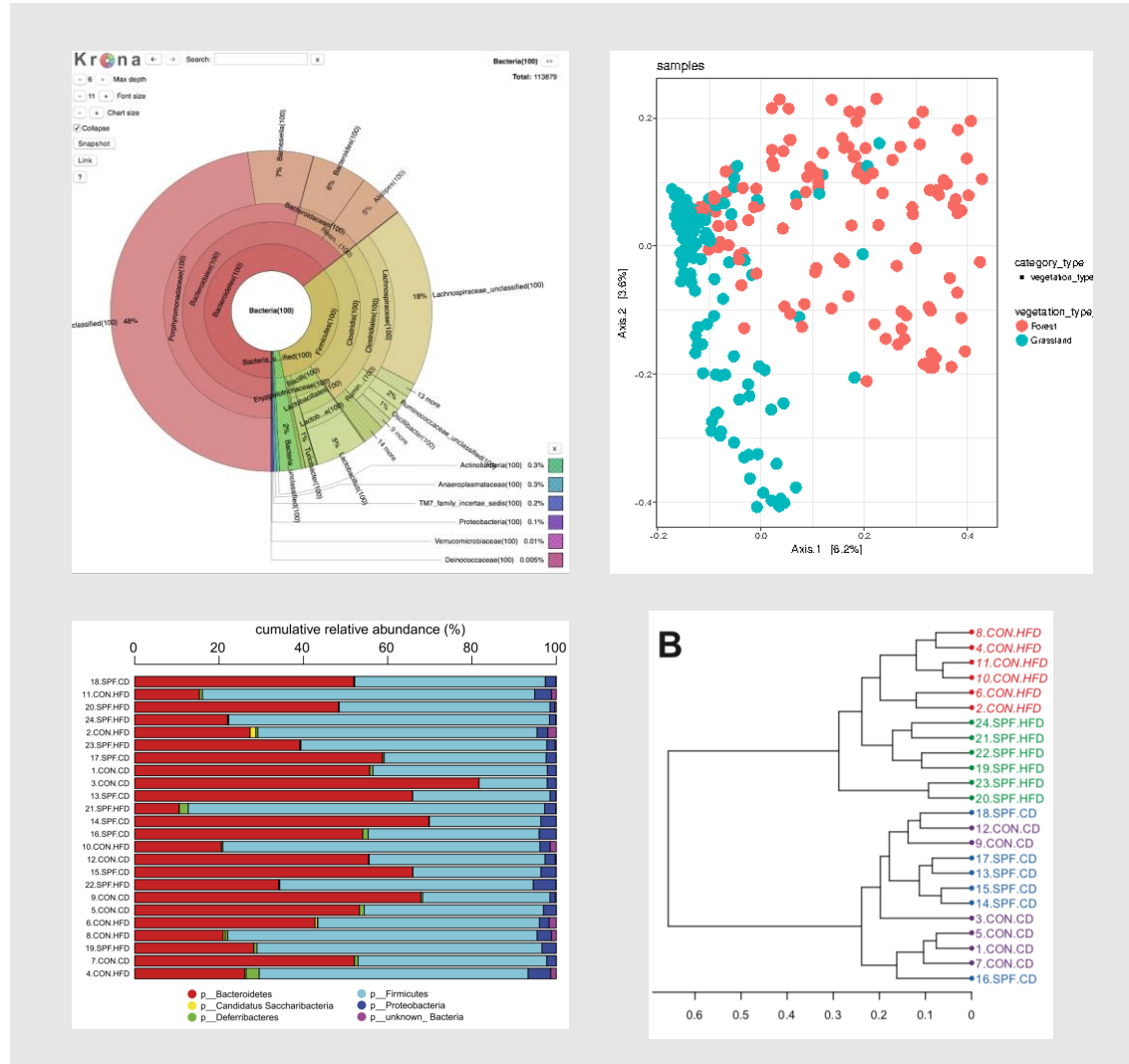
- Short-list of tools/functions to add (sequence search, bar plots, bubble plots)

- Development of on-line training material
  (to be available via EcoED and Galaxy!Training portals)

- One three-hour training session planned - Nov 14th via EMBL-ABR 'hybrid' delivery model at several sites (currently: SA, Tas, NSW, Vic, Qld)

Status (II)

- New commitment from CSIRO, BPA, Cth Dept Environment (Director of National Parks), iMOS) to develop a comprehensive national soil, marine and freshwater microbiome survey

- Ongoing and critical need to make the information within metagenomics data accessible to wide range of user communities:
  - Researchers
  - Land/Water Managers
  - National Parks and Conservation Organisations
  - Farmers/Fisheries

- The system developed represents the initial foundations for such a cloud-based analysis tool for environmental metagenomics data

Development

Jeff Christiansen
Derek Benson
Gareth Price
Mike Thang

Grahame Bowland
Tamas Szabo
Samuel Chang

Simon Gladman
Anna Syme

Andrew Bissett
Sophie Mazard

Michael Hope
Corinna Paeper

https://bioscience-rdc.blogspot.com

Funding