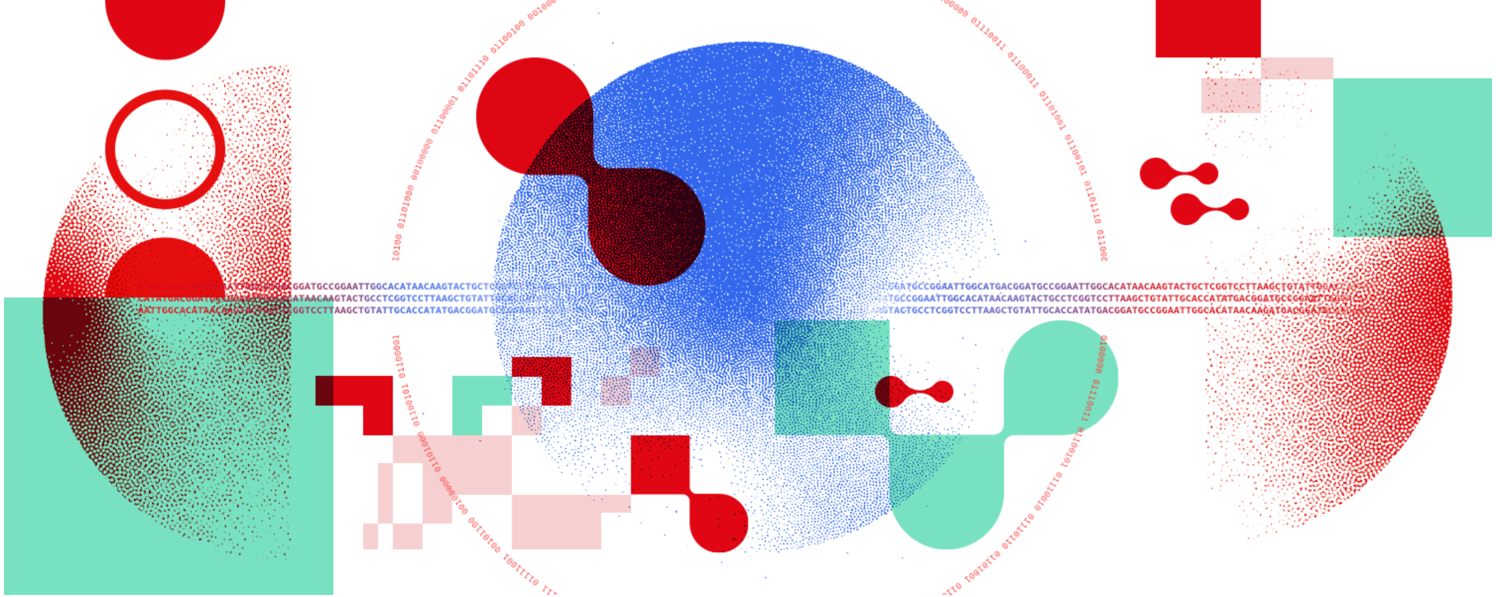

Snakemake for reproducible research

Additional advanced concepts

Antonin Thiébaut

antonin.thiebaut@unil.ch

UNIL | Université de Lausanne

Swiss Institute of Bioinformatics

Running snakemake on clusters and cloud

- Built-in support for Kubernetes / Google cloud and AWS

- Snakemake can make use of a scheduler (slurm, SGE, LFS …) to execute jobs on a cluster with almost no changes (runtime, memory…) to the rules

- Syntax: snakemake --cluster "sbatch" (qsub, sbatch …)

- Advanced syntax: command can take job information from rule definition

```
snakemake --jobs 12 --cluster “sbatch --cpus-per-task={threads}”
```

- Specify the maximum number of jobs to submit with **“-j / --jobs”**

Execution profiles

- Execution profiles are presets of execution parameter values (-j <N>, --use-conda, --resources mem_mb=100…)

- Implemented as directory and stored in *~/.config/snakemake/<profile_name>/*
  Minimum: config.yaml with syntax <run_option>: <value>

- Profiles can be extended a lot, especially for HPC environments
  - Scripts to submit jobs
  - Scripts to check job status
  - ➢ Advanced customization

- Collection of official profiles on Github
  - Custom profile for Slurm developed by our lab

Working with remote inputs

- Snakemake implements remote file access for many protocols

- Idea:
  - Import module for the remote access protocol
  - Initiate remote provider instance in snakefile’s body
  - Access remote files within a rule

- Files are downloaded to a sub-dir of the current working directory and deleted after the job requiring them is completed

- Amazon Simple Storage Service (AWS S3)
- Google Cloud Storage (GS)
- Microsoft Azure Blob Storage
- File transfer over SSH (SFTP)
- Read-only web (HTTP[S])
- File transfer protocol (FTP)
- Dropbox
- XRootD
- GenBank / NCBI Entrez
- WebDAV
- GFAL
- GridFTP
- iRODS
- EGA
- Zenodo

Reminder on best practices

- One repository = one workflow
- Use Conda environments / Docker containers when possible
- Break out large workflow into modules with extension “.smk”
- Specify parameters in a config file located in a ‘config’ folder
- If you have many samples with information, use a sample sheet located in the ‘config’ folder
- Follow the official directory structure
- Use explicit rule and variable names
- Comment to explain your workflow; use docstring comments in rules