Fully reproducible data analysis with Snakemake and Bioconda

Johannes Köster
Centrum Wiskunde & Informatica, Amsterdam
Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston

The needs of data analysis

- Scalability
  - Handle tens to thousands of samples via parallelization
  - Avoid redundant computations when changing datasets or parameters

- Reproducibility
  - Document parameters, tools, versions
  - Execute and deploy without manual intervention

Define workflows via generic rules

```bash
rule mytask:
  input:
    "reference.fasta",
    "reads/{dataset}.fastq"
  output:
    "mapped/{dataset}.bam"
  environment:
    "software.yaml"
  resources:
    mem=4
  shell:
    "bwa mem {input} | samtools view -b {output}"
```

- Use shell commands, scripts (R, Python), and tool wrappers
- Dependencies between rules are determined automatically
- Implicit parallelization to compute servers and clusters

Bioinformatics software installation is heterogeneous

- Over 1500 packages
- Over 100 maintainers

By combining Snakemake and Bioconda, data analyses become reproducible with minimal effort

- Clone workflow repository
  - $ git clone https://github.com/user/workflow

- Install Snakemake
  - $ conda install snakemake

- Execute workflow
  - $ snakemake

Snakemake formalizes, documents, and executes data analyses

- Used by various high-impact studies
  - Learn more

Bioconda normalizes software installation via easy to create package recipes

- Over 1500 packages
- Over 100 maintainers

Works for any language (R, Python, C++, Rust, Perl, ...)

Define isolated software environments per rule

- channels:
  - bioconda
- dependencies:
  - bwa ==0.7.4
  - samtools ==1.1

- Isolation allows conflicting versions on the same system
- Exact versions ensure full reproducibility

Fully reproducible data analysis with Snakemake and Bioconda

Johannes Köster
Centrum Wiskunde & Informatica, Amsterdam
Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston