PathogenDx: Automated Analysis of Whole Genome Sequencing Data for the Identification and Analysis of Pathogen Populations

Zachary S.L. Foster, Fernanda I. Bocardo, Hung Phan, Marina Witherell, Alexandra J. Weisberg, Melodie L. Putnam, Jeff H. Chang, Niklaus J. Grünwald

Summary:

Automated and rapid pipelines are needed to leverage the increasing availability of whole genome sequencing data. We have devloped a pipeline for use in diagnostic clinics to automate the characterization of populations of pathogens. The pipeline is built with Nextflow, which provides a foundation for reproducible and scalable analyses. The pipeline accepts the paths to raw reads for one or more organisms and creates interactive HTML reports or PDF documents. Significant features include the ability to analyze unidentified eukaryotic and prokaryotic samples, creation of reports for multiple user-defined groupings of samples, automated discovery and downloading of reference assemblies from NCBI RefSeq, and rapid initial identification based on k-mer sketches followed by a more robust core genome phylogeny.

Features implemented:

- The entire pipline is automated and runs with a single command
- Genome assembly and annotation (Prokaryotes only)
- Core genome phylogeny with RefSeq genomes for context (Prokaryotes only)
- Variant calling with a user-defined reference or one selected from RefSeq
- Minimum spanning network and SNP phylogeny from variant data
- Reports for each user-defined group of samples as HTML reports with interactive figures or static PDF documents (in progress)

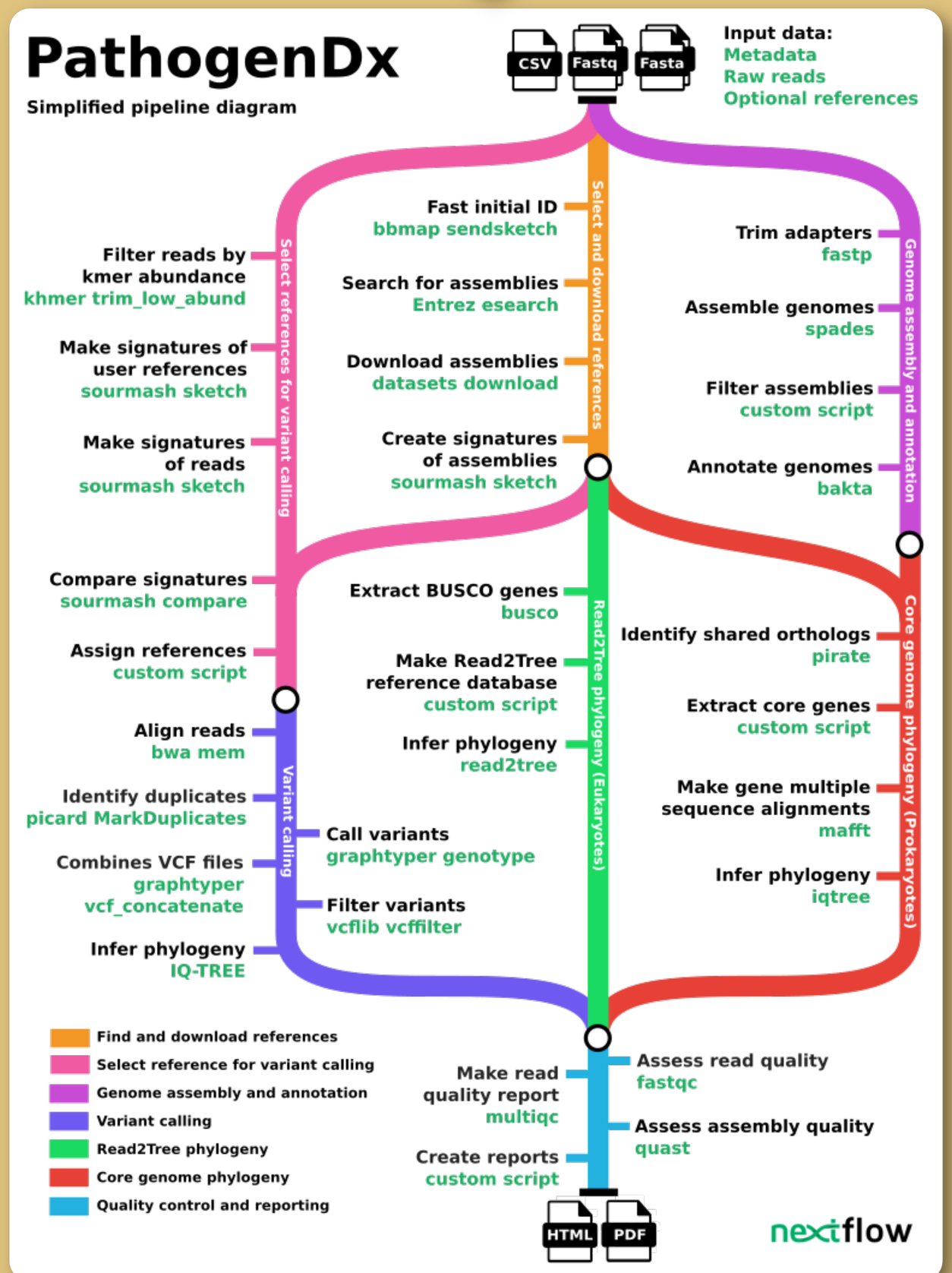
Features planned:

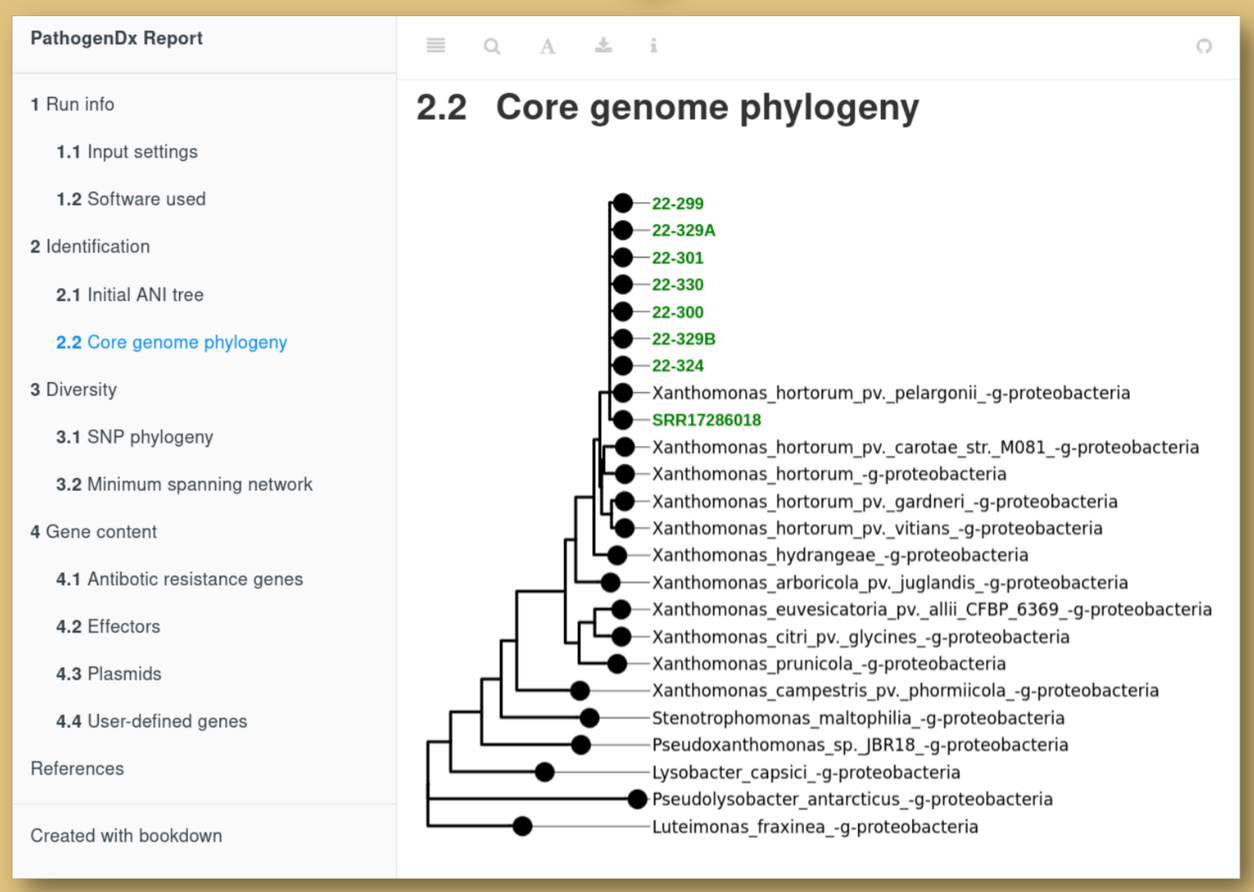
- Ability to use long reads (e.g. Nanopore, PacBio)
- A phylogeny of BUSCO genes extracted from raw reads with Read2Tree (Eukaryotes only)
- Detection of genes such as antibiotic resistance loci and effectors
- A map of sample locations annotated with analysis results
- Analysis of samples with sequences from a host mixed in
- Identification of viruses or other pathogenic organisms in environmental samples

Features provided by Nextflow:

- All programs needed to run the pipeline are installed automatically
- Processes are run in parallel, allowing for analysis of massive data sets quickly if sufficient computing resources are available
- Runs on personal computers, high performance clusters, or commercial cloud services such as AWS
- Inputs and outputs can be stored anywhere on the internet
- Samples can be added without rerunning the entire pipeline
- The pipeline can pick up where it left off if it is interrupted
- Each process runs in its own docker/singularity/conda environment, enabling reproducibility

	A	B	C	D	E	F
1	sample	fastq_1	fastq_2	reference	reference_id	report_group
2	22-299	22-299_R1.fastq.gz	22-299_R2.fastq.gz			xan_test;subgroup
3	22-300	22-300_R1.fastq.gz	22-300_R2.fastq.gz			xan_test;subgroup
4	22-301	22-301_R1.fastq.gz	22-301_R2.fastq.gz			xan_test;subgroup
5	22-324	22-324_R1.fastq.gz	22-324_R2.fastq.gz			xan_test;subgroup
6	22-329A	22-329A_R1.fastq.gz	22-329A_R2.fastq.gz	reference-22-331.fna	22_331_assembly	xan_test
7	22-329B	22-329B_R1.fastq.gz	22-329B_R2.fastq.gz	reference-22-331.fna	22_331_assembly	xan_test
8	22-330	22-330_R1.fastq.gz	22-330_R2.fastq.gz	reference-22-331.fna	22_331_assembly	xan_test
9	SRR17286018	SRR17286018_R1.fastq.gz	SRR17286018_R2.fastq.gz	reference-22-331.fna	22_331_assembly	xan_test
10	pram1	S13_L006_R1_001.fastq.gz	S13_L006_R2_001.fastq.gz	PR-102_v3.1.fasta.gz	PR-102_v3.1	pram_test
11	pram2	S8_L006_R1_001.fastq.gz	S8_L006_R2_001.fastq.gz	PR-102_v3.1.fasta.gz	PR-102_v3.1	pram_test





USDA Agricultural Research Service

Oregon State University OSU

This work was supported by the National Institutes of Food and Agriculture, US Department of Agriculture award 2021-67021-34343 to NJG and JHC.