

Snakemake: how to install and use


Jun Duan
2021-01-21



Reproducibility issue is a great challenge in bioinformatics analysis

- Many difficulties of reproducibility in analysis
 - What software version?
 - What arguments?
 - How is this recorded?
 - How can someone else replicate the analysis
- Scaling up analysis is difficult
 - Debugging scripts
 - Ensuring cluster jobs run appropriately
- Many things to consider, before you start thinking about the bioinformatics analysis

Snakemake—a scalable bioinformatics workflow engine

[J Köster](#), [S Rahmann](#) - Bioinformatics, 2012 - academic.oup.com  Paperpile

Snakemake is a workflow engine that provides a readable Python-based workflow definition language and a powerful execution environment that scales from single-core workstations to compute clusters without modifying the workflow. It is the first system to support the use of automatically inferred multiple named wildcards (or variables) in input and output filenames. Availability: <http://snakemake.googlecode.com>. Contact: johannes.koester@uni-due.de

☆  Cited by 1095  Related articles All 14 versions Import into BibTeX

[HTML] Nextflow enables reproducible computational workflows

[P Di Tommaso](#), [M Chatzou](#), [EW Floden](#), [PP Barja...](#) - Nature ..., 2017 - nature.com  Paperpile

The increasing complexity of readouts for omics analyses goes hand-in-hand with concerns about the reproducibility of experiments that analyze 'big data'^{1, 2, 3}. When analyzing very large data sets, the main source of computational irreproducibility arises from a lack of good ...

☆  Cited by 442  Related articles All 6 versions Import into BibTeX



Advantages of Snakemake

- It's free, open-source, and conda-installable
- Snakemake is written using Python, but supports bash and R code as well.
- Snakemake works cross-platform (Windows, MacOS, Linux) and is compatible with all HPC schedulers.

Resources

https://github.com/Public-Health-Bioinformatics/snakemake_demo/blob/main/Intro_snakemake.ipynb

```
git clone --recursive https://github.com/Public-Health-Bioinformatics/snakemake_demo.git
```

References:

- Documentation: <https://snakemake.readthedocs.io/>
- Wrappers: <https://snakemake-wrappers.readthedocs.io/>
- Snakemake-Workflows project: <https://github.com/snakemake-workflows/docs>
- Snakemake-Profiles project: <https://github.com/snakemake-profiles/doc>

Set up snakemake

Recommended way of installation is using conda

Install conda

If you haven't had conda yet, you can install miniconda like this:


```
wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh  
sh Miniconda3-latest-Linux-x86_64.sh
```

Install Snakemake through conda

```
conda create -c bioconda -c conda-forge -n snakemake_demo snakemake-minimal -y
```

Or you can mount your existing conda environment, then install like this:

```
conda install snakemake-minimal
```



Tips: If installing snakemake with conda is taking very long, a solution can be to use **mamba**

How to verify the installation

:

```
$ snakemake --version
```

Activate the conda environment:

```
source activate snakemake_demo
```



Examples

Please see find the at https://github.com/Public-Health-Bioinformatics/snakemake_demo/blob/main/Intro_snake_make.ipynb