Reproducible data analysis with Snakemake

Johannes Köster

2019

https://koesterlab.github.io

Snakemake is popular

150k downloads since 2015

497 citations (+150 in 2019)

https://snakemake.readthedocs.io

Concise DSL

```
rule mytask:
    input:
        "data/{sample}.txt"
    output:
        "result/{sample}.txt"
    shell:
        "some-tool {input} > {output}"
```

https://snakemake.readthedocs.io

Python scripts

```
rule mytask:
    input:
        "data/{sample}.txt"
    output:
        "result/{sample}.txt"
    script:
        "scripts/mytask.py"
```

R scripts

```
rule mytask:
    input:
        "data/{sample}.txt"
    output:
        "result/{sample}.txt"
    script:
        "scripts/mytask.R"
```

Julia scripts

```
rule mytask:
    input:
        "data/{sample}.txt"
    output:
        "result/{sample}.txt"
    script:
        "scripts/mytask.jl"
```

No boilerplate

```
rule mytask:
    input:
    output: "data/{sample}.txt"
    script: "result/{sample}.txt"
            "scripts/mytask.py"
```

```
import matplotlib.pyplot as plt
import pandas as pd

d = pd.read_table(snakemake.input[0])

d.hist(bins=snakemake.config["hist-bins"])

plt.savefig(snakemake.output[0])
```

https://snakemake.readthedocs.io

No boilerplate

```
rule mytask:
    input:
    output: "data/{sample}.txt"
    script: "result/{sample}.txt"
            "scripts/mytask.py"
```

```
import matplotlib.pyplot as plt
import pandas as pd

d = pd.read_table(snakemake.input[0])

d.hist(bins=snakemake.config["hist-bins"])

plt.savefig(snakemake.output[0])
```

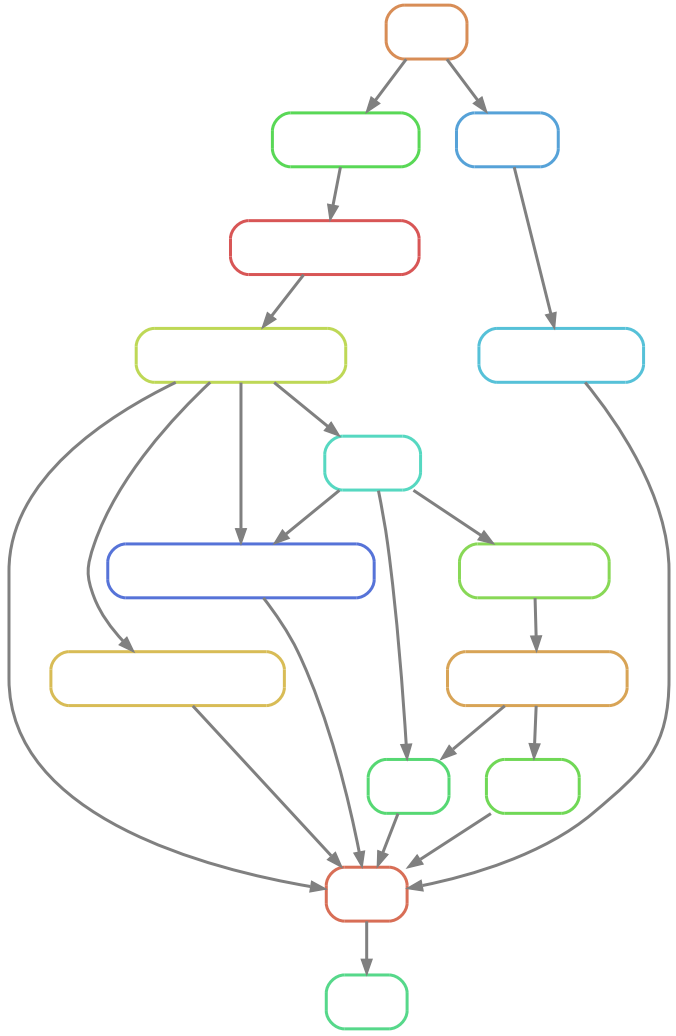
https://snakemake.readthedocs.io

Implicit dependencies

```
rule mytask:
  input:
        "path/to/{dataset}.txt"
   output:
        "result/{dataset}.txt"
  script:
        "scripts/myscript.R"


rule myfiltration:
   input:
        "result/{dataset}.txt"
    output:
        "result/{dataset}.filtered.txt"
   shell:
        "mycommand {input} > {output}"


rule aggregate:
  input:
        "results/dataset1.filtered.txt",
        "results/dataset2.filtered.txt"
   output:
        "plots/myplot.pdf"
  script:
        "scripts/myplot.R"
```



https://snakemake.readthedocs.io

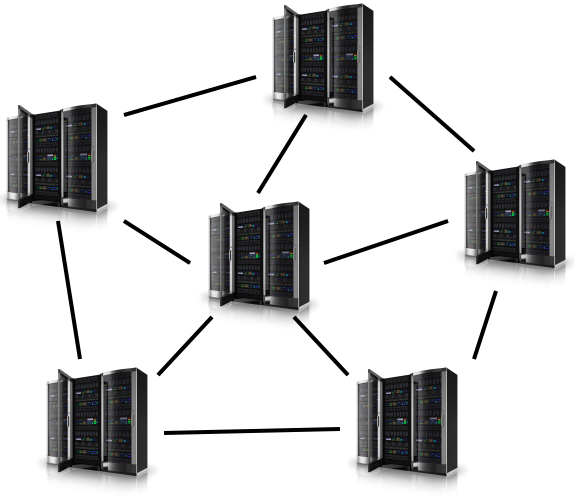
Scalability

workstation

compute server

cluster

grid computing



cloud computing

kubernetes

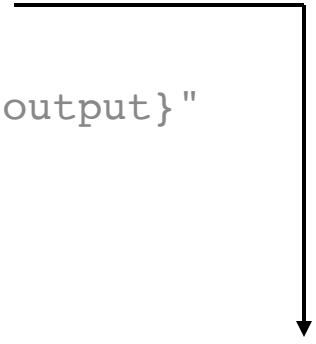
amazon web services

Google Cloud Platform

https://snakemake.readthedocs.io

Conda integration

```
rule mytask:
    input:
        "path/to/{dataset}.txt"
    output:
        "result/{dataset}.txt"
    conda:
        "envs/mycommand.yaml"
    shell:
        "mycommand {input} > {output}"
```



```
channels:
  - bioconda
  - conda-forge
dependencies:
  -mycommand =2.3.1
```

https://snakemake.readthedocs.io

Singularity integration

```
rule mytask:
    input:
        "path/to/{dataset}.txt"
    output:
        "result/{dataset}.txt"
    singularity:
        "docker://some/container"
    shell:
        "mycommand {input} > {output}"
```

Singularity + Conda

```
rule mytask:
    input:
        "path/to/{dataset}.txt"
    output:
        "result/{dataset}.txt"
    conda:
        "envs/mycommand.yaml"
    singularity:
        "docker://some/os"
    shell:
        "mycommand {input} > {output}"
```

https://snakemake.readthedocs.io

Snakemake

Snakemake Report

Fri Jul 19 16:58:33 2019 CET
Snakemake 5.5.3+6.g668fec20

- Workflow
- Statistics
- Configuration
- Rules

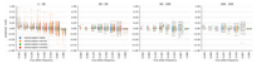
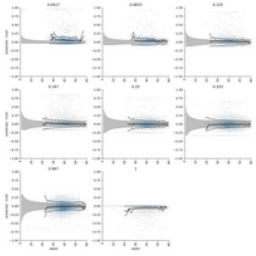
RESULTS

- Allele Frequency Estimation
- Concordance
- FDR Control
- Precision and Recall
- Score Distribution

Results

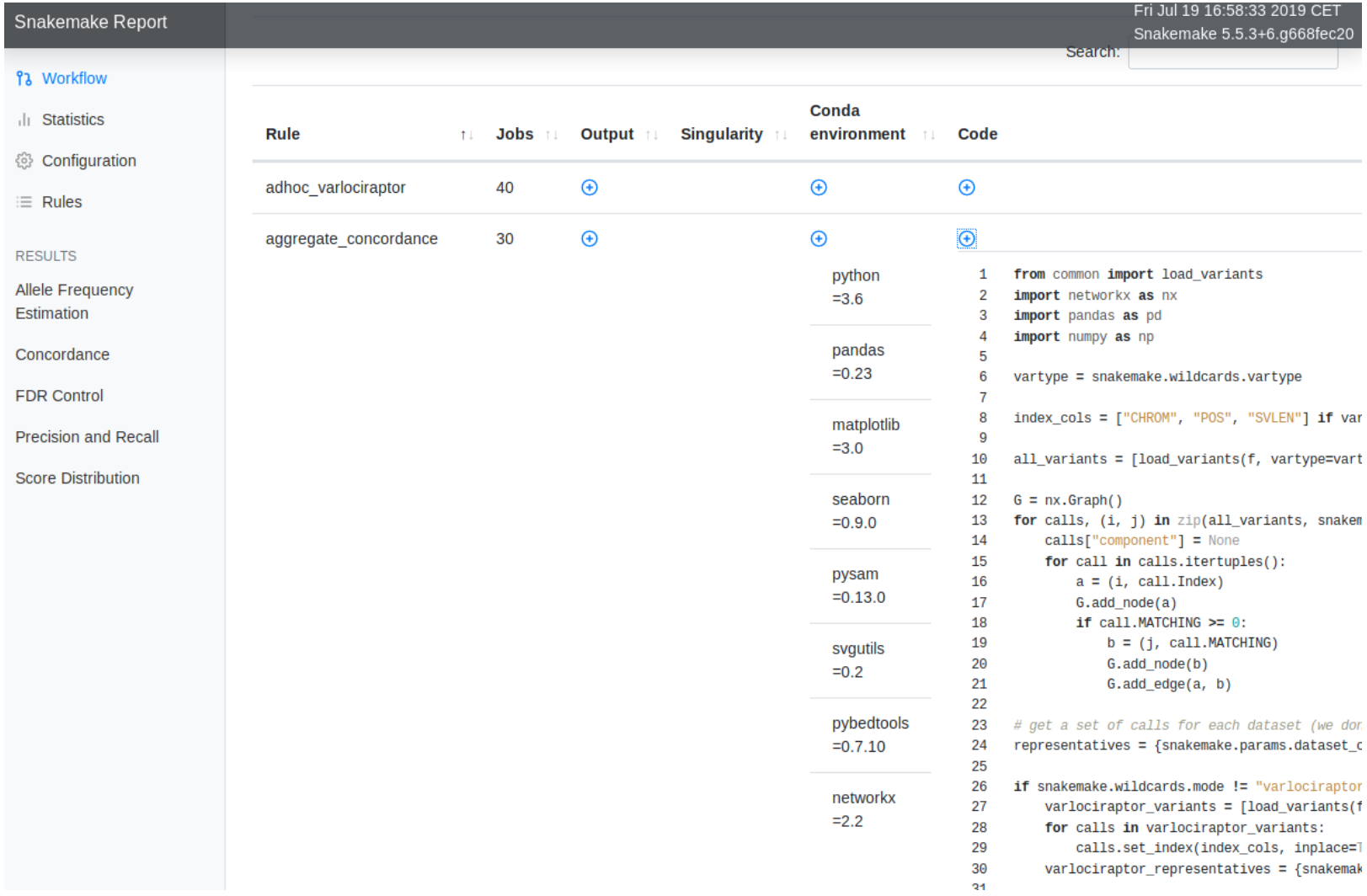
Allele Frequency Estimation

Show 10 entries Search:

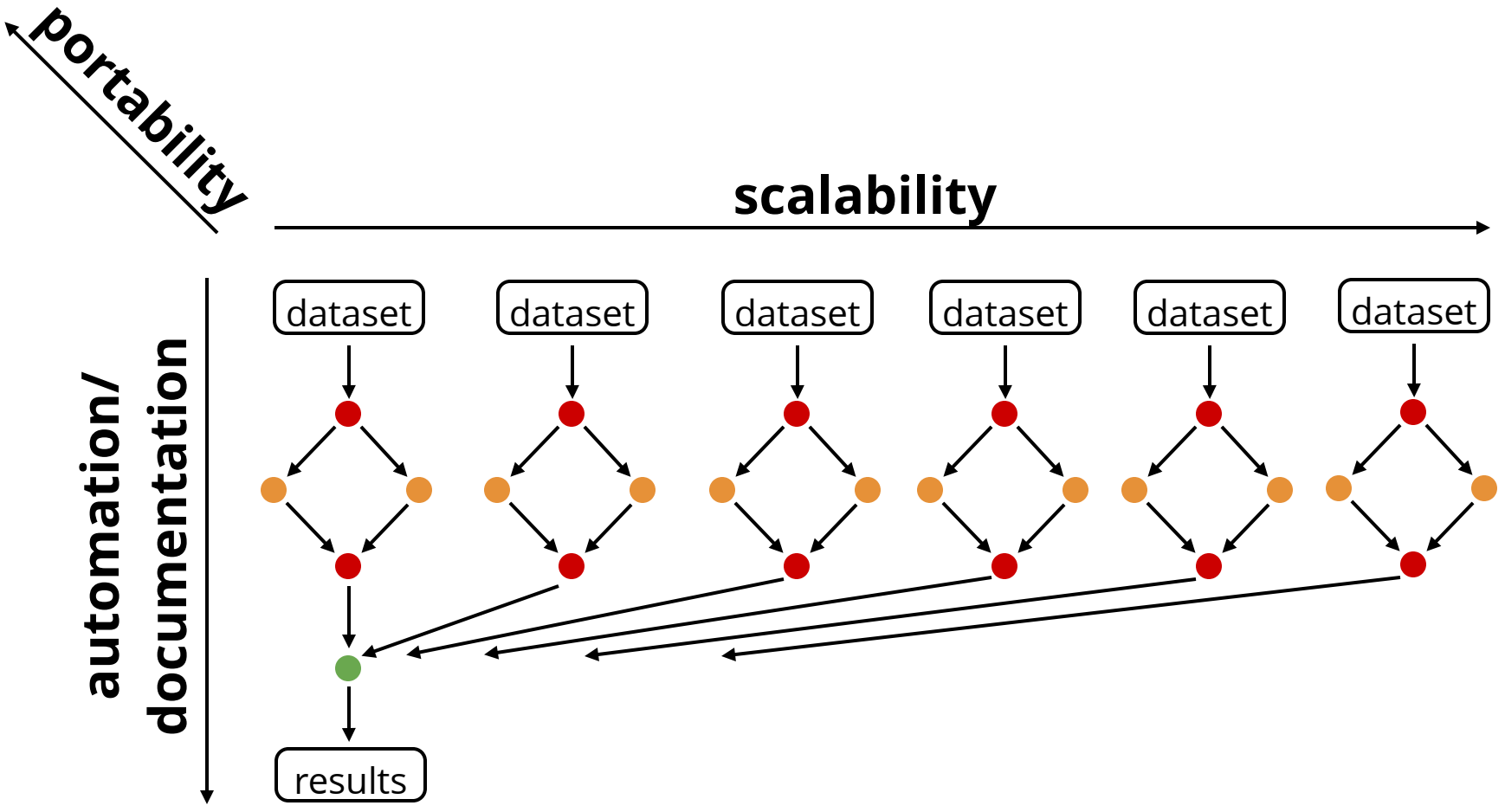
File	Size	Description	Job properties	
simulated-bwa.DEL.svg	328.3 kB	Allele frequency estimation error vs. true allele frequency on simulated data.	**Rule** plot_allelefreq **Wildcards** run=simulated-bwa, vartype=DEL **Params** varlociraptor_callers=['delly', 'lancet', 'manta', 'strelka'], len_ranges=[[1, 30], [30, 50], [50, 100], [100, 250]]	
simulated-bwa.DEL.svg	1.9 MB	Allele frequency estimation error vs. true allele frequency on simulated data. The dashed lines depict the standard deviation, solid line depicts the mean. The grey area shows the standard	**Rule** plot_allelefreq_scatter **Wildcards** run=simulated-bwa, vartype=DEL **Params** depth_ranges=[[1, 20], [20, 40]], callers=['delly', 'lancet', 'manta', 'strelka']	

https://snakemake.readthedocs.io

Snakemake



https://snakemake.readthedocs.io



https://snakemake.readthedocs.io