Leveraging Nextflow and Bactopia for Automated Plant Pathogen Characterization

Zachary S.L. Foster, Martha Sudermann, Nicholas C. Cauldron, Fernanda I. Bocardo, Hung Phan, Jeff H. Chang, Niklaus J. Grünwald

Abstract

Automated and rapid pipelines are needed to analyze abundant biological genome sequences. Nextflow provides a foundation for reproducible and scalable analyses. One application that uses Nextflow is Bactopia, which provides workflows for genome-scale bacterial sequence analysis. However, such general-use pipelines must be adapted to specific cases. We are adapting and extending Bactopia for use in diagnostic plant clinics to automate pathogen detection and characterization. Significant additions include the ability to analyze eukaryotic pathogens using mitochondrial or diploid genomes, automated pathogen identification, and separation of host and pathogen sequences. This will enable rapid detection of unknown pathogens or variants infecting a known host. We will be testing our pipeline using whole genome sequences of the Sudden Oak Death pathogen Phytophthora ramorum. Any general-use Nextflow modules or workflows developed will be offered as contributions to nf-core and Bactopia.

Motivation

The Plant Clinic at Oregon State University provides plant disease diagnosis and plant pathogen identification services. The use of whole genome Illumina sequencing of infected plant material and isolates would enhance the effectiveness of this service, but tools that can be used by lab technicians with minimal experience with bioinformatics are required. We are designing a pipeline to automate the analysis of plant pathogen genome sequence data, including fungal, protist, and prokaryotic pathogens, potentially infecting a known host plant. This pipeline will enable lab technicians to run analyses using a simple GUI and produce PDF/HTML reports. Select portions of the results can be passed on to clients of the OSU Plant Clinic withing days of submitting a sample. The tool will be made open source and freely available, allowing for widespread use of next generation sequencing data for plant pathogen identification and tracking.

Goals

Create pipeline to identify species of unknown samples

- Remove host reads by read mapping to nearest reference
- Assign Life Identification Number (LIN) codes to identify prokaryote species
- Local assembly of barcode sequences to identify eukaryotes

Create pipeline to analyze and track plant pathogen species

- One instance of the pipeline per species
- Add new samples over time with minimal recompute
- Track geographic and temporal distribution
- Track the presence of genes of interest
- Infer population structure using variants
- Infer phylogenetic relationships using variants or MSLA

Simple GUI to allow pipeline to be run by lab technicians

- Allow entry of raw data location, metadata, and pipeline settings
- Execute analyses and monitors progress
- Record history of analyses and links to result location

Output PDF or HTML reports of results

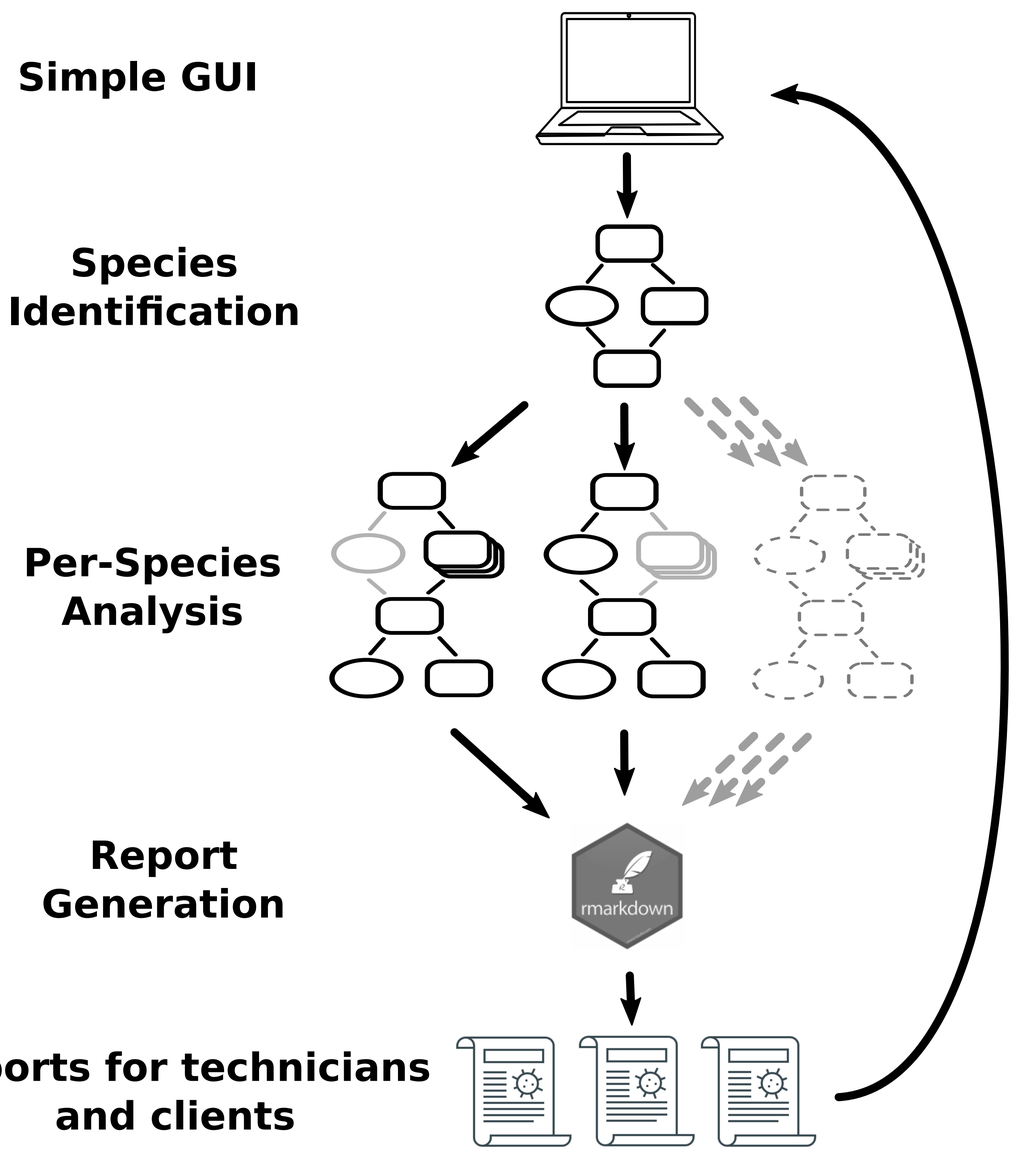
- Provide data and figures for lab technicians
- Provide simplified data and figures for sharing with clients

Manage results and intermediate data

- Manage limited storage space by deleting parts of Nextflow work directory as needed
- Share data common to all pipelines
- Archive past results

Methods

- Use Nextflow to construct and run pipelines
- Use Nextflow Tower, R shiny, or a Python GUI toolkit to create a simple GUI
- Use R Markdown to generate PDF/HTML reports from pipeline output
- Use and adapt Nextflow workflows from Bactopia and nf-core to the extent possible
- Submit all novel Nextflow workflows and modules as contributions to nf-core and Bactopia as appropriate



USDA Agricultural Research Service

This work was supported by the National Institutes of Food and Agriculture, US Department of Agriculture award 2021-67021-34343 to NJG and JHC. We would like to thank the creators of Nextflow, Bactopia, and other open source software used in this project.

Oregon State University OSU