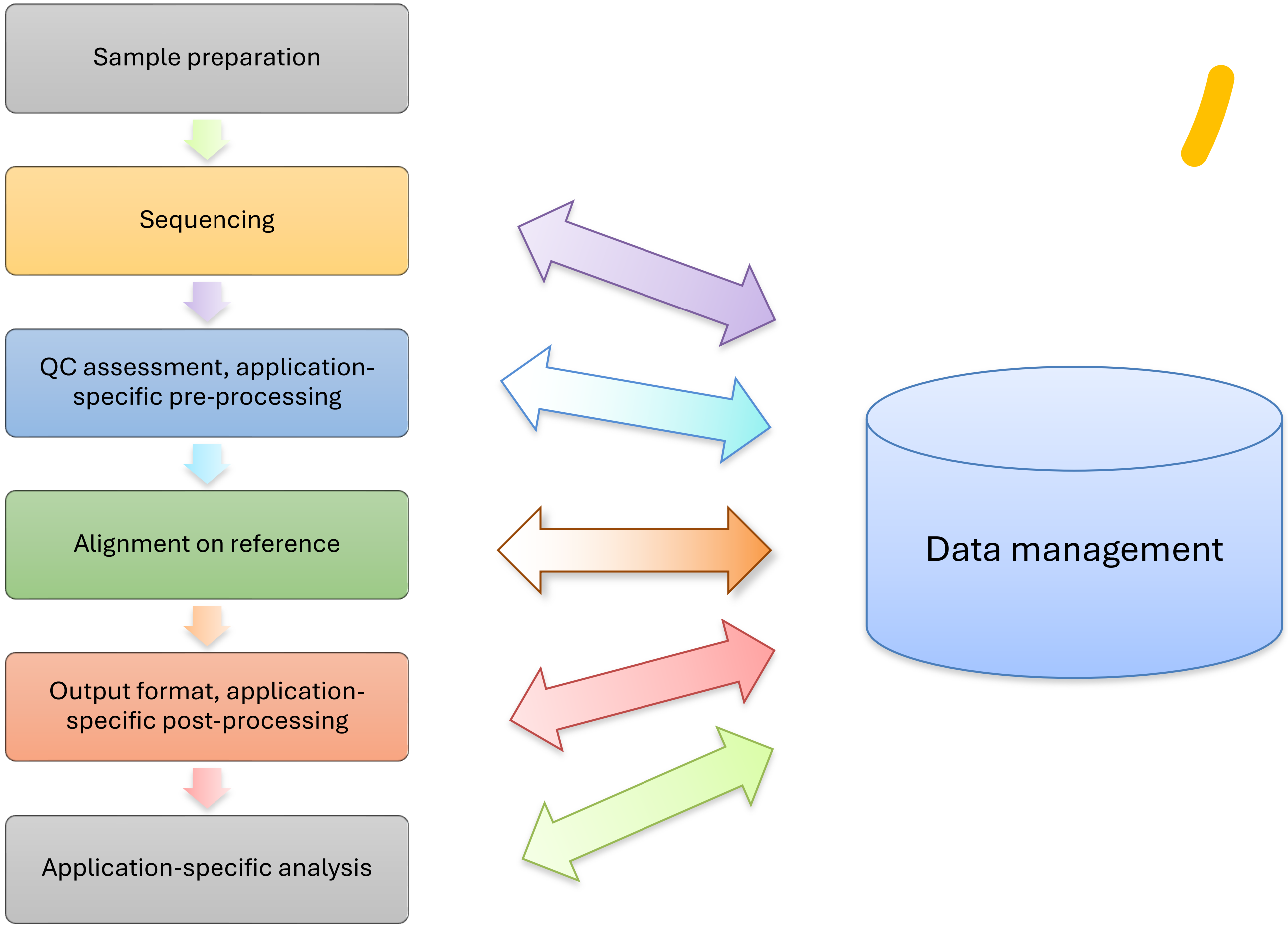
Setting up an automated framework for bioinformatic and multiomic analyses on IBiSCo

IBiSCo Workshop – 19/04/2024

“Omic” Sequencing applications

Whole exome sequencing (WES)	Whole genome sequencing (WGS)	Whole-genome Bisulfite Sequencing (WGBS)	Reduced Representation Bisulfite Sequencing (RRBS)
3’DGE RNA sequencing	Full length RNA sequencing	Small / microRNA sequencing	Single Cell RNA sequencing
ChIP sequencing	ATAC sequencing	MeDIP sequencing	Sequential Analysis of MacroMolecules accessibility sequencing

Experiment outline



Sample preparation

Sequencing

QC assessment, application-specific pre-processing

Alignment on reference

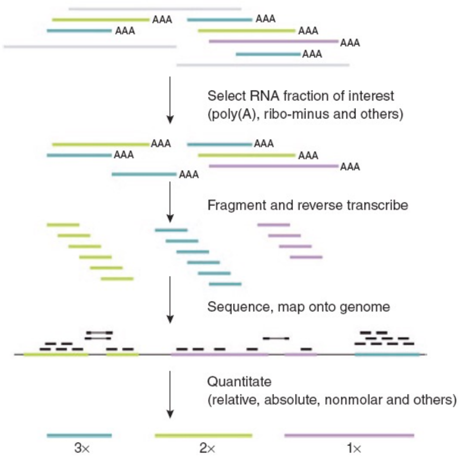
Output format, application-specific post-processing

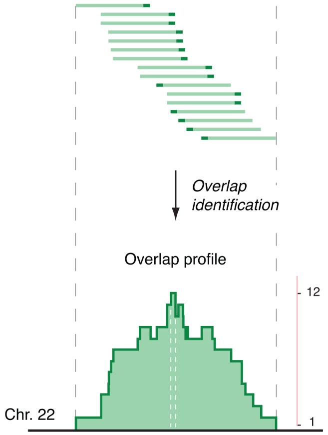
Application-specific analysis

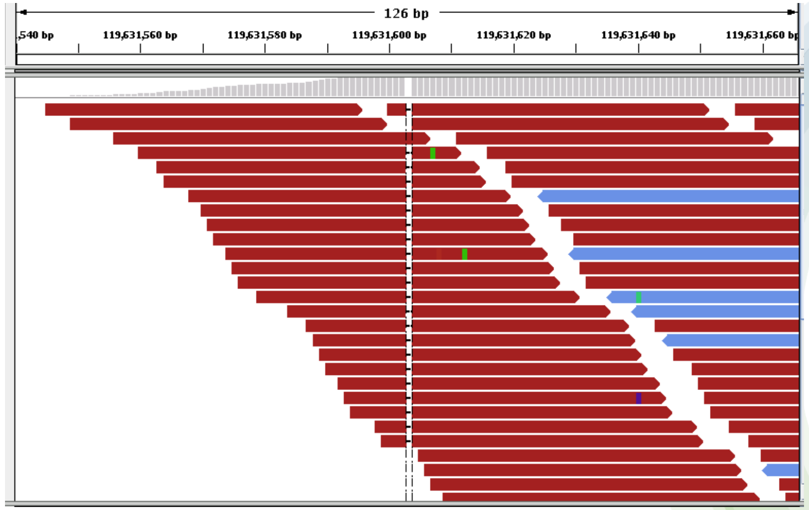
Data management

Application specific processing

- RNA-Seq
  - Count reads for each feature
  - Transcript isoform identification
- ChIP-Seq / ATAC-Seq / SAMMY-Seq
  - Signal quantification base by base
  - Peak calling / relative region
- Variant discovery
  - Refine alignment
  - Quantify alleles by position





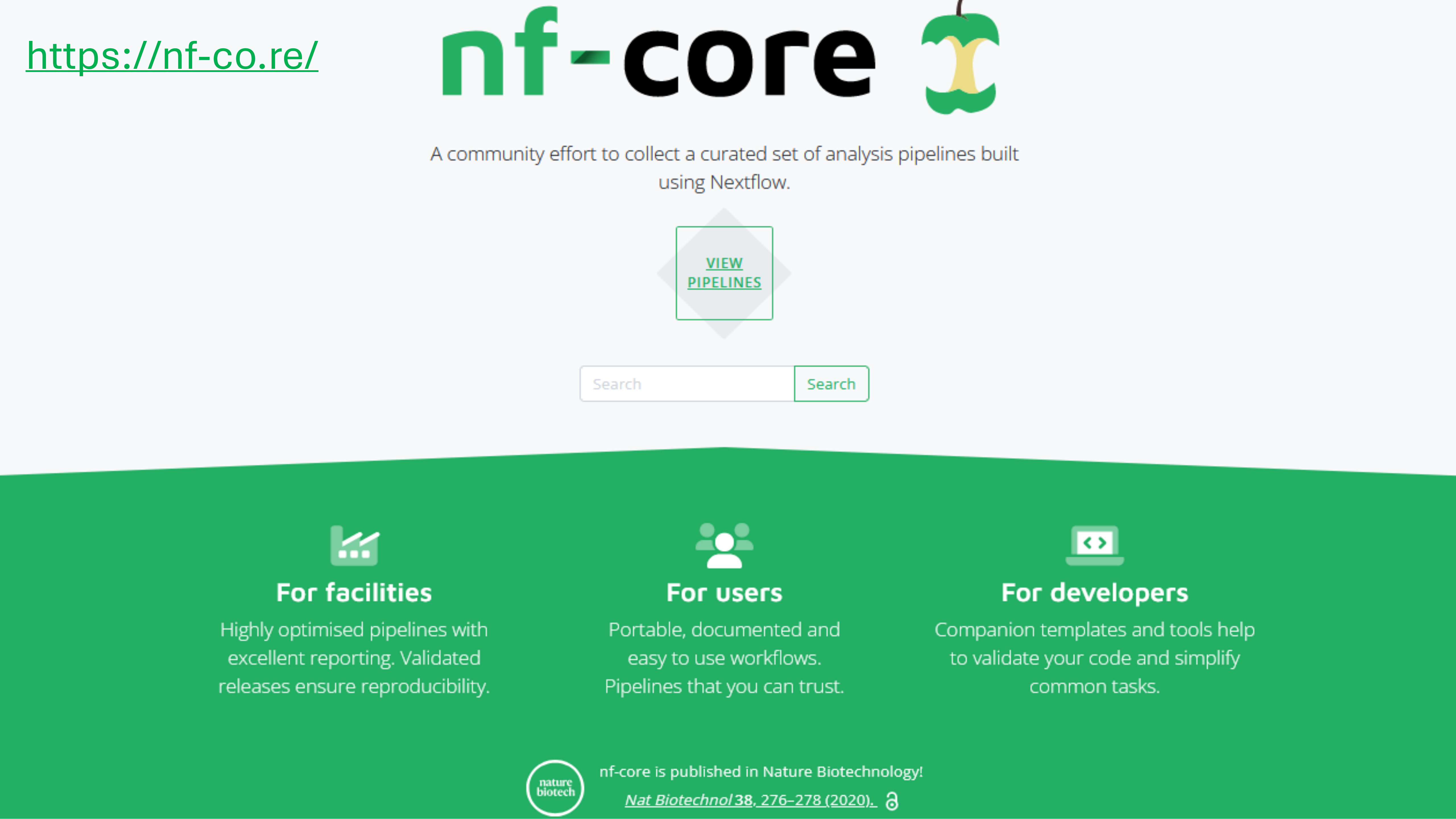


A workflow management system: nextflow

- ✓Framework that allows to chain commands seamlessly
- ✓Modular
- ✓Portable: the same script can run on a pc, HPC clusters with different queue managers or cloud systems
- ✓Reproducible: natively implements the use of containers (docker, singularity
- ✓Can combine existing scripts written in different languages
- ✓Widely used, with a large and active community of users
- ✓Includes parallelization, analysis progress monitoring and technical report
- ✓Resume function

Nat Biotechnol **35**, 316–319 (2017).

https://nf-co.re/



A community effort to collect a curated set of analysis pipelines built using Nextflow.

For facilities

Highly optimised pipelines with excellent reporting. Validated releases ensure reproducibility.

For users

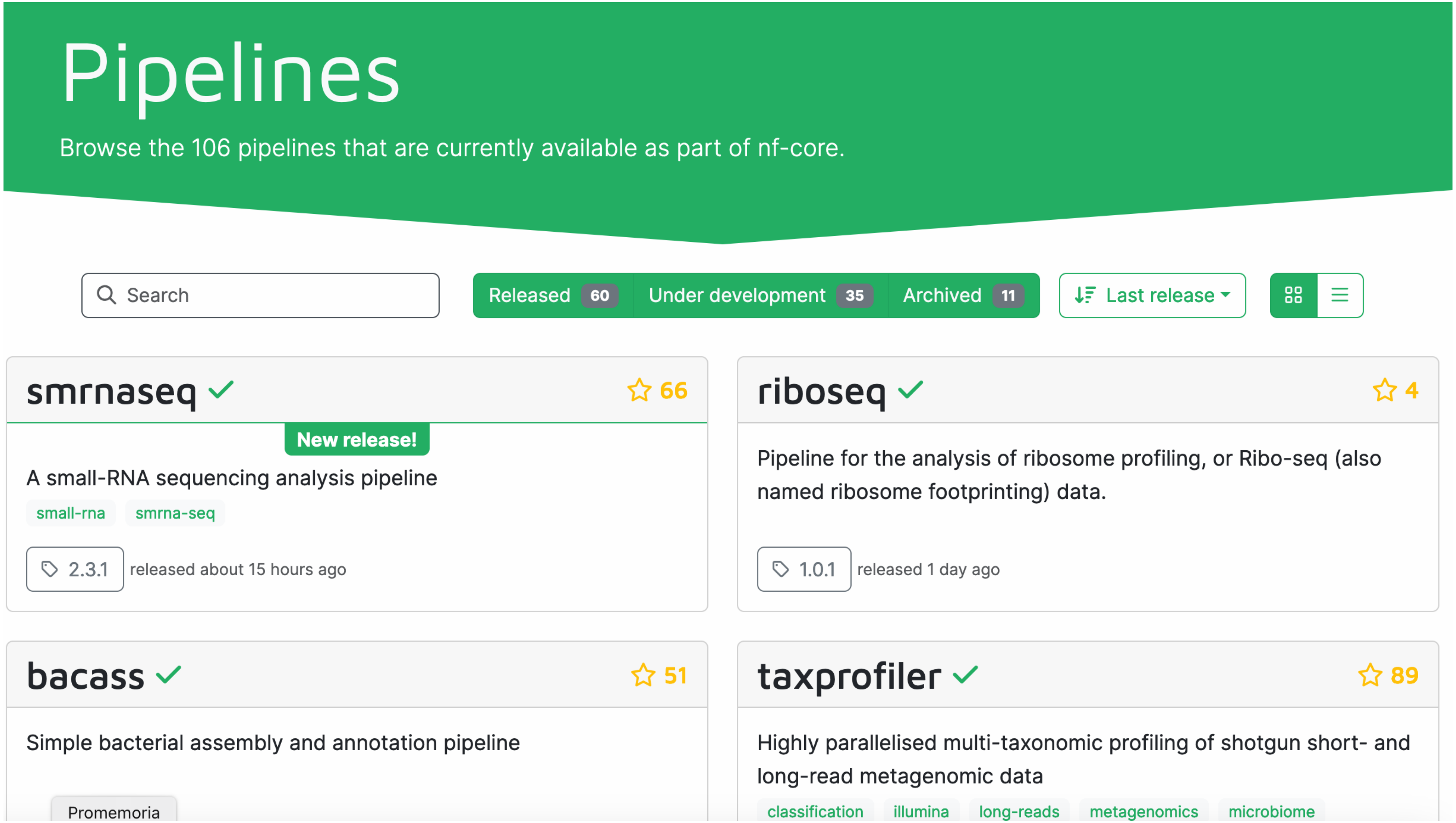
Portable, documented and easy to use workflows. Pipelines that you can trust.

For developers

Companion templates and tools help to validate your code and simplify common tasks.

nf-core is published in Nature Biotechnology!

Nat Biotechnol 38, 276–278 (2020).



Pipelines

Browse the 106 pipelines that are currently available as part of nf-core.

smrnaseq ✓ ☆ 66

New release!

A small-RNA sequencing analysis pipeline

2.3.1 released about 15 hours ago

riboseq ✓ ☆ 4

Pipeline for the analysis of ribosome profiling, or Ribo-seq (also named ribosome footprinting) data.

1.0.1 released 1 day ago

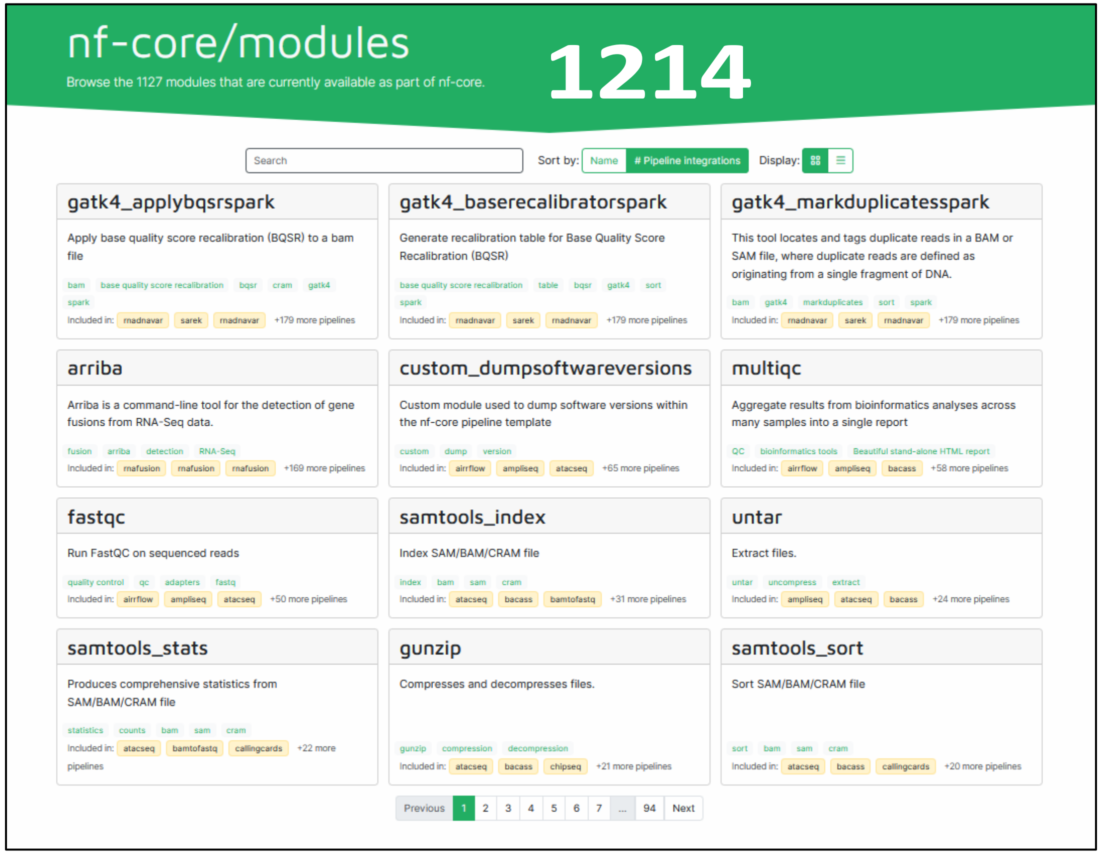
bacass ✓ ☆ 51

Simple bacterial assembly and annotation pipeline

taxprofiler ✓ ☆ 89

Highly parallelised multi-taxonomic profiling of shotgun short- and long-read metagenomic data

nf-core/modules & subworkflows



nf-core/modules

1214

Browse the 1127 modules that are currently available as part of nf-core.

gatk4_applybqsrspark

Apply base quality score recalibration (BQSR) to a bam file

bam · base quality score recalibration · bqsr · cram · gatk4 · spark

Included in: rnadnavar sarek rnadnavar +179 more pipelines

gatk4_baserecalibratorspark

Generate recalibration table for Base Quality Score Recalibration (BQSR)

base quality score recalibration · table · bqsr · gatk4 · sort · spark

Included in: rnadnavar sarek rnadnavar +179 more pipelines

gatk4_markduplicatesspark

This tool locates and tags duplicate reads in a BAM or SAM file, where duplicate reads are defined as originating from a single fragment of DNA.

bam · gatk4 · markduplicates · sort · spark

Included in: rnadnavar sarek rnadnavar +179 more pipelines

arriba

Arriba is a command-line tool for the detection of gene fusions from RNA-Seq data.

fusion · arriba · detection · RNA-Seq

Included in: rnafusion rnafusion rnafusion +169 more pipelines

custom_dumpsoftwareversions

Custom module used to dump software versions within the nf-core pipeline template

custom · dump · version

Included in: airrflow ampliseq atacseq +65 more pipelines

multiqc

Aggregate results from bioinformatics analyses across many samples into a single report

QC · bioinformatics tools · Beautiful stand-alone HTML report

Included in: airrflow ampliseq bacass +58 more pipelines

fastqc

Run FastQC on sequenced reads

quality control · qc · adapters · fastq

Included in: airrflow ampliseq atacseq +50 more pipelines

samtools_index

Index SAM/BAM/CRAM file

index · bam · sam · cram

Included in: atacseq bacass bamtofastq +31 more pipelines

untar

Extract files.

untar · uncompress · extract

Included in: ampliseq atacseq bacass +24 more pipelines

samtools_stats

Produces comprehensive statistics from SAM/BAM/CRAM file

statistics · counts · bam · sam · cram

Included in: atacseq bamtofastq callingcards +22 more pipelines

gunzip

Compresses and decompresses files.

gunzip · compression · decompression

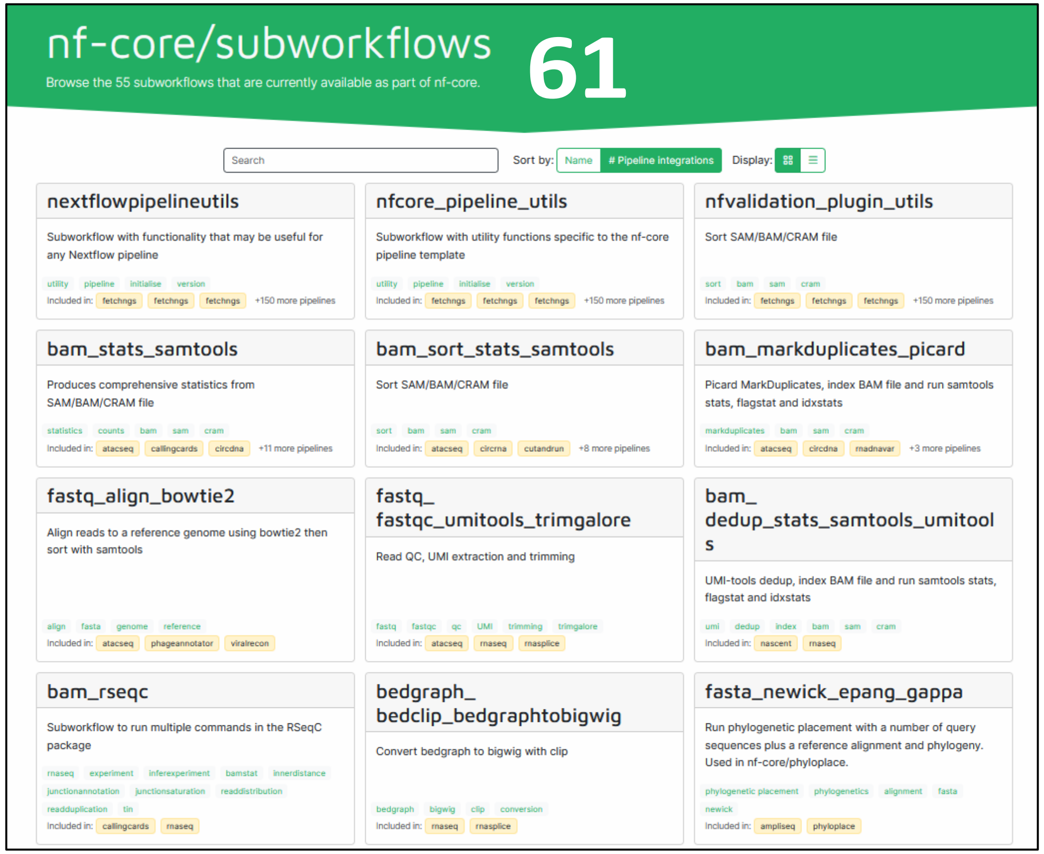
Included in: atacseq bacass chipseq +21 more pipelines

samtools_sort

Sort SAM/BAM/CRAM file

sort · bam · sam · cram

Included in: atacseq bacass callingcards +20 more pipelines



nf-core/subworkflows

61

Browse the 55 subworkflows that are currently available as part of nf-core.

nextflowpipelineutils

Subworkflow with functionality that may be useful for any Nextflow pipeline

utility · pipeline · initialise · version

Included in: fetchngs fetchngs fetchngs +150 more pipelines

nfcore_pipeline_utils

Subworkflow with utility functions specific to the nf-core pipeline template

utility · pipeline · initialise · version

Included in: fetchngs fetchngs fetchngs +150 more pipelines

nfvalidation_plugin_utils

Sort SAM/BAM/CRAM file

sort · bam · sam · cram

Included in: fetchngs fetchngs fetchngs +150 more pipelines

bam_stats_samtools

Produces comprehensive statistics from SAM/BAM/CRAM file

statistics · counts · bam · sam · cram

Included in: atacseq callingcards circdna +11 more pipelines

bam_sort_stats_samtools

Sort SAM/BAM/CRAM file

sort · bam · sam · cram

Included in: atacseq circrna cutandrun +8 more pipelines

bam_markduplicates_picard

Picard MarkDuplicates, index BAM file and run samtools stats, flagstat and idxstats

markduplicates · bam · sam · cram

Included in: atacseq circdna rnadnavar +3 more pipelines

fastq_align_bowtie2

Align reads to a reference genome using bowtie2 then sort with samtools

align · fasta · genome · reference

Included in: atacseq phageannotator viralrecon

fastq_fastqc_umitools_trimgalore

Read QC, UMI extraction and trimming

fastq · fastqc · qc · UMI · trimming · trimgalore

Included in: atacseq rnaseq rnasplice

bam_dedup_stats_samtools_umitools

UMI-tools dedup, index BAM file and run samtools stats, flagstat and idxstats

umi · dedup · index · bam · sam · cram

Included in: nascent rnaseq

bam_rseqc

Subworkflow to run multiple commands in the RSeqC package

rnaseq · experiment · inferexperiment · bamstat · innerdistance · junctionannotation · junctionsaturation · readdistribution · readduplication · tin

Included in: callingcards rnaseq

bedgraph_bedclip_bedgraphtobigwig

Convert bedgraph to bigwig with clip

bedgraph · bigwig · clip · conversion

Included in: rnaseq rnasplice

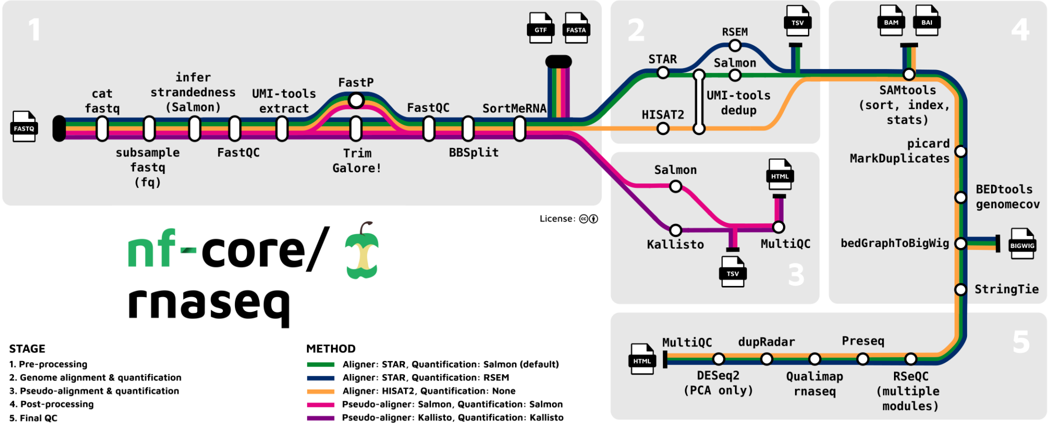
fasta_newick_epang_gappa

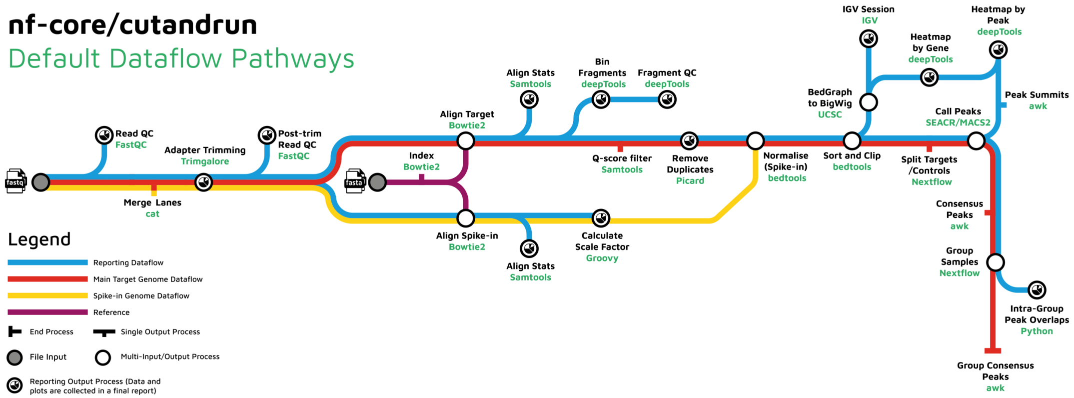
Run phylogenetic placement with a number of query sequences plus a reference alignment and phylogeny. Used in nf-core/phyloplace.

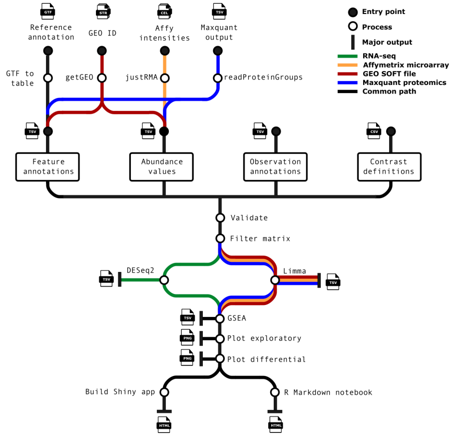
phylogenetic placement · phylogenetics · alignment · fasta · newick

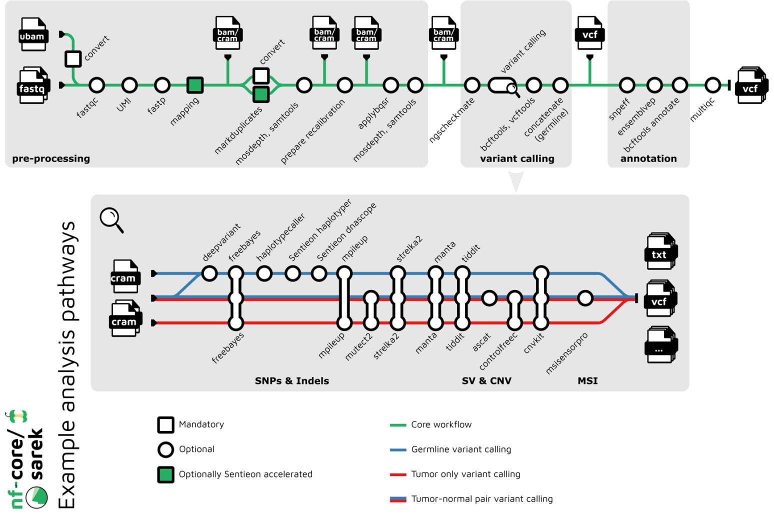
Included in: ampliseq phyloplace

Pipeline workflow schemes





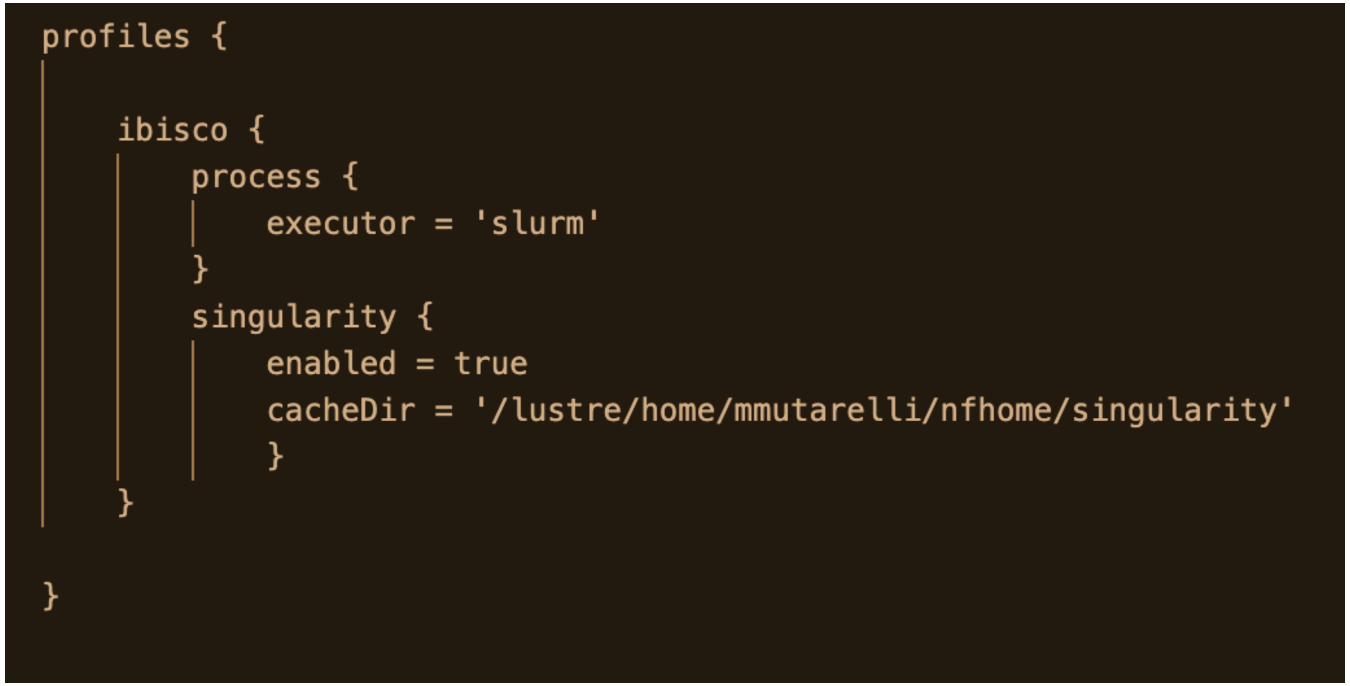




Adapted from: Fellows Yates, James A., et al. PeerJ 9 (2021).

Setting up nexflow pipelines on IBIScO

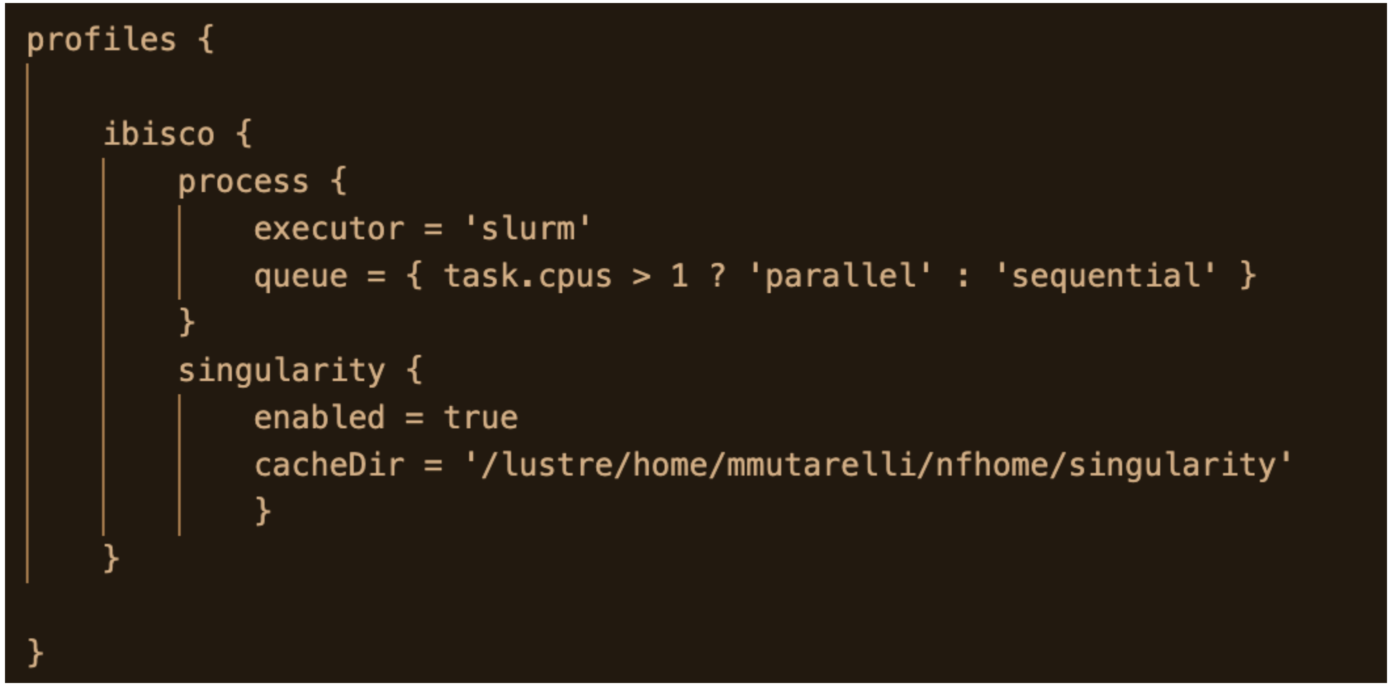
- config file



```
profiles {

    ibisco {
        process {
            executor = 'slurm'
        }
        singularity {
            enabled = true
            cacheDir = '/lustre/home/mmutarelli/nfhome/singularity'
            }
    }

}
```

Setting up nexflow pipelines on IBIScO

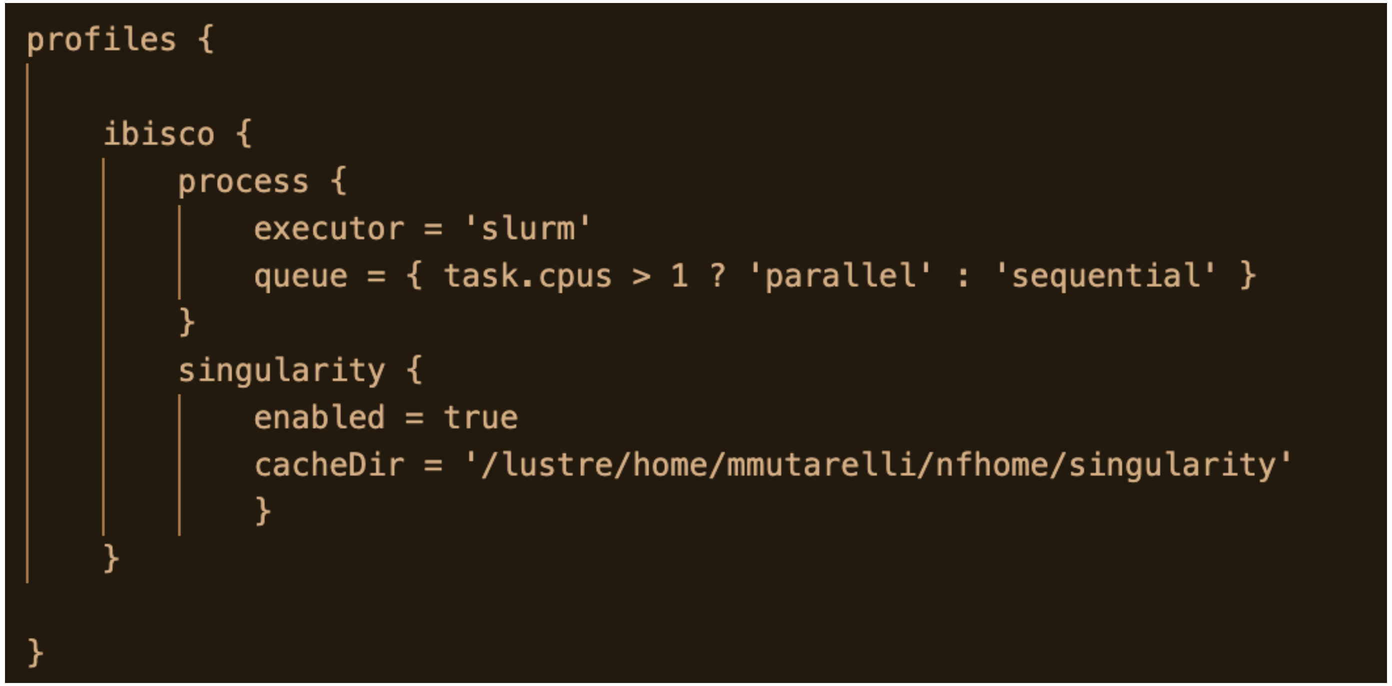
- config file



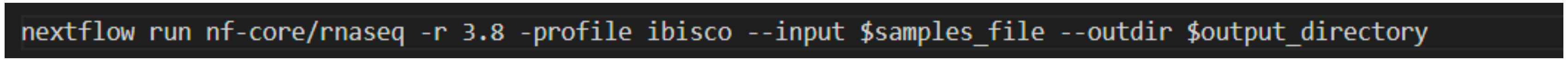
```
profiles {

  ibisco {
    process {
      executor = 'slurm'
      queue = { task.cpus > 1 ? 'parallel' : 'sequential' }
    }
    singularity {
      enabled = true
      cacheDir = '/lustre/home/mmutarelli/nfhome/singularity'
      }
  }

}
```

Setting up nexflow pipelines on IBIScO

- config file



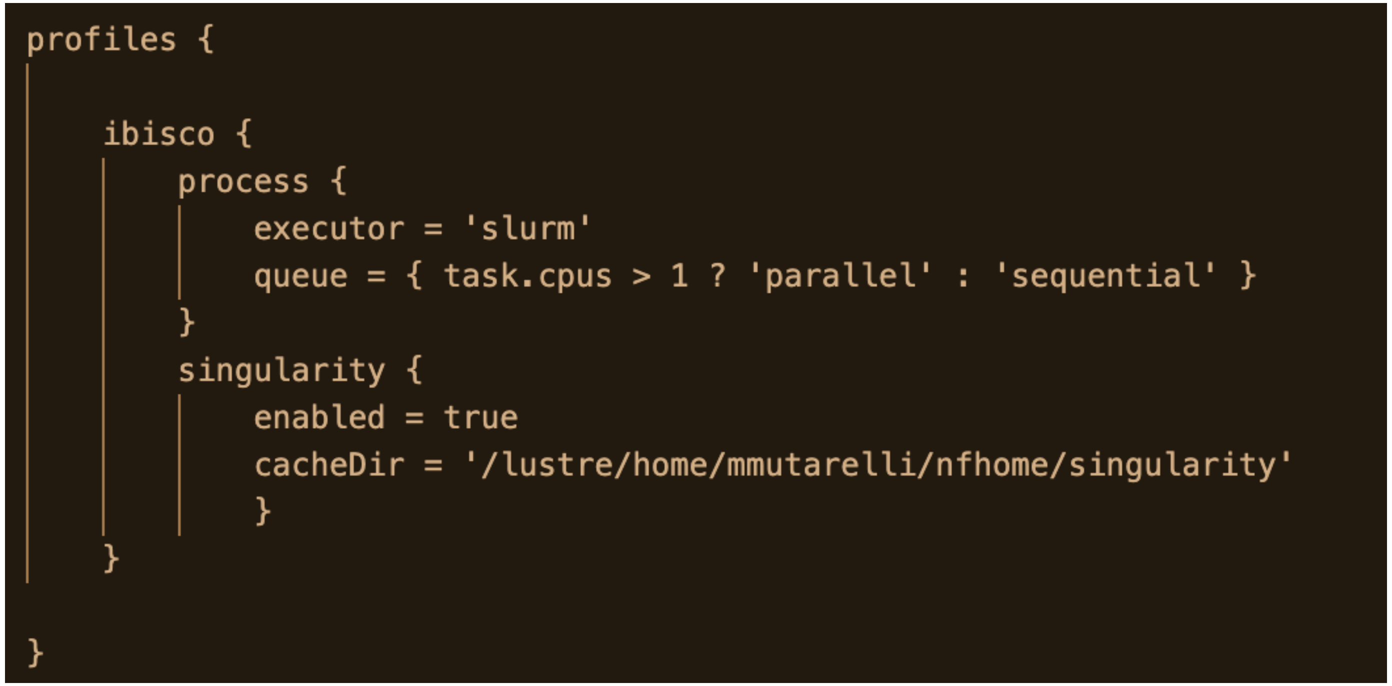
```
profiles {

    ibisco {
        process {
            executor = 'slurm'
            queue = { task.cpus > 1 ? 'parallel' : 'sequential' }
        }
        singularity {
            enabled = true
            cacheDir = '/lustre/home/mmutarelli/nfhome/singularity'
            }
    }

}
```



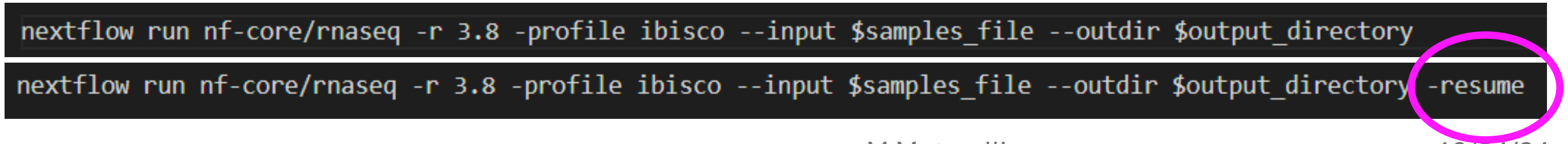
```
nextflow run nf-core/rnaseq -r 3.8 -profile ibisco --input $samples_file --outdir $output_directory
```

Setting up nexflow pipelines on IBIScO

- config file



```
profiles {

    ibisco {
        process {
            executor = 'slurm'
            queue = { task.cpus > 1 ? 'parallel' : 'sequential' }
        }
        singularity {
            enabled = true
            cacheDir = '/lustre/home/mmutarelli/nfhome/singularity'
            }
    }

}
```

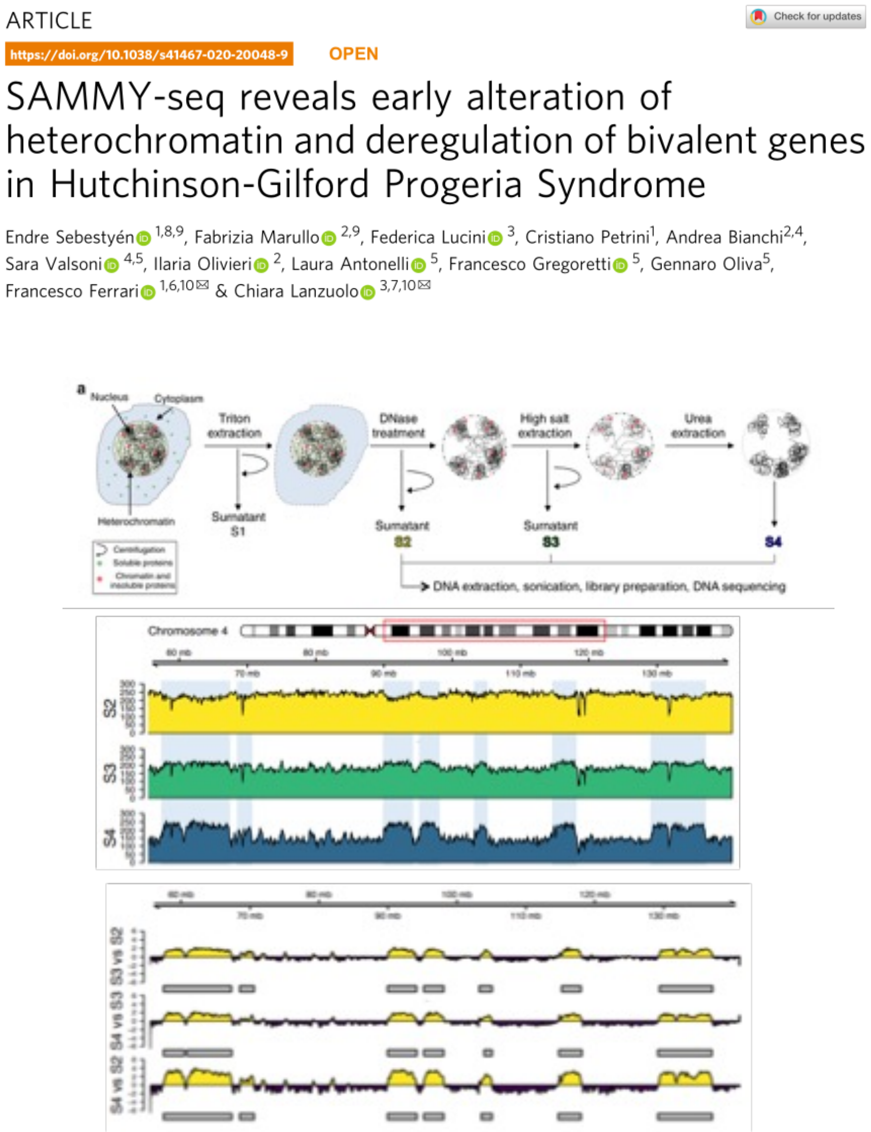


```
nextflow run nf-core/rnaseq -r 3.8 -profile ibisco --input $samples_file --outdir $output_directory
```

```
nextflow run nf-core/rnaseq -r 3.8 -profile ibisco --input $samples_file --outdir $output_directory -resume
```

Setting up nexflow pipelines on IBIScO

fetchngs	☑	~ 60 samples, public
rnaseq	☑	~ 150 samples, 6 projects
differentialabundance	☑	~ 75 samples, 4 projects
cutandrun	☑	~ 15 samples, 1 projects
chipseq	☑	~ 5 samples, 1 project
sarek	☑	~ 55 samples, 4 projects
rnavar	☑	~ 45 samples, 3 projects
rnafusion	☑	~ 45 samples, 3 projects
methylseq	☑	~ 17 samples, 2 projects
sammyseq	☑	~ 190 samples, 6 project



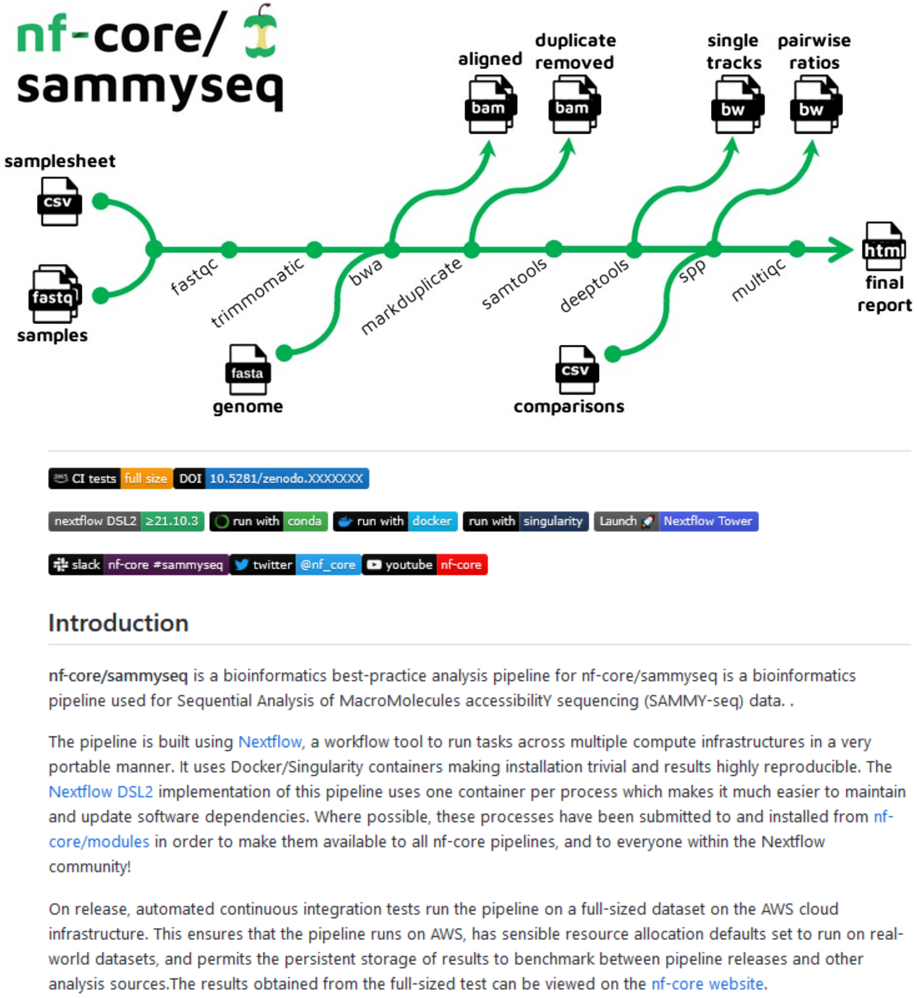
ARTICLE

https://doi.org/10.1038/s41467-020-20048-9 OPEN

Check for updates

SAMMY-seq reveals early alteration of heterochromatin and deregulation of bivalent genes in Hutchinson-Gilford Progeria Syndrome

Endre Sebestyén[1,8,9], Fabrizia Marullo[2,9], Federica Lucini[3], Cristiano Petrini[1], Andrea Bianchi[2,4], Sara Valsoni[4,5], Ilaria Olivieri[2], Laura Antonelli[5], Francesco Gregoretti[5], Gennaro Oliva[5], Francesco Ferrari[1,6,10]✉ & Chiara Lanzuolo[3,7,10]✉



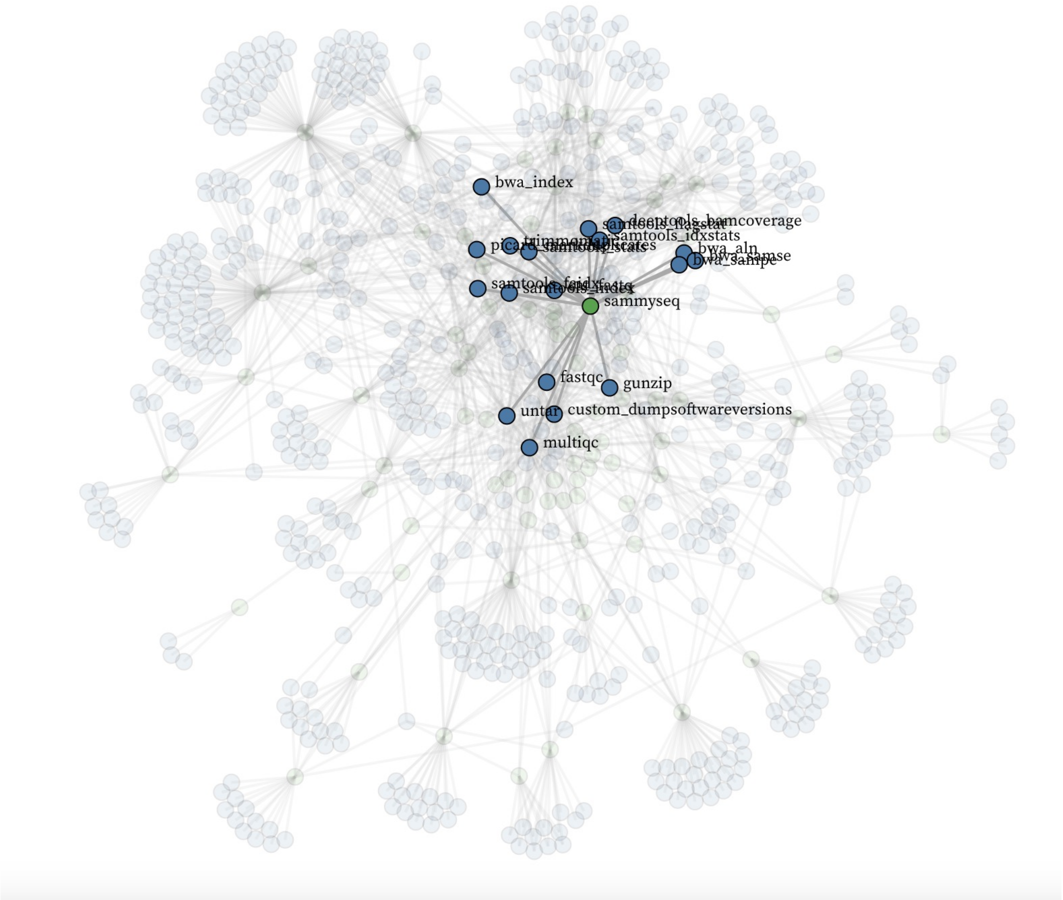
Introduction

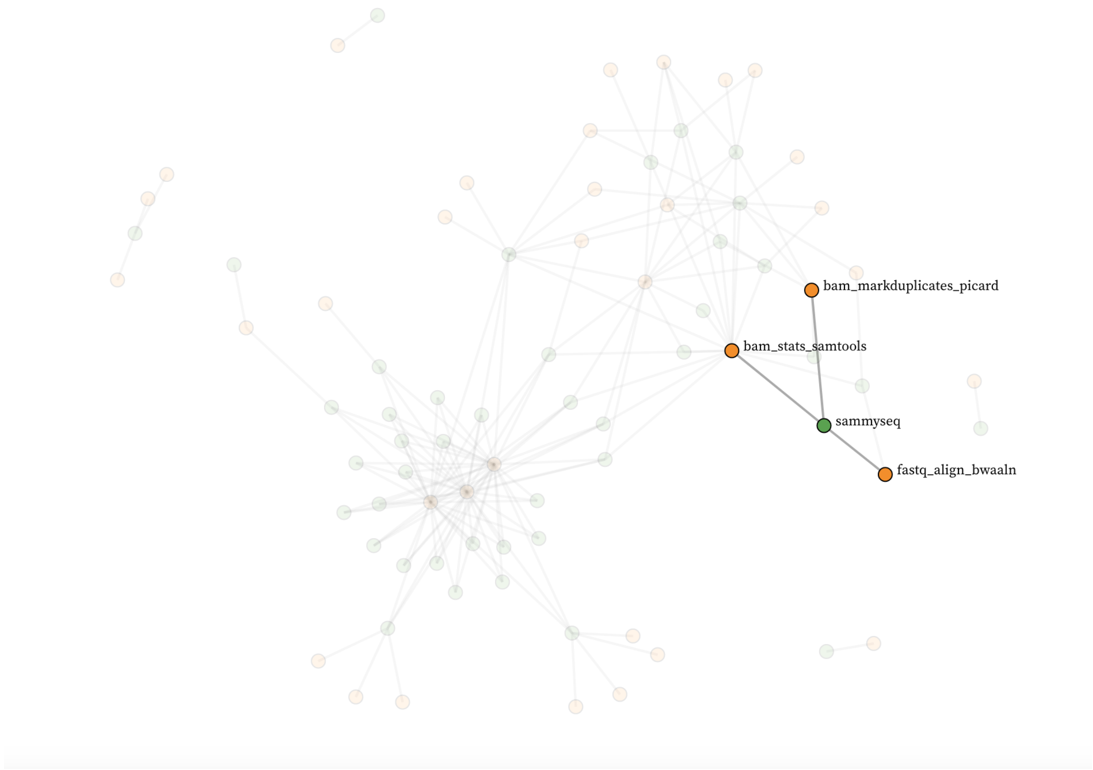
nf-core/sammyseq is a bioinformatics best-practice analysis pipeline for nf-core/sammyseq is a bioinformatics pipeline used for Sequential Analysis of MacroMolecules accessibilitY sequencing (SAMMY-seq) data. .

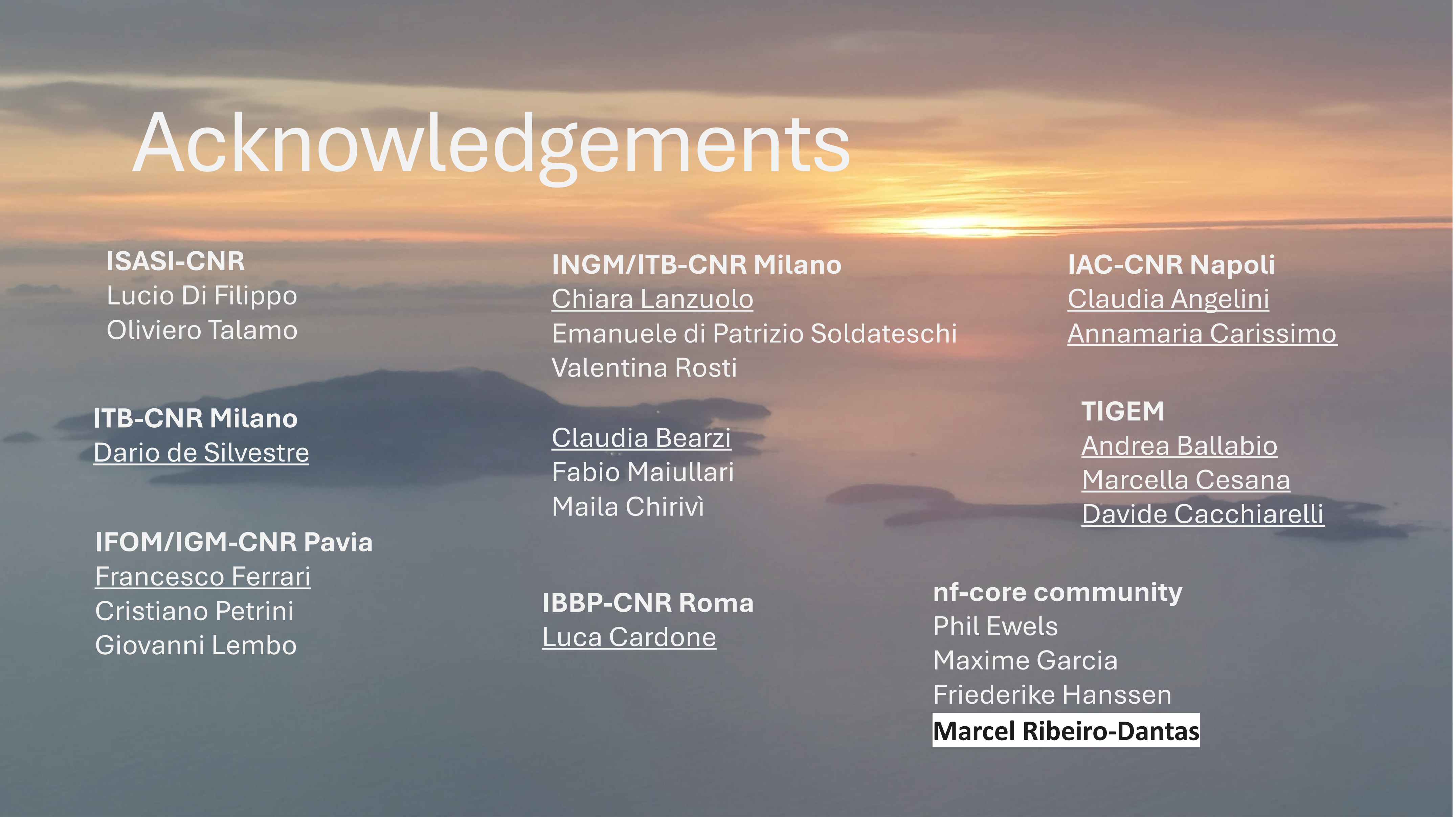
The pipeline is built using Nextflow, a workflow tool to run tasks across multiple compute infrastructures in a very portable manner. It uses Docker/Singularity containers making installation trivial and results highly reproducible. The Nextflow DSL2 implementation of this pipeline uses one container per process which makes it much easier to maintain and update software dependencies. Where possible, these processes have been submitted to and installed from nf-core/modules in order to make them available to all nf-core pipelines, and to everyone within the Nextflow community!

On release, automated continuous integration tests run the pipeline on a full-sized dataset on the AWS cloud infrastructure. This ensures that the pipeline runs on AWS, has sensible resource allocation defaults set to run on real-world datasets, and permits the persistent storage of results to benchmark between pipeline releases and other analysis sources.The results obtained from the full-sized test can be viewed on the nf-core website.

nf-core modules and subworkflow network







Acknowledgements

ISASI-CNR
Lucio Di Filippo
Oliviero Talamo

ITB-CNR Milano
Dario de Silvestre

IFOM/IGM-CNR Pavia
Francesco Ferrari
Cristiano Petrini
Giovanni Lembo

INGM/ITB-CNR Milano
Chiara Lanzuolo
Emanuele di Patrizio Soldateschi
Valentina Rosti

Claudia Bearzi
Fabio Maiullari
Maila Chirivì

IBBP-CNR Roma
Luca Cardone

IAC-CNR Napoli
Claudia Angelini
Annamaria Carissimo

TIGEM
Andrea Ballabio
Marcella Cesana
Davide Cacchiarelli

nf-core community
Phil Ewels
Maxime Garcia
Friederike Hanssen
Marcel Ribeiro-Dantas

Powered by **IBIScO**