Snakemake

https://github.com/nichollskc/IMPC_analysis

https://snakemake.readthedocs.io/en/stable/index.html

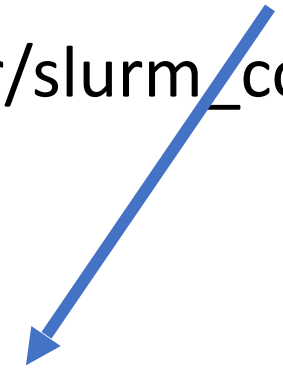
Idea of snakemake – defines dependencies

- Python version of Make
- Rules define inputs, outputs and commands to go from inputs to outputs
- Re-runs a rule if the inputs are older than the outputs

```
configfile: "config.yml"

rule fetch_fastq:
    params:
        url=lambda wildcards: config['FASTQ_URLS'][wildcards.ID]
    output:
        temp("data/fastq/{ID}.fastq.gz")
    shell:
        "wget -O {output} --no-verbose {params.url} 2>&1"

rule run_kallisto:
    input:
        "data/fastq/{ID}.fastq.gz"
    output:
        "data/kallisto/{ID}/abundance.tsv"
    shell:
        "kallisto quant --index={config[INDEX_FILE]} --output-dir=$(dir
name {output}) --single -l 50 -s 2 {input} 2>&1"

rule generate_count_matrix:
    input:
        expand("data/kallisto/{ID}/abundance.tsv",
               ID=list(config['FASTQ_URLS'].keys()))
    output:
        df="data/tpm.tsv"
    run:
        import worker
        df = worker.generat
        df.to_csv(output.df
```


```
snakemake data/tpm.tsv
```

snakemake --cluster “<COMMAND>” --jobs 100
--cluster-config cluster/slurm_config.json



Nice features

- Temporary/protected outputs
- Special status for log files
- Job groups – run jobs in one SLURM job
- R/python scripts can access snakemake variables e.g. list of inputs
- Different cluster settings for different jobs

- Excellent documentation

https://snakemake.readthedocs.io/en/stable/index.html