Workforce

Theo Portlock[1] Justin O'Sullivan[1]

[1]The University of Auckland

03-04-2024

Science

Science

- Repeatablilty

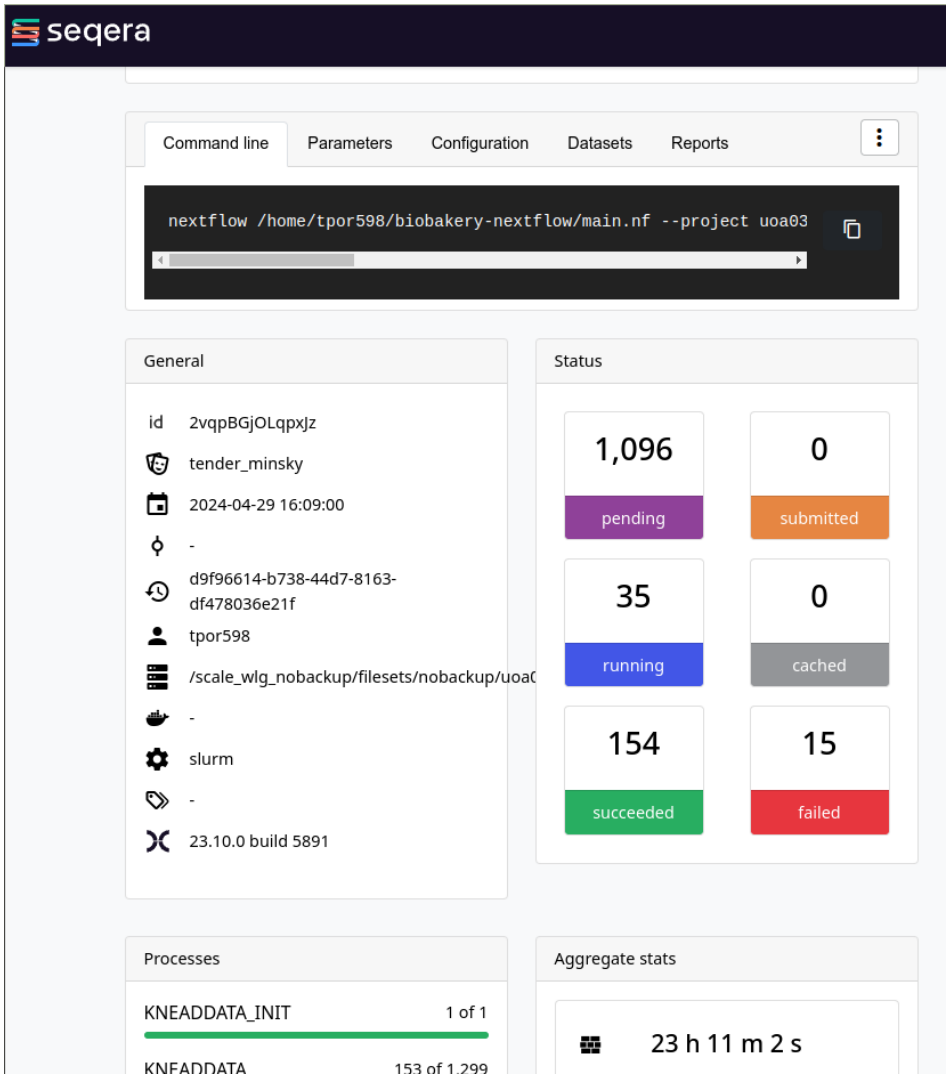
Requirements of reproducibility

- Easy to run/use
  - Good documentation
  - Detailed methods (version control)
  - Multiple platforms/OS
  - Cheap - maximises resources
  - Scaleable - as datasets get bigger, this becomes more important
- Easy to understand - (also makes it easier to upgrade)
  - Good tutorials
- Easy to troubleshoot/test
  - Able to resume from previous steps
  - Errors are clear
- Good data hygeine

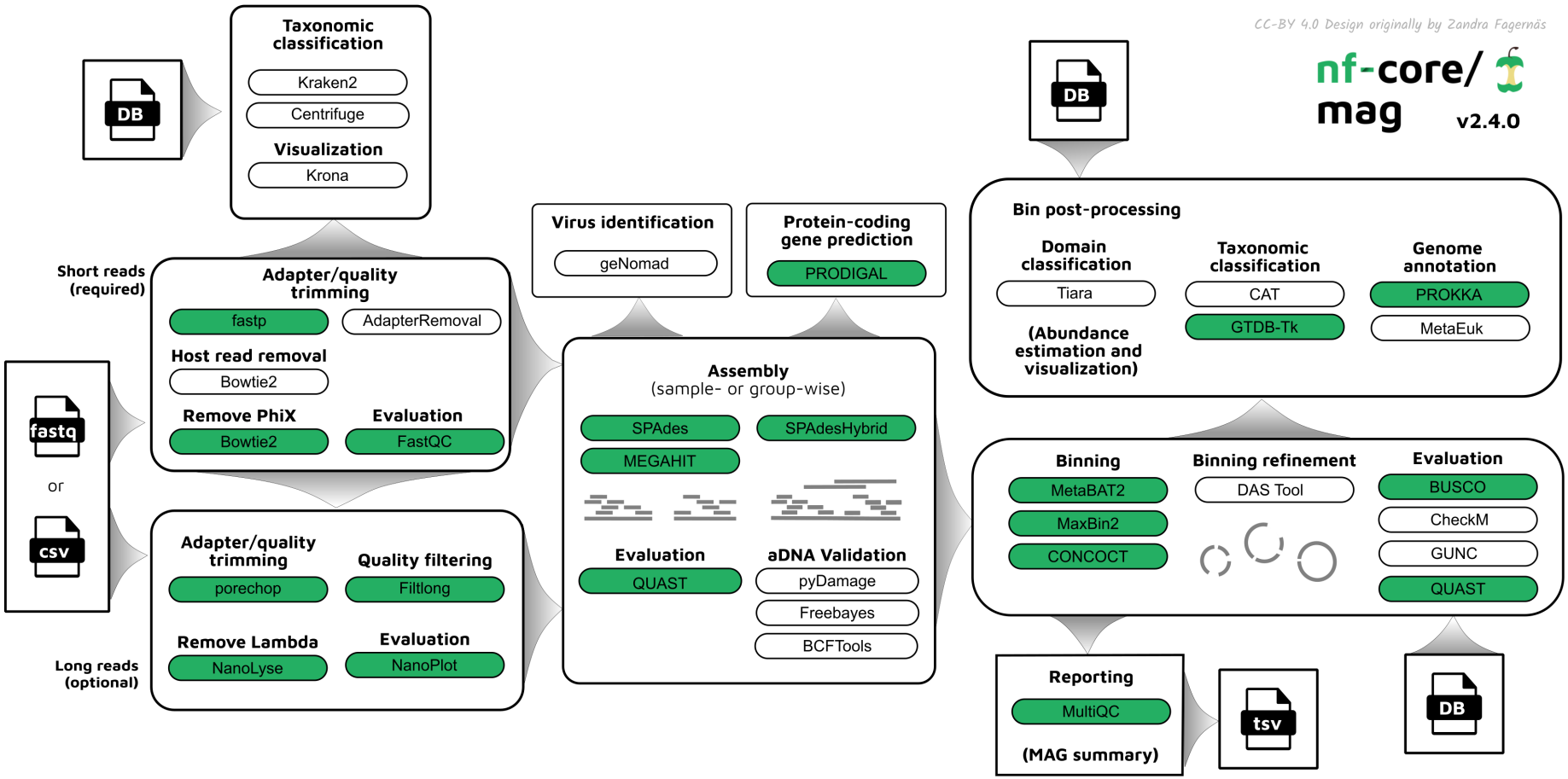
Nextflow

- Pipeline writing language
- Nf-core
- Sequera - Phil Ewels
- Meets *most* of these requirements
- Containerised, version controlled, and can run on multiple platforms
- Steep learning curve

Nextflow


seqera
Command line
Parameters
Configuration
Datasets
Reports
nextflow /home/tpor598/biobakery-nextflow/main.nf --project uoa03
General
id 2vqpBGjOLqpxJz
tender_minsky
2024-04-29 16:09:00
-
d9f96614-b738-44d7-8163-df478036e21f
tpor598
/scale_wlg_nobackup/filesets/nobackup/uoa0
-
slurm
-
23.10.0 build 5891
Status
1,096
pending
0
submitted
35
running
0
cached
154
succeeded
15
failed
Processes
KNEADDATA_INIT 1 of 1
KNEADDATA 153 of 1,299
Aggregate stats
23 h 11 m 2 s

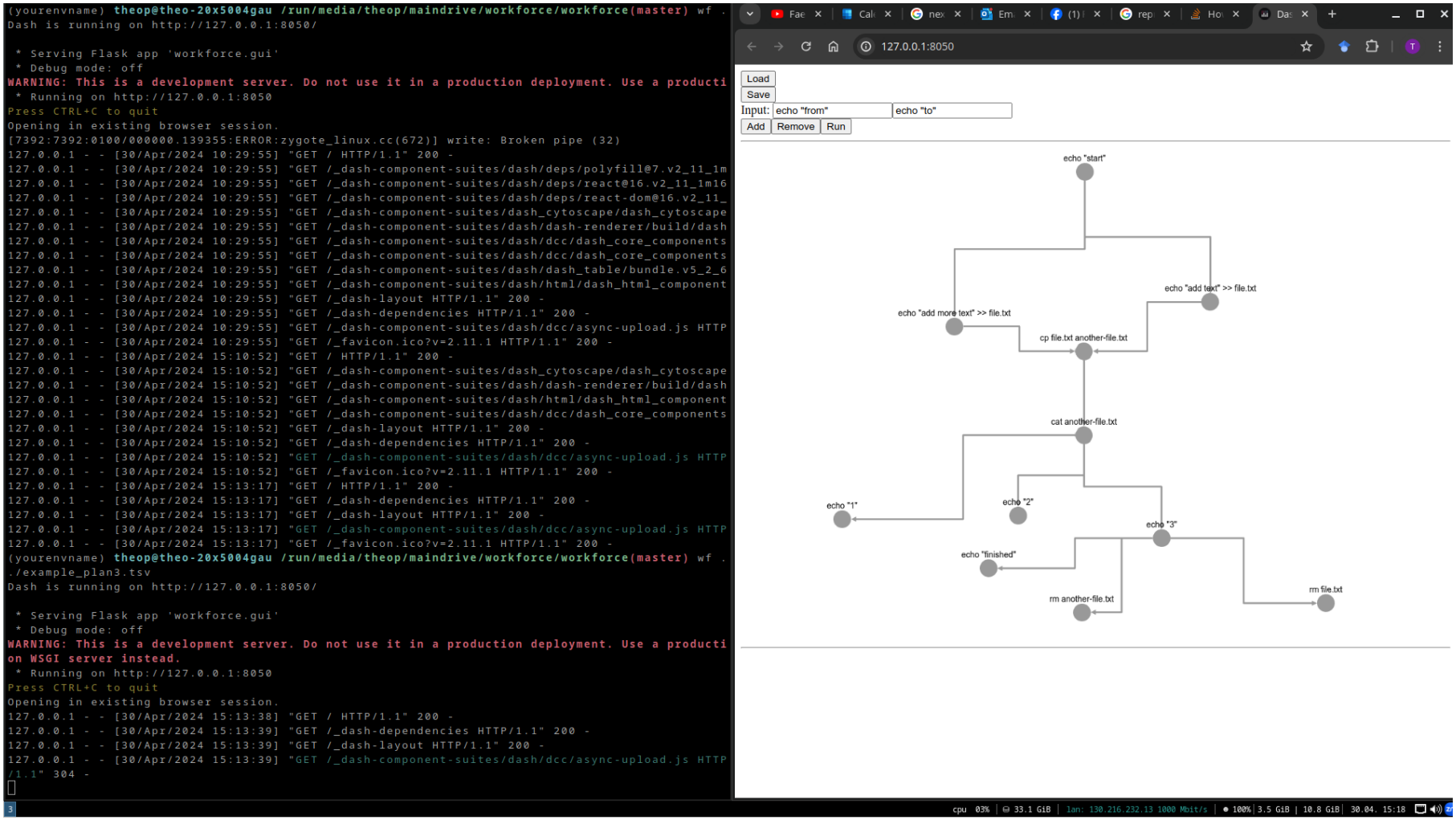
Nextflow


CC-BY 4.0 Design originally by Zandra Fagernäs
nf-core/mag v2.4.0
DB
Taxonomic classification
Kraken2
Centrifuge
Visualization
Krona
Short reads (required)
Adapter/quality trimming
fastp
AdapterRemoval
Host read removal
Bowtie2
Remove PhiX
Bowtie2
Evaluation
FastQC
fastq
or
csv
Long reads (optional)
Adapter/quality trimming
porechop
Quality filtering
Filtlong
Remove Lambda
NanoLyse
Evaluation
NanoPlot
Virus identification
geNomad
Protein-coding gene prediction
PRODIGAL
Assembly
(sample- or group-wise)
SPAdes
SPAdesHybrid
MEGAHIT
Evaluation
QUAST
aDNA Validation
pyDamage
Freebayes
BCFTools
DB
Bin post-processing
Domain classification
Tiara
(Abundance estimation and visualization)
Taxonomic classification
CAT
GTDB-Tk
Genome annotation
PROKKA
MetaEuk
Binning
MetaBAT2
MaxBin2
CONCOCT
Binning refinement
DAS Tool
Evaluation
BUSCO
CheckM
GUNC
QUAST
Reporting
MultiQC
(MAG summary)
tsv
DB

Downstream analysis

- 12,000 line R script, python file, or SPSS file
- Collection of random CUSTOM scripts released on github
- Slow
- Often difficult to install correct versions of packages
- Documentation limited to README or methods section, often split between
- Only run on one platform/OS
- ALL THE SAME REQUIREMENTS APPLY!

Workforce


(yourenvname) theop@theo-20x5004gau /run/media/theop/maindrive/workforce/workforce(master) wf .
Dash is running on http://127.0.0.1:8050/
* Serving Flask app 'workforce.gui'
* Debug mode: off
WARNING: This is a development server. Do not use it in a production deployment. Use a producti
* Running on http://127.0.0.1:8050
Press CTRL+C to quit
Opening in existing browser session.
[7392:7392:0100/000000.139355:ERROR:zygote_linux.cc(672)] write: Broken pipe (32)
127.0.0.1 - - [30/Apr/2024 10:29:55] "GET / HTTP/1.1" 200 -
127.0.0.1 - - [30/Apr/2024 10:29:55] "GET /_dash-component-suites/dash/deps/polyfill@7.v2_11_1m
127.0.0.1 - - [30/Apr/2024 10:29:55] "GET /_dash-component-suites/dash/deps/react@16.v2_11_1m16
127.0.0.1 - - [30/Apr/2024 10:29:55] "GET /_dash-component-suites/dash/deps/react-dom@16.v2_11_
127.0.0.1 - - [30/Apr/2024 10:29:55] "GET /_dash-component-suites/dash_cytoscape/dash_cytoscape
127.0.0.1 - - [30/Apr/2024 10:29:55] "GET /_dash-component-suites/dash/dash-renderer/build/dash
127.0.0.1 - - [30/Apr/2024 10:29:55] "GET /_dash-component-suites/dash/dcc/dash_core_components
127.0.0.1 - - [30/Apr/2024 10:29:55] "GET /_dash-component-suites/dash/dcc/dash_core_components
127.0.0.1 - - [30/Apr/2024 10:29:55] "GET /_dash-component-suites/dash/dash_table/bundle.v5_2_6
127.0.0.1 - - [30/Apr/2024 10:29:55] "GET /_dash-component-suites/dash/html/dash_html_component
127.0.0.1 - - [30/Apr/2024 10:29:55] "GET /_dash-layout HTTP/1.1" 200 -
127.0.0.1 - - [30/Apr/2024 10:29:55] "GET /_dash-dependencies HTTP/1.1" 200 -
127.0.0.1 - - [30/Apr/2024 10:29:55] "GET /_dash-component-suites/dash/dcc/async-upload.js HTTP
127.0.0.1 - - [30/Apr/2024 10:29:55] "GET /_favicon.ico?v=2.11.1 HTTP/1.1" 200 -
127.0.0.1 - - [30/Apr/2024 15:10:52] "GET / HTTP/1.1" 200 -
127.0.0.1 - - [30/Apr/2024 15:10:52] "GET /_dash-component-suites/dash_cytoscape/dash_cytoscape
127.0.0.1 - - [30/Apr/2024 15:10:52] "GET /_dash-component-suites/dash/dash-renderer/build/dash
127.0.0.1 - - [30/Apr/2024 15:10:52] "GET /_dash-component-suites/dash/html/dash_html_component
127.0.0.1 - - [30/Apr/2024 15:10:52] "GET /_dash-component-suites/dash/dcc/dash_core_components
127.0.0.1 - - [30/Apr/2024 15:10:52] "GET /_dash-layout HTTP/1.1" 200 -
127.0.0.1 - - [30/Apr/2024 15:10:52] "GET /_dash-dependencies HTTP/1.1" 200 -
127.0.0.1 - - [30/Apr/2024 15:10:52] "GET /_dash-component-suites/dash/dcc/async-upload.js HTTP
127.0.0.1 - - [30/Apr/2024 15:10:52] "GET /_favicon.ico?v=2.11.1 HTTP/1.1" 200 -
127.0.0.1 - - [30/Apr/2024 15:13:17] "GET / HTTP/1.1" 200 -
127.0.0.1 - - [30/Apr/2024 15:13:17] "GET /_dash-dependencies HTTP/1.1" 200 -
127.0.0.1 - - [30/Apr/2024 15:13:17] "GET /_dash-layout HTTP/1.1" 200 -
127.0.0.1 - - [30/Apr/2024 15:13:17] "GET /_dash-component-suites/dash/dcc/async-upload.js HTTP
127.0.0.1 - - [30/Apr/2024 15:13:17] "GET /_favicon.ico?v=2.11.1 HTTP/1.1" 200 -
(yourenvname) theop@theo-20x5004gau /run/media/theop/maindrive/workforce/workforce(master) wf .
./example_plan3.tsv
Dash is running on http://127.0.0.1:8050/
* Serving Flask app 'workforce.gui'
* Debug mode: off
WARNING: This is a development server. Do not use it in a production deployment. Use a producti
on WSGI server instead.
* Running on http://127.0.0.1:8050
Press CTRL+C to quit
Opening in existing browser session.
127.0.0.1 - - [30/Apr/2024 15:13:38] "GET / HTTP/1.1" 200 -
127.0.0.1 - - [30/Apr/2024 15:13:39] "GET /_dash-dependencies HTTP/1.1" 200 -
127.0.0.1 - - [30/Apr/2024 15:13:39] "GET /_dash-layout HTTP/1.1" 200 -
127.0.0.1 - - [30/Apr/2024 15:13:39] "GET /_dash-component-suites/dash/dcc/async-upload.js HTTP
/1.1" 304 -
127.0.0.1:8050
Load
Save
Input:
echo "from"
echo "to"
Add
Remove
Run
echo "start"
echo "add text" >> file.txt
echo "add more text" >> file.txt
cp file.txt another-file.txt
cat another-file.txt
echo "1"
echo "2"
echo "3"
echo "finished"
rm another-file.txt
rm file.txt

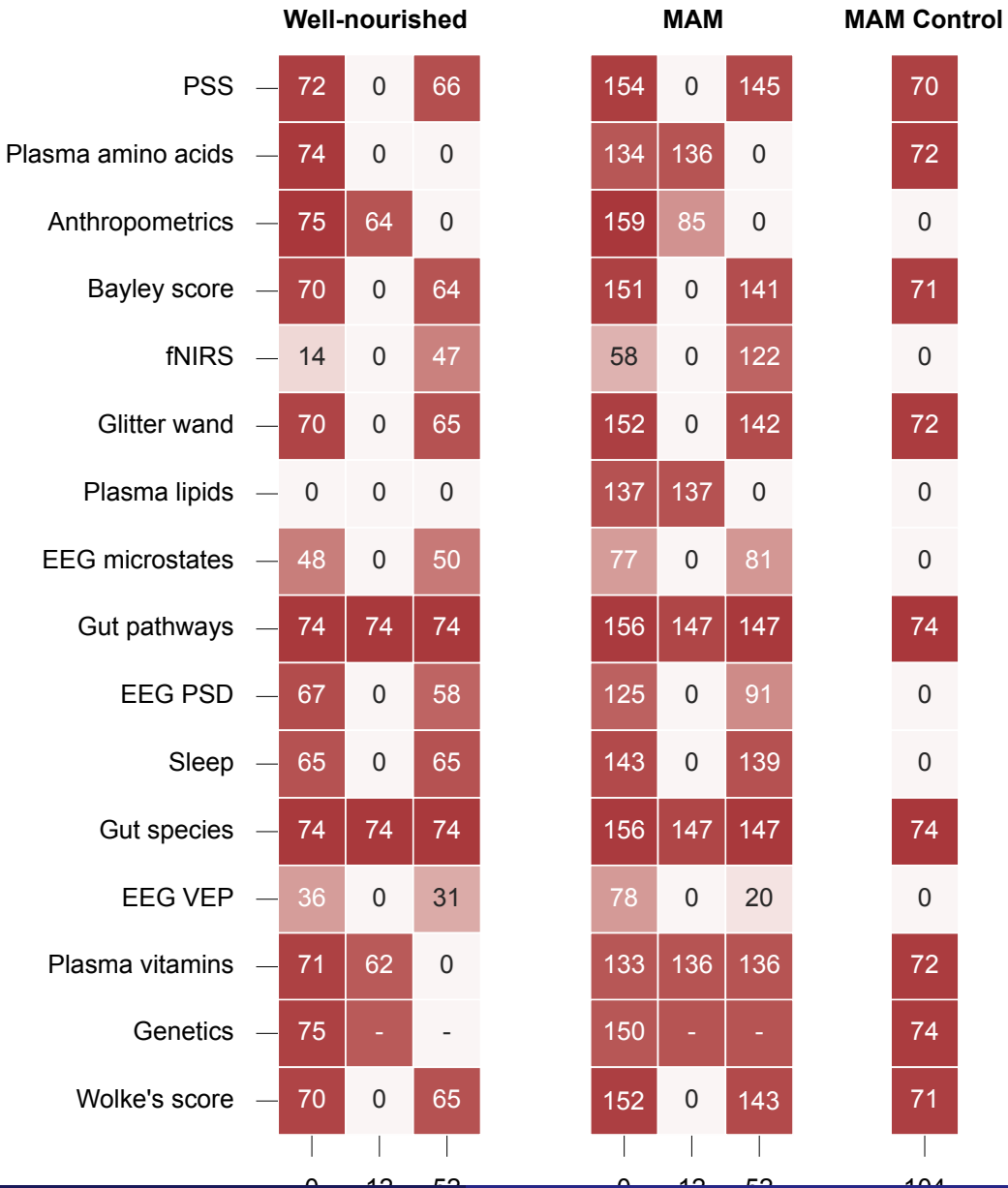
Workforce

- A DCG (Directed Cyclic graph) based pipeline designer for downstream analysis
- Installed with ‘pip install workforce’
- Integrates into existing tools
- Uses less code
- Process tracking
- Install with pip on any operating system
- Can run over multiple servers in parallel
- Leans on existing command line commands
- Simple interface
- Can interact with a run while it’s still running
- Collaberative development

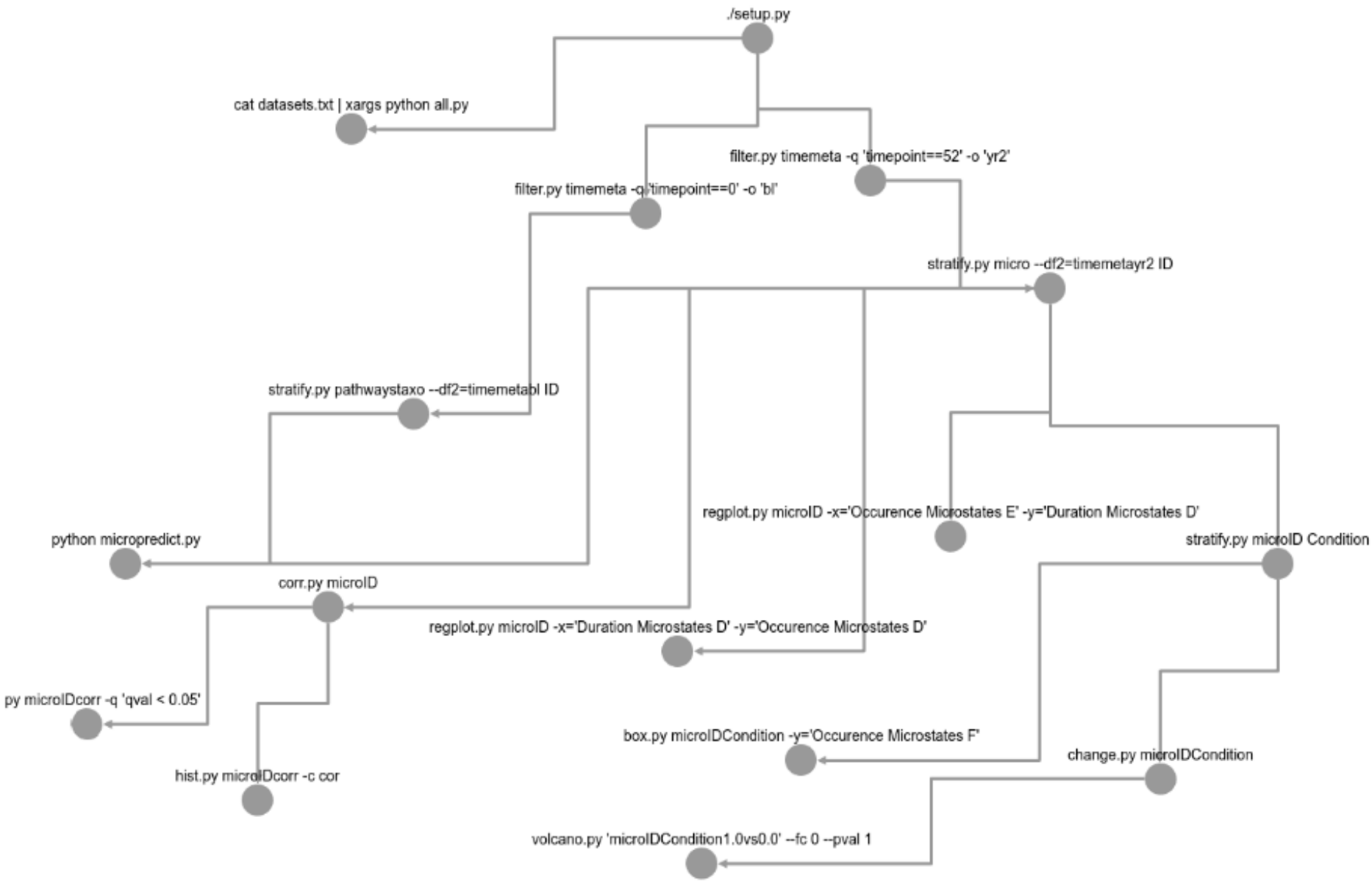
Workforce

- networkx
- subprocess
- multiprocessing
- dash cytoscape

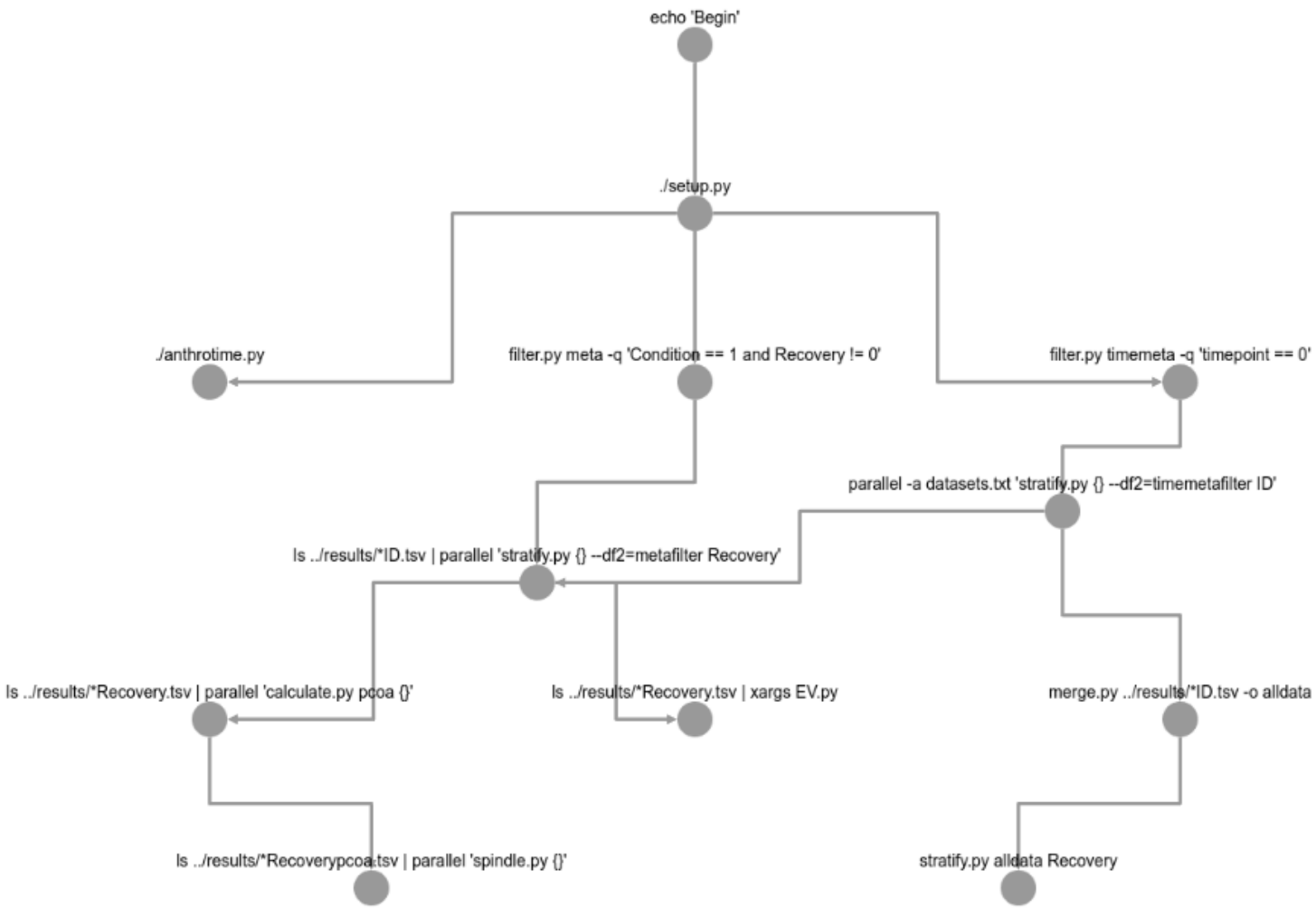
Workforce - case study



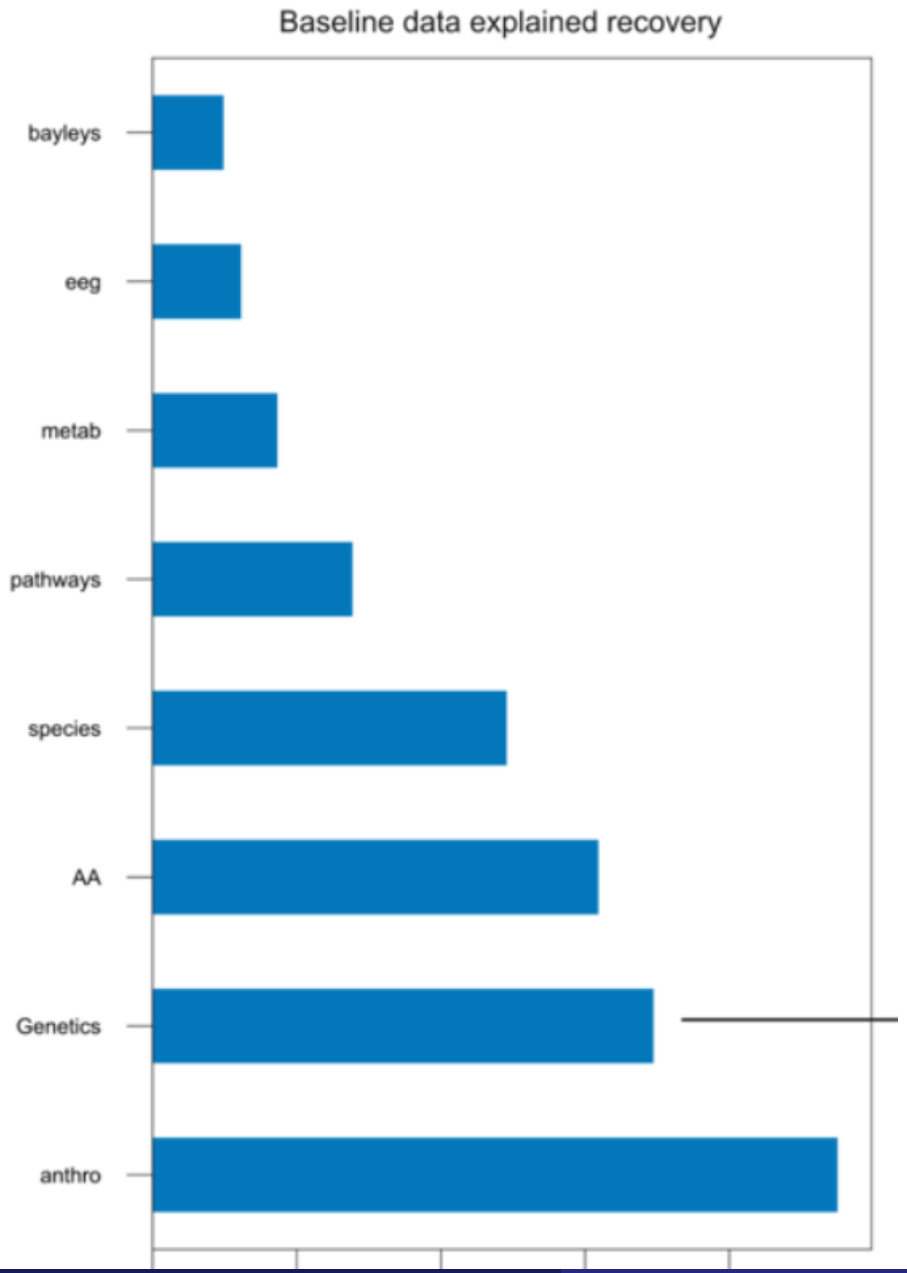
Workforce - case study



Workforce - case study



Workforce - case study



Future work

- Continuous workflows for data recording

Final thoughts

- Workforce is a pipeline designer for downstream analysis
- It runs processes in parallel and across servers
- Much more testing is required!