wellcome
connecting
science

Bioinformatics Essentials

Dr. Jacqui Keane

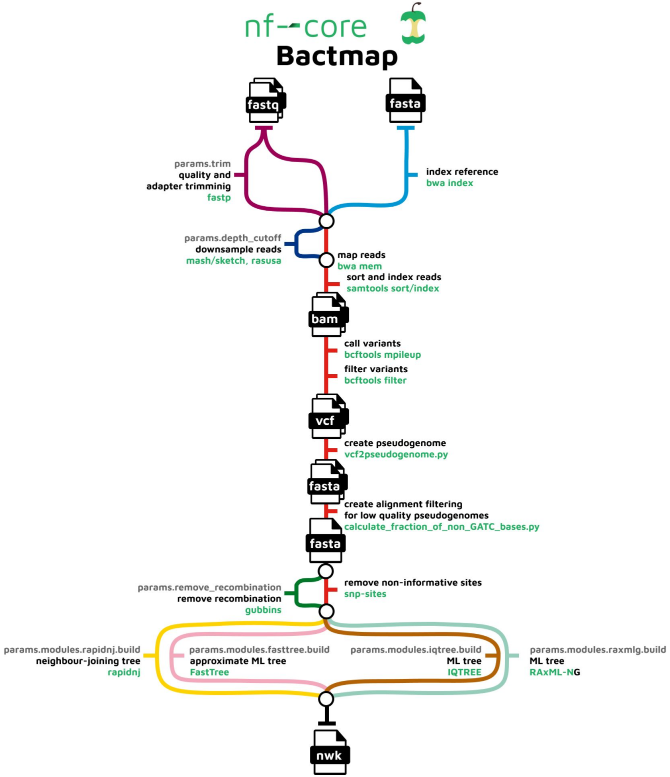
🦋 @drjkeane
drjkeane@gmail.com

Overview

- Bioinformatics Software

Bioinformatics Databases, Software, and Tools
FlyBase
swissprot
MEGA
Molecular Evolutionary Genetics Analysis
DDBJ
DNA Data Bank of Japan
RCSB PDB
PROTEIN DATA BANK
BLAST
Basic Local Alignment Search Tool
NCBI
W2
CLUSTAL

- Bioinformatics Pipelines



nf-core
Bactmap

fastq
fasta

params.trim
quality and
adapter trimminig
fastp

index reference
bwa index

params.depth_cutoff
downsample reads
mash/sketch, rasusa

map reads
bwa mem

sort and index reads
samtools sort/index

bam

call variants
bcftools mpileup

filter variants
bcftools filter

vcf

create pseudogenome
vcf2pseudogenome.py

fasta

create alignment filtering
for low quality pseudogenomes
calculate_fraction_of_non_GATC_bases.py

fasta

remove non-informative sites
snp-sites

params.remove_recombination
remove recombination
gubbins

params.modules.rapidnj.build
neighbour-joining tree
rapidnj

params.modules.fasttree.build
approximate ML tree
FastTree

params.modules.iqtree.build
ML tree
IQTREE

params.modules.raxmlg.build
ML tree
RAxML-NG

nwk

Learning Outcomes

- You can expect to be able to:
  - Describe what a software package manager is and why are useful
  - Install a software package manager (conda/mamba)
  - Install bioinformatics software with the conda package manager
  - Describe what a workflow manager is and why useful
  - Install Nextflow and run a Nextflow pipeline

Overview

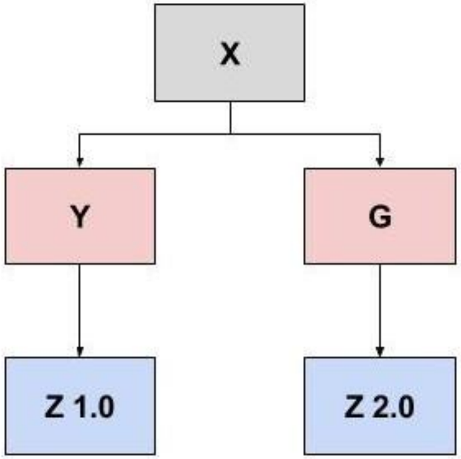
- Bioinformatics Software

Bioinformatics Databases, Software, and Tools
FlyBase
swissprot
M E G A
Molecular Evolutionary Genetics Analysis
DDBJ
DNA Data Bank of Japan
RCSB PDB
PROTEIN DATA BANK
BLAST
Basic Local Alignment Search Tool
NCBI
W2
CLUSTAL

Bioinformatics Software

- Many bioinformatics software
- Complex to install and manage
  - - e.g. conflicting dependencies
- Use software package manager

Dependencies Tree



Conda

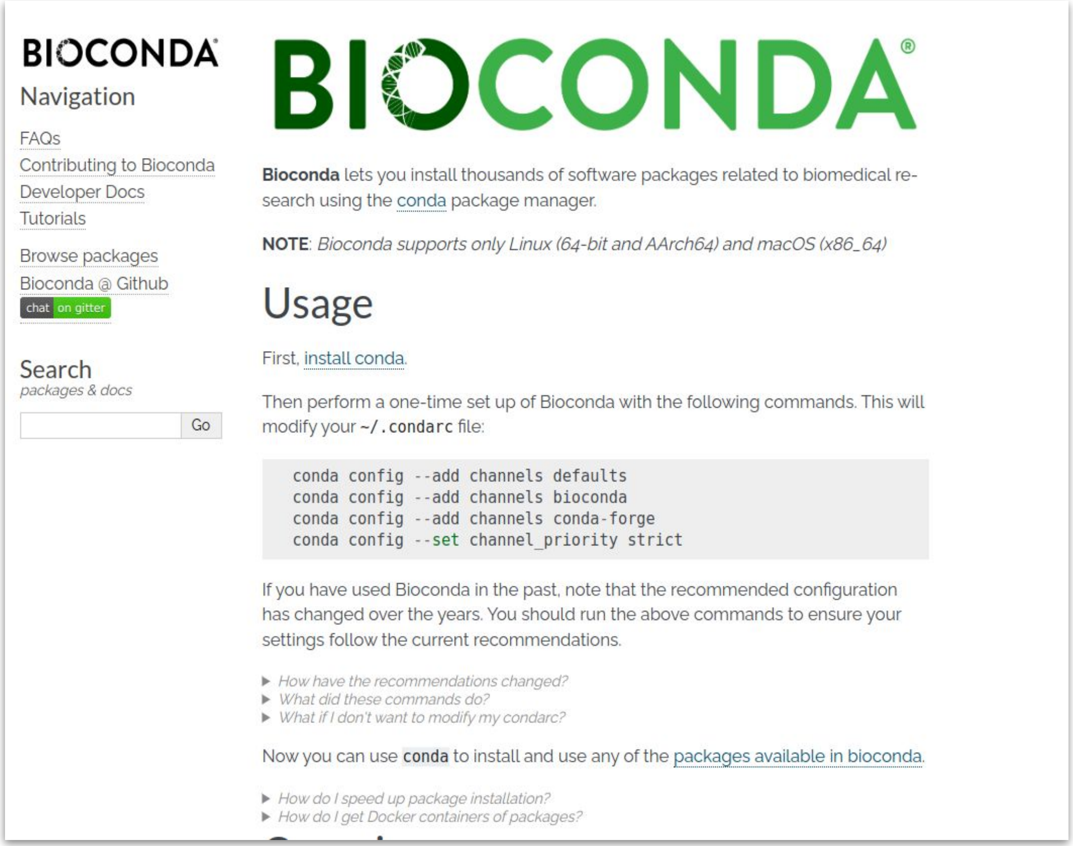
- Automates the process of installing software and their dependencies

- Software applications available via channels

- A channel online locations of where packages are stored

e.g. conda-forge, r, bioconda

Bioconda channel



BIOCONDA

Navigation

- FAQs
- Contributing to Bioconda
- Developer Docs
- Tutorials
- Browse packages
- Bioconda @ Github

Search
packages & docs

Bioconda lets you install thousands of software packages related to biomedical research using the conda package manager.

NOTE: *Bioconda supports only Linux (64-bit and AArch64) and macOS (x86_64)*

Usage

First, install conda.

Then perform a one-time set up of Bioconda with the following commands. This will modify your `~/.condarc` file:

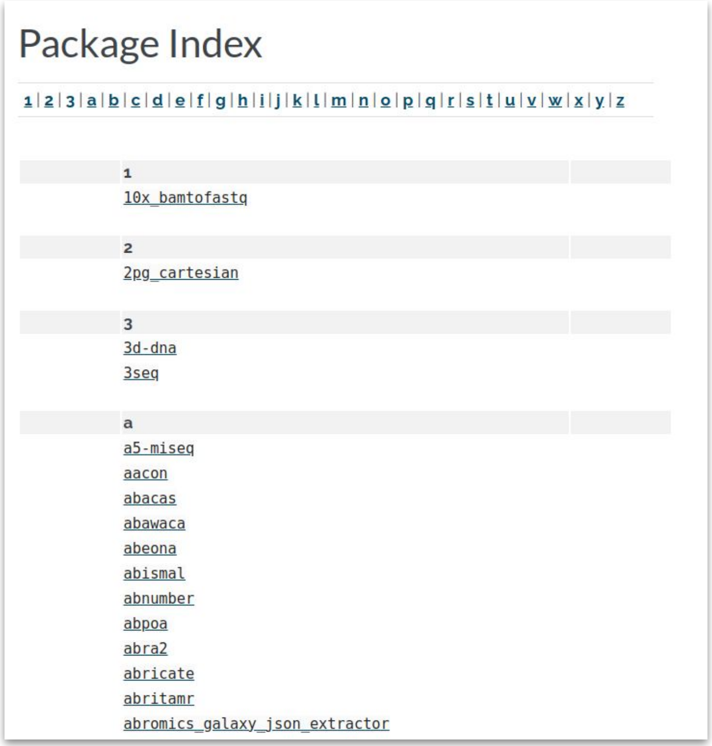
```
conda config --add channels defaults
conda config --add channels bioconda
conda config --add channels conda-forge
conda config --set channel_priority strict
```

If you have used Bioconda in the past, note that the recommended configuration has changed over the years. You should run the above commands to ensure your settings follow the current recommendations.

- ▶ *How have the recommendations changed?*
- ▶ *What did these commands do?*
- ▶ *What if I don't want to modify my condarc?*

Now you can use `conda` to install and use any of the packages available in bioconda.

- ▶ *How do I speed up package installation?*
- ▶ *How do I get Docker containers of packages?*



Package Index

1 | 2 | 3 | a | b | c | d | e | f | g | h | i | j | k | l | m | n | o | p | q | r | s | t | u | v | w | x | y | z

1

10x_bamtofastq

2

2pg_cartesian

3

3d-dna
3seq

a

a5-miseq
aacon
abacas
abawaca
abeona
abismal
abnumber
abpoa
abra2
abricate
abritamr
abromics_galaxy_json_extractor

Conda environments

- A separate space (box) where you install specific versions of software/set of softwares

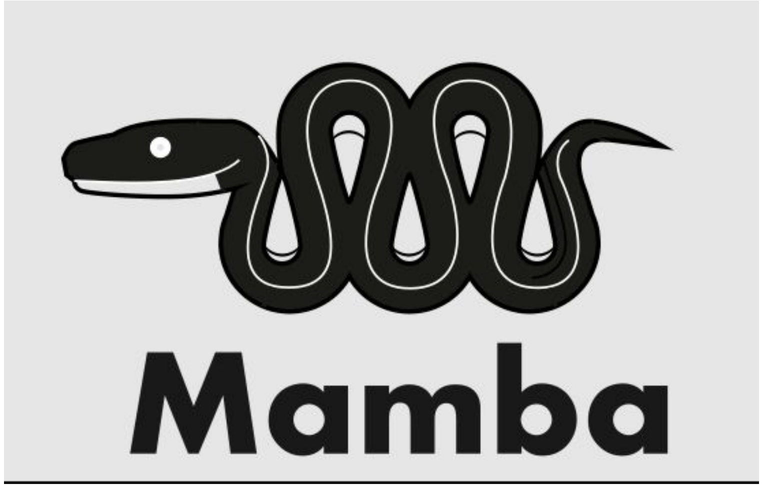
- Prevent conflicts between different softwares by allowing you to manage dependencies separately

- Create, activate, and switch between environments easily

CONDA

Mamba

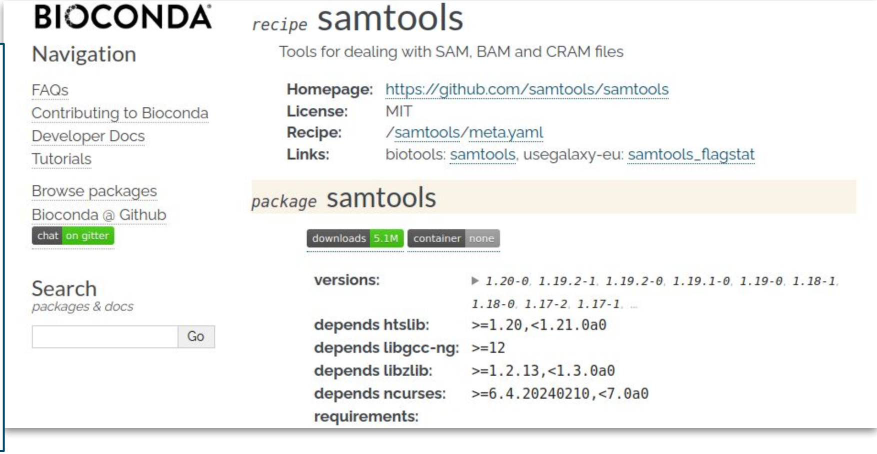
- A software package manager similar to Conda
- Faster and uses less memory, good for installing software faster and using less computer resources
- Fully compatible with conda



Demo

- Bioinformatics Software
  - Install conda from: https://docs.conda.io/projects/conda/en/latest/user-guide/install/linux.html
  - Install mamba from: https://mamba.readthedocs.io/en/latest/installation/mamba-installation.html
  - Use conda to install samtools

```
$ conda info --envs
$ conda list
$ samtools
$ conda create -n samtools-1.21 samtools=1.21.0
$ conda info -envs
$ conda activate samtools-1.21
$ conda list
$ samtools
$ conda deactivate
```



recipe samtools

Tools for dealing with SAM, BAM and CRAM files

Homepage: https://github.com/samtools/samtools
License: MIT
Recipe: /samtools/meta.yaml
Links: biotools: samtools, usegalaxy-eu: samtools_flagstat

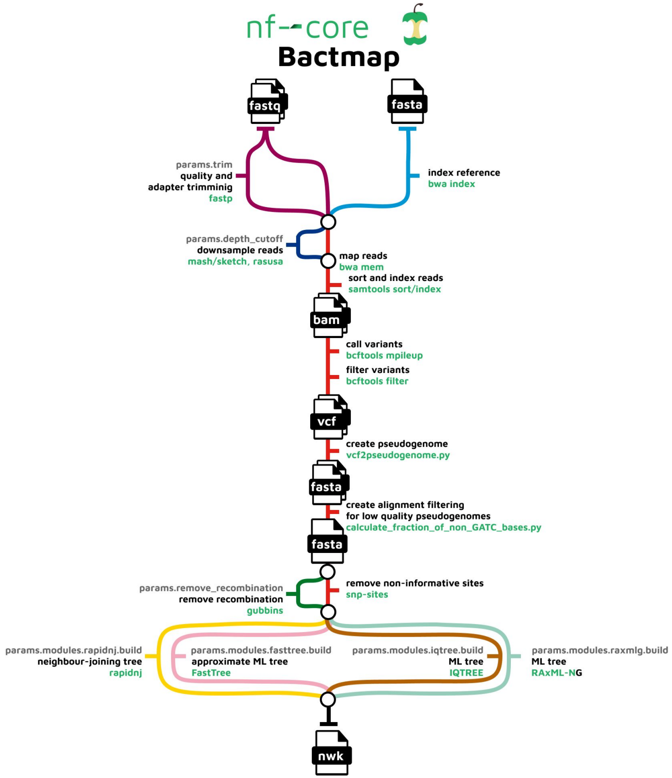
package samtools

downloads 5.1M container none

versions: ▸ 1.20-0, 1.19.2-1, 1.19.2-0, 1.19.1-0, 1.19-0, 1.18-1, 1.18-0, 1.17-2, 1.17-1, ...
depends htslib: >=1.20,<1.21.0a0
depends libgcc-ng: >=12
depends libzlib: >=1.2.13,<1.3.0a0
depends ncurses: >=6.4.20240210,<7.0a0
requirements:

Overview

- Bioinformatics Pipelines

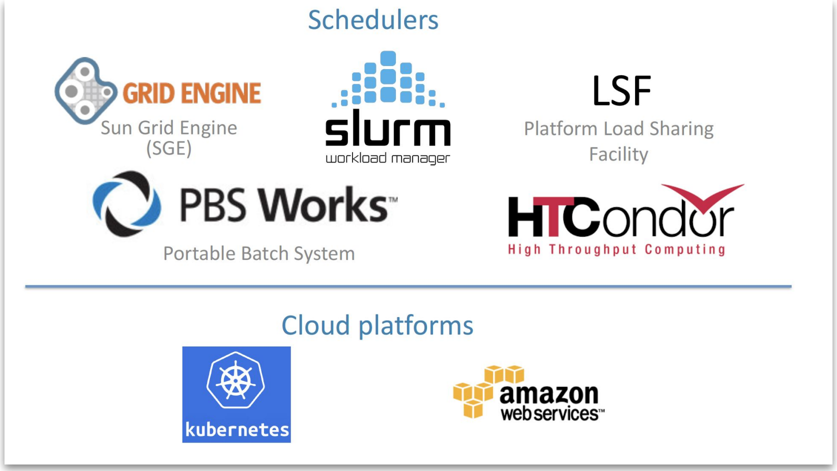


Bioinformatics Pipelines

- Automate analysis using a pipeline
  - e.g. bash script

- Re-write the pipeline for different platforms

- Use workflow/pipeline management system



Nextflow

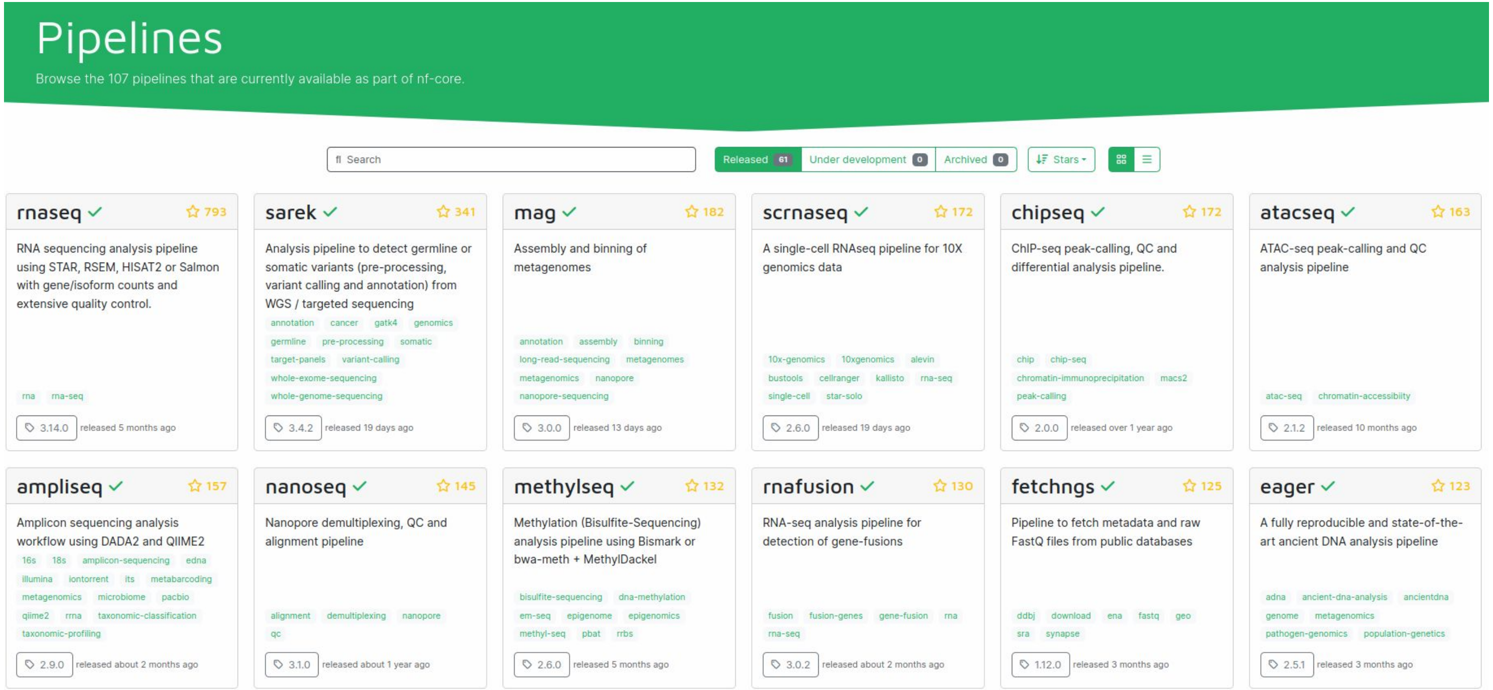
- Use it to write and run data-intensive workflows/pipelines

- Nextflow consists of a
  - ***language*** to write pipelines
  - ***engine*** to run pipelines

- Write a pipeline once and run everywhere!

nf-core

- Community effort to collect a set of best practice analysis pipelines built using Nextflow



Pipelines

Browse the 107 pipelines that are currently available as part of nf-core.

rnaseq ✓ ☆ 793
RNA sequencing analysis pipeline using STAR, RSEM, HISAT2 or Salmon with gene/isoform counts and extensive quality control.
rna rna-seq
3.14.0 released 5 months ago

sarek ✓ ☆ 341
Analysis pipeline to detect germline or somatic variants (pre-processing, variant calling and annotation) from WGS / targeted sequencing
annotation cancer gatk4 genomics germline pre-processing somatic target-panels variant-calling whole-exome-sequencing whole-genome-sequencing
3.4.2 released 19 days ago

mag ✓ ☆ 182
Assembly and binning of metagenomes
annotation assembly binning long-read-sequencing metagenomes metagenomics nanopore nanopore-sequencing
3.0.0 released 13 days ago

scrnaseq ✓ ☆ 172
A single-cell RNAseq pipeline for 10X genomics data
10x-genomics 10xgenomics alevin bustools cellranger kallisto rna-seq single-cell star-solo
2.6.0 released 19 days ago

chipseq ✓ ☆ 172
ChIP-seq peak-calling, QC and differential analysis pipeline.
chip chip-seq chromatin-immunoprecipitation macs2 peak-calling
2.0.0 released over 1 year ago

atacseq ✓ ☆ 163
ATAC-seq peak-calling and QC analysis pipeline
atac-seq chromatin-accessibilty
2.1.2 released 10 months ago

ampliseq ✓ ☆ 157
Amplicon sequencing analysis workflow using DADA2 and QIIME2
16s 18s amplicon-sequencing edna illumina iontorrent its metabarcoding metagenomics microbiome pacbio qiime2 rrna taxonomic-classification taxonomic-profiling
2.9.0 released about 2 months ago

nanoseq ✓ ☆ 145
Nanopore demultiplexing, QC and alignment pipeline
alignment demultiplexing nanopore qc
3.1.0 released about 1 year ago

methylseq ✓ ☆ 132
Methylation (Bisulfite-Sequencing) analysis pipeline using Bismark or bwa-meth + MethylDackel
bisulfite-sequencing dna-methylation em-seq epigenome epigenomics methyl-seq pbat rrbs
2.6.0 released 5 months ago

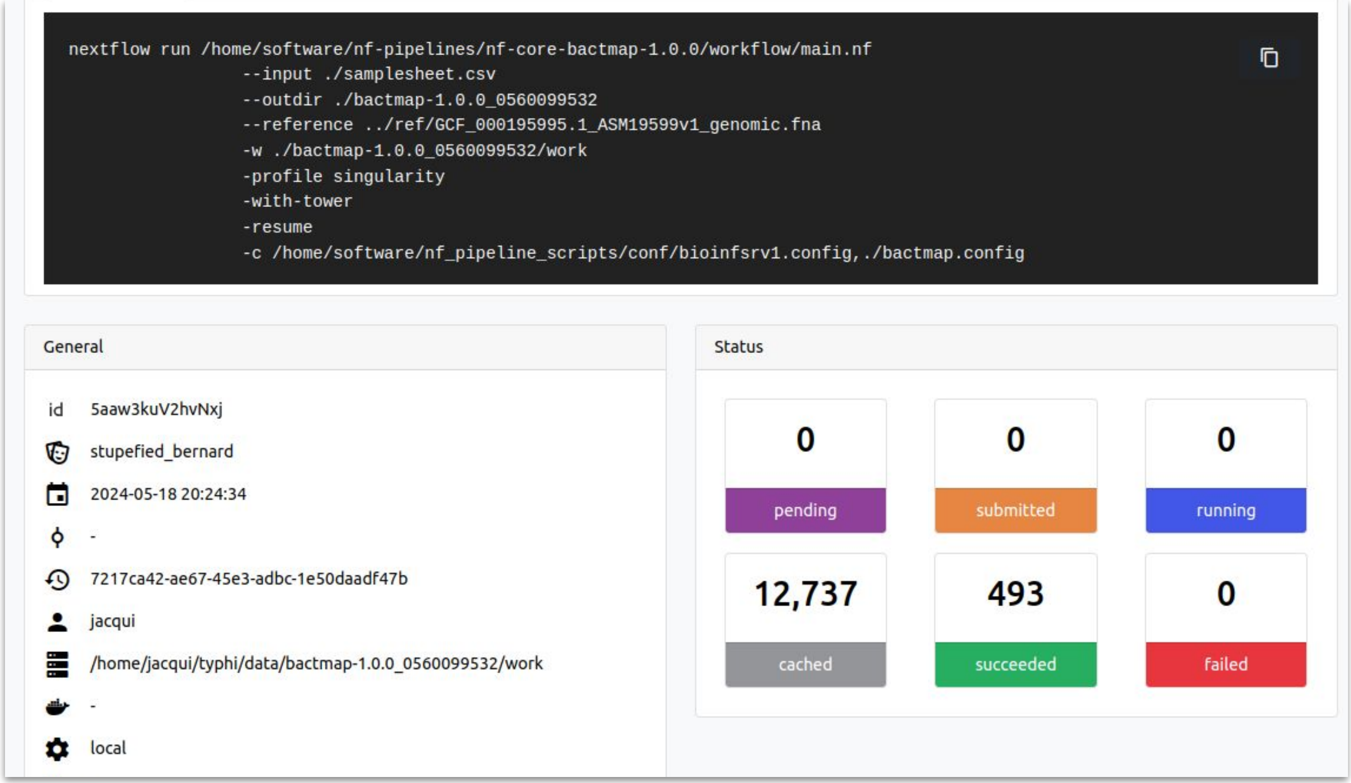
rnafusion ✓ ☆ 130
RNA-seq analysis pipeline for detection of gene-fusions
fusion fusion-genes gene-fusion rna rna-seq
3.0.2 released about 2 months ago

fetchngs ✓ ☆ 125
Pipeline to fetch metadata and raw FastQ files from public databases
ddbj download ena fastq geo sra synapse
1.12.0 released 3 months ago

eager ✓ ☆ 123
A fully reproducible and state-of-the-art ancient DNA analysis pipeline
adna ancient-dna-analysis ancientdna genome metagenomics pathogen-genomics population-genetics
2.5.1 released 3 months ago

Sequera Platform (Nextflow Tower)

- Manages and tracks the executions of Nextflow pipelines



```
nextflow run /home/software/nf-pipelines/nf-core-bactmap-1.0.0/workflow/main.nf
                 --input ./samplesheet.csv
                 --outdir ./bactmap-1.0.0_0560099532
                 --reference ../ref/GCF_000195995.1_ASM19599v1_genomic.fna
                 -w ./bactmap-1.0.0_0560099532/work
                 -profile singularity
                 -with-tower
                 -resume
                 -c /home/software/nf_pipeline_scripts/conf/bioinfsrv1.config,./bactmap.config
```

General

id 5aaw3kuV2hvNxj

stupefied_bernard

2024-05-18 20:24:34

-

7217ca42-ae67-45e3-adbc-1e50daadf47b

jacqui

/home/jacqui/typhi/data/bactmap-1.0.0_0560099532/work

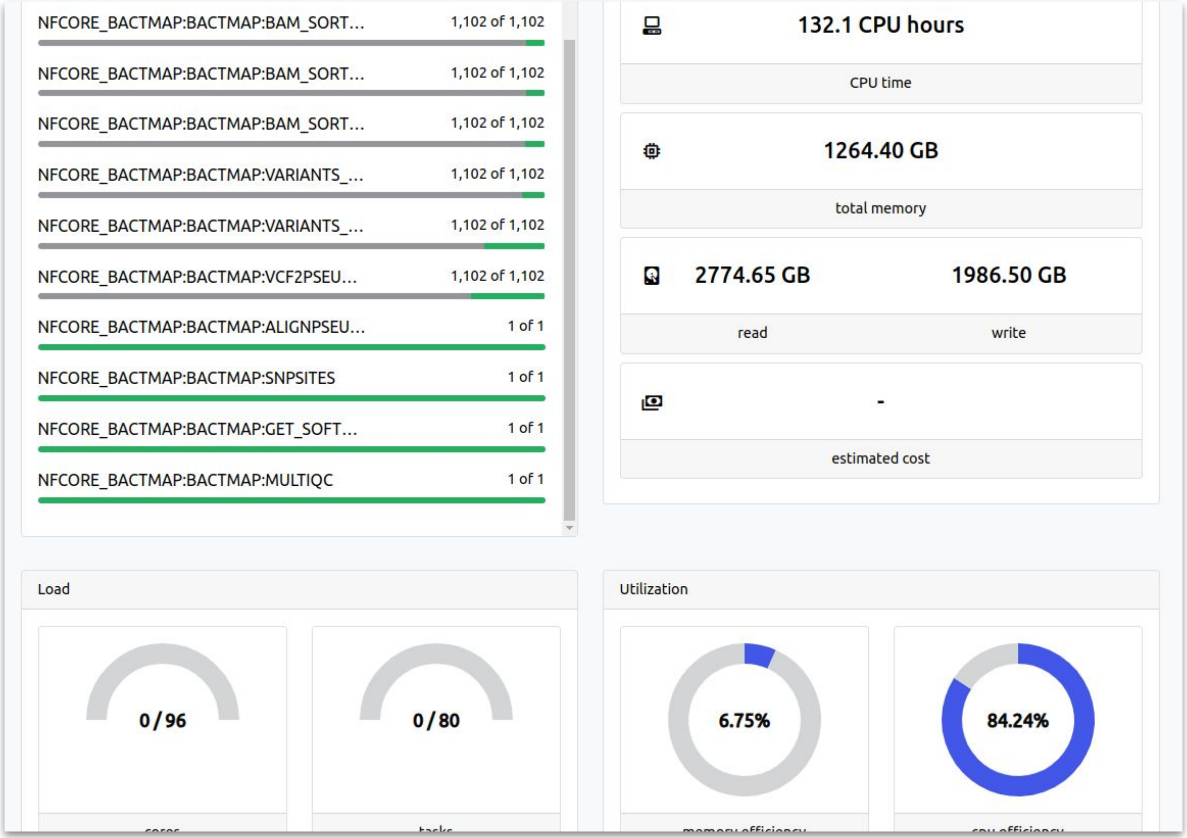
-

local

Status

0	0	0
pending	submitted	running
12,737	493	0
cached	succeeded	failed

Sequera Platform (Nextflow Tower)



NFCORE_BACTMAP:BACTMAP:BAM_SORT...	1,102 of 1,102
NFCORE_BACTMAP:BACTMAP:BAM_SORT...	1,102 of 1,102
NFCORE_BACTMAP:BACTMAP:BAM_SORT...	1,102 of 1,102
NFCORE_BACTMAP:BACTMAP:VARIANTS_...	1,102 of 1,102
NFCORE_BACTMAP:BACTMAP:VARIANTS_...	1,102 of 1,102
NFCORE_BACTMAP:BACTMAP:VCF2PSEU...	1,102 of 1,102
NFCORE_BACTMAP:BACTMAP:ALIGNPSEU...	1 of 1
NFCORE_BACTMAP:BACTMAP:SNPSITES	1 of 1
NFCORE_BACTMAP:BACTMAP:GET_SOFT...	1 of 1
NFCORE_BACTMAP:BACTMAP:MULTIQC	1 of 1

132.1 CPU hours

CPU time

1264.40 GB

total memory

2774.65 GB read

1986.50 GB write

-

estimated cost

Load

0 / 96

cores

0 / 80

tasks

Utilization

6.75%

memory efficiency

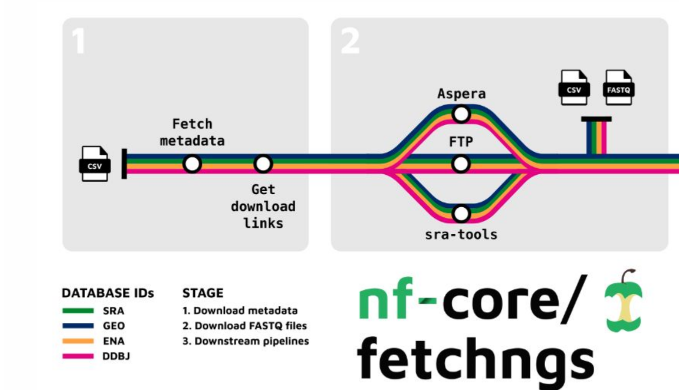
84.24%

cpu efficiency

Demo

- Bioinformatics Pipelines
  - Install Nextflow and nf-core
  - Install nf-core pipeline
    - fetchngs
  - Run the fetchngs pipeline

```
$ conda create -n nf-pipelines
$ conda activate nf-pipelines
$ conda install nextflow nf-core
$ nf-core pipelines list
$ nf-core pipelines list | grep -i
rna
```



```
$ wget
https://github.com/sylabs/singularity/releases/download/v4.2.2/singularity-ce_4.2.2-noble_amd64.deb
$ sudo dpkg -i  singularity-ce_4.2.2-noble_amd64.deb
$ nextflow run nf-core/fetchngs -profile singularity –input ids.csv –outdir .
```

Questions?