



wellcome
connecting
science

Bioinformatics Essentials

Dr. Jacqui Keane

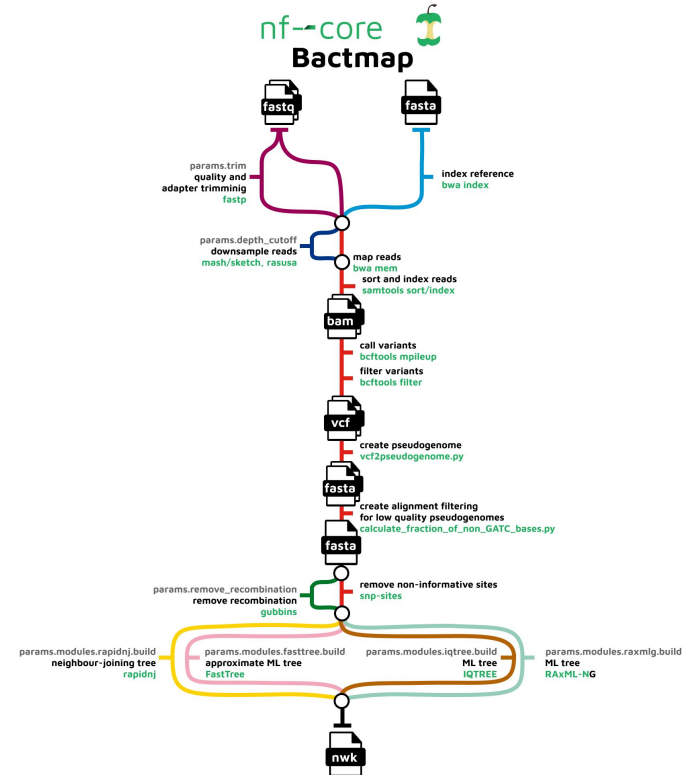


Overview

► Bioinformatics Software



► Bioinformatics Pipelines



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

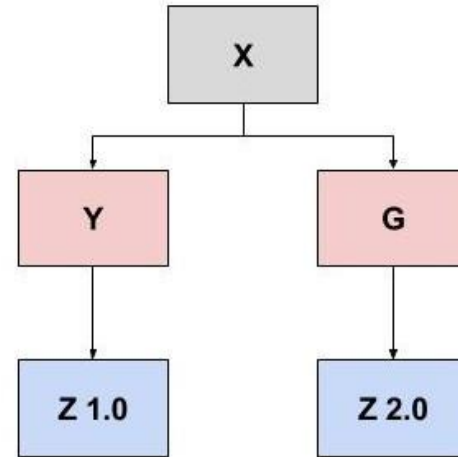
- Bioinformatics Software



Bioinformatics 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](#)

[chat on gitter](#)

Search

[packages & docs](#)

BIOCONDA®

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



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



Mamba

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>

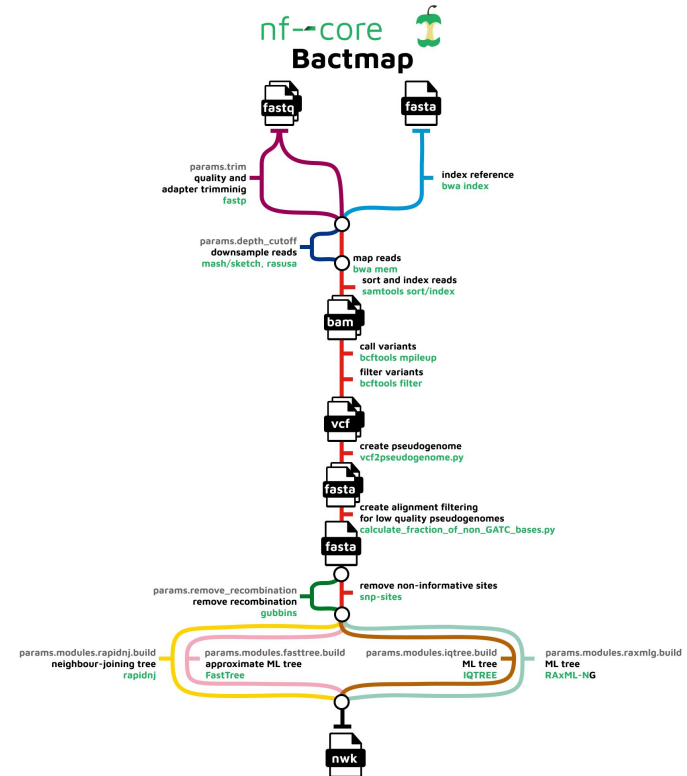
► Use conda to install samtools

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

The screenshot shows the Bioconda website interface. On the left, there is a navigation menu with links for FAQs, Contributing to Bioconda, Developer Docs, Tutorials, Browse packages, and Bioconda @ Github. Below this is a search bar with a 'Go' button. The main content area displays the 'recipe' for the 'samtools' package. It includes the homepage URL (https://github.com/samtools/samtools), the license (MIT), the recipe path (/samtools/meta.yaml), and links to biotools, samtools, and usegalaxy-eu. Below the recipe section, there is a 'package' section for 'samtools' showing download statistics (5.1M), container availability (none), and a list of 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, ...). The package details section lists dependencies: htlib (>=1.20, <1.21.0a0), libgcc-ng (>=12), libzlib (>=1.2.13, <1.3.0a0), and ncurses (>=6.4.20240210, <7.0a0).

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!



- 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.

Released 41

Under development 0

Archived 0

Stars 47

32

<div>rnaseq ✓ ☆ 793</div> <div>RNA sequencing analysis pipeline using STAR, RSEM, HISAT2 or Salmon with gene/isoform counts and extensive quality control.</div> <div>annotation cancer gatk4 genomics germline pre-processing somatic target-panels variant-calling whole-exome-sequencing whole-genome-sequencing</div> <div>rna rna-seq</div> <div>3.14.0 released 5 months ago</div>	<div>sarek ✓ ☆ 341</div> <div>Analysis pipeline to detect germline or somatic variants (pre-processing, variant calling and annotation) from WGS / targeted sequencing</div> <div>annotation cancer gatk4 genomics germline pre-processing somatic target-panels variant-calling whole-exome-sequencing whole-genome-sequencing</div> <div>3.4.2 released 19 days ago</div>	<div>mag ✓ ☆ 182</div> <div>Assembly and binning of metagenomes</div> <div>annotation assembly binning long-read-sequencing metagenomes metagenomics nanopore nanopore-sequencing</div> <div>3.0.0 released 13 days ago</div>	<div>scrnaseq ✓ ☆ 172</div> <div>A single-cell RNAseq pipeline for 10X genomics data</div> <div>10x-genomics 10xgenomics alevin bustools cellranger kallisto rna-seq single-cell star-solo</div> <div>2.6.0 released 19 days ago</div>	<div>chipseq ✓ ☆ 172</div> <div>ChIP-seq peak-calling, QC and differential analysis pipeline.</div> <div>chip chip-seq chromatin-immunoprecipitation macs2 peak-calling</div> <div>2.0.0 released over 1 year ago</div>	<div>atacseq ✓ ☆ 163</div> <div>ATAC-seq peak-calling and QC analysis pipeline</div> <div>atac-seq chromatin-accessibility</div> <div>2.1.2 released 10 months ago</div>
<div>ampliseq ✓ ☆ 157</div> <div>Amplicon sequencing analysis workflow using DADA2 and QiIME2</div> <div>16s 18s amplicon-sequencing edna illumina iontorrent its metabarcoding metagenomics microbiome pacbio qiime2 rna taxonomic-classification taxonomic-profiling</div> <div>2.9.0 released about 2 months ago</div>	<div>nanoseq ✓ ☆ 145</div> <div>Nanopore demultiplexing, QC and alignment pipeline</div> <div>alignment demultiplexing nanopore qc</div> <div>3.1.0 released about 1 year ago</div>	<div>methyseq ✓ ☆ 132</div> <div>Methylation (Bisulfite-Sequencing) analysis pipeline using Bismark or bwa-meth + MethylDackel</div> <div>bisulfite-sequencing dna-methylation em-seq epigenome epigenomics methyl-seq pbat rrbis</div> <div>2.6.0 released 5 months ago</div>	<div>rnafusion ✓ ☆ 130</div> <div>RNA-seq analysis pipeline for detection of gene-fusions</div> <div>fusion fusion-genes gene-fusion ma rna-seq</div> <div>3.0.2 released about 2 months ago</div>	<div>fetchngs ✓ ☆ 125</div> <div>Pipeline to fetch metadata and raw FastQ files from public databases</div> <div>dbj download ena fastq geo sra synapse</div> <div>1.12.0 released 3 months ago</div>	<div>eager ✓ ☆ 123</div> <div>A fully reproducible and state-of-the-art ancient DNA analysis pipeline</div> <div>adna ancient-dna-analysis ancientdna genome metagenomics pathogen-genomics population-genetics</div> <div>2.5.1 released 3 months ago</div>

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/bioinfsv1.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

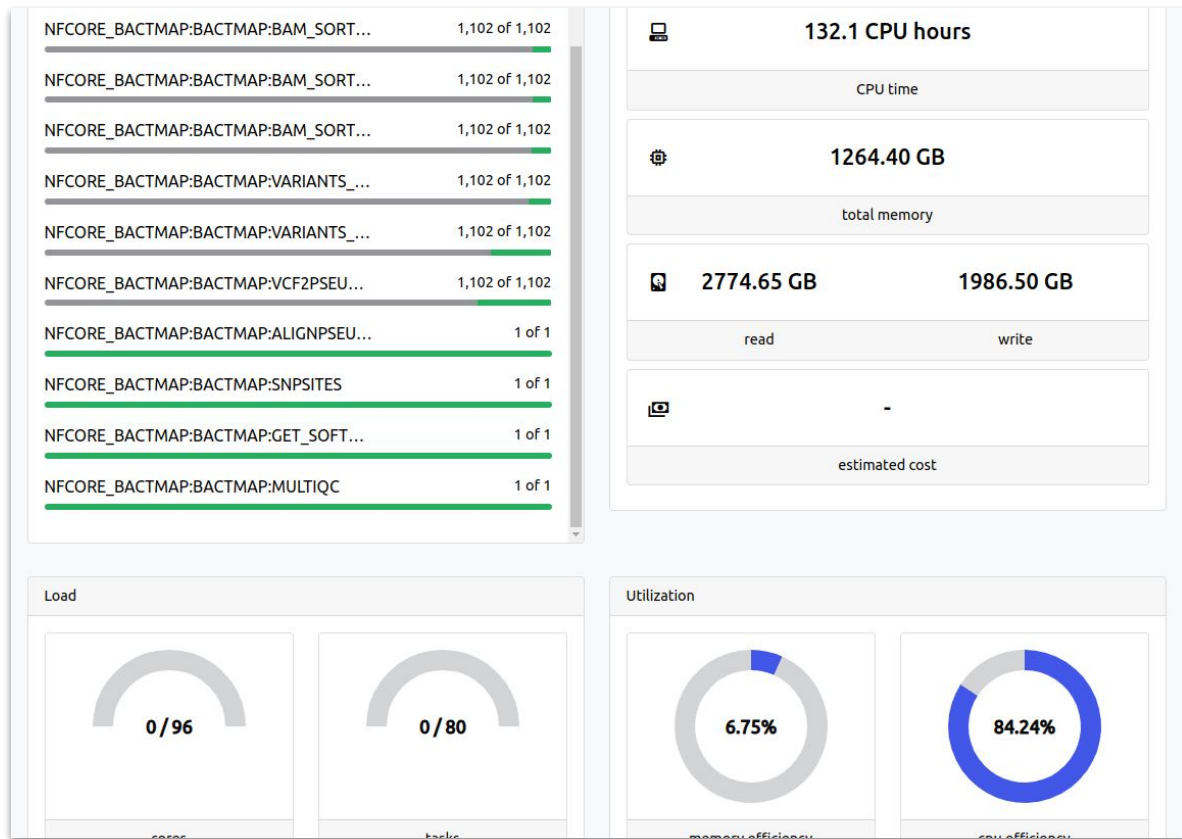
-

local

Status

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

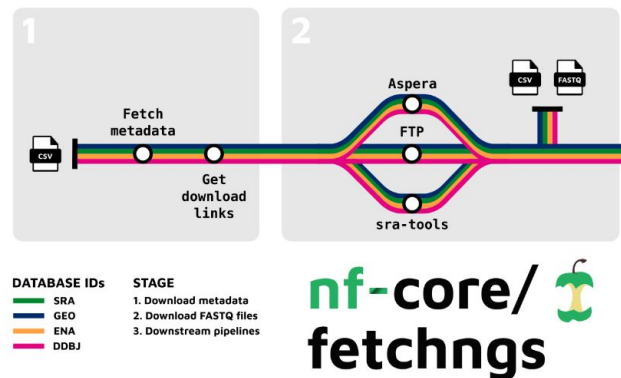
Sequera Platform (Nextflow Tower)



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 list  
$ nf-core list | grep -i rna
```



```
$ wget  
https://github.com/sylabs/singularity/releases/download/v4.1.2/singularity-Ce_4.1.2-jammy_amd64.deb  
$ sudo dpkg -i singularity-ce_4.1.2-jammy_amd64.deb  
$ nextflow run nf-core/fetchngs -profile singularity -input ids.csv -outdir .
```

Questions?

