Bioinformatic pipelines with UNIX

Libor Mořkovský, Václav Janoušek

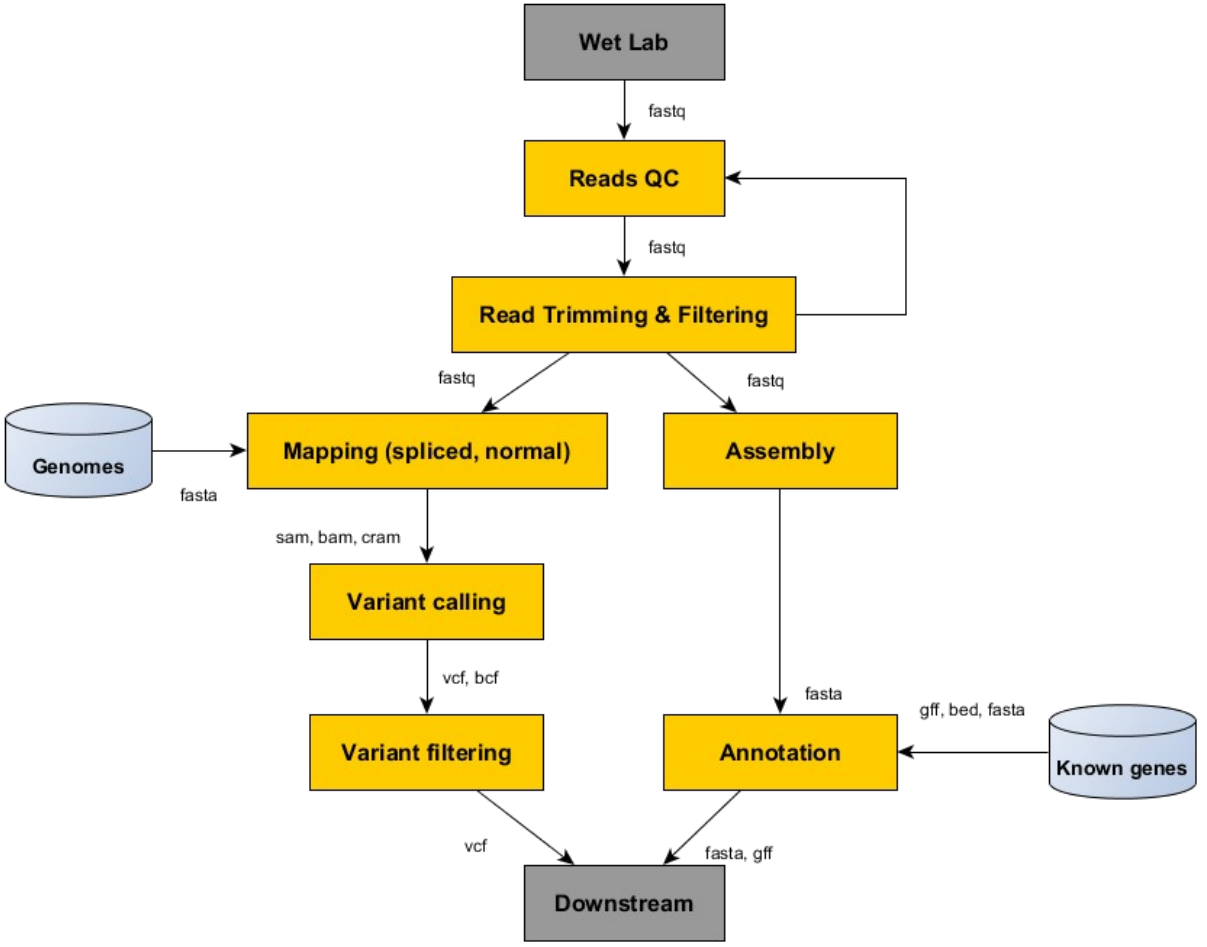
Why use pipelines?

- Reproducibility
- Easy reusability
- Repeated runs with varying parameter values
- Combination of different tools and custom scripts in main Shell script
- Makefile can be used to set up the environment for easy transferability
- Versioning using Git

NGS pipeline

Pipeline to carry the NGS data processing to obtain **high quality variants** or **genome assemblies**

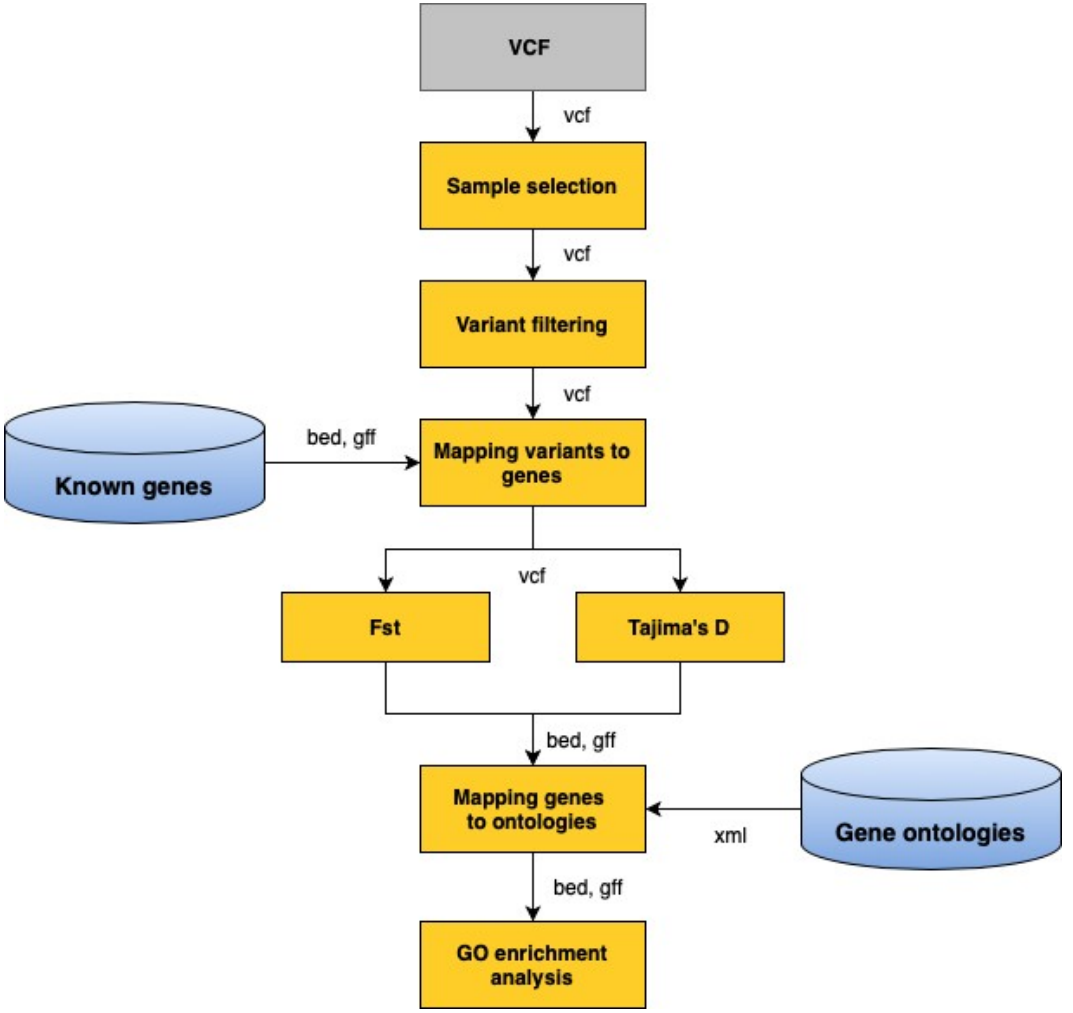
Pipeline can be rerun with different parameter values such as trimming parameter values, k-mer size, phred thresholds to compare resulting quality of variants or assemblies



Data analysis pipeline

Pipeline to carry **Gene Ontology enrichment analysis** according to widely used population genetics metrics

Pipeline can be easily versioned and rerun with different samples and filtering parameters



Use of Git

Github can be use to backup, version and share the code with collaborators

Easy **transferability** when code developed at multiple platforms (local machine, MetaCentrum, commercial cloud services)

Repository can be made **public** to enable reproducibility of the work when published

Your collaborators and other scientists interested in your work will appreciate that!

Possible repository structure:

data
docs
results
src
.gitignore
README.md
workflow.sh | pipeline.sh
instal.sh

Scaling up computational resources

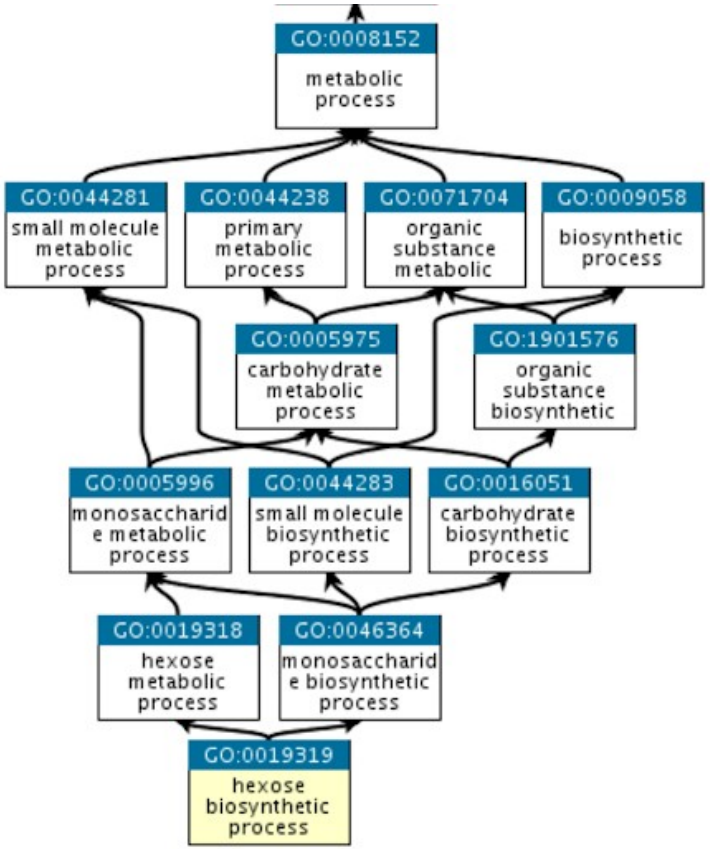
The size of the NGS data from many samples can grow to **several terabytes**

To process and analyze such data needs computational resources of **100s of gigabytes of RAM**

Having properly written code in the Git repository **enables to transfer** easily necessary code **to the cloud** to use machines with appropriate computational capacity

MetaCentrum provides computing resources to academics and researchers for free

Exercise: Mouse Gene Ontology enrichment analysis pipeline



- unifying representations of gene and gene product attributes across all species
- maintain and develop controlled vocabulary of gene and gene product attributes in three main domains:
  - biological processes
  - molecular functions
  - cellular componenets

GENEONTOLOGY
Unifying Biology

http://geneontology.org

Exercise: Mouse Gene Ontology enrichment analysis pipeline

- Gene Ontology enrichment analysis for low and high divergence genes among two house mouse subspecies
- SNP variants for two mouse strains PWD/PhJ and WSB/EiJ represent *Mus musculus musculus* and *Mus musculus domesticus* subspecies