Bioinformatics pipelines with Unix

Libor Mořkovský, Václav Janoušek

Why use pipelines for bioinformatics tasks?

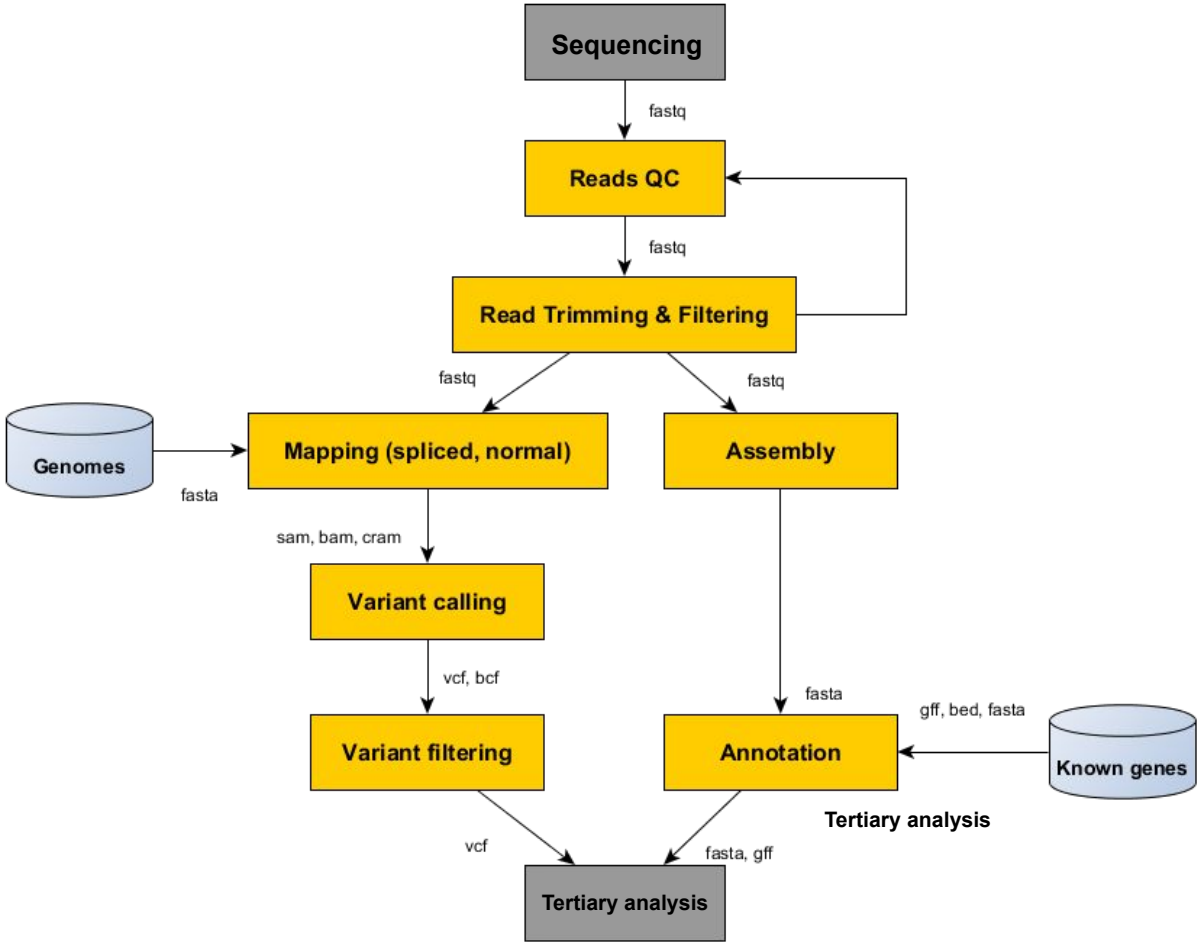
- Reproducibility (anyone can re-run the analysis and check the code)
- Reusability (varying parameter values)
- Combination of different tools and custom scripts in main shell script
- Shell script/makefile can be used to set up the environment for easy transferability
- Versioning using Git and sharing via GitHub

NGS pipeline

(secondary analysis)

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

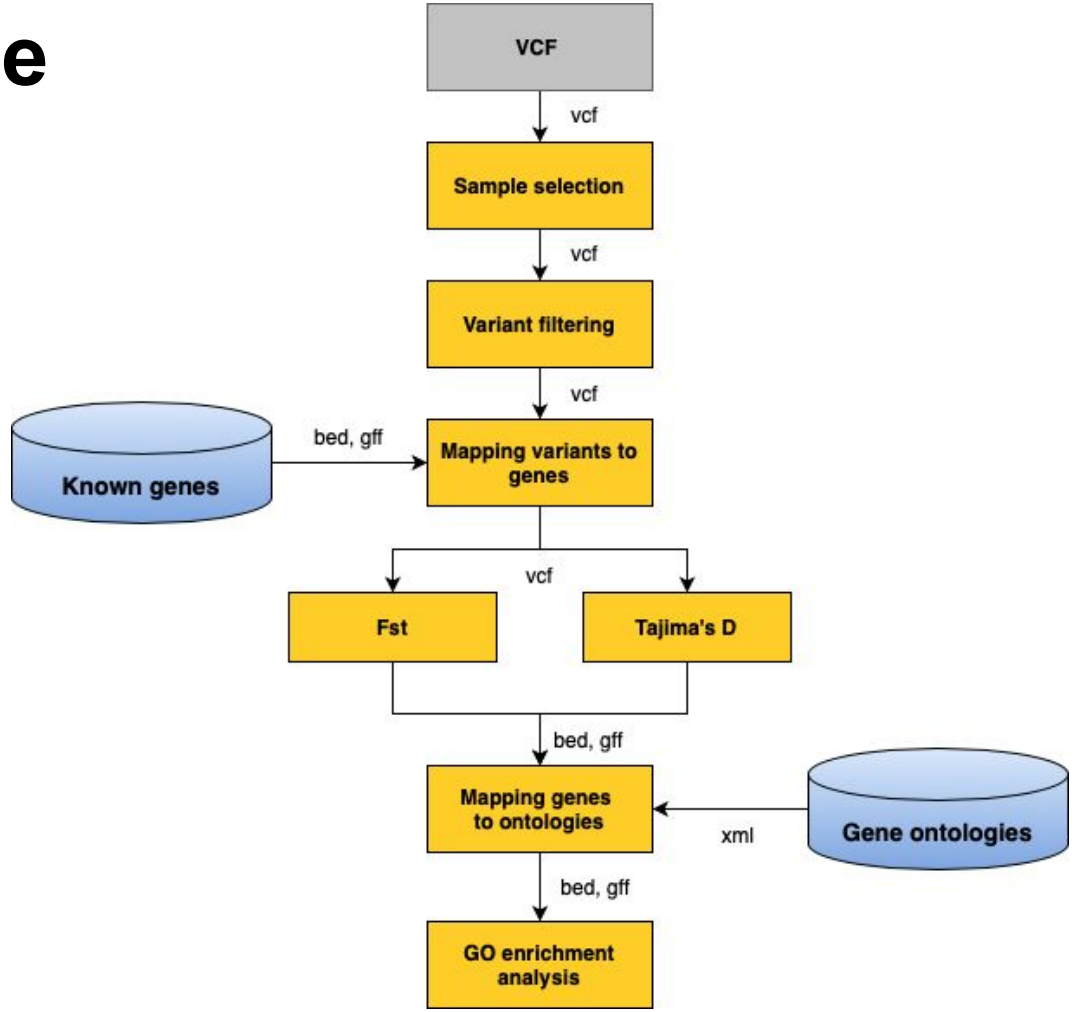


Data analysis pipeline

(tertiary analysis)

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 GitHub

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

Do not upload data to Git/GitHub!!

Possible repository structure:

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

Scaling up computational resources

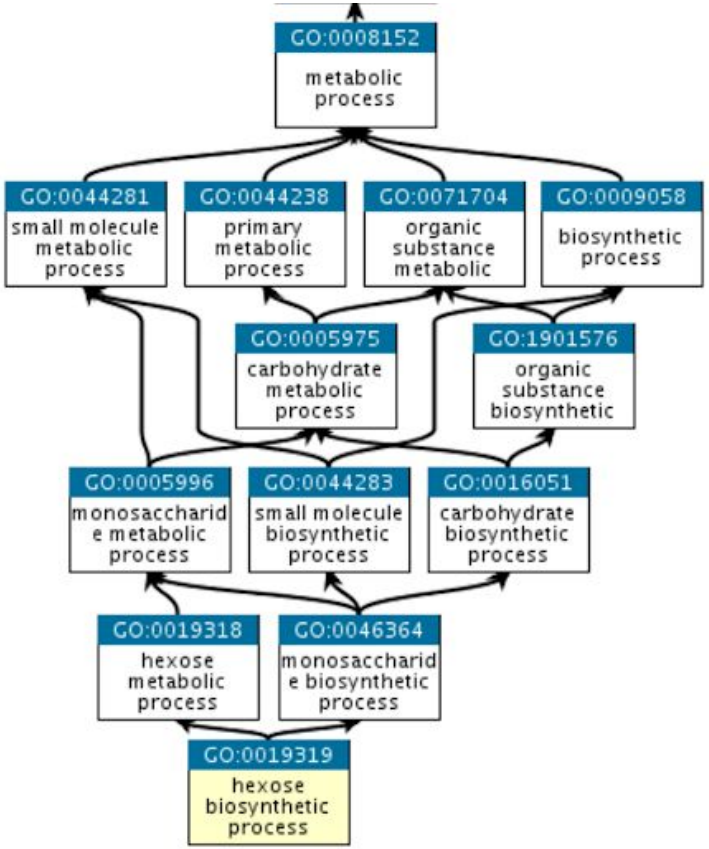
The size of the NGS data from many samples can grow to **several terabytes**

Processing and analysis of such data needs computational resources of **100s of gigabytes of RAM and high CPU/GPU**

Having properly written code in the Git repository **enables to transfer** easily necessary code **in 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



Gene Ontology

- Unification of gene and gene product attributes across all species into curated vocabulary (ontology)
- Vocabulary of gene and gene product attributes in three main domains:
  - biological processes
  - molecular functions
  - cellular components

GENEONTOLOGY
Unifying Biology

http://geneontology.org

Exercise:
Mouse Gene Ontology enrichment analysis pipeline

- SNP variants for two mouse strains PWD/PhJ and WSB/EiJ represent *Mus musculus musculus* and *Mus musculus domesticus* subspecies
- Gene Ontology gene enrichment analysis for low and high divergence genes among two house mouse subspecies
- See ngs-course.readthedocs.io Session 8