The Galaxy service pilot in CSIRO

A collaboration between science and IT

Steve McMahon, Philippe Moncuquet, Sean Li, Ondrej Hlinka, Josh Bowden, Sean McWilliam and Annette McGrath

CSIRO BIOINFORMATICS CORE, INFORMATION MANAGEMENT & TECHNOLOGY
www.csiro.au

Galaxy (1) (2) (3) is a web based bioinformatics toolkit with the ability to perform, reproduce and share complete analyses.

A Galaxy service pilot was set up in CSIRO (Australia's Commonwealth Scientific and Industrial Research Organisation) for the benefit of biologists and bioinformaticians. The service pilot was implemented as a collaboration between CSIRO's Information Management and Technology staff (IM&T) and the CSIRO Bioinformatics Core (CBC). This made the best use of the IT infrastructure and service delivery expertise (through IM&T) and the bioinformatics domain expertise (through CBC).

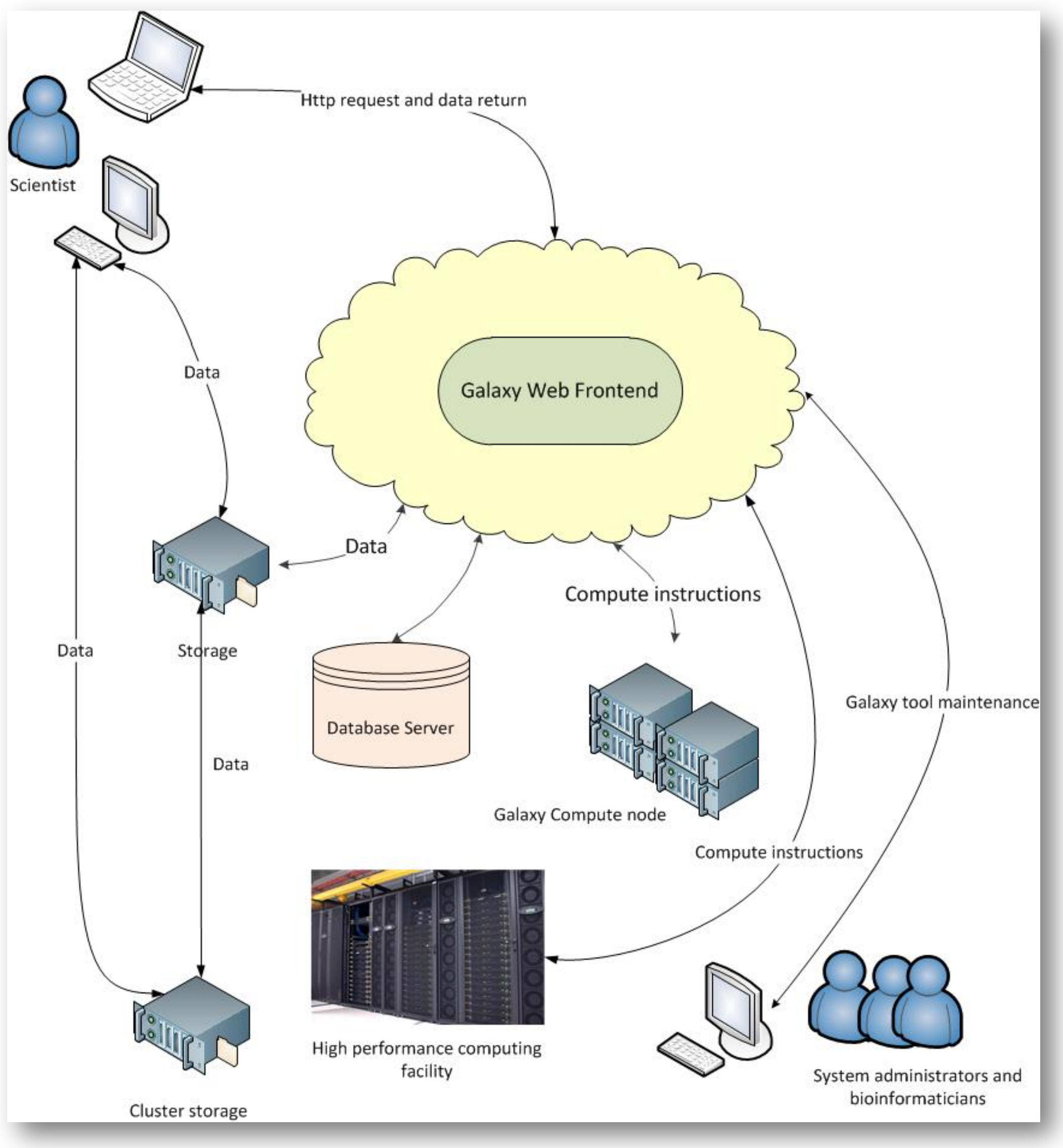
This presentation outlines Galaxy, the way it was implemented in CSIRO as a service pilot and some of the outcomes and related experiences.

Problem

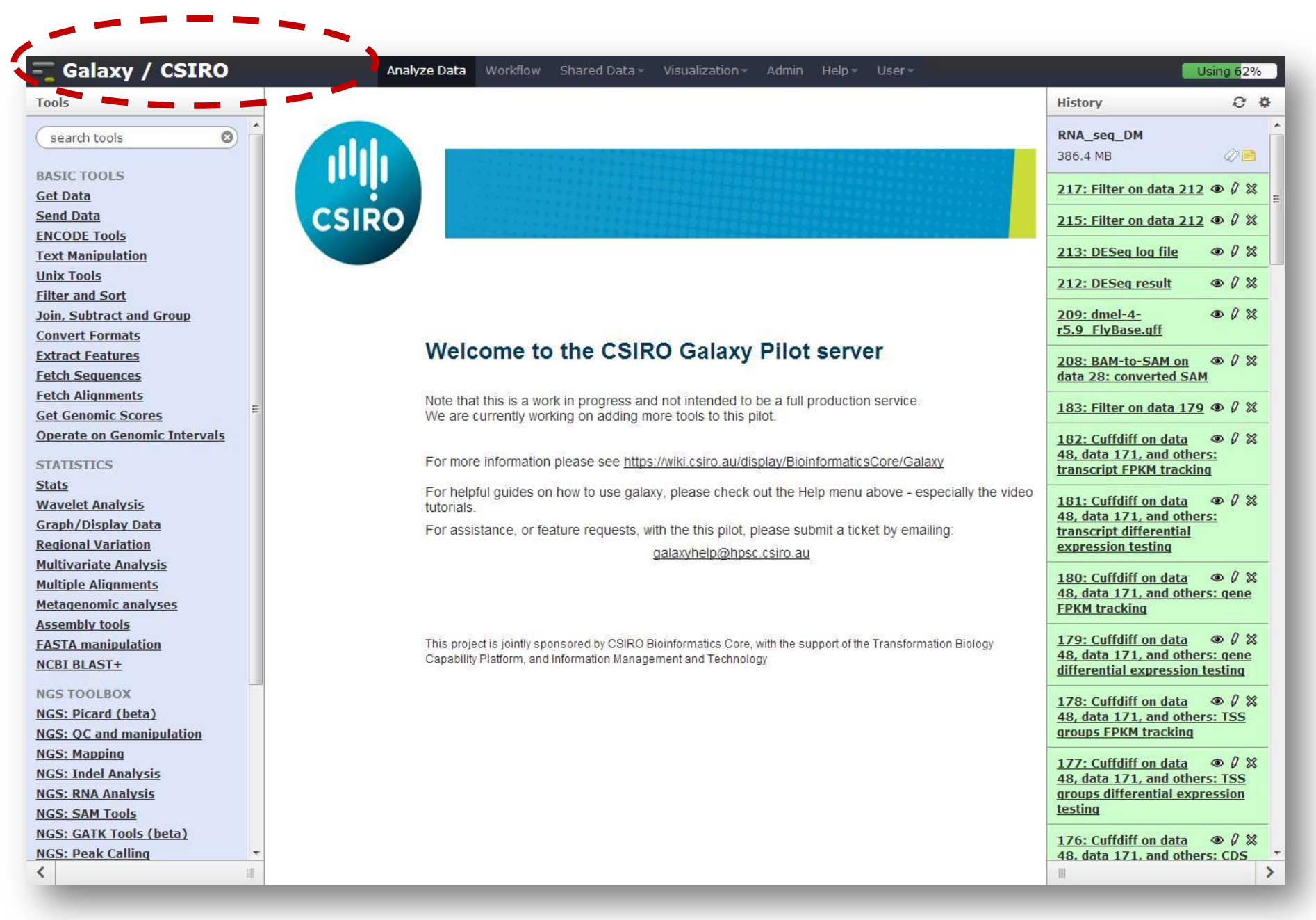
As next generation sequencing becomes more affordable, more experiments requiring bioinformatic analysis are designed by biologists. In CSIRO they had been relying on a limited number of skilled bioinformaticians to carry out this analysis. Alternatively they may have chosen to educate themselves in these same skills which is time consuming. Analysis was not always performed on the best computing resources and not always in the most optimal way. It was proposed that a Galaxy instance would **empower biologists** to perform their own data analysis by providing easy access to bioinformatics analysis tools. The success of this project relied on both scalable hardware solutions and bioinformatics knowledge and support.

Architecture

An effective bioinformatics platform needs to be scalable as computational power and storage needs are variable. The CSIRO Galaxy pilot project was designed to service approximately **100 users**. As a pilot it had to be capable enough to be useful whilst not being too expensive.



The infrastructure consisted of 2 physical machines – one to host the Galaxy database and the other to run Galaxy and be the general compute engine. There were also **2 virtual machines** running the web front ends for a production instance and a development instance of Galaxy. The compute machine had **32 cores** and **192 GB RAM** and ran the Torque batch system.



Welcome to the CSIRO Galaxy Pilot server

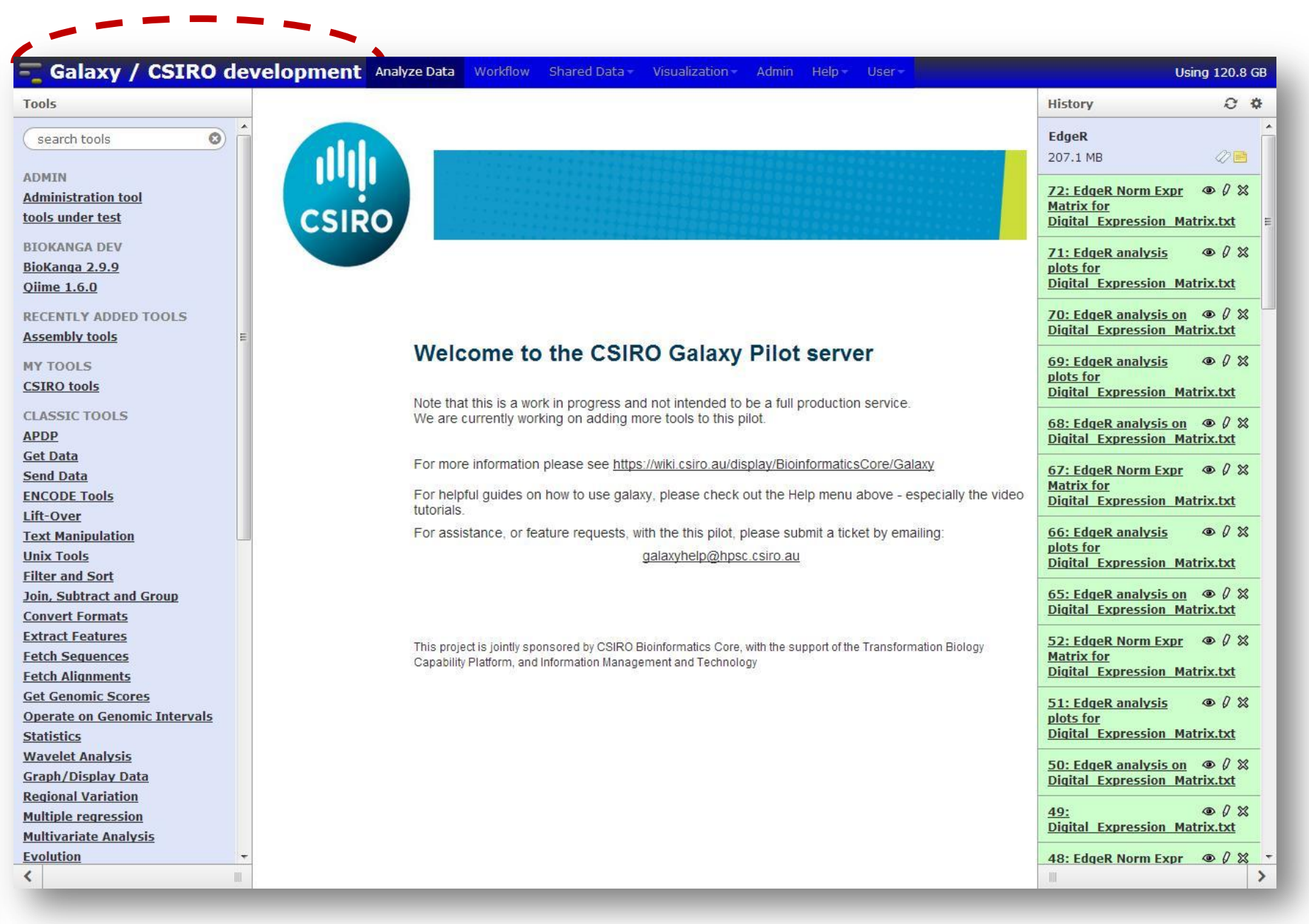
Note that this is a work in progress and not intended to be a full production service.
We are currently working on adding more tools to this pilot.

For more information please see https://wiki.csiro.au/display/BioinformaticsCore/Galaxy

For helpful guides on how to use galaxy, please check out the Help menu above - especially the video tutorials.

For assistance, or feature requests, with the this pilot, please submit a ticket by emailing:
galaxyhelp@hpsc.csiro.au

This project is jointly sponsored by CSIRO Bioinformatics Core, with the support of the Transformation Biology Capability Platform, and Information Management and Technology


Galaxy / CSIRO development
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 120.8 GB
Tools
search tools
ADMIN
Administration tool
tools under test
BIOKANGA DEV
BioKanga 2.9.9
Qiime 1.6.0
RECENTLY ADDED TOOLS
Assembly tools
MY TOOLS
CSIRO tools
CLASSIC TOOLS
APDP
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Unix Tools
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Wavelet Analysis
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
History
EdgeR
207.1 MB
72: EdgeR Norm Expr Matrix for Digital_Expression_Matrix.txt
71: EdgeR analysis plots for Digital_Expression_Matrix.txt
70: EdgeR analysis on Digital_Expression_Matrix.txt
69: EdgeR analysis plots for Digital_Expression_Matrix.txt
68: EdgeR analysis on Digital_Expression_Matrix.txt
67: EdgeR Norm Expr Matrix for Digital_Expression_Matrix.txt
66: EdgeR analysis plots for Digital_Expression_Matrix.txt
65: EdgeR analysis on Digital_Expression_Matrix.txt
52: EdgeR Norm Expr Matrix for Digital_Expression_Matrix.txt
51: EdgeR analysis plots for Digital_Expression_Matrix.txt
50: EdgeR analysis on Digital_Expression_Matrix.txt
49: Digital_Expression_Matrix.txt
48: EdgeR Norm Expr

Welcome to the CSIRO Galaxy Pilot server

Note that this is a work in progress and not intended to be a full production service.
We are currently working on adding more tools to this pilot.

For more information please see https://wiki.csiro.au/display/BioinformaticsCore/Galaxy

For helpful guides on how to use galaxy, please check out the Help menu above - especially the video tutorials.

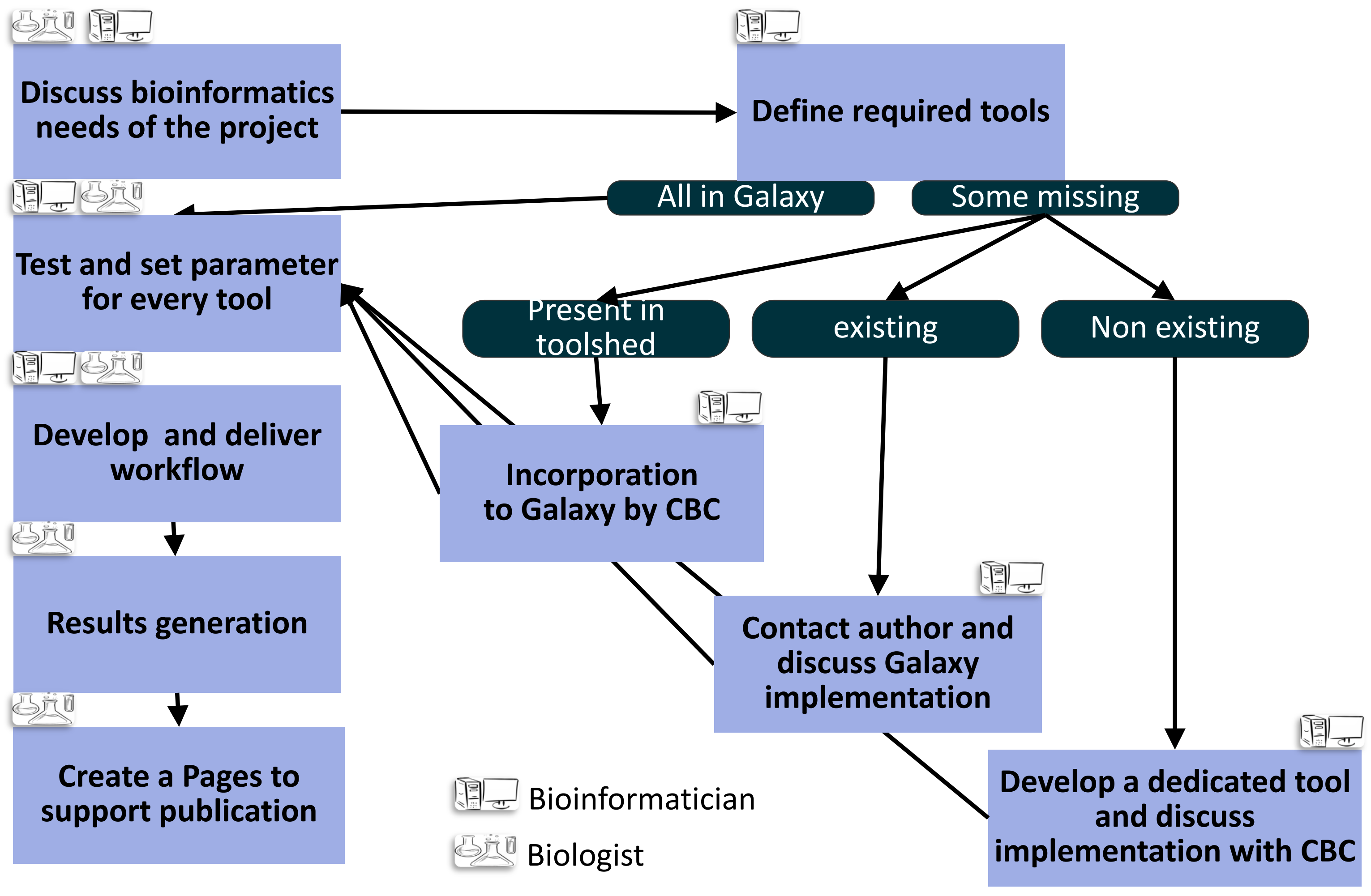
For assistance, or feature requests, with the this pilot, please submit a ticket by emailing:
galaxyhelp@hpsc.csiro.au

This project is jointly sponsored by CSIRO Bioinformatics Core, with the support of the Transformation Biology Capability Platform, and Information Management and Technology

As the pilot was being developed the storage allocated to it went from about 1TB to 5TB and then later to 16TB. The available storage seemed to be a key factor for whether some users made use of the pilot.

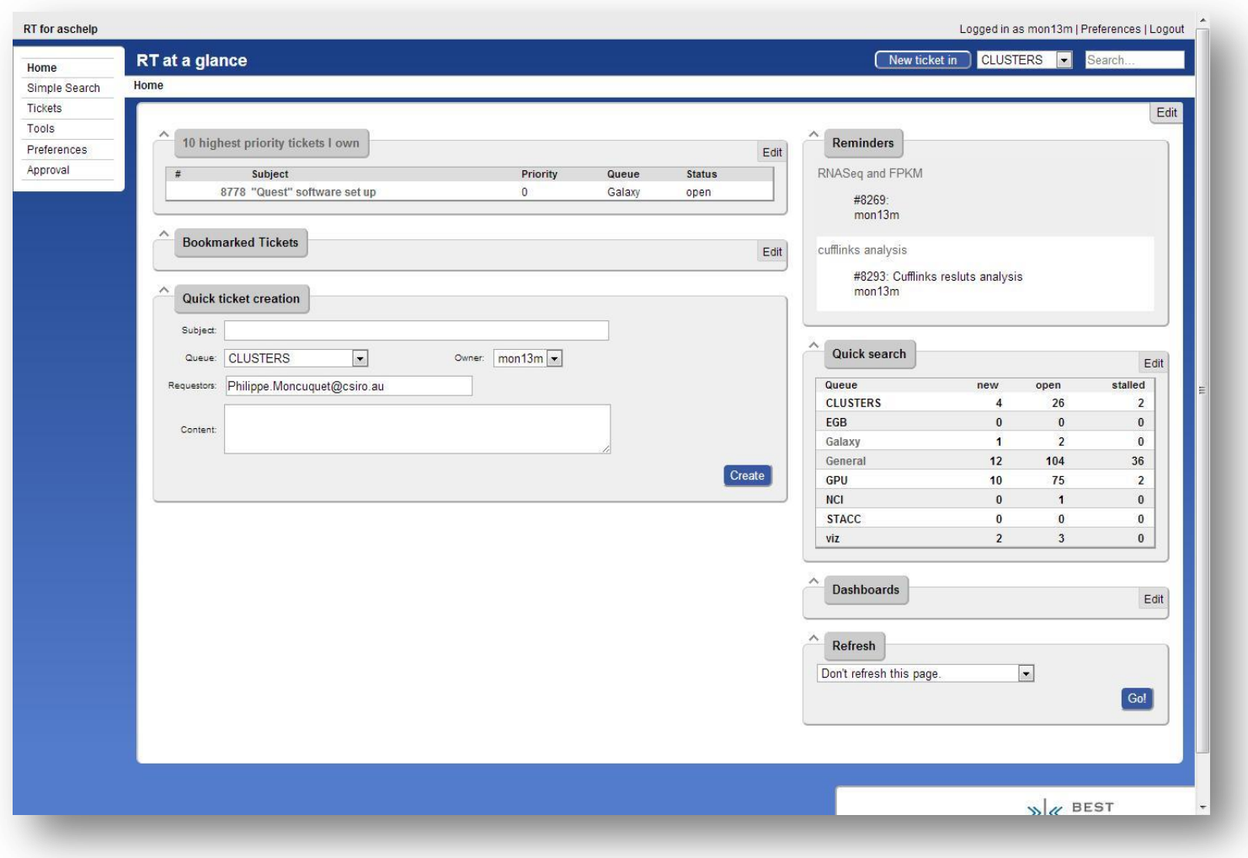
A typical NGS project

A good collaboration between biologist and bioinformatician is essential for any successful NGS experiment. Galaxy offers the means to design the analysis of an experiment from beginning to end. Galaxy workflows are a great option to **tackle reproducibility issues** and enhance sharing of the scientific work.


Discuss bioinformatics needs of the project
Define required tools
All in Galaxy
Some missing
Test and set parameter for every tool
Present in toolshed
existing
Non existing
Develop and deliver workflow
Incorporation to Galaxy by CBC
Results generation
Contact author and discuss Galaxy implementation
Create a Pages to support publication
Develop a dedicated tool and discuss implementation with CBC
Bioinformatician
Biologist

Support, development and user education are key

As for all web-based platforms, development, maintenance, **support and user training** play a central role in our Galaxy production instance. Maintenance and development were dealt with by both bioinformaticians and IT staff. Development followed an "agile" process with issues managed in Jira – web based software suited to this process. User initiated issues were submitted to RT (Request Tracker), another issue tracker. The vast majority of the user support queries were related to the correct choice and use of bioinformatics tools.



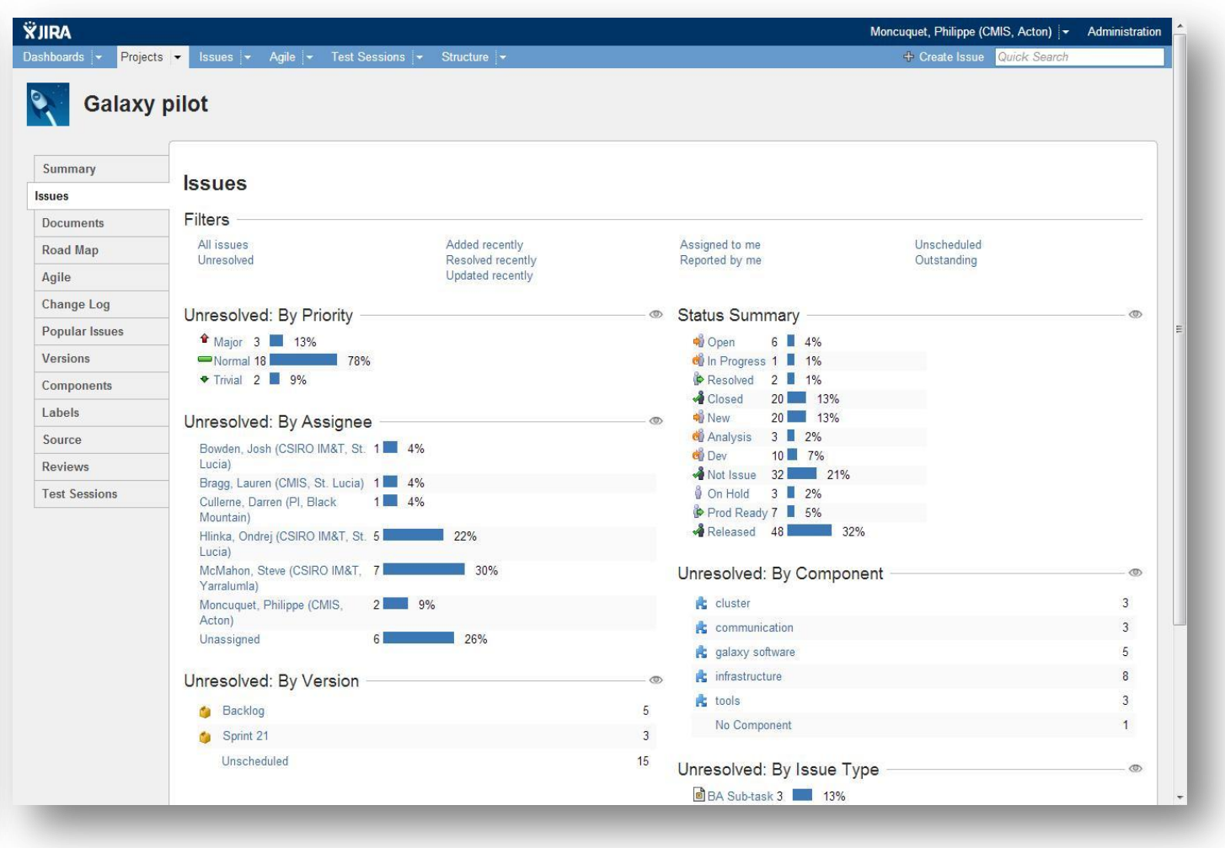
10 highest priority tickets I own

#	Subject	Priority	Queue	Status
8778	"Quest" software set up	0	Galaxy	open

RNASeq and FPKM
#8269: mon13m

cufflinks analysis
#8293: Cufflinks resluts analysis mon13m

Queue	new	open	stalled
CLUSTERS	4	26	2
EGB	0	0	0
Galaxy	1	2	0
General	12	104	36
GPU	10	75	2
NCI	0	1	0
STACC	0	0	0
viz	2	3	0



Unresolved: By Priority

Major	3	13%
Normal	18	78%
Trivial	2	9%

Unresolved: By Assignee

Bowden, Josh (CSIRO IM&T, St. Lucia)	1	4%
Bragg, Lauren (CMIS, St. Lucia)	1	4%
Cullerne, Darren (PI, Black Mountain)	1	4%
Hlinka, Ondrej (CSIRO IM&T, St. Lucia)	5	22%
McMahon, Steve (CSIRO IM&T, Yarralumla)	7	30%
Moncuquet, Philippe (CMIS, Acton)	2	9%
Unassigned	6	26%

Unresolved: By Version

Backlog	5
Sprint 21	3
Unscheduled	15

Status Summary

Open	6	4%
In Progress	1	1%
Resolved	2	1%
Closed	20	13%
New	20	13%
Analysis	3	2%
Dev	10	7%
Not Issue	32	21%
On Hold	3	2%
Prod Ready	7	5%
Released	48	32%

Unresolved: By Component

cluster	3
communication	3
galaxy software	5
infrastructure	8
tools	3
No Component	1

Unresolved: By Issue Type

BA Sub-task	3	13%

User training was carried out via live demos. There are also plans for hands-on tutorials.

Pilot results

The pilot gathered **113 users**. More than 300 tools have been made available in the platform and over **1000 jobs** are launched each month. The pilot has been used in 1 submitted article so far. 19 workflows are shared with users as well as some useful datasets.

Since the announcement of the pilot last October, more than 60 support tickets have been created and resolved.

The pilot was considered a great success showing how CSIRO IT and science staff could work together to achieve project goals and a full production service is being planned.

FOR FURTHER INFORMATION

Steve McMahon
steve.mcmahon@csiro.au
Philippe Moncuquet
philippe.moncuquet@csiro.au
or visit www.csiro.au

REFERENCES

1. *Galaxy: a web-based genome analysis tool for experimentalists.* **Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J.** January 2010, Current Protocols in Molecular Biology, pp. 1-21.

2. *Galaxy: a platform for interactive large-scale genome analysis.* **Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A.** 10, October 2005, Genome Research, Vol. 15, pp. 1451-5.

3. *Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.* **Goecks, J, Nekrutenko, A, Taylor, J and The Galaxy Team.** 8, August 2010, Genome Biol., Vol. 11.