

Background concepts

- **The command line** (a.k.a. the terminal, bash, the console, etc)

- Most scientific software runs on the command line

- Good for automation, reproducibility, flexibility, and ease of development

.Pipelines

- Collections of independent applications connected together

- More convenient than manually running applications, but less flexible

.Dependencies

- The bane of complicated software and pipelines (a.k.a. dependency hell)

Nextflow

- .A framework for writing and running pipelines
- .Dependencies are installed automatically as needed
- .Each step of a pipeline runs in an isolated environment
- .Automatically runs in parallel when possible
- .Caches intermediate results
- .Interrupted analyses can be restarted
- .Samples can be added to existing analyses
- .Intermediate results from other analyses can be used when appropriate

Pathogensurveillance

- .A nextflow pipeline for identification and population genomics of pathogens

- .Inputs:

- .A sample data spreadsheet (CSV) with one row per sample

- .(optional) A reference data spreadsheet with one row per reference

- .Outputs:

- .A human report with interactive plots and tables

- .A directory of intermediate results

- .Can use raw Nanopore, Pacbio, or Illumina reads as input