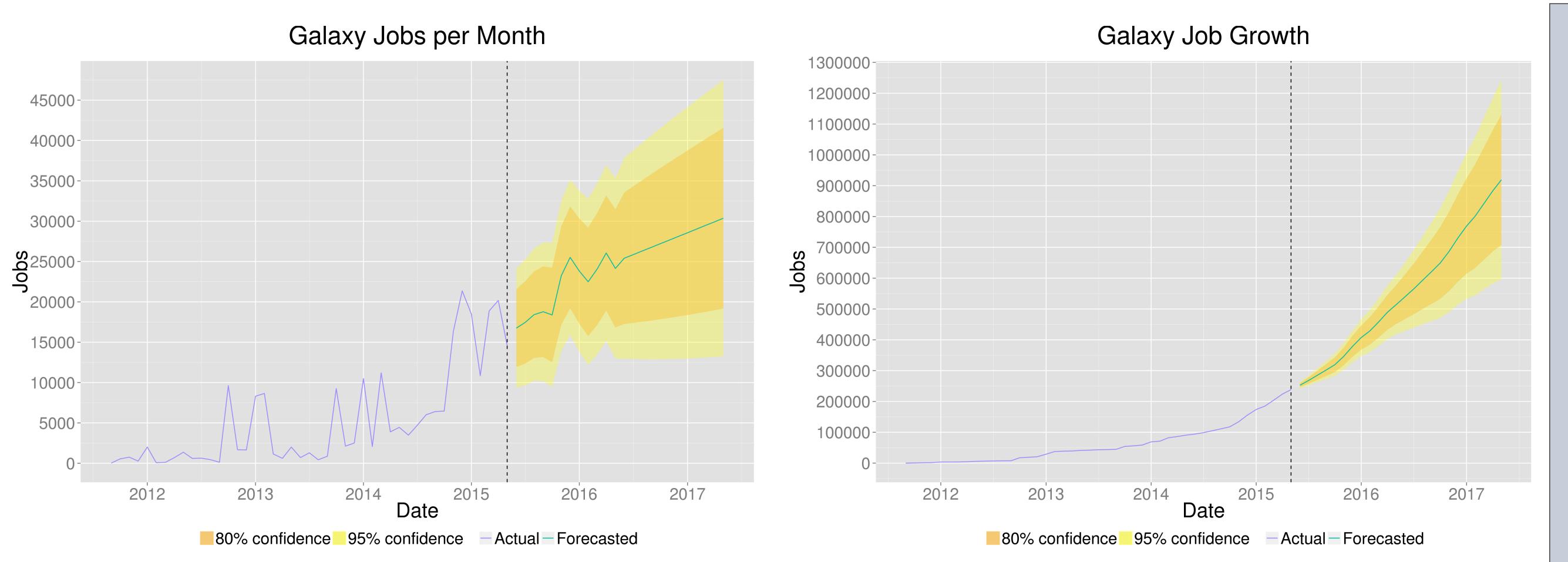
Bioinformatics Evolving at Canada's National Microbiology Laboratory

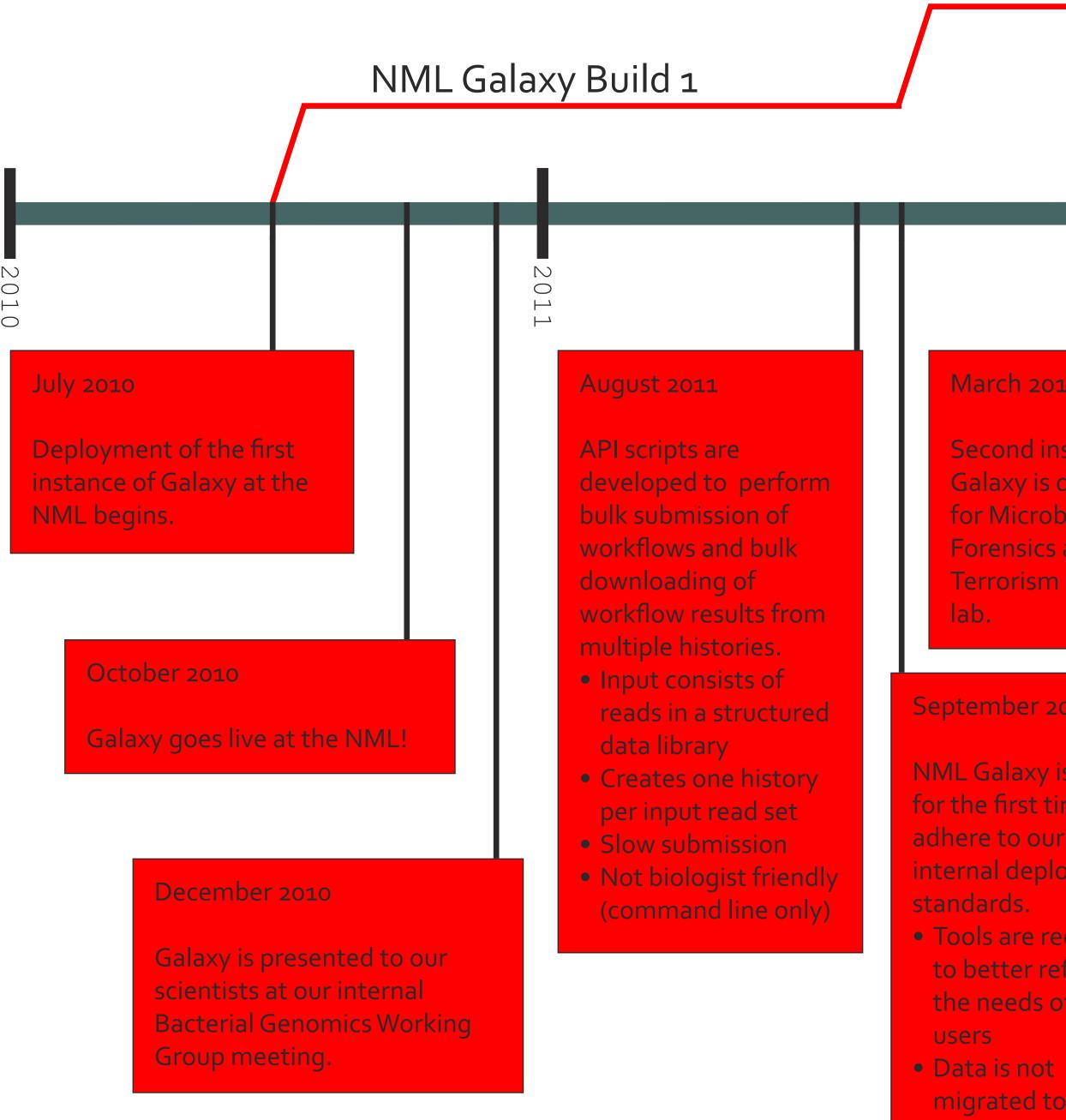
Eric Enns¹, Philip Mabon¹, Jennifer Cabral^{1,2}, Mariam Iskander^{1,2}, Cameron Sieffert¹, Natalie Knox¹, Heather Kent¹, Shane Thiessen¹, Josh Adam¹, Aaron Petkau¹, Thomas Matthews¹, Franklin Bristow¹, Gary Van Domselaar^{1,2} ¹National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg, MB, Canada, ²Department of Computer Science, University of Manitoba, Winnipeg, MB, Canada

Abstract

The National Microbiology Laboratory (NML) is Canada's leading public health laboratory, responsible for the identification, control and prevention of infectious diseases. The bioinformatics core facility at the NML deployed our first instance of Galaxy^a in 2010. The introduction of the Galaxy platform has revolutionized bioinformatics at the NML by bridging the gap between bioinformaticians and biologists.

Prior to Galaxy, most of our in-house tools and pipelines required an extensive background in UNIX command line and high performance computing to operate. This requirement demanded that bioinformaticians be intimately involved in projects with significant computational requirements. Galaxy was selected to be the bioinformatics analysis platform at the NML, as it made our tools and pipelines accessible to biologists. Bioinformaticians are able to focus more time on tool and pipeline development, as their project involvement has been reduced. Biologists are able to perform analyses on their own as Galaxy lowers the barrier to carrying out complex analyses in a high performance computing environment. As a result NGS (Next-generation sequencing) projects are able to progress at a much faster rate. Moving forward, we are developing a Galaxy-powered infectious disease analysis platform for our standardized analyses while retaining our traditional Galaxy environment for ad hoc pipeline development and bioinformatics analysis.





Public Health Agency of Canada Agence de la santé publique du Canada

NML Galaxy Build 2

March 2012

Second instance o[.] Galaxy is deployed for Microbial Forensics and Bio-Terrorism Defence

20

September 2011

NML Galaxy is rebuilt for the first time to adhere to our new nternal deployment

Tools are reduced to better reflect the needs of our

migrated to the new build

ugust 2012

Development of Core SNV Single Nucleotide Variant) hylogenomics Pipeline begins.

- Generates a whole genome SNV-based phylogeny to
- depict the evolutionary relationship among a set of isolates or strains

NML Microbial Whole Genome Workshop now includes a Galaxy component.

- Labs first focus on command line usage, followed by Galaxy
- Bulk submission is still a barrier for large scale

projects

April 2012

uly 2013

First publication using our Core SNV

- Phylogenomics Pipeline.
- Evolutionary Dynamics of Vibrio cholera O1 following a Single-
- Source Introduction to Haitid

November 2013

IRIDA^b (Integrated Rapid Infectious) Disease Anlaysis) is deployed at the NML for storage and management of sequencing runs.

- Replaces API scripts for bulk uploading sequencing runs in Data Libraries
- Automatically performs quality assessment using FastQC^c

NML Galaxy Build 3

IRIDA

May 2014

Focus of NML Microbial Whole Genome Workshop switches from command e labs and tutorials to Galaxy labs and tutorials.

June 2014

Philip Mabon (NML) discovers dataset collections as developed by John Chilton, and starts assisting in testing and bug discovery.

October 2014

NML Galaxy is rebuilt for the second time, due to issues using MySOL.

- Previous version could no longer be upgraded
- Database engine is changed from MySQL to PostgreSQL to alleviate all our previous issues
- Data is not migrated to the new build

April 2015

Galaxy powered analysis engine switched on in IRIDA. Includes an Assembly and Annotation Worklow, and our Core SNV

Phylogenomics workflow

une 2015

Galaxv. • 1005 - 10005 of jobs per

April 2015

y default.

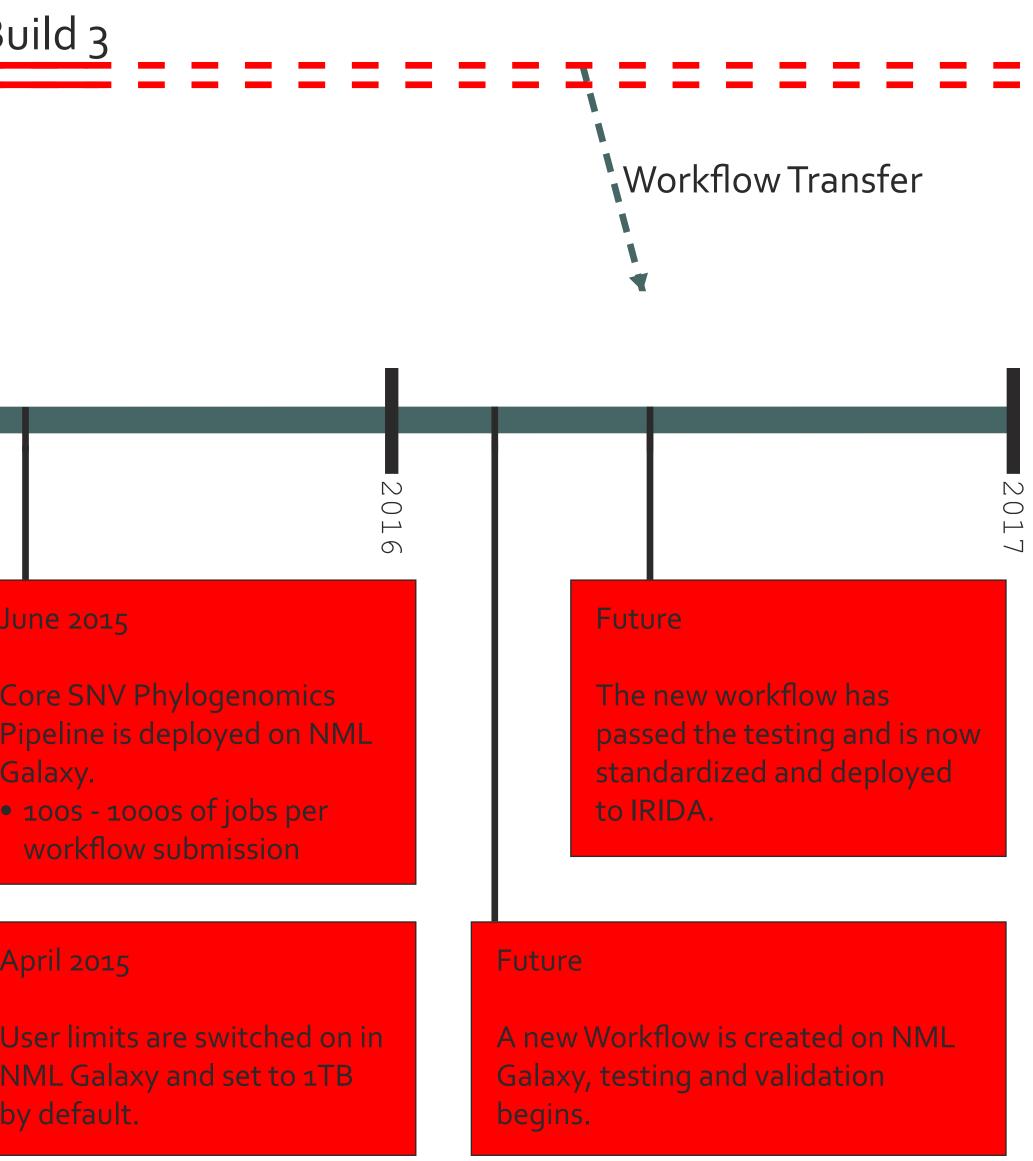
References

^bIRIDA: http://irida.ca

Conclusion

Until recently, the adoption of Galaxy at the NML has been slow due to an initial lack of focus on internal training and certain operational requirements not met by Galaxy. The two most important features implemented in Galaxy that have allowed us to rapidly push Galaxy into production are dataset collections and workflows.

We routinely perform large-scale analyses on hundreds of bacterial isolates in a single, batch job. Before the dataset collections feature was added to Galaxy, we resorted to writing custom API scripts for bulk submission of data and downloading of workflow results. These scripts were not biologist friendly because they could only be run from the command line. With the introduction of dataset collections, we were able to retire the API scripts for batch submission and allow biologists to solely use the Galaxy web interface for managing their data. We were also able to increase our focus on user training, further allowing biologists at the NML to take on their own bioinformatics analyses. By allowing biologists to manage their own data and analyses, the bioinformaticians at the NML are further able to concentrate their time on tool and pipeline development.



- ^aGalaxy: https://galaxyproject.org
- ^cFastQC: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ ^d Katz, L. S. et al. 2013. Evolutionary dynamics of Vibrio cholerae O1 following a single-source introduction to Haiti. MBio, 4, pp. eoo398–eoo413.
 - Canada