

# Bioinformatics Evolving at Canada's National Microbiology Laboratory

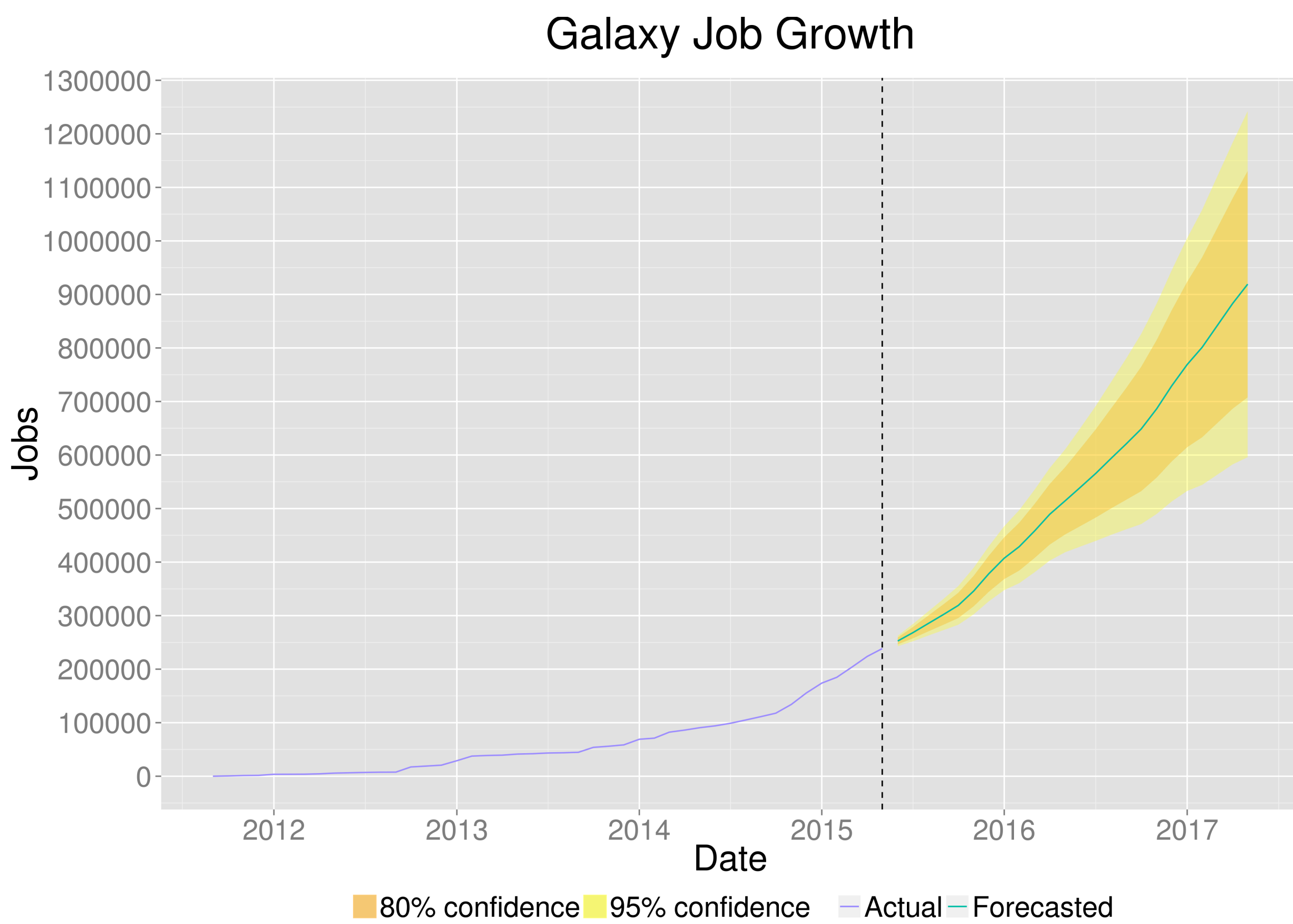
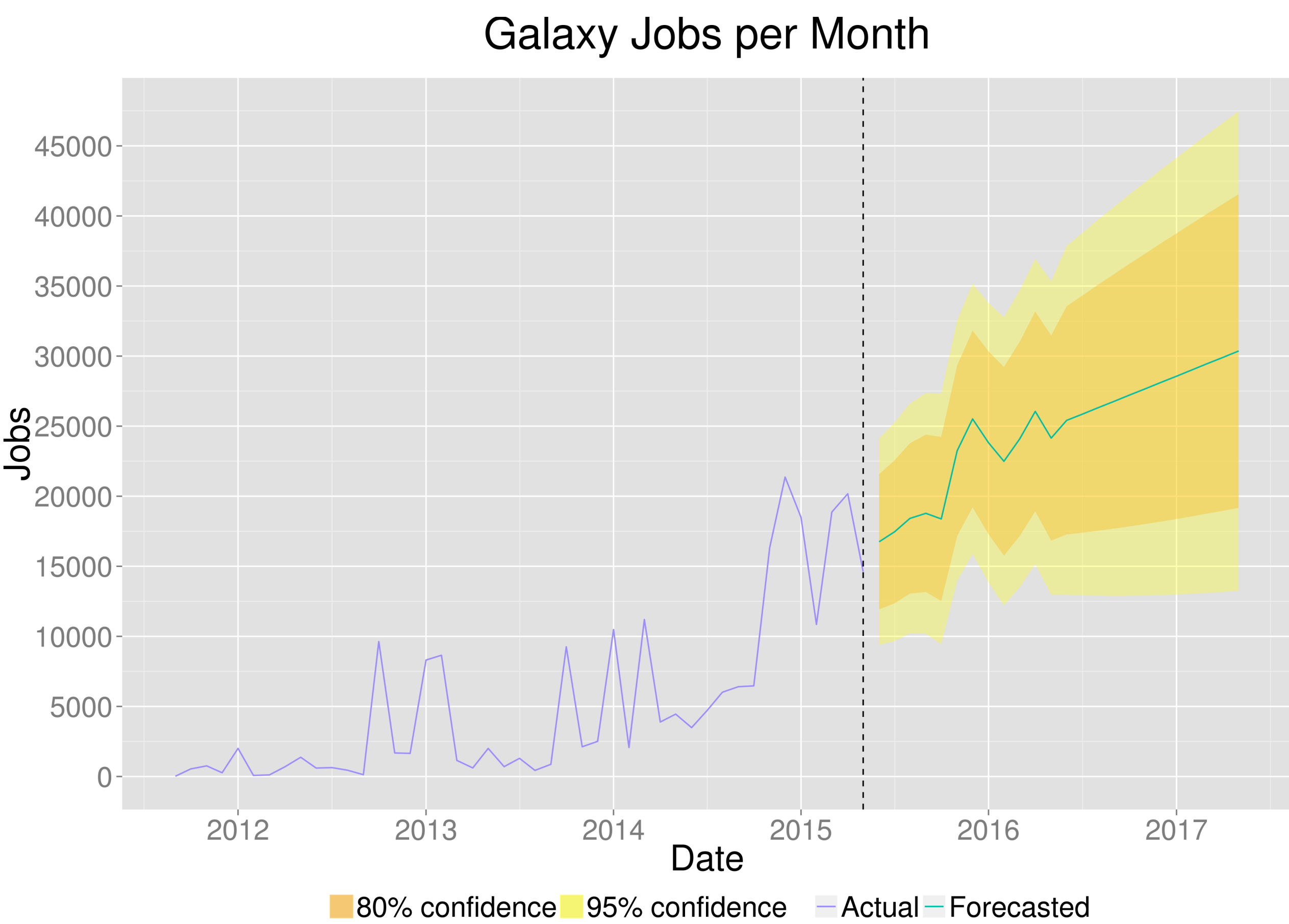
Eric Enns<sup>1</sup>, Philip Mabon<sup>1</sup>, Jennifer Cabral<sup>1,2</sup>, Mariam Iskander<sup>1,2</sup>, Cameron Sieffert<sup>1</sup>, Natalie Knox<sup>1</sup>, Heather Kent<sup>1</sup>, Shane Thiessen<sup>1</sup>, Paul Williams<sup>1</sup>, Brian Yeo<sup>2</sup>, Joel Thiessen<sup>1,2</sup>, Josh Adam<sup>1</sup>, Aaron Petkau<sup>1</sup>, Thomas Matthews<sup>1</sup>, Franklin Bristow<sup>1</sup>, Gary Van Domselaar<sup>1,2</sup>

<sup>1</sup>National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg, MB, Canada, <sup>2</sup>Department 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<sup>a</sup> in 2010. The introduction of the Galaxy platform has revolutionized bioinformatics at the NML by bridging the gap between bioinformaticians and biologists.

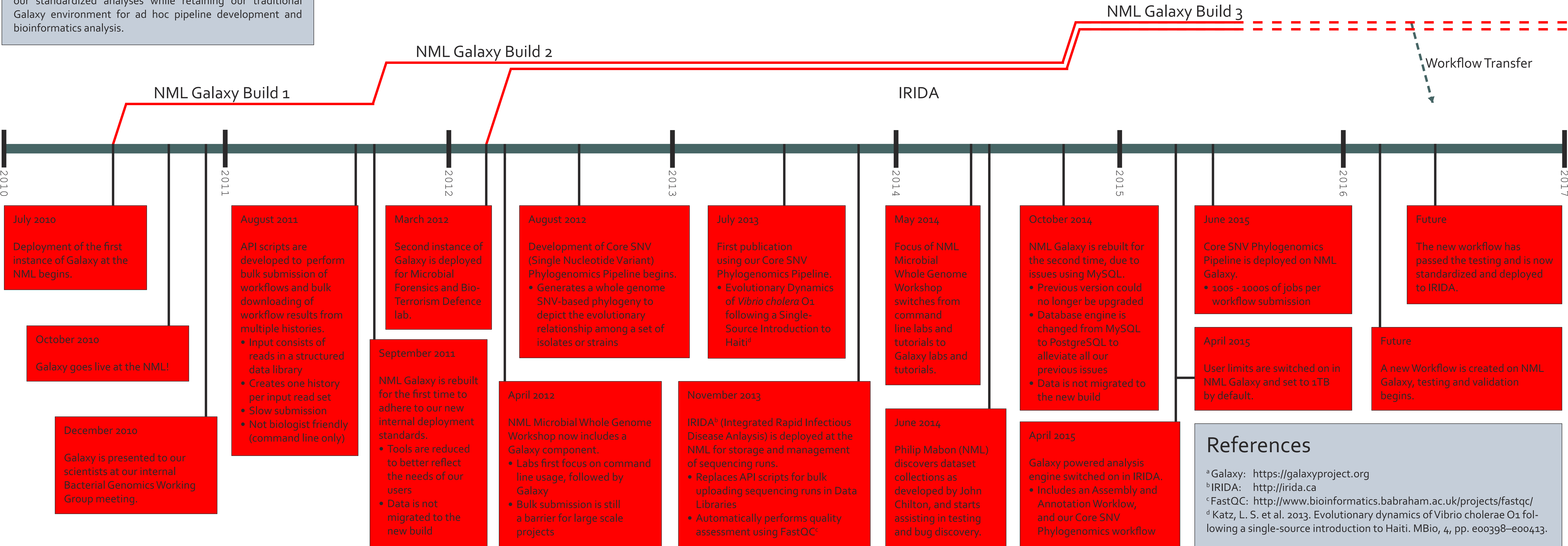
Prior to Galaxy, most of our in-house tools and pipelines required an extensive background in UNIX command line and high performance computing to operate. This requirement demanded that bioinformaticians be intimately involved in projects with significant computational requirements. Galaxy was selected 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.



## 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.



## References

- <sup>a</sup> Galaxy: <https://galaxyproject.org>
- <sup>b</sup> IRIDA: <http://irida.ca>
- <sup>c</sup> FastQC: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- <sup>d</sup> Katz, L. S. et al. 2013. Evolutionary dynamics of *Vibrio cholerae* O1 following a single-source introduction to Haiti. MBio, 4, pp. e00398–e00413.