Agriculture and Agri-Food Canada | Agriculture et Agroalimentaire Canada

Developing a Genome Annotation Workflow Within Galaxy

Iyad Kandalaft[1], Michael Li, Jeff Cullis, Hai D.T. Nguyen, Christopher T. Lewis[2], Keith A. Seifert

Eastern Cereal Oilseed Research Centre (ECORC), Agriculture and Agri-Food Canada (AAFC), Ottawa, Ontario, Canada.

[1]Correspondending author <Iyad.Kandalaft@Agr.gc.ca>, [2]Presenting author

Abstract

Current DNA sequencing technologies enable routine generation of microbial genomes. While automated annotation tools exist, they are subject to regular improvement, and the effort required to install, update, and execute these tools in a systematic way poses a barrier for many researchers. To facilitate rigorous and reproducible annotation, we demonstrate a Galaxy-driven genome annotation workflow. In the process, we documented proposed IT best practices for sustainable tool and workflow development within Galaxy.

Background

Routine generation of microbial genomes using next generation sequencing technologies places high demand on bioinformatics for whole-genome annotation. The emergence of annotation tools and automated pipelines coupled with their continual improvement and availability of new data makes rigorous annotation and re-annotation useful for genome analysis. However, the effort required to install and update these tools in a systematic way poses a barrier for many researchers, particularly while complying with information management and reproducibility standards and best practices. We aim to facilitate these efforts by leveraging Galaxy, an open web-based workflow platform, and developing workflows that tie in several existing annotation pipelines and tools.[1-3]

Methodology

We begin by enabling genome annotation software in Galaxy that is not currently available to the galaxy community. This involves developing the web-based user interface (wrappers) and creating tool-dependency packages to build and install the software. The tools are tied together into a galaxy-driven genome annotation workflow that: 1) heuristically and empirically annotates genes with Maker[4], 2) identifies gene clusters with AntiSMASH 2[5], 3) identifies gene function using InterProScan 5[8], 4) locates microsatellites with MISA, 5) detects SNPs via Mauve[6], and 6) designs flanking primers by leveraging Primer3[7] (Figure 1B).

Our Galaxy-tailored systems development life cycle facilitates development efforts while ensuring a stable production environment through several checkpoints, as shown in Figure 1A.

Impact & Discussion

The workflow leverages Galaxy's ability to install and systematically execute annotation software on whole genomes, thereby ensuring minimal effort rigorous genome annotations. Furthermore, Galaxy permits simple updates, amendments, and drop-in replacement of tools in the workflow as required. Galaxy's ability to make multiple tool versions available concurrently ensures annotation reproducibility by allowing the utilization of historical workflow and tool versions. In addition, we promote the Generic Feature Format (GFF) version 3[9] as the unified annotation file type, through developing format conversion tools, to simplify importing genomes and annotations into a genome browser for manual curation and publication. By combining the Tool Shed and SCM repositories during development, the developer easily pushes modifications to the Tool Shed for testing and utilize the SCM to revert or commit changes quickly and as needed. This results in the developer consistently testing the installation of the repository into Galaxy from the Tool Shed, which can be automated using the Galaxy API.

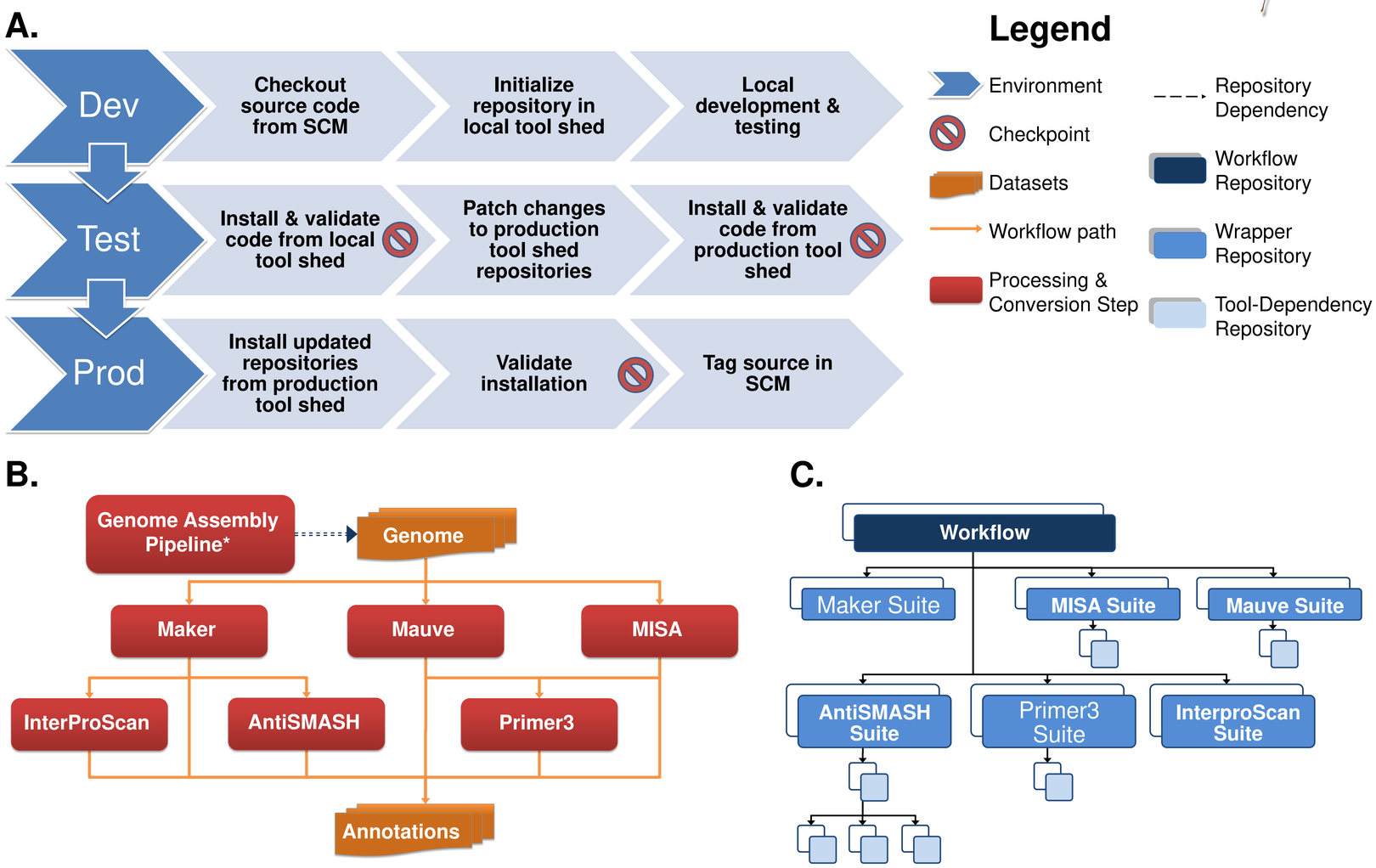

Figure 1. Technical architecture of the genome annotation workflow & development flow

A) The figure displays the development environment flow, which involves checking out a repository from the SCM and instantiating one or more mercurial repositories in a local Galaxy Tool Shed. Utilizing the svnignore/.gitignore and .hgignore paradigms, Galaxy-specific files are isolated from the development and documentation material and allowing to commit changes to either system independently and easily. Following local development and testing, changes are pushed to a testing Galaxy environment from the local tool shed for validation and subsequently pushed to the production environment. Several checkpoints safeguard the production environment against unforeseen issues at prior development stages. B) The outlined annotation workflow representation displays the flow from genome sequences to annotation data through a set of annotation tools and format conversion steps. C) The representation portrays the dependency hierarchy between the workflow, tool wrapper, and tool-dependency repositories. This modular approach reduces reliance on system installed software by allowing Galaxy to build the software required for each component of the workflow. In addition, it simplifies management of workflows within galaxy and distribution of workflows and associated tools to collaborators.

*Refer to Cullis et al., "Assessing and Improving In-house Genome and Transcriptome Assembly Solutions: A Case Study". ISMB (2014)

Conclusion & Future Work

The developed workflow permits rigorous genome annotations with minimal effort reproducibility within Galaxy. The outlined development flow promotes rapid tool development while maintaining best practices. We present these to the Galaxy community for discussion. The source code is available via GitHub (http://github.com/AAFC-MBB) and the workflow and wrappers will be made available to the community shortly.

Acknowledgements

This initiative is supported and funded by the Genomics Research & Development Initiative (GRDI) Shared Priorities Project on Quarantine and Invasive Species (QIS) and the Growing Forward 2 project titled Genome annotation, real-time PCR assays, and mycotoxigenic potential of ochratoxin A producing Penicillium and Aspergillus species on Canadian grain and wine grapes.

References

1. Goecks, Jeremy, et al. "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences." Genome Biol 11.8 (2010): R86.
2. Blankenberg, Daniel, et al. "Galaxy: a web-based genome analysis tool for experimentalists." Current protocols in molecular biology (2010): 19-10.
3. Giardine, Belinda, et al. "Galaxy: a platform for interactive large-scale genome analysis." Genome research 15.10 (2005): 1451-1455.
4. Cantarel, Brandi L., et al. "MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes." Genome research 18.1 (2008): 188-196.
5. Blin, Kai, et al. "antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers." Nucleic acids research (2013): gkt449.
6. Darling, Aaron E., et al. "progressivSeMauve: multiple genome alignment with gene gain, loss and rearrangement." PloS one 5.6 (2010): e11147.
7. Koressaar, Triinu, and Maido Remm. "Enhancements and modifications of primer design program Primer3." Bioinformatics 23.10 (2007): 1289-1291.
8. Jones, Philip, et al. "InterProScan 5: genome-scale protein function classification." Bioinformatics 30.9 (2014): 1236-1240.
9. The Sequence Ontology: A tool for the unification of genome annotations. Eilbeck K., Lewis S.E., Mungall C.J., Yandell M., Stein L., Durbin R., Ashburner M. Genome Biology (2005) 6:R44

Canada