UNICAMP

Instituto de Biologia

laboratório IMUNO INFLAMAÇÃO

FAPESP
FUNDAÇÃO DE AMPARO À PESQUISA DO ESTADO DE SÃO PAULO

NIH
National Institutes of Health

Introduction to Linux, Nextflow and nf-core

Vinícius de Rezende Rodovalho

Contents

- **Part 1:** Linux:
  - Basic commands;
  - Running scripts and programs.
- **Part 2:** Nextflow;
- **Part 3:** nf-core;
- **Part 4:** nf-core pipelines:
  - nf-core/rnaseq;
  - nf-core/differentialabundance;
  - nf-core/ampliseq.
- **Part 5:** concluding remarks:
  - Troubleshooting and resources.

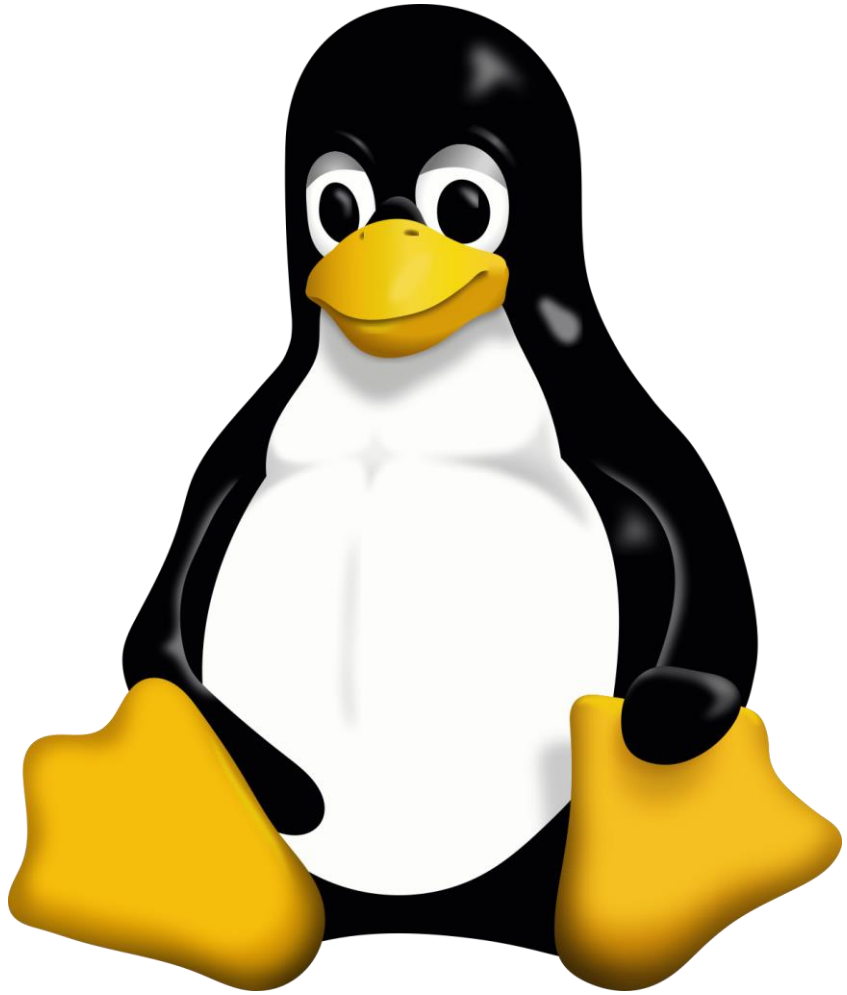
PART 1

Linux

Linux

- **Linux** is a term popularly used to refer to **operating systems** that utilize the **Linux kernel**.
  - Initially developed and used by groups of enthusiasts in personal computers,
  - Now it has the collaboration of **large companies** (IBM, Sun Microsystems, HP, Red Hat, Novell, Oracle, Google, Microsoft and Canonical).

- Linux is one of the most prominent examples of **free** and **open source** software collaboration.
  - Its source code is available for anyone to use, study, modify and distribute freely according to the terms of the license.



Tux the penguin, the mascot of Linux.

Linux

- **Linus** Benedict **Torvalds**, born 28 December 1969, a Finnish-American software engineer, is the creator and lead developer of the Linux kernel.

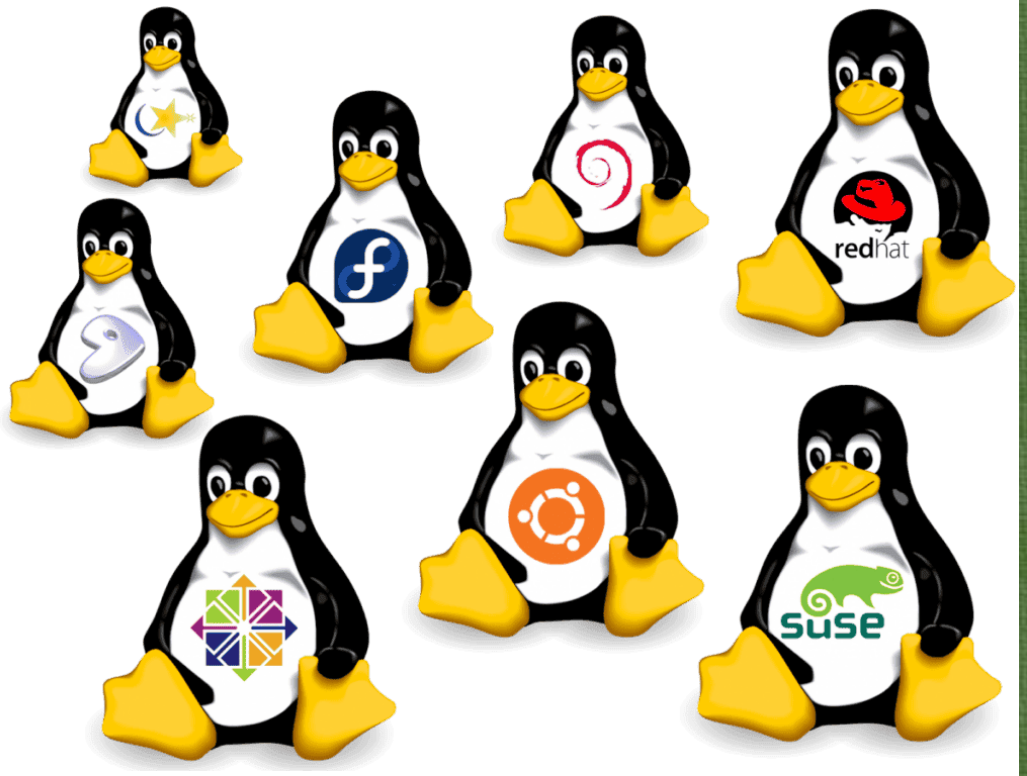
- In an interview Linus commented on the **penguin bite**:
  - *“I've been to Australia several times (…) But my first trip — and the one when I was bitten by a ferocious fairy penguin: you really should keep those things locked up! — was in 93 or so, talking about Linux for the Australian Unix Users Group.”*

Linus Benedict Torvalds
28 December 1969 (age 54)
Helsinki, Finland

Linux distributions

- Normally, Linux is found in a distribution:
  - A distribution includes the Linux kernel, libraries and utilities, and a collection of applications, such as the LibreOffice suite and web browsers.

- Examples of popular Linux distributions:
  - **Ubuntu**: A user-friendly and widely popular distribution;
  - **Debian**: A stable and reliable distribution;
  - **Fedora**: A cutting-edge distribution that prioritizes innovation;
  - **Linux Mint**: A beginner-friendly distribution based on Ubuntu;
  - **Red Hat Enterprise Linux** (RHEL): A commercial distribution;
  - **CentOS**: A community-supported enterprise-class distribution derived from Red Hat Enterprise Linux;
  - **SUSE Linux Enterprise Server** (SLES);
  - **Debian Server.**

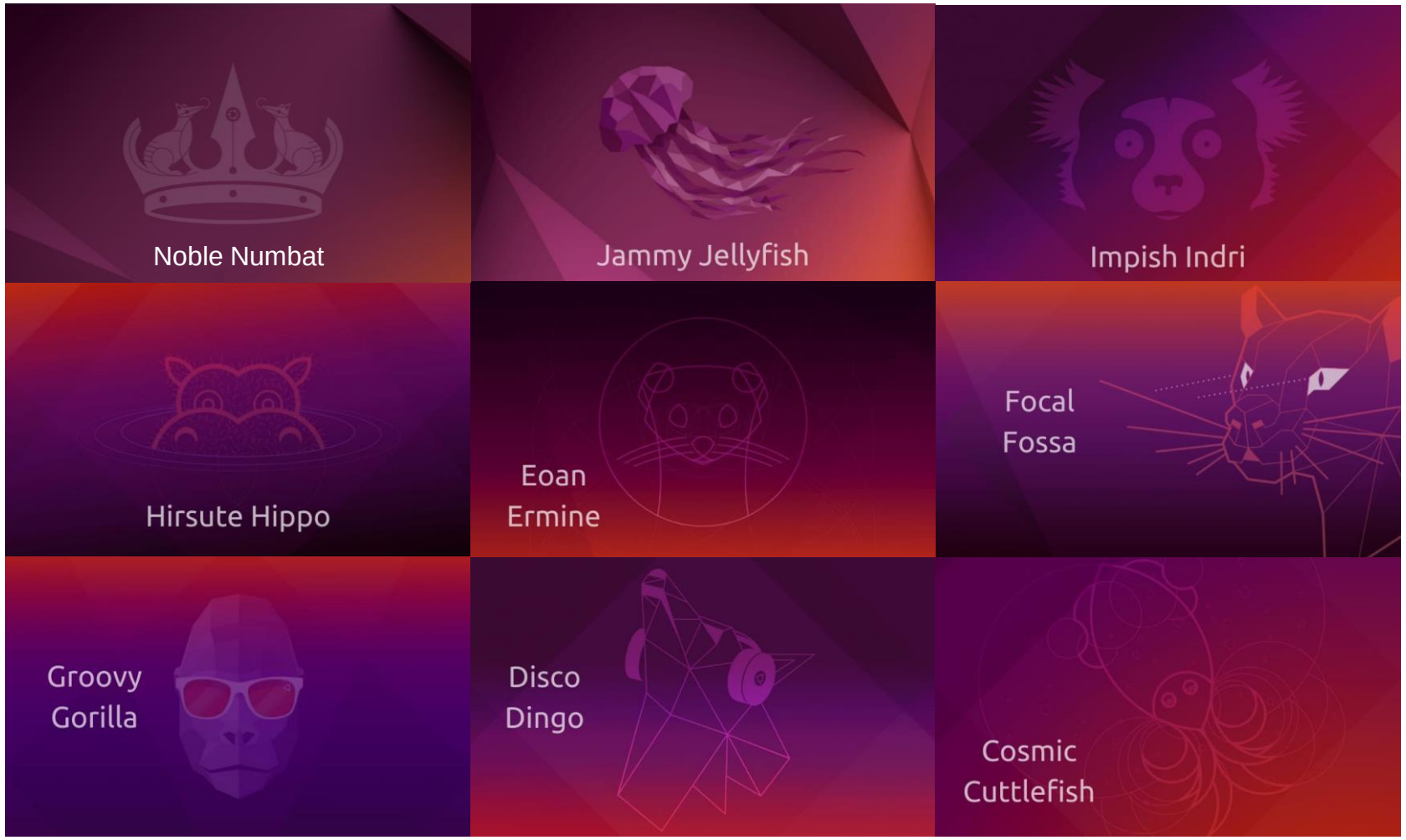


Ubuntu

- Ubuntu (meaning humanity in Bantu) describes a set of closely related African-origin value systems that emphasize the interconnectedness of individuals with their surrounding societal and physical worlds.
  - “I am because we are” ;
  - “I am because you are”;
  - “humanity towards others”.

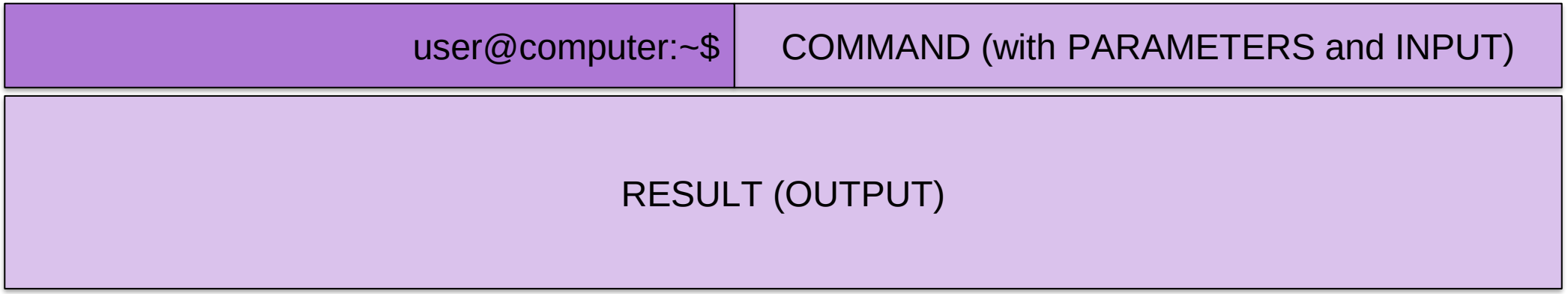
ubuntu

Mascots of Ubuntu Releases



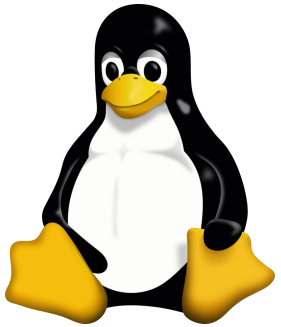
Commands

- **Commands**: Instructions passed via the terminal.
- **Parameters** (hyphen followed by letters: -a, -l…): Options added to modify the command.
- **Input/Output**: Whatever is passed to or produced by a command.
- **Pipes** (|): Connection between commands, passing an output as input for the next command.



Practice: get to your Linux terminal!

Open Ubuntu (WSL)



Just open a terminal

If you installed Ubuntu, but couldn't do the rest

- Download:
  - **wget** https://github.com/conda-forge/miniforge/releases/latest/download/Miniforge3-Linux-x86_64.sh
- Install:
  - **bash Miniforge3-Linux-x86_64.sh**
  - *Enter, More..., yes, Enter, yes.*
- Close terminal and open again.
- Create a new environment:
  - **mamba create -n training pandas -c conda-forge**
- Activate your environment:
  - **mamba activate training**

If Ubuntu (WSL) is not working in your computer

- Go to GitPod:
  - https://gitpod.io/workspaces
  - Continue with GitHub and log in.
- Create a new Workspace for:
  - https://github.com/vrrodovalho/nf-core-training
  - Choose **Large** configuration.
- Proceed with the instructions in README.md.

Commands: *pwd*

- pwd - print name of current/working directory

rodovalhovr@IS-2346:~$	pwd
/home/rodovalhovr	

Commands: *ls*

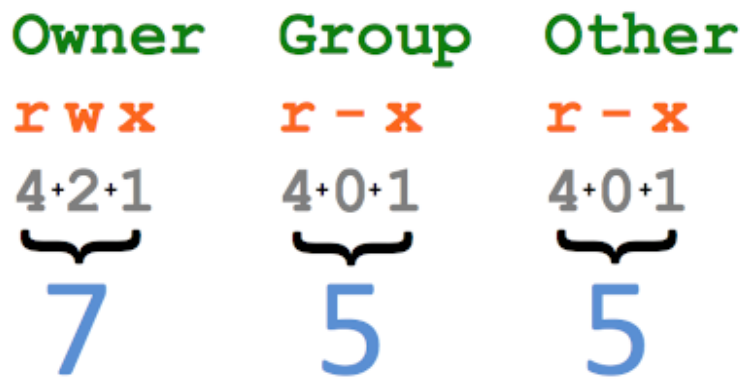
- ls - list directory contents
  - -l use a long listing format
  - -h with -l and -s, print sizes like 1K 234M 2G etc.
  - -a do not ignore entries starting with .

rodovalhovr@IS-2346:~$	ls -lh

```
total 99M
-rw-r--r--  1 rodovalhovr rodovalhovr  99M Jul 13  2023 Miniconda3-latest-Linux-x86_64.sh
drwxr-xr-x 19 rodovalhovr rodovalhovr 4.0K Oct  9  2023 miniconda3
drwxr-xr-x  2 rodovalhovr rodovalhovr 4.0K Feb  9 09:53 test
```

Commands: *chmod*

- chmod - change file mode bits




```
rodovalhovr@IS-2346:~$ chmod 775 Miniconda3-latest-Linux-x86_64.sh
```

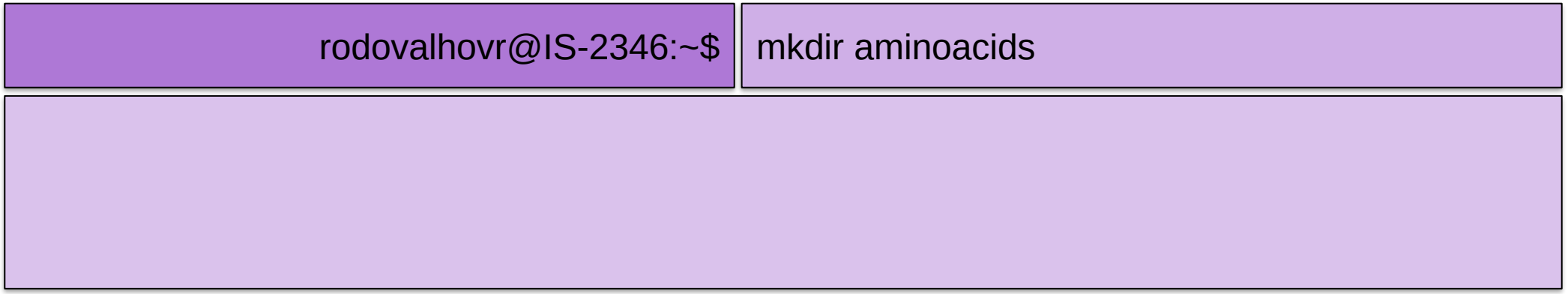
Commands: *chmod + ls -lh*

rodovalhovr@IS-2346:~$	ls -lh

```
total 99M
-rwxrwxr-x  1 rodovalhovr rodovalhovr  99M Jul 13  2023 Miniconda3-latest-Linux-x86_64.sh
drwxr-xr-x 19 rodovalhovr rodovalhovr 4.0K Oct  9  2023 miniconda3
drwxr-xr-x  2 rodovalhovr rodovalhovr 4.0K Feb  9 09:53 test
```

Commands: *mkdir*

- mkdir - make directories



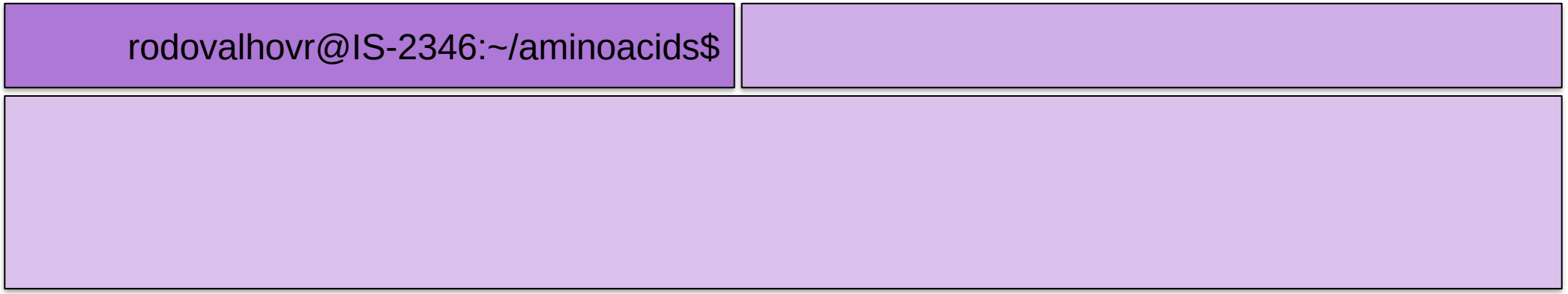
Commands: *cd*

- cd - change directory

rodovalhovr@IS-2346:~$	cd aminoacids

Commands: *cd*

- cd - change directory



Commands: *nano*

- nano is a small and friendly editor.

rodovalhovr@IS-2346:~/aminoacids$	nano seq1.fasta
>seq1 MACAMCLKLTGGA **(ctrl+X, Yes)**	

Commands: *cat, less, more*

- cat - concatenate files and print on the standard output
- more - file perusal filter for crt viewing
- less - opposite of more

```
rodovalhovr@IS-2346:~/aminoacids$ cat seq1.fasta
>seq1
MACAMCLKLTGGA
```

Commands: *head, tail*

- head - output the first part of files
- tail - output the last part of files

```
rodovalhovr@IS-2346:~/aminoacids$ head -n 1 seq1.fasta
>seq1
```

Commands: *grep*

- grep - print lines that match patterns

rodovalhovr@IS-2346:~/aminoacids$	cat seq1.fasta | grep M
MACA**M**CLKLTGGA	

Commands: *cp*

- cp - copy files and directories

```
rodovalhovr@IS-2346:~/aminoacids$ cp seq1.fasta seq1_copy.fasta
```

Commands: *mv*

- mv - move (rename) files

```
rodovalhovr@IS-2346:~/aminoacids$ mv seq1_copy.fasta seq1_new_name.fasta
```

Commands: *rm*

- rm - remove files or directories

```
rodovalhovr@IS-2346:~/aminoacids$ rm seq1.fasta
```

Commands: *nano + chmod + bash*

rodovalhovr@IS-2346:~/aminoacids$	nano script_test.sh
echo “The script is running...”	**(ctrl+X, Yes)**

Commands: *nano + chmod + bash*

rodovalhovr@IS-2346:~/aminoacids$	chmod 775 script_test.sh

Commands: *nano + chmod + bash*

rodovalhovr@IS-2346:~/aminoacids$	bash script_test.sh
“The script is running...”	

Test your knowledge!

- Test your knowledge!
  - Create a directory called **sequences**;
  - Inside it, create the files with the specified **names**
  - and *content*:
    - **1.fasta:**

      @hahaha

      AGTCGTAGAC
    - **2.fasta:**

      @hghghg

      ATGCATCGCAT
    - **script.sh:**

```
#!/bin/bash
for filename in `ls *.fasta`
do
        search="@"
        replace=">"
        sed -i "s/$search/$replace/" $filename
done
```

Test your knowledge!

- Test your knowledge!
  1. Check the content of the **sequences** directory;
  2. Check the content of **1.fasta** and **2.fasta** files;
  3. Change the permissions of the **script.sh** file to allow execution;
  4. Check again the content of the **sequences** directory;
  5. Execute the script;
  6. Check again the content of **1.fasta** and **2.fasta** files.

  Did anything change?

Now, if you haven't done it yet, let's create a virtual workspace!

- Go to GitPod:
  - https://gitpod.io/workspaces
  - Continue with GitHub and log in.
- Create a new Workspace for:
  - https://github.com/vrrodovalho/nf-core-training
  - Choose **Large** configuration.
- Proceed with the instructions in README.md.

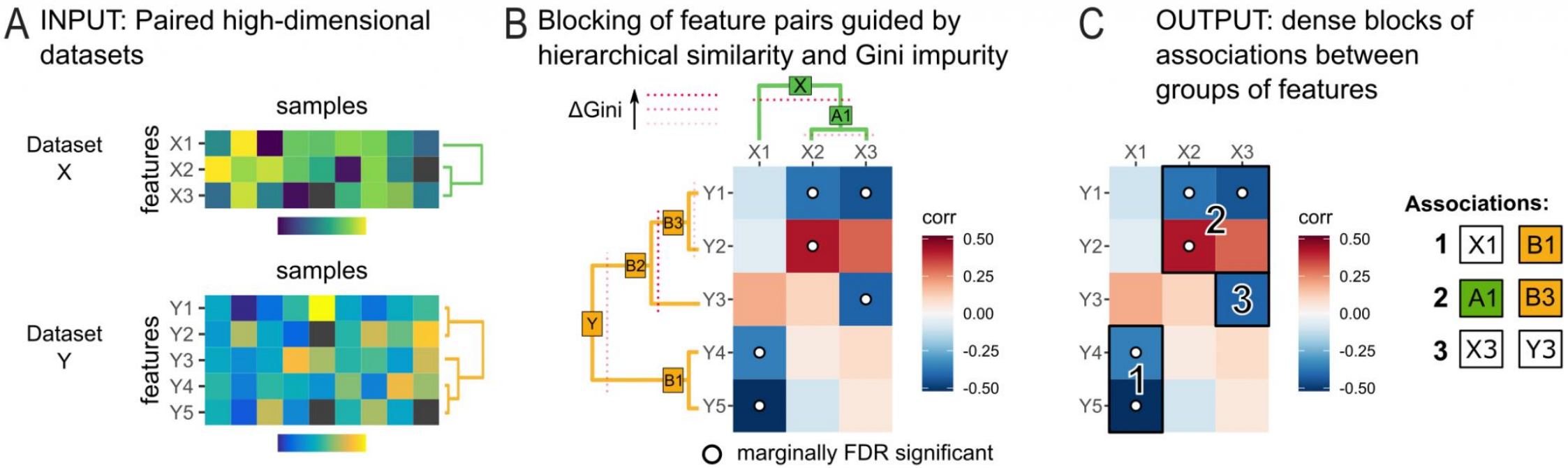
Running generate_samplesheet.py

- This is a script that can quickly generate sample sheets with your sample IDs and file paths.
  - To see the documentation and usage, use the command: **generate_samplesheet.py –h**

- Examples:
  - Single-end:
    - **python** generate_samplesheet.py **--input** path/to/data/ **--pattern** "([a-zA-Z0-9_]+).effective.(.fastq.gz)" **--paired** False
  - Paired-end:
    - **python** generate_samplesheet.py **--input** path/to/data/ **--pattern** "([a-zA-Z0-9_]+)_xxx.R1.(.fastq.gz)" **--paired** True **--replace** R1,R2

- Try to do it for the data in the virtual machine!

Running Halla (Hierarchical All-against-All association)



- HAllA is a method for finding blocks of associated features in high-dimensional datasets measured from a common set of samples.

- Installation:
  - (bash): **pip install halla**
  - (R): **install.packages("eva")**
  - (R): **install.packages("XICOR")**

Running Halla (Hierarchical All-against-All association)

- Run Halla for metabolomics.csv and taxonomy.csv files!

```
#!/bin/bash

halla_input1=metabolomics.csv
halla_input2=taxonomy.csv

echo -ne "\n\n>>> Computing ${halla_input1} x ${halla_input2}
associations...\n\n“
halla \
    -y ${halla_input1} \
    -x ${halla_input2} \
    -m spearman \
    --hallagram \
    --clustermap \
    --fdr_method fdr_bh \
    --fdr_alpha 0.05 \
    -n -1 \
    --diagnostic_plot \
    -o output_${halla_input1}_x_${halla_input2}
```

Commands: *halla*

rodovalhovr@IS-2346:~$	bash halla_script.sh
rodovalhovr@IS-2346:~$	halla -y halla_input1.tsv –x halla_input2.tsv -m spearman --hallagram --clustermap --fdr_method fdr_bh --fdr_alpha 0.05 -n -1 --diagnostic_plot -o output_name

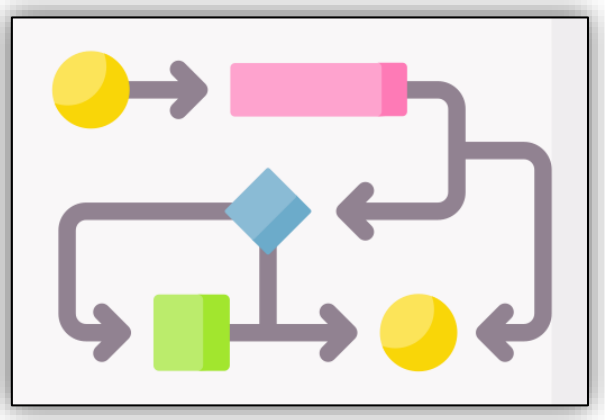
PART 2

Nextflow

Bioinformatics workflows

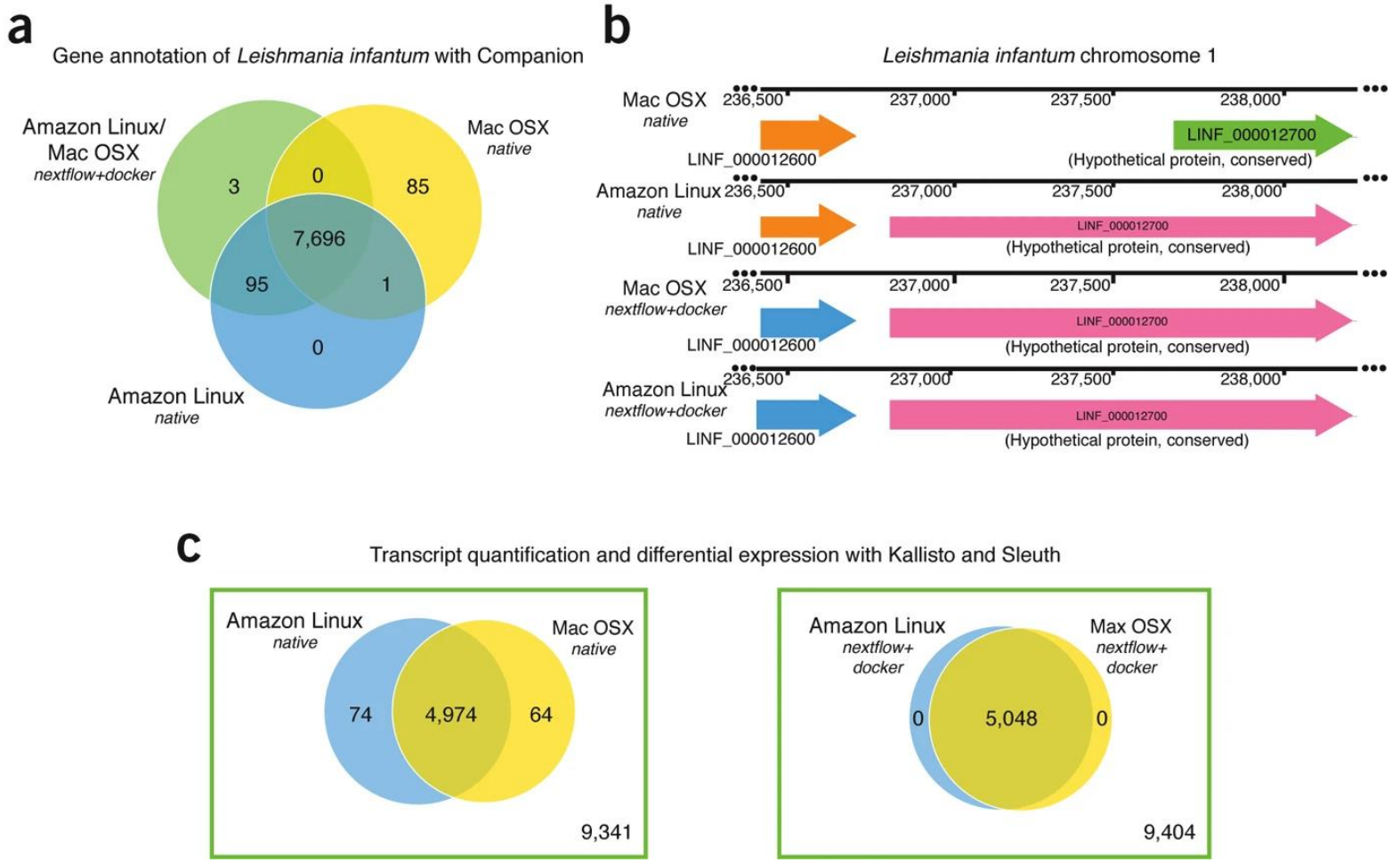
Many databases

Diversity of tools



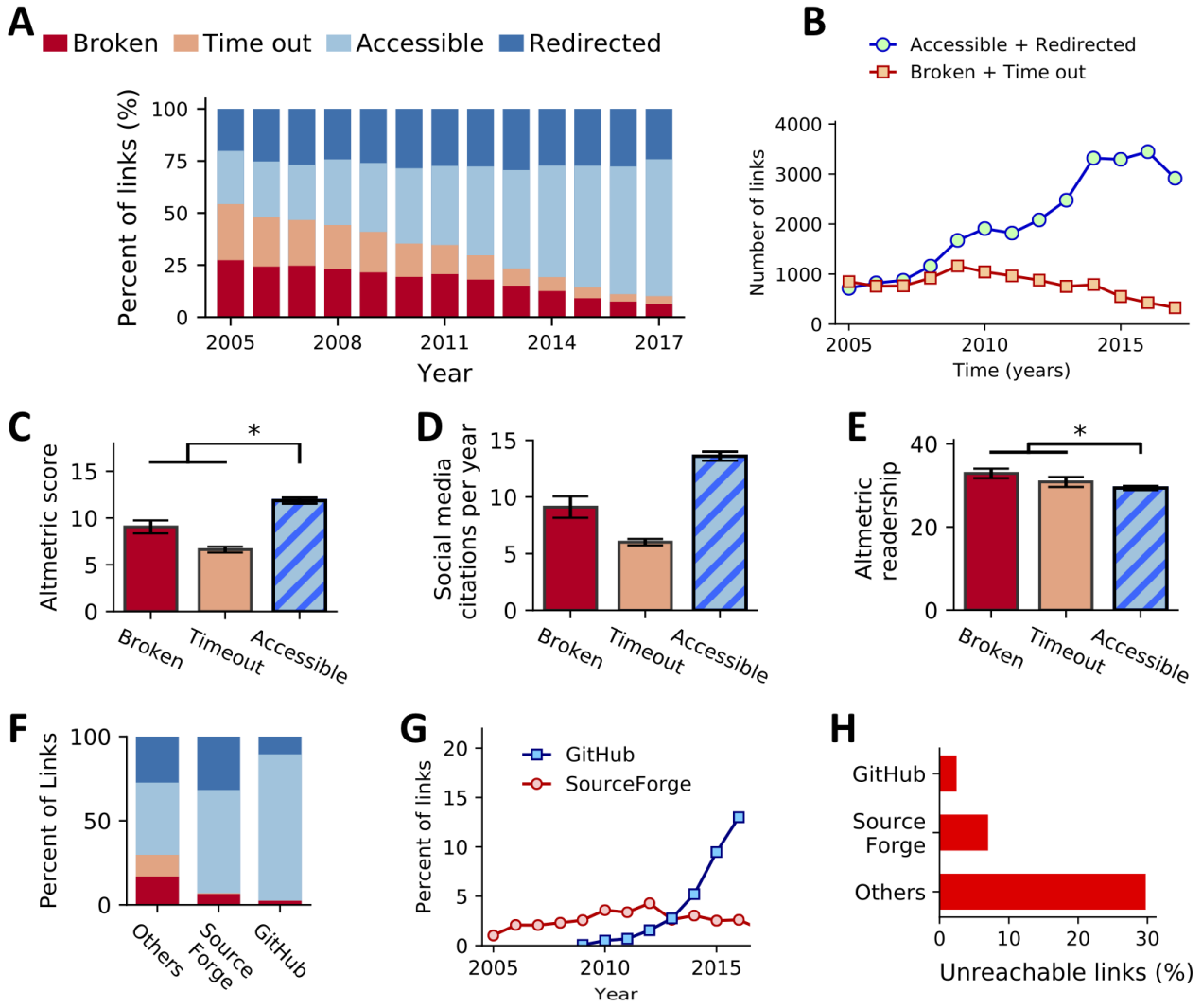
Dynamic pipelines

Platform differences on genome annotation and transcript quantification



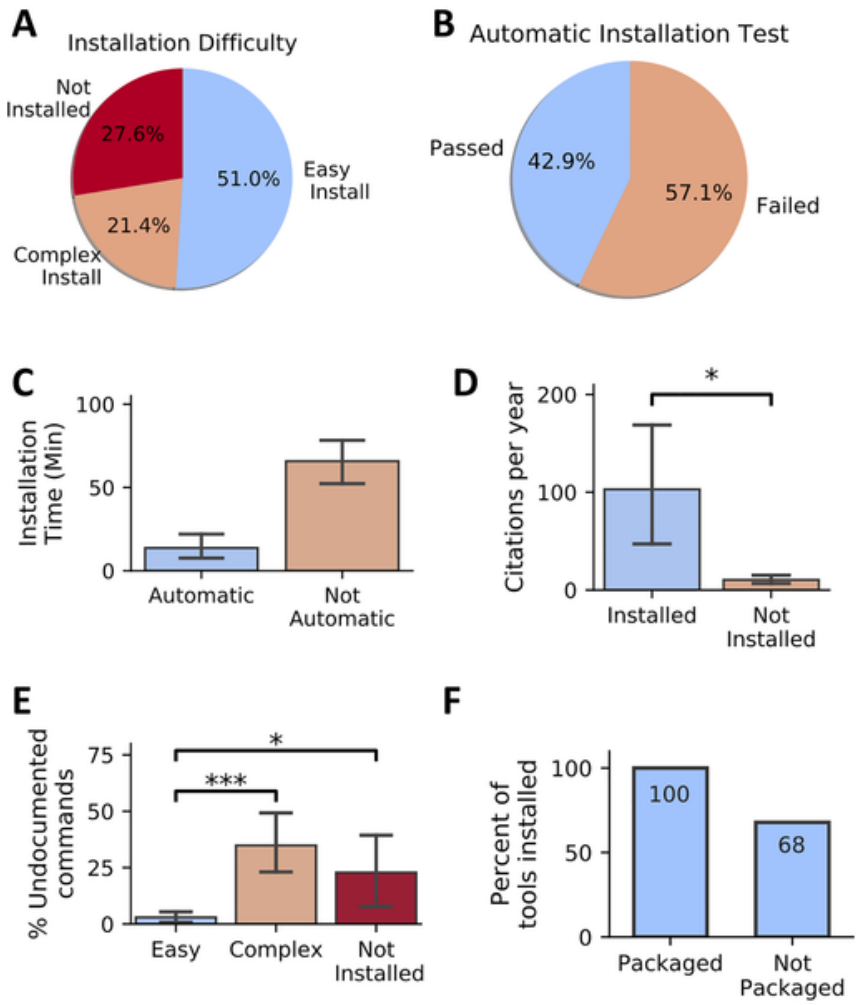
(Di Tommaso et al. Nat Biotechnol. 2017)

Archival stability of published computational tools and resources



(Mangul et al. PLoS Biol. 2019)

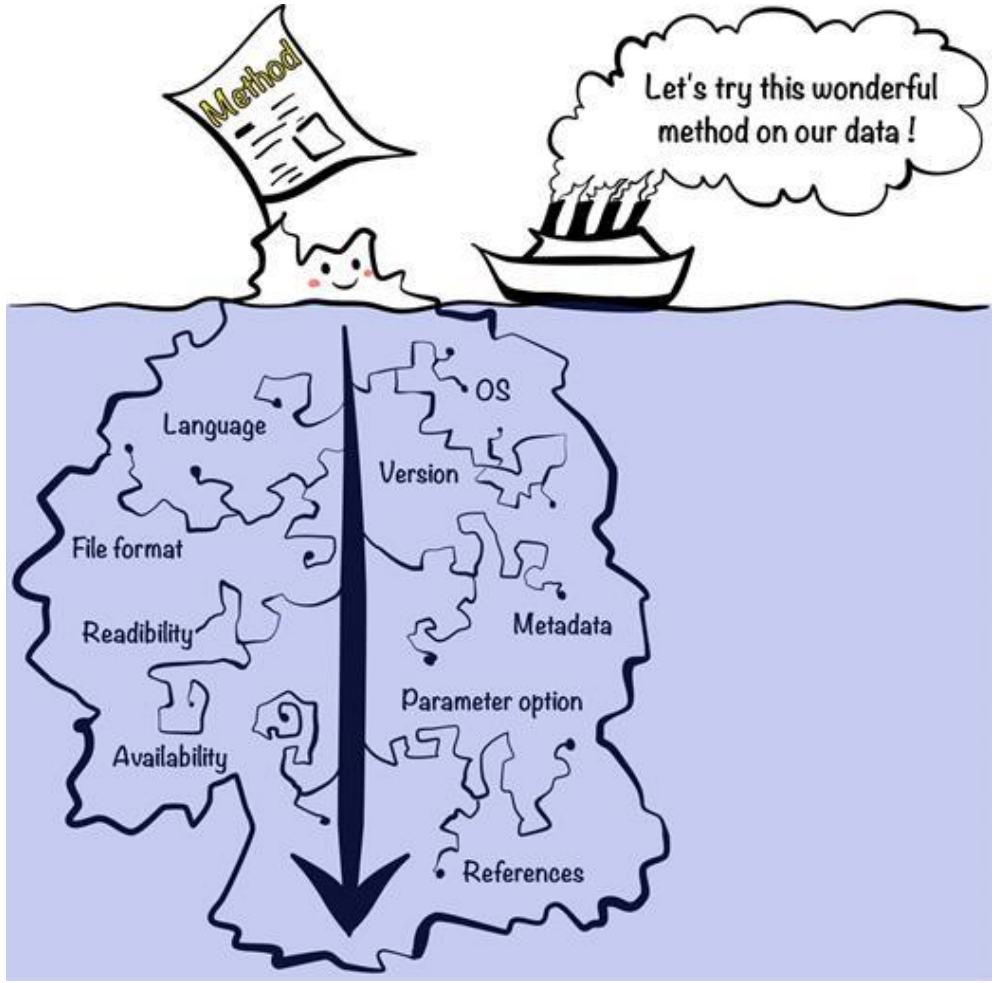
Installability of published software tools



(Mangul et al. PLoS Biol. 2019)

Hidden reproducibility issues are like an underwater iceberg

- There is a significant gap between:
  - Apparent executable work
    - (i.e., above water portion of iceberg);
  - And necessary effort in practice
    - (i.e., the full iceberg).
- The main barrier to reproducibility is the lack of compatibility among:
  - environments,
  - programming languages,
  - software versions, and the like.
- At the individual level, the key is research practices such as:
  - well-written, tested, and documented code;
  - well-curated data;
  - and the use of online repositories and collaborative tools.



(Kim et al. Gigascience 2018)

What is Nextflow?

- **Nextflow** is a:
  - **Workflow orchestration engine**;
  - **Domain-Specific Language** (DSL) that makes it easy to write data-intensive computational workflows.

- Designed around the idea that the **Linux** platform is the lingua franca of data science.
  - Powerful command-line and scripting tools that facilitate complex data manipulations.

- Nextflow's core features are:
  - Workflow **portability** and **reproducibility**;
  - **Scalability** of **parallelization** and deployment;
  - **Integration** of existing tools, systems, and industry standards.

What is Nextflow?

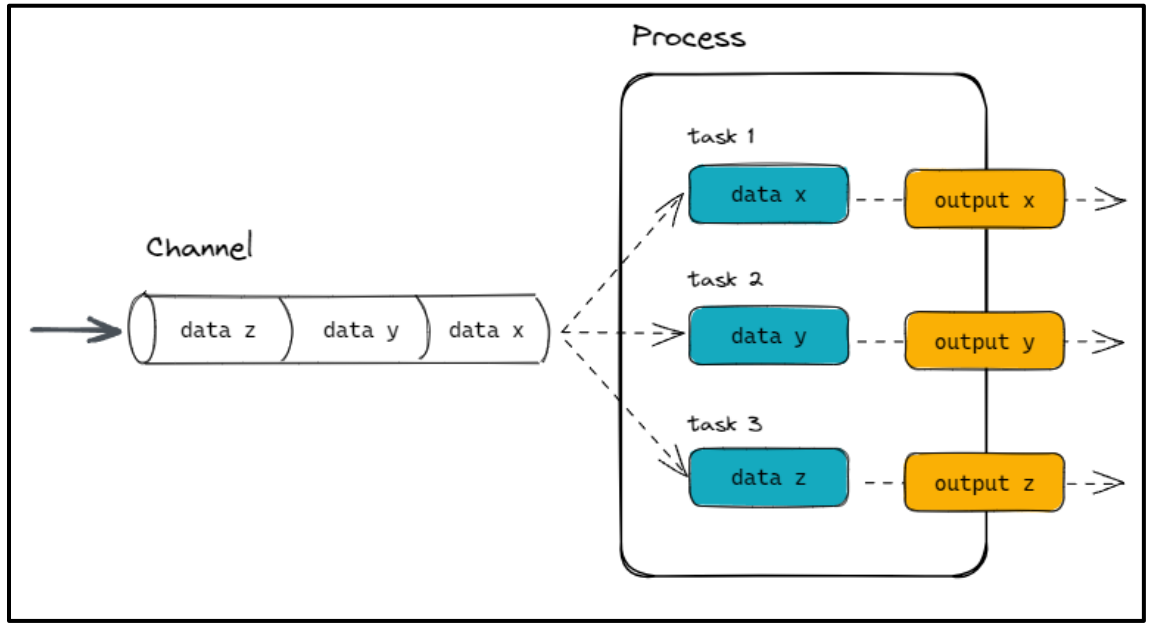
nextflow

- Language
- Pipeline description
- Process, Channel, Workflow

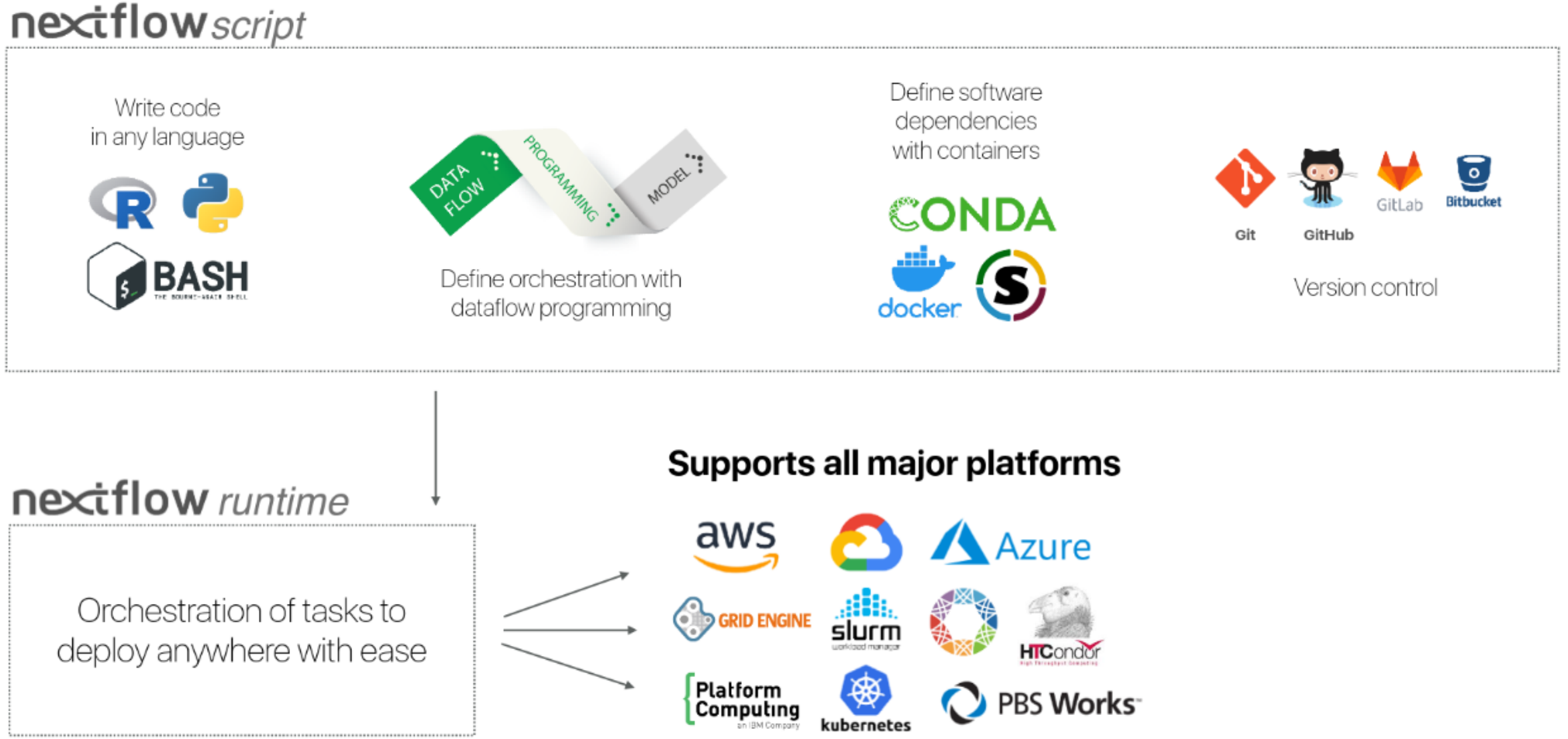
```
process sayHello {
  input:
    val cheers
  output:
    stdout

  """
  echo $cheers
  """
}

workflow {
  channel.of('Ciao','Hello','Hola') | sayHello | view
}
```

Workflows, Processes, Channels and Executors

- A Nextflow **workflow** is made by joining together different **processes**:
  - Each process can be written in any scripting **language** that can be executed by the Linux platform (Bash, Perl, Ruby, Python, etc.);
  - Processes are executed **independently**;
  - A process defines **what** command or script has to be executed.
- Processes communicate via asynchronous first-in, first-out (FIFO) queues, called **channels**.
- The **executor** determines how that script is run in the target platform.
  - If not otherwise specified, processes are executed on the **local** computer.
  - For real-world computational workflows a high-performance computing (**HPC**) is often needed.



Workflows, Processes, Channels and Executors



Nextflow installation

- Check requirements:

  java -version

  (Java 11 or later, up to 22)

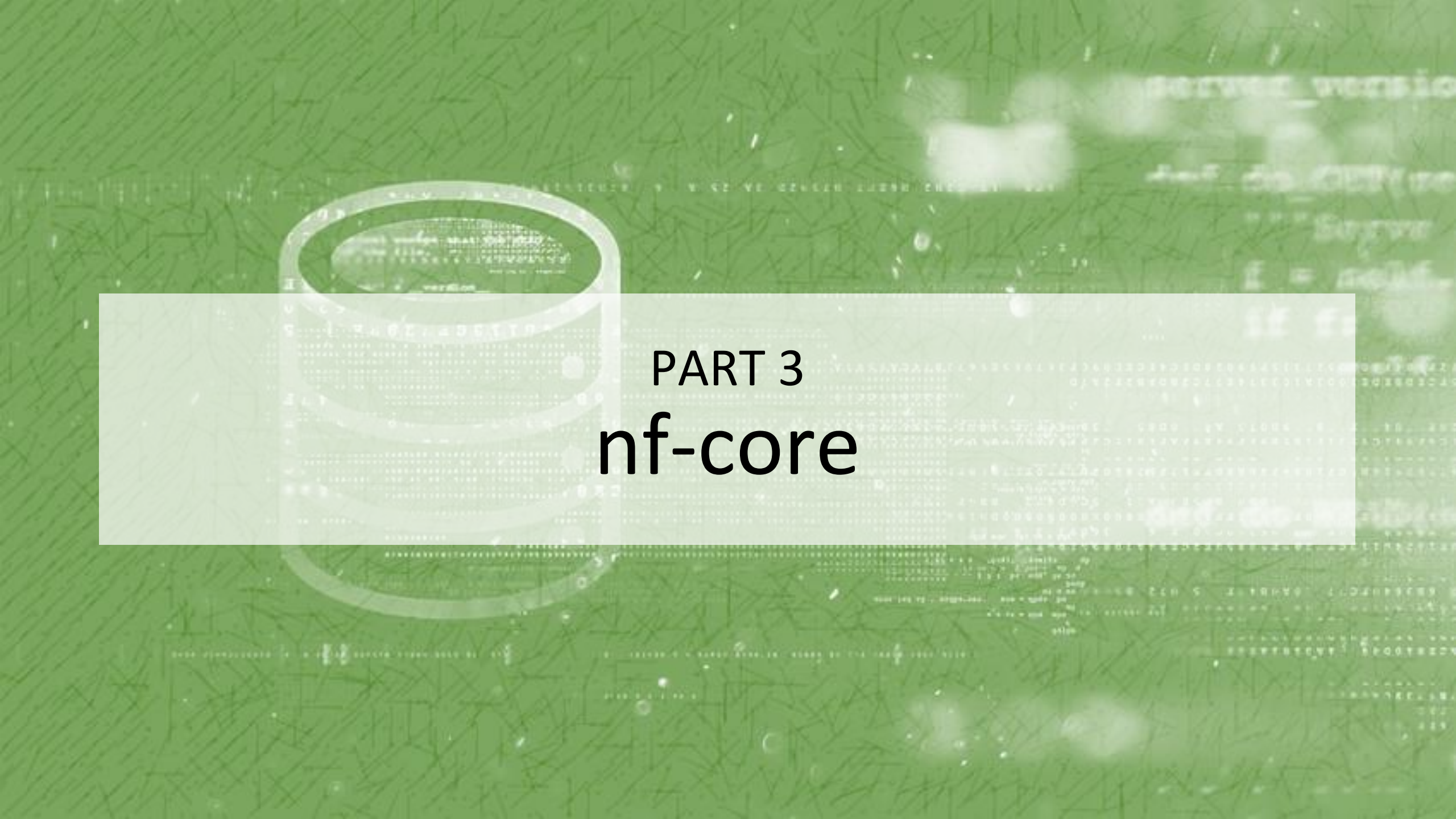
- If not satisfied:
  - **curl -s https://get.sdkman.io | bash**
  - *(Open new terminal)*
  - **sdk install java 17.0.10-tem**
  - **java -version**

- Install:
  - **curl -s https://get.nextflow.io | bash**
  - **chmod +x nextflow**
  - **sudo mv nextflow /usr/local/bin**
  - **nextflow info**

- Update:
  - **nextflow self-update**

Nextflow's key features

- Reproducible — between executions
- Portable — between systems
- Scalable — anywhere



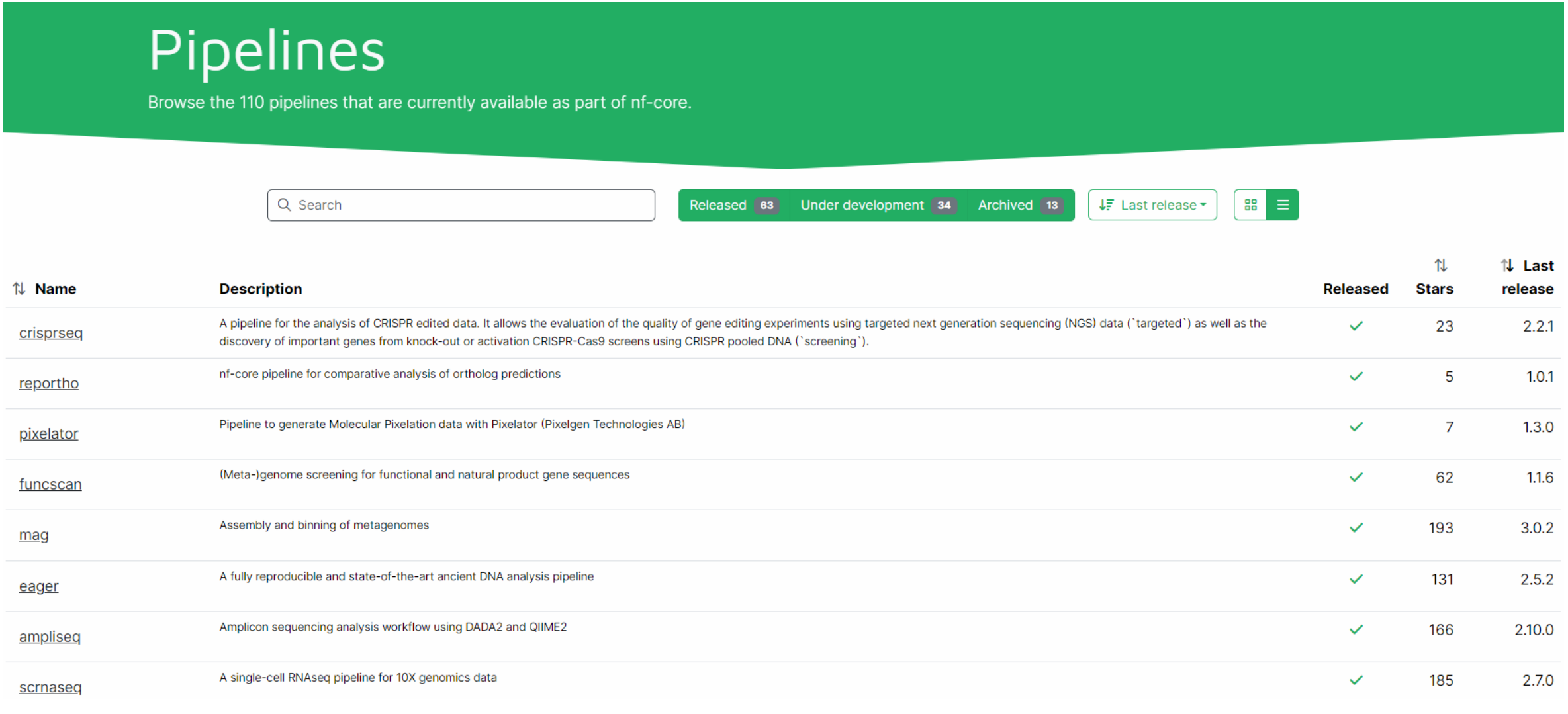
PART 3
nf-core

nf-core

- A **community** effort to collect a **curated** set of analysis **pipelines** built using Nextflow.
- nf-core has three target audiences:
  - **Facilities**, it provides highly automated and optimised pipelines that guarantee reproducibility of results for users.
  - **Single users** benefit from portable, documented, and easy-to-use workflows.
  - **Developers** can write Nextflow pipelines with nf-core's ready-made templates and tools.

(Ewels, et al. Nat Biotechnol. 2020)

Pipelines



Name	Description	Released	Stars	Last release
crisprseq	A pipeline for the analysis of CRISPR edited data. It allows the evaluation of the quality of gene editing experiments using targeted next generation sequencing (NGS) data (`targeted`) as well as the discovery of important genes from knock-out or activation CRISPR-Cas9 screens using CRISPR pooled DNA (`screening`).	✓	23	2.2.1
reportho	nf-core pipeline for comparative analysis of ortholog predictions	✓	5	1.0.1
pixelator	Pipeline to generate Molecular Pixelation data with Pixelator (Pixelgen Technologies AB)	✓	7	1.3.0
funcscan	(Meta-)genome screening for functional and natural product gene sequences	✓	62	1.1.6
mag	Assembly and binning of metagenomes	✓	193	3.0.2
eager	A fully reproducible and state-of-the-art ancient DNA analysis pipeline	✓	131	2.5.2
ampliseq	Amplicon sequencing analysis workflow using DADA2 and QIIME2	✓	166	2.10.0
scrnaseq	A single-cell RNAseq pipeline for 10X genomics data	✓	185	2.7.0

Modules

nf-core/modules

Browse the 1291 modules that are currently available as part of nf-core.

Search | Name

Name	Description	Keywords	# Pipeline integrations
abacas	contiguate draft genome assembly	genome assembly contiguate	1
abricate_run	Screen assemblies for antimicrobial resistance against multiple databases	bacteria assembly antimicrobial resistance	1
abricate_summary	Screen assemblies for antimicrobial resistance against multiple databases	bacteria assembly antimicrobial reistance	
abritamr_run	A NATA accredited tool for reporting the presence of antimicrobial resistance genes in bacterial genomes	bacteria fasta antibiotic resistance	
adapterremoval	Trim sequencing adapters and collapse overlapping reads	trimming adapters merging fastq	2
adapterremovalfixprefix	Fixes prefixes from AdapterRemoval2 output to make sure no clashing read names are in the output. For use with DeDup.	adapterremoval fastq dedup	
admixture	ADMIXTURE is a program for estimating ancestry in a model-based manner from large autosomal SNP genotype datasets, where the individuals are unrelated (for example, the individuals in a case-control association study).	ancestry population genetics admixture reference panels gwas	
affy_justrma	Read CEL files into an ExpressionSet and generate a matrix	affy microarray expression matrix	1

Subworkflows (modules sets)

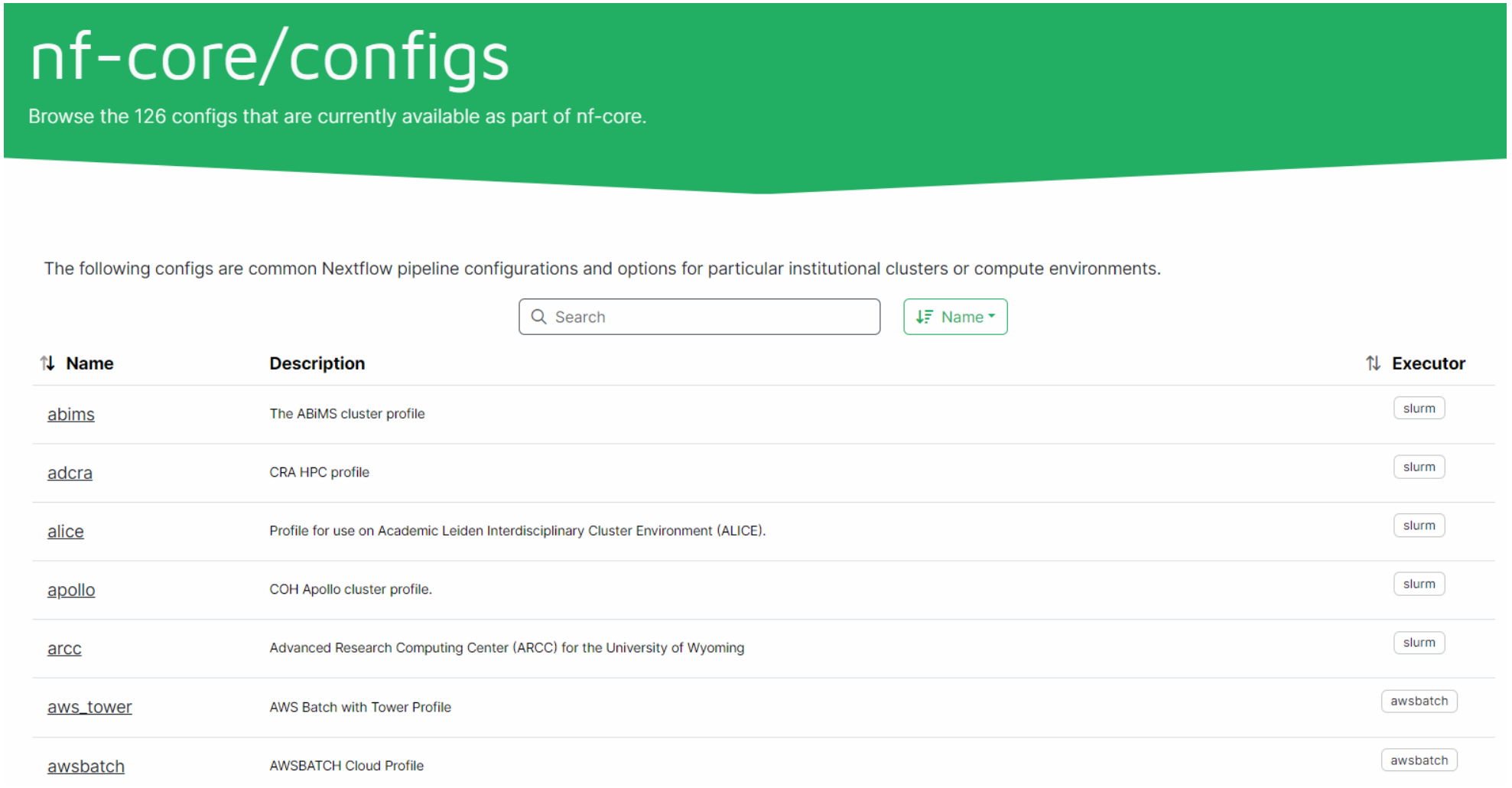
nf-core/subworkflows

Browse the 65 subworkflows that are currently available as part of nf-core.

Search | Name

Name	Description	Keywords	# Pipeline integrations
bam_cnv_wisecondorx	A subworkflow for calling CNVs using WisecondorX	cnv bam bed cram plots genomics	
bam_create_som_pon_gatk	Perform variant calling on a set of normal samples using mutect2 panel of normals mode. Group them into a genomicsdbworkspace using genomicsdbimport, then use this to create a panel of normals using createsomaticpanelofnormals.	gatk4 mutect2 genomicsdbimport createsomaticpanelofnormals variant_calling genomicsdb_workspace panel_of_normals	1
bam_dedup_stats_samtools_umicollapse	umicollapse, index BAM file and run samtools stats, flagstat and idxstats	umi dedup index bam sam cram	
bam_dedup_stats_samtools_umitools	UMI-tools dedup, index BAM file and run samtools stats, flagstat and idxstats	umi dedup index bam sam cram	3
bam_docounts_contamination_angs	Calculate contamination of the X-chromosome with ANGSD	angsd bam contamination docounts	

Shared configs



nf-core/configs

Browse the 126 configs that are currently available as part of nf-core.

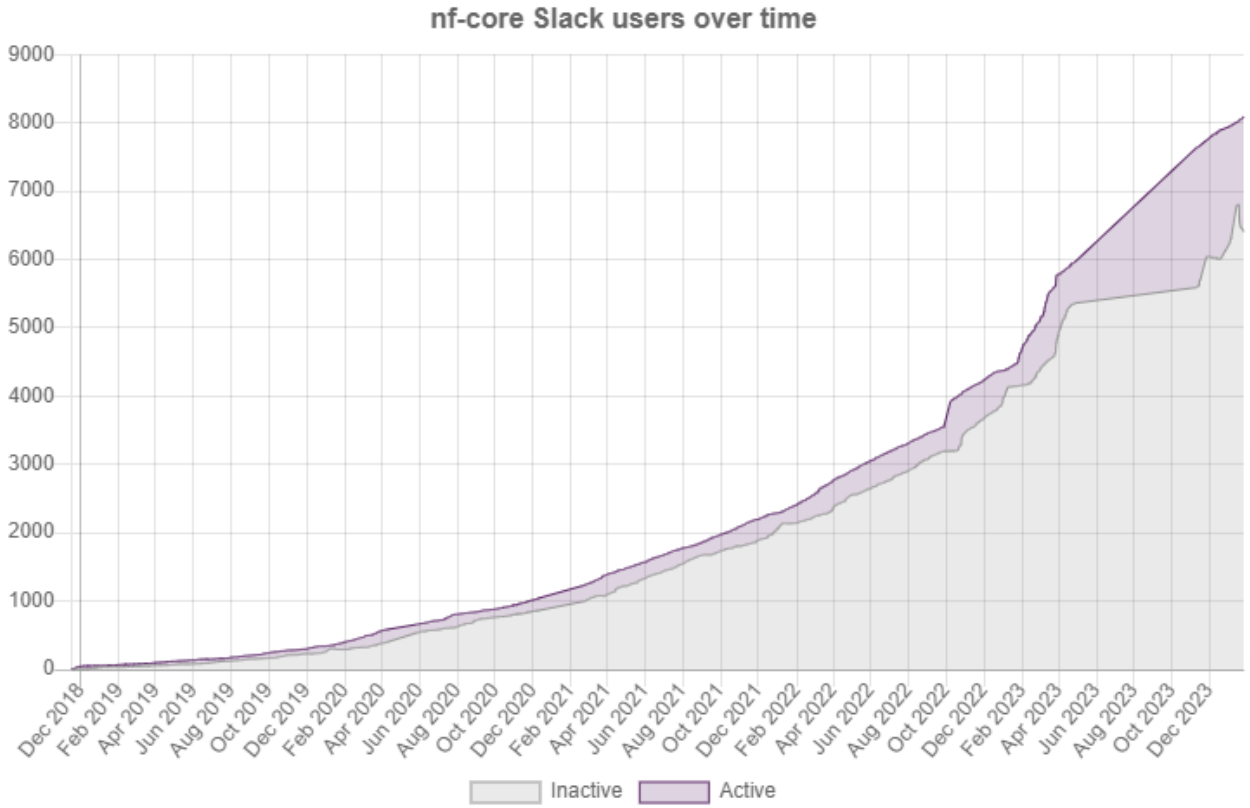
The following configs are common Nextflow pipeline configurations and options for particular institutional clusters or compute environments.

Search

Name

Name	Description	Executor
abims	The ABiMS cluster profile	slurm
adcra	CRA HPC profile	slurm
alice	Profile for use on Academic Leiden Interdisciplinary Cluster Environment (ALICE).	slurm
apollo	COH Apollo cluster profile.	slurm
arcc	Advanced Research Computing Center (ARCC) for the University of Wyoming	slurm
aws_tower	AWS Batch with Tower Profile	awsbatch
awsbatch	AWSBATCH Cloud Profile	awsbatch

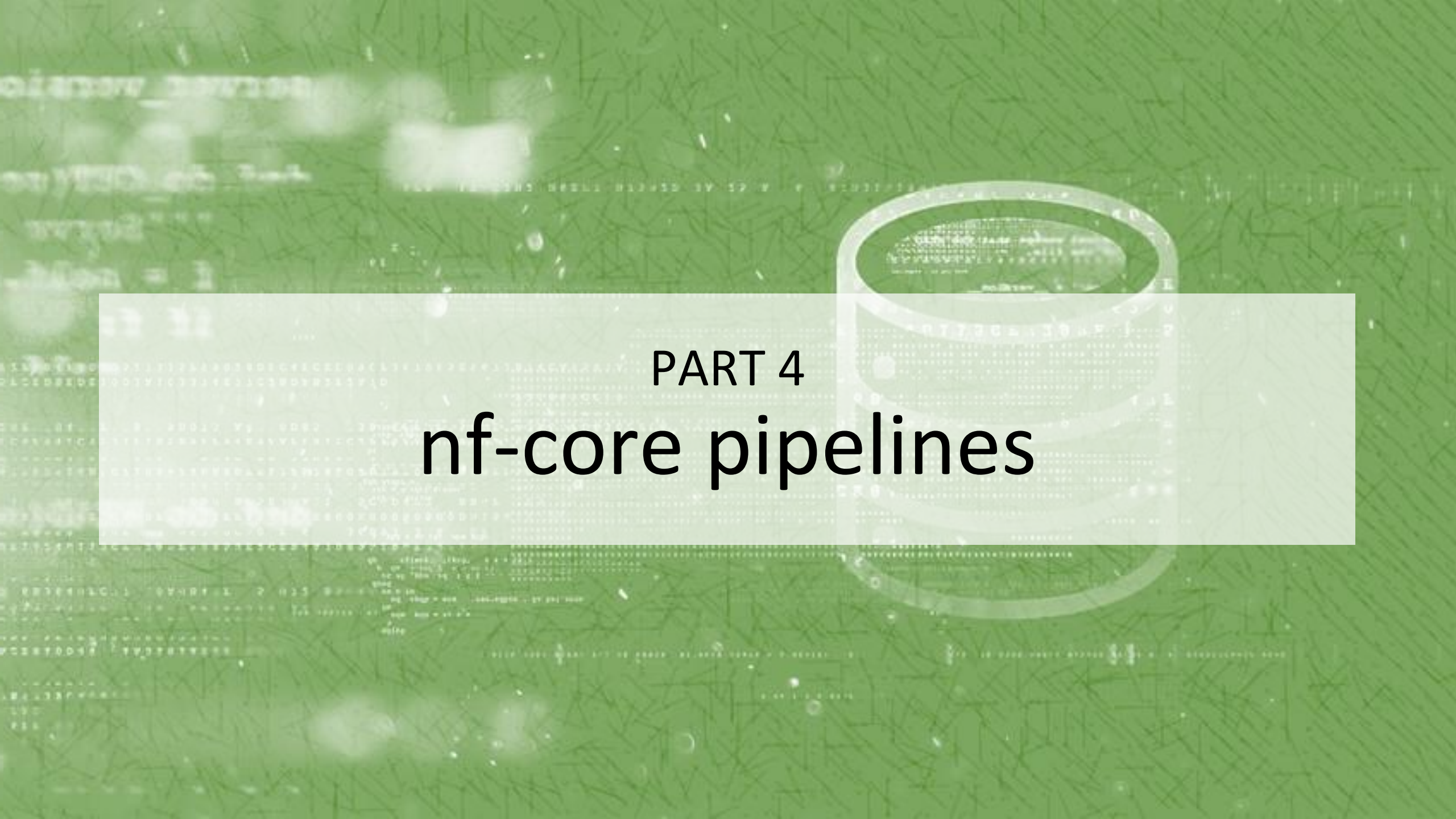
nf-core in numbers



8090 Slack users	916 GitHub organisation members
2401 GitHub contributors	3705 Twitter followers
224 Repositories	18.33K Pull Requests
43.29K Commits	8.21K Issues

nf-core installation

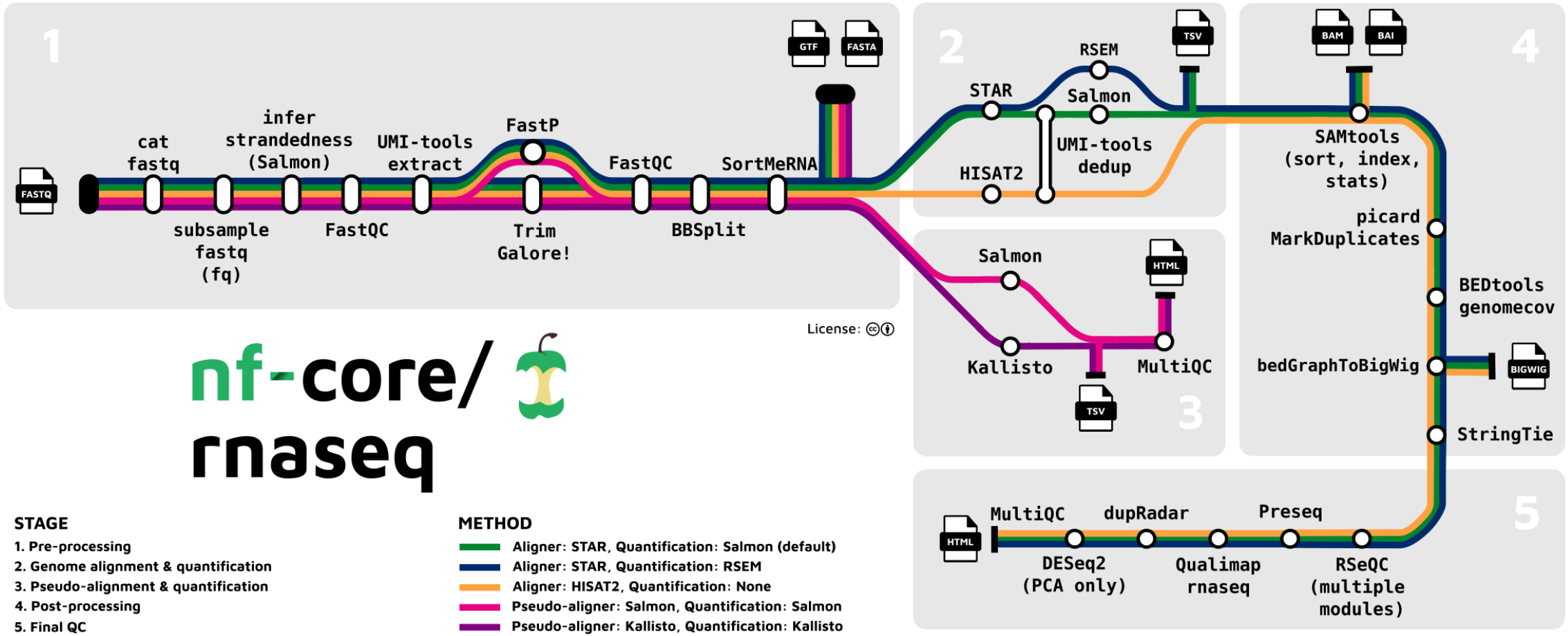
- Install nf-core:
  - **mamba create --name nf-core python=3.12 nf-core nextflow -c bioconda**
  - **mamba activate nf-core**
  - **nf-core list**



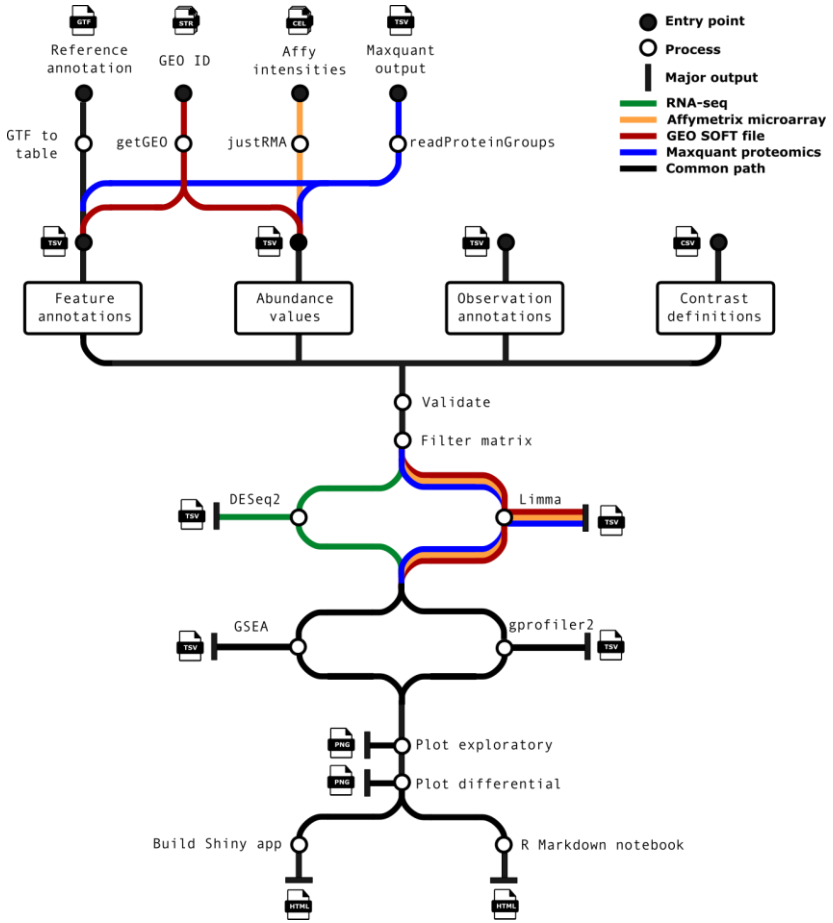
PART 4

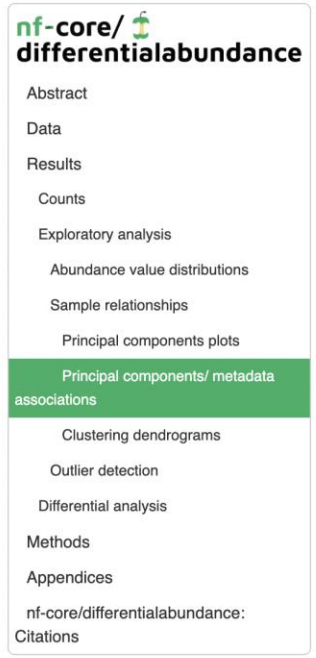
nf-core pipelines

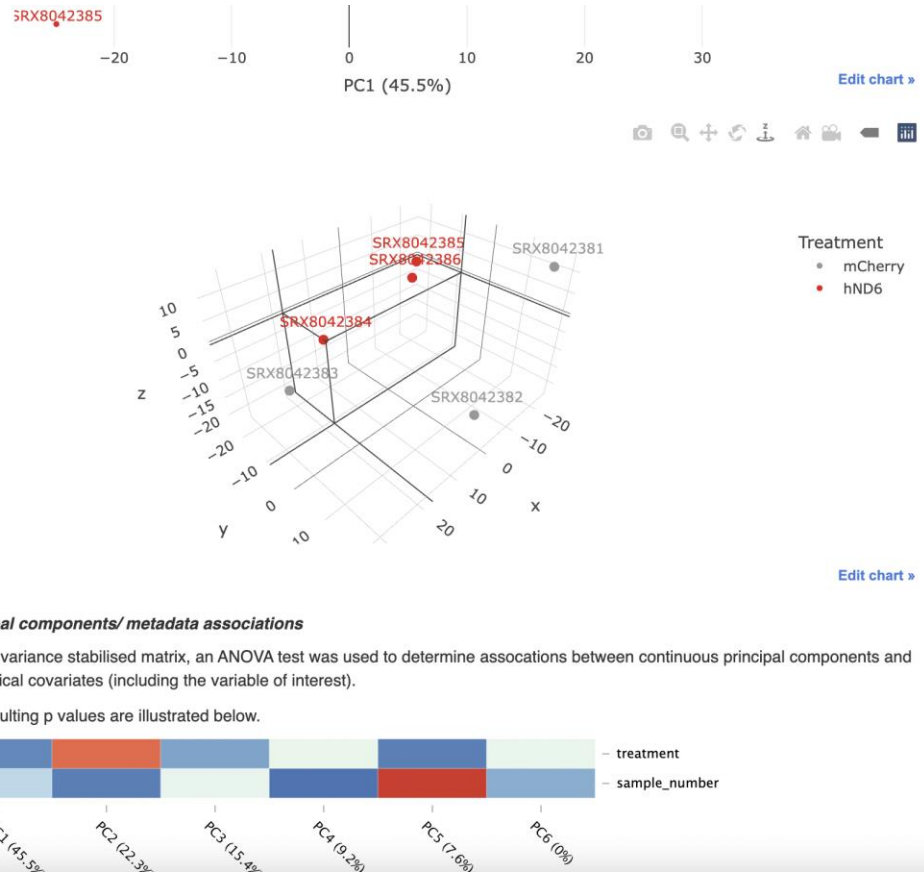
nf-core/rnaseq



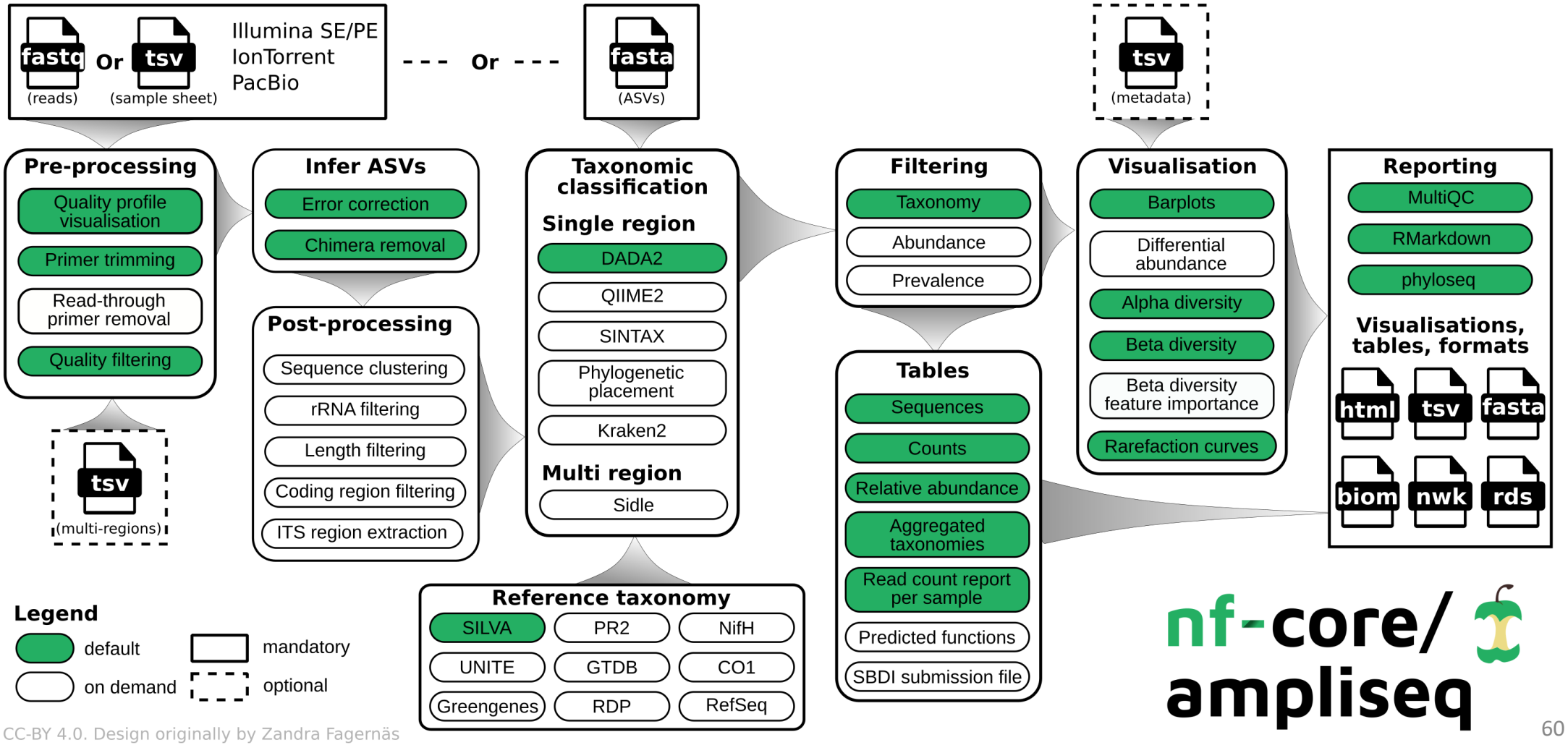
nf-core/differentialabundance







nf-core/ampliseq



nf-core/ampliseq

nf-core/ampliseq

Edit

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

Launch version 2.10.0

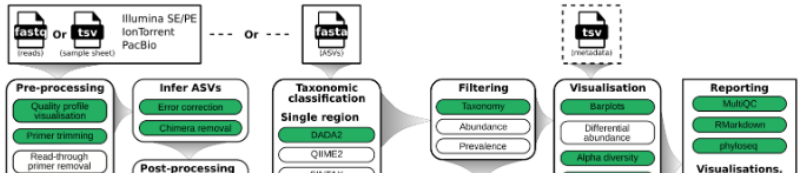
https://github.com/nf-core/ampliseq

Introduction | Usage | Parameters | Output | Results | Releases | 2.10.0

Introduction

nfcore/ampliseq is a bioinformatics analysis pipeline used for amplicon sequencing, supporting denoising of any amplicon and supports a variety of taxonomic databases for taxonomic assignment including 16S, ITS, CO1 and 18S. Phylogenetic placement is also possible. Multiple region analysis such as 5R is implemented. Supported is paired-end Illumina or single-end Illumina, PacBio and IonTorrent data. Default is the analysis of 16S rRNA gene amplicons sequenced paired-end with Illumina.

A video about relevance, usage and output of the pipeline (version 2.1.0; 26th Oct. 2021) can also be found in YouTube and billibilli, the slides are deposited at figshare.



run with

```
nf-core launch nf-core/ampliseq
```

nf-core | Nextflow | Seqera Platform

subscribers	stars
146	166
open issues	open PRs
21	1
last release	last update
about 1 month ago	about 1 month ago

get help

Ask a question on Slack

Open an issue on GitHub

How to run these pipelines?

samplesheet.tsv

sampleID	forwardReads
HB_1	fastq/HB1_1.fq.gz
HB_2	fastq/HB2_1.fq.gz
HB_3	fastq/HB3_1.fq.gz
N_1	fastq/N1_1.fq.gz
N_2	fastq/N2_1.fq.gz
N_3	fastq/N3_1.fq.gz

metadata.tsv

sampleID	condition
HB_1	HB
HB_2	HB
HB_3	HB
N_1	N
N_2	N
N_3	N

params.yaml

```
input: './samplesheet.tsv'
metadata: './metadata.tsv'
outdir: './results'
dada_ref_taxonomy: 'silva=138'
picrust: true
metadata_category: "condition"
```

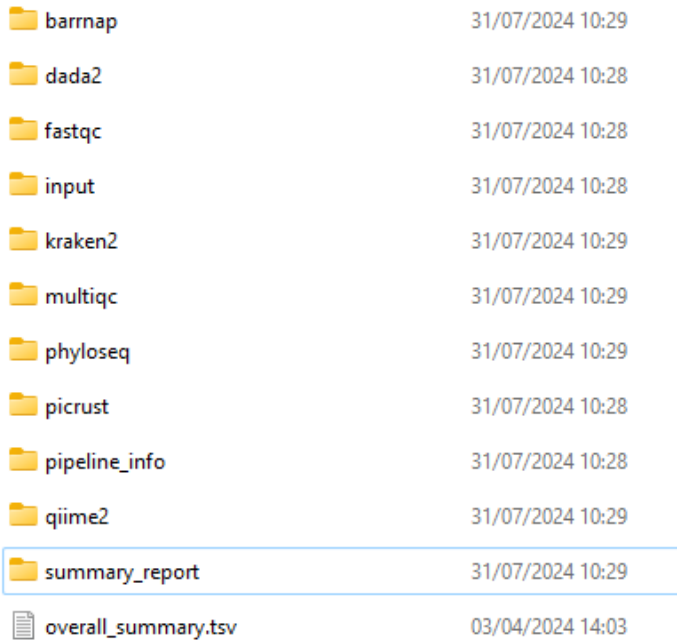
run_nextflow.sh

```
#!/bin/bash

nextflow run nf-core/ampliseq \
    -params-file params.yaml \
    -profile docker \
    -r 2.8.0
```

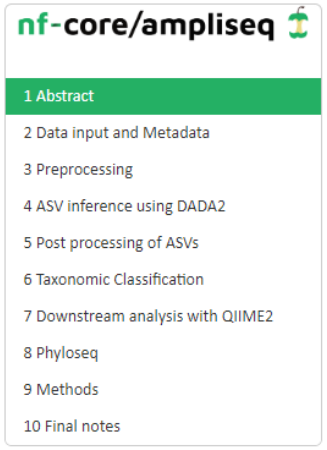
Let's practice!

- Login to GitHub.
- Go to GitPod:
  - https://gitpod.io/workspaces
- Create a new Workspace for:
  - https://github.com/vrrodovalho/nf-core-training
- If you couldn't run everything, see the results here:
  - https://gitpod.io#snapshot/6240c4f4-a127-4d78-9a80-808f7cf7cb25

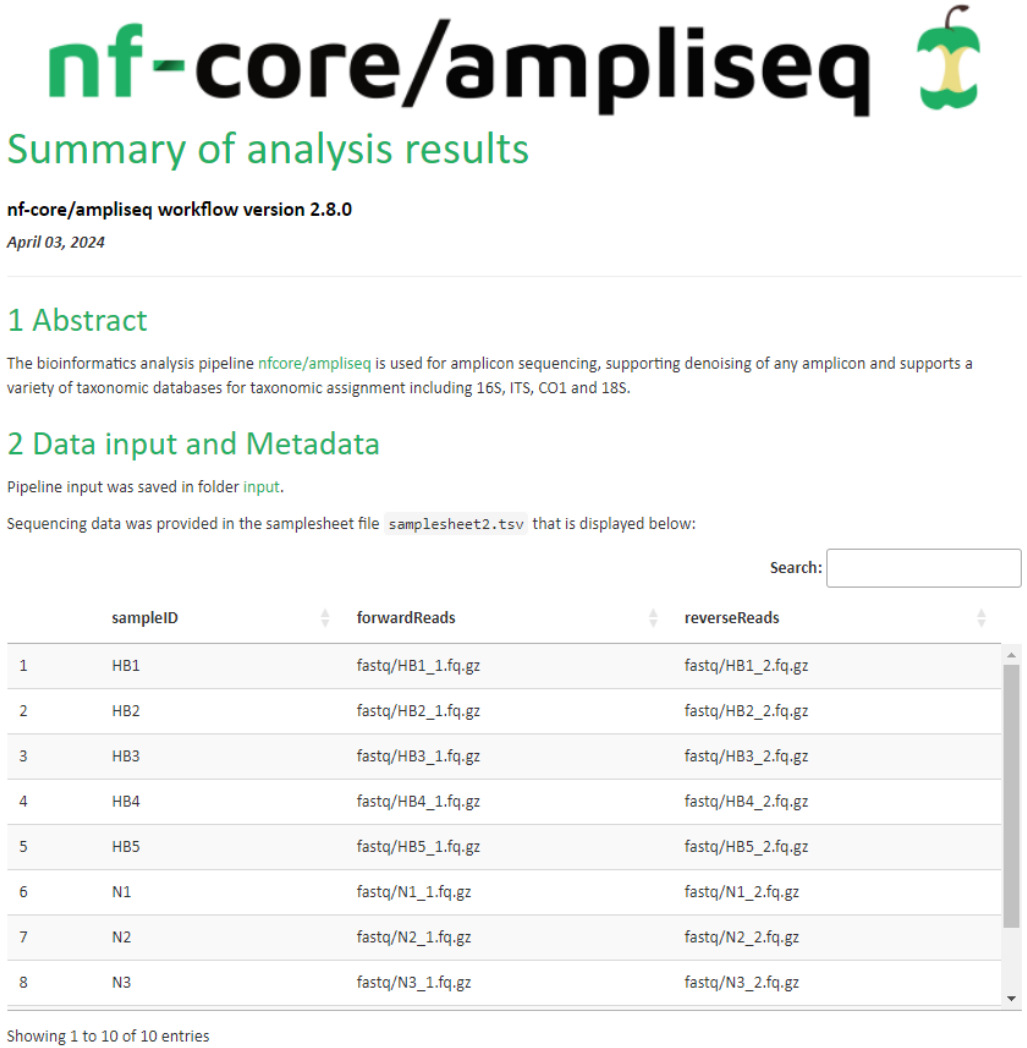
Examples of ampliseq results



barrnap	31/07/2024 10:29
dada2	31/07/2024 10:28
fastqc	31/07/2024 10:28
input	31/07/2024 10:28
kraken2	31/07/2024 10:29
multiqc	31/07/2024 10:29
phyloseq	31/07/2024 10:29
picrust	31/07/2024 10:28
pipeline_info	31/07/2024 10:28
qiime2	31/07/2024 10:29
summary_report	31/07/2024 10:29
overall_summary.tsv	03/04/2024 14:03



nf-core/ampliseq



nf-core/ampliseq

Summary of analysis results

nf-core/ampliseq workflow version 2.8.0

April 03, 2024

1 Abstract

The bioinformatics analysis pipeline nfcore/ampliseq is used for amplicon sequencing, supporting denoising of any amplicon and supports a variety of taxonomic databases for taxonomic assignment including 16S, ITS, CO1 and 18S.

2 Data input and Metadata

Pipeline input was saved in folder input.

Sequencing data was provided in the samplesheet file `samplesheet2.tsv` that is displayed below:

Search:

	sampleID	forwardReads	reverseReads
1	HB1	fastq/HB1_1.fq.gz	fastq/HB1_2.fq.gz
2	HB2	fastq/HB2_1.fq.gz	fastq/HB2_2.fq.gz
3	HB3	fastq/HB3_1.fq.gz	fastq/HB3_2.fq.gz
4	HB4	fastq/HB4_1.fq.gz	fastq/HB4_2.fq.gz
5	HB5	fastq/HB5_1.fq.gz	fastq/HB5_2.fq.gz
6	N1	fastq/N1_1.fq.gz	fastq/N1_2.fq.gz
7	N2	fastq/N2_1.fq.gz	fastq/N2_2.fq.gz
8	N3	fastq/N3_1.fq.gz	fastq/N3_2.fq.gz

Showing 1 to 10 of 10 entries

Examples of ampliseq results

3 Preprocessing

3.1 FastQC

FastQC gives general quality metrics about your sequenced reads. It provides information about the quality score distribution across your reads, per base sequence content (%A/T/G/C), adapter contamination and overrepresented sequences. The sequence quality was checked using FastQC and resulting data was aggregated using the FastQC module of MultiQC. For more quality controls and per sample quality checks you can check the full MultiQC report, which can be found in multiqc/multiqc_report.html.

Examples of ampliseq results

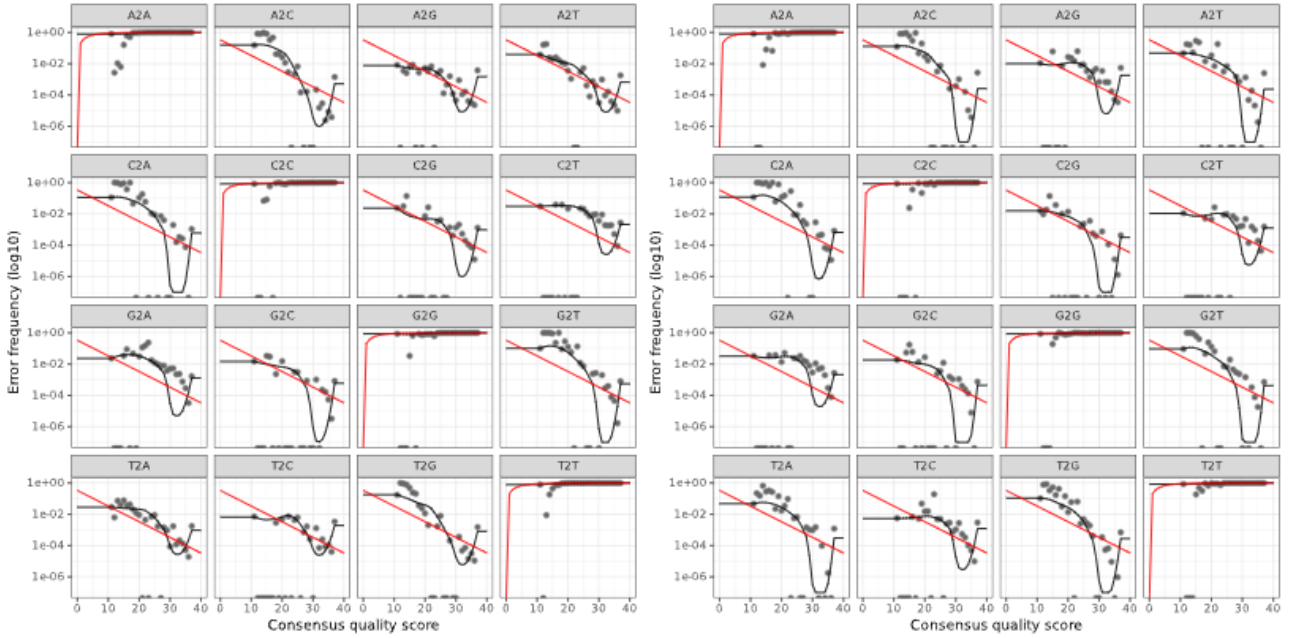
4 ASV inference using DADA2

DADA2 performs fast and accurate sample inference from amplicon data with single-nucleotide resolution. It infers exact amplicon sequence variants (ASVs) from amplicon data with fewer false positives than many other methods while maintaining high sensitivity.

DADA2 reduces sequence errors and dereplicates sequences by quality filtering, denoising, read pair merging (for paired end Illumina reads only) and PCR chimera removal.

4.1 Error correction

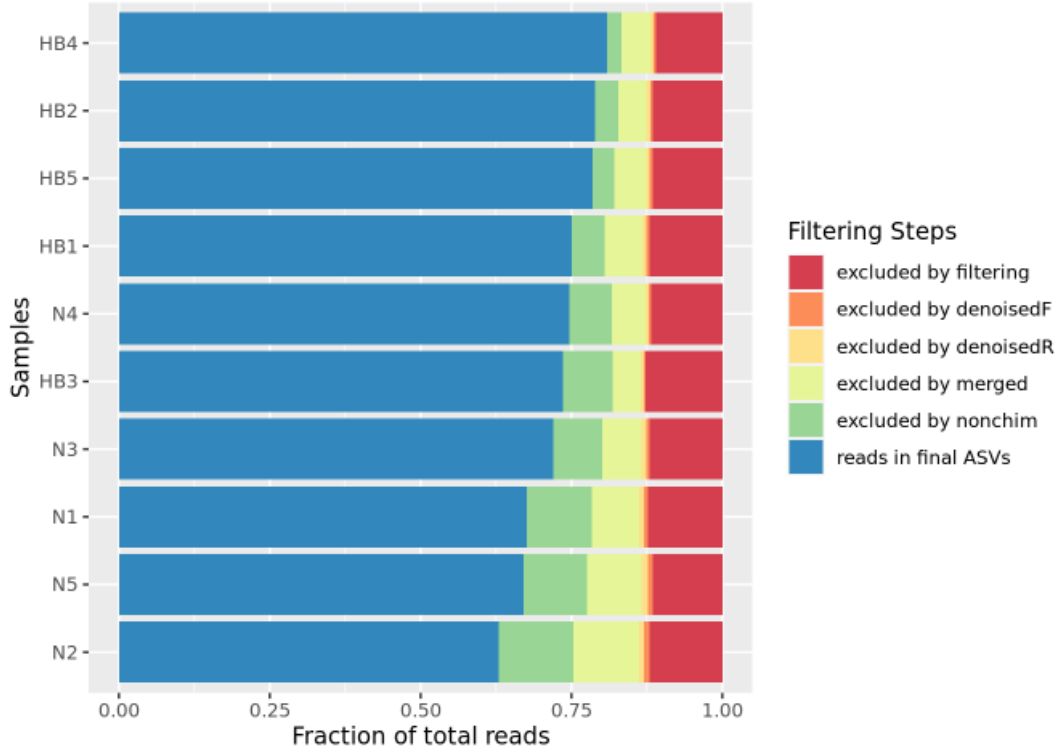
Read error correction was performed using estimated error rates, visualized below. Error rates for forward reads are at the left side and reverse reads are at the right side.



Examples of ampliseq results

Samples with unusual low reads numbers relative to the number of expected ASVs should be treated cautiously, because the abundance estimate will be very granular and might vary strongly between (theoretical) replicates due to high impact of stochasticity.

Following, the numbers of the table above are shown in stacked barcharts as percentage of DADA2 input reads. Stacked barchart of read pair numbers (denoisedF & denoisedR halfed, because each pair is split) per sample and processing stage:



Between 62.96% and 80.92% reads per sample (average 73.1%) were retained for analysis within DADA2 steps.

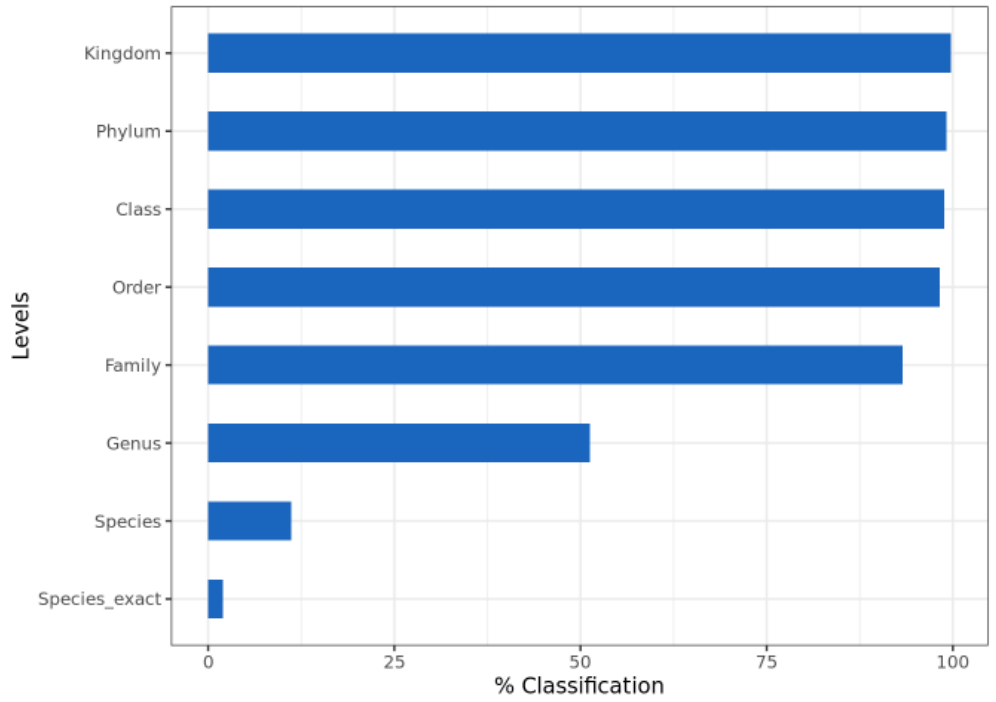
Examples of ampliseq results

6 Taxonomic Classification

6.1 DADA2

The taxonomic classification was performed by DADA2 using the database: `Silva 138.1 prokaryotic SSU`. More details about the reference taxonomy database can be found in the 'Methods section'.

DADA2 classified 99.77 % ASVs at Kingdom level, 99.09 % ASVs at Phylum level, 98.83 % ASVs at Class level, 98.26 % ASVs at Order level, 93.2 % ASVs at Family level, 51.22 % ASVs at Genus level, 11.16 % ASVs at Species level, 1.96 % ASVs at Species_exact level.



DADA2 taxonomy assignments can be found in folder dada2 in files `ASV_tax_*.tsv`.

Examples of ampliseq results

7.6 Diversity analysis

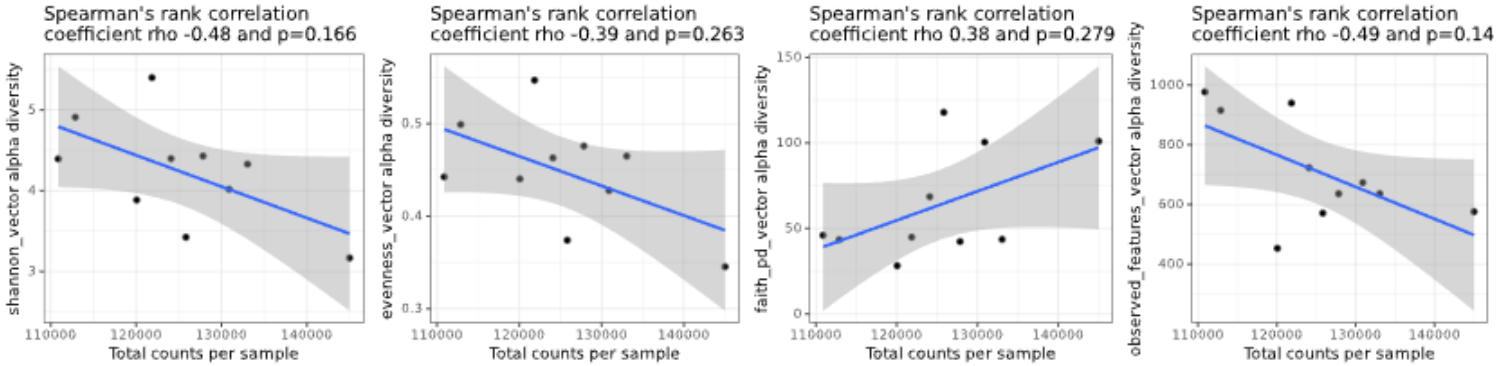
Diversity measures summarize important sample features (alpha diversity) or differences between samples (beta diversity). Diversity calculations are based on sub-sampled data rarefied to 110879 counts.

7.6.1 Alpha diversity indices

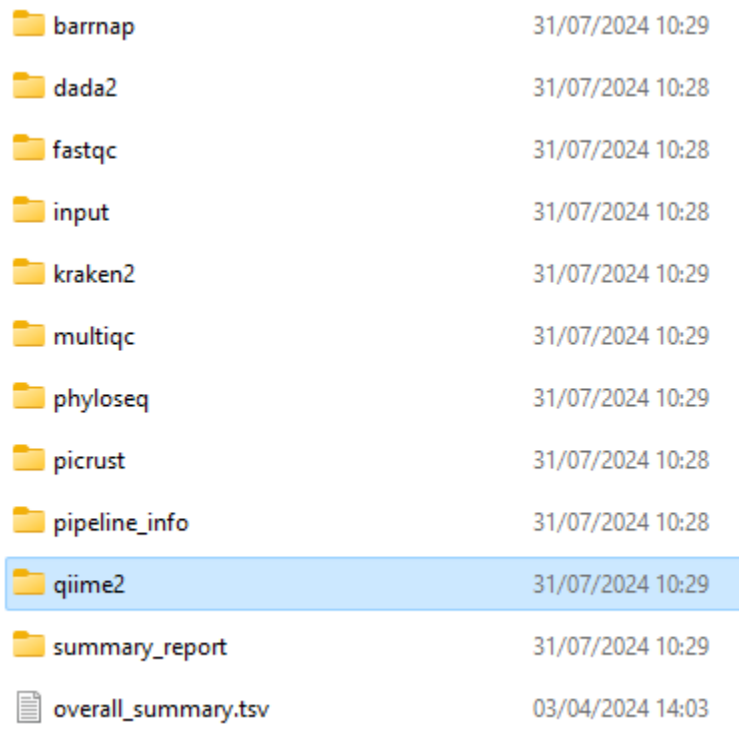
Alpha diversity measures the species diversity within samples. Please note that ASVs were inferred for each sample independently, that can make alpha diversity indices a poor estimate of true diversity. This step calculates alpha diversity using various methods and performs pairwise comparisons of groups of samples. It is based on a phylogenetic tree of all ASV sequences. Folder qiime2/diversity/alpha_diversity contains the alpha-diversity data:

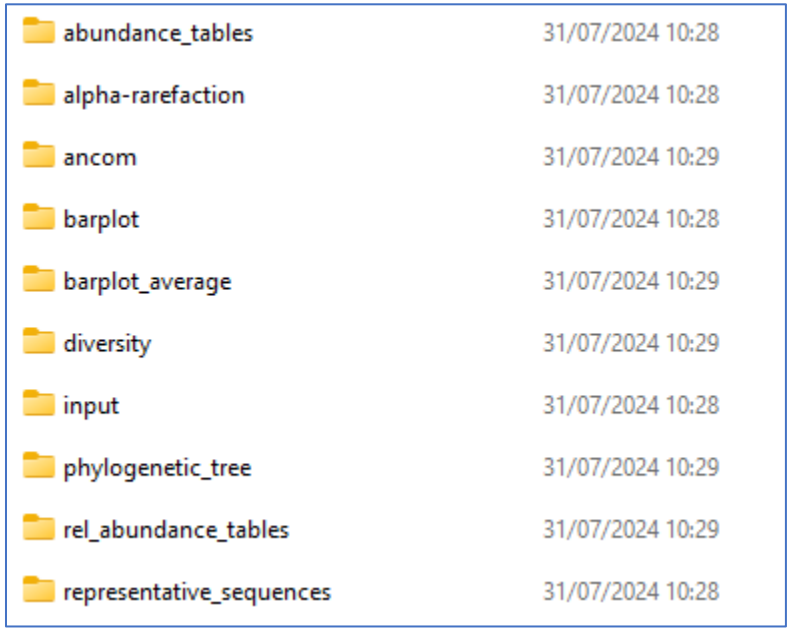
- Shannon's diversity index (quantitative): qiime2/diversity/alpha_diversity/shannon_vector/index.html
- Pielou's Evenness: qiime2/diversity/alpha_diversity/evenness_vector/index.html
- Faith's Phylogenetic Diversity (qualitiative, phylogenetic) qiime2/diversity/alpha_diversity/faith_pd_vector/index.html
- Observed OTUs (qualitative): qiime2/diversity/alpha_diversity/observed_features_vector/index.html

Alpha diversity is considered not trustworthy when it correlates positively with sequencing depth. Spearman's rank correlation was calculated for total counts per sample after all filtering steps (in folder qiime2/abundance_tables) with alpha diversity measures. No significant positive correlation was found between alpha diversity and sample counts:



Examples of ampliseq results



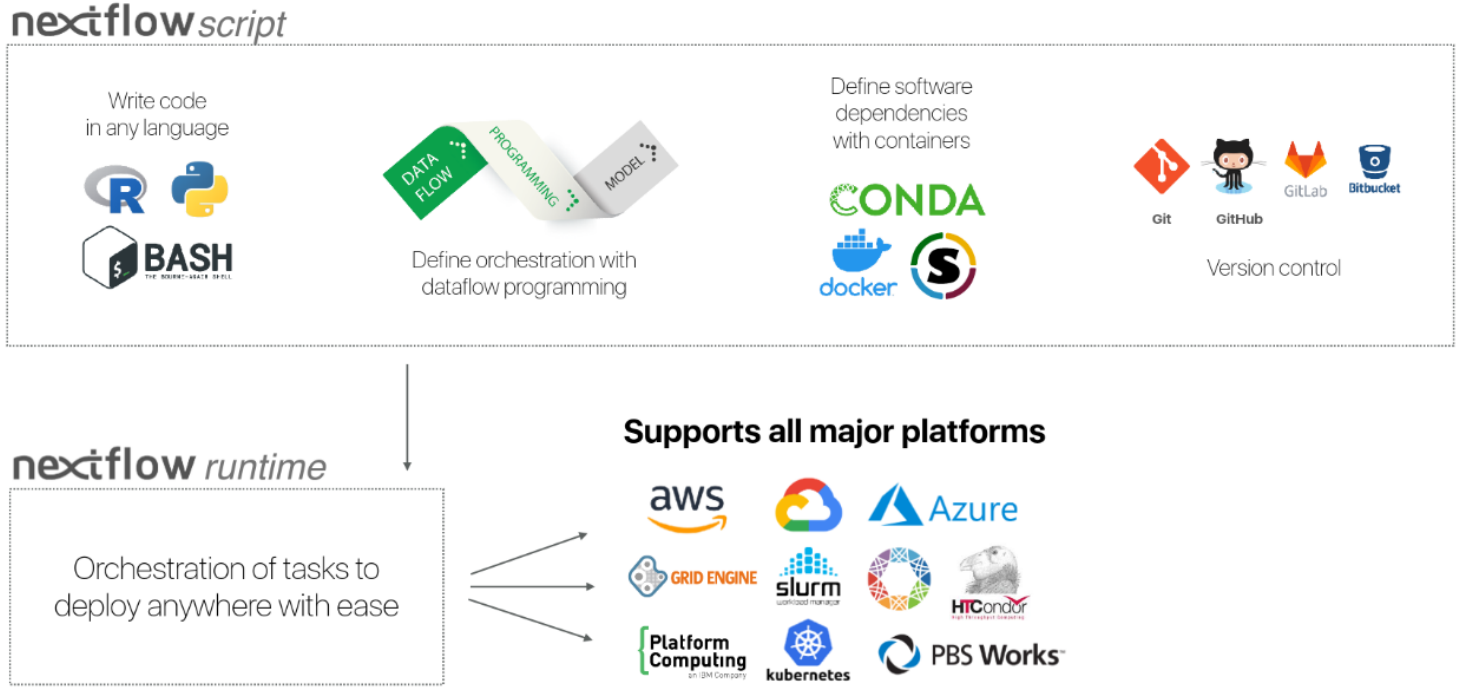


PART 5

Concluding remarks

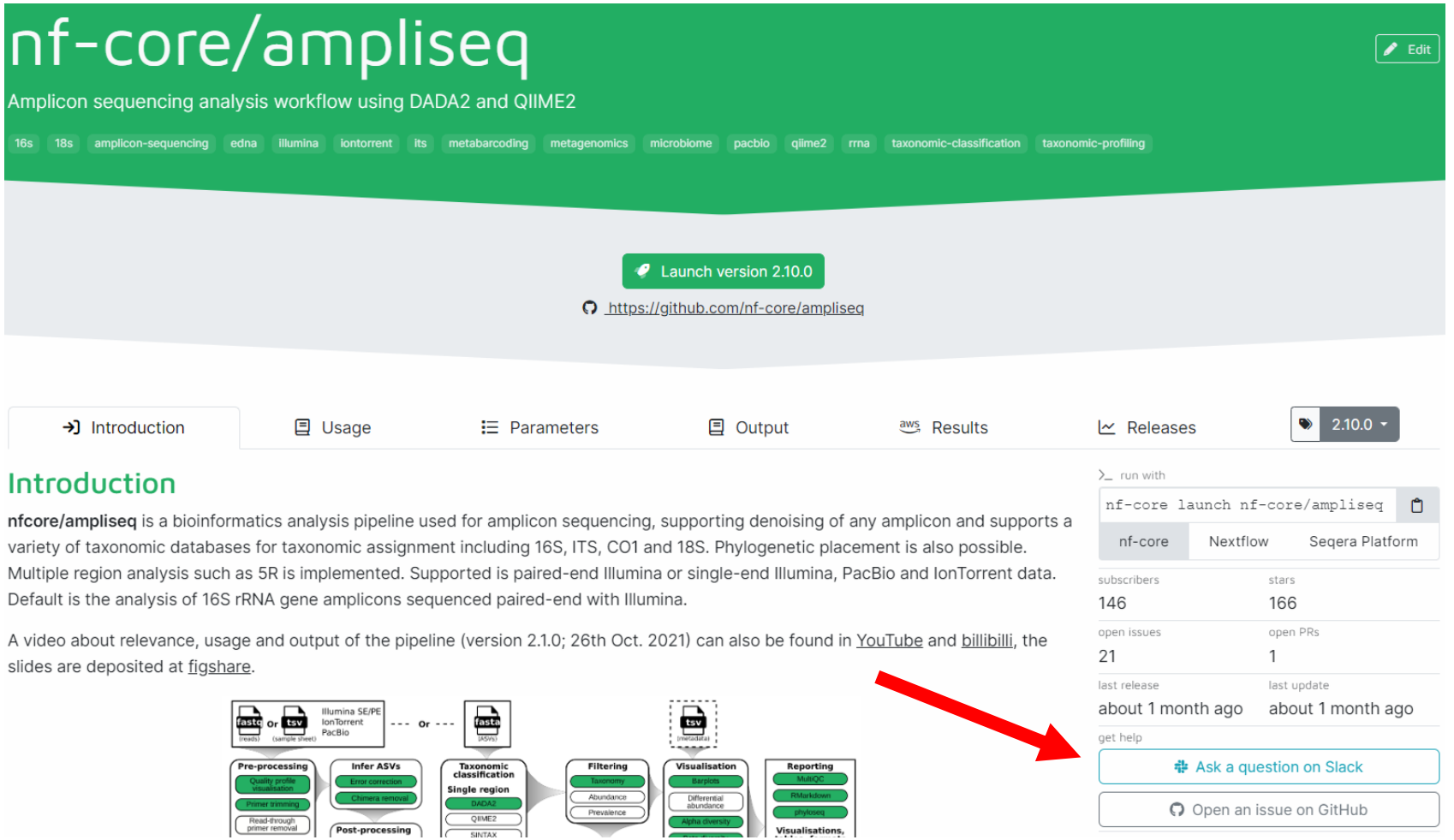
Now what?

- Start running local.
- Proceed to study configuration settings for other systems:
  - HPC (slurm, SGE…);
  - Cloud computing.
- Which container engines are more suitable?
- **Advanced**: Learn the language to create your own workflows.

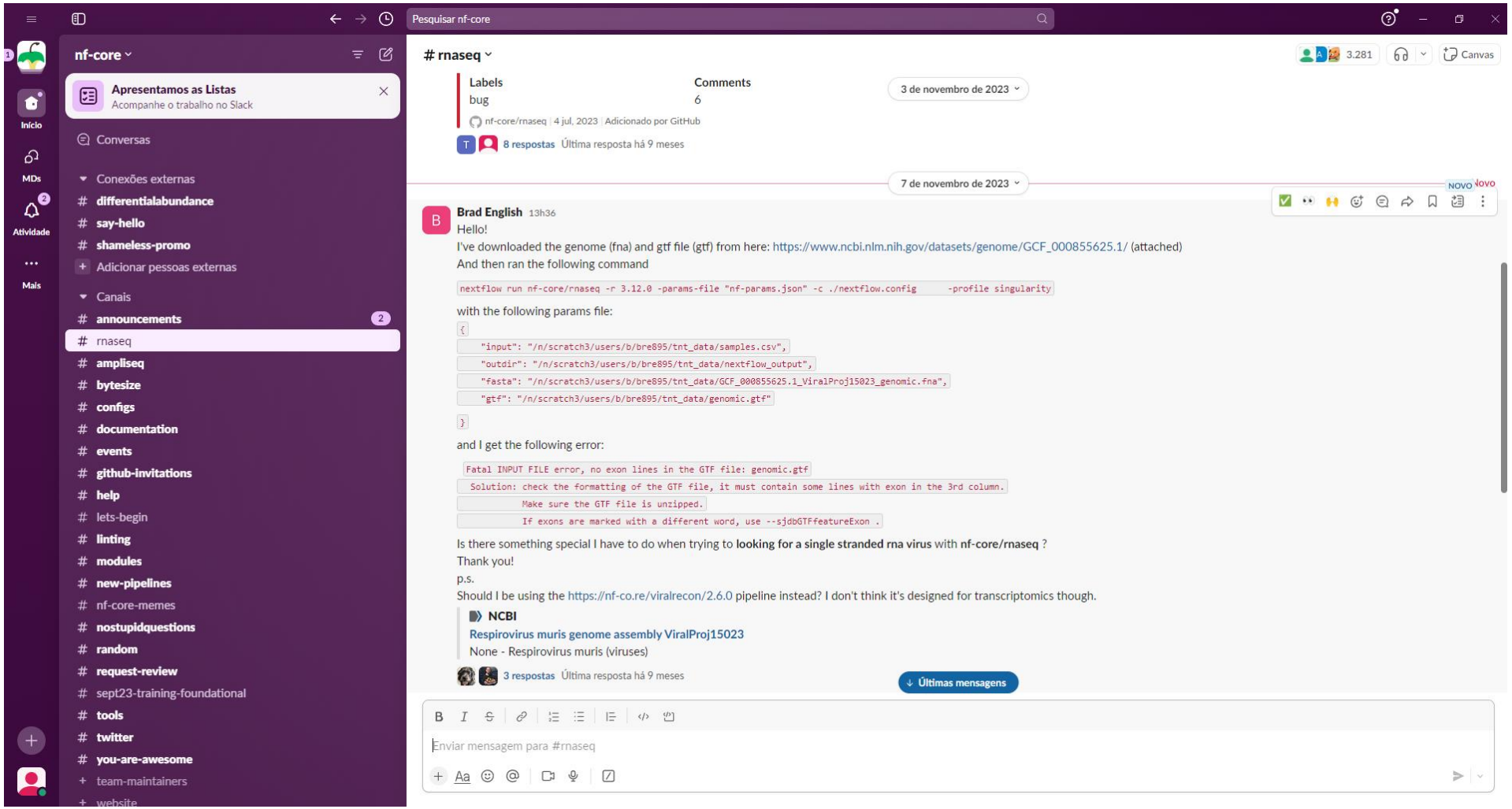


What if I have problems?

- Troubleshooting:
  - https://nf-co.re/docs/usage/troubleshooting/basics
  - Slack!
  - Mind resources and storage limitations!

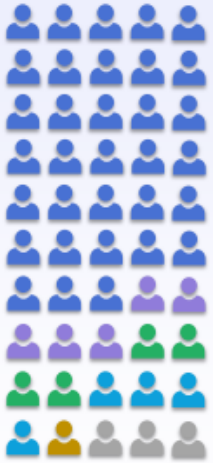


You have a lot of friends at Slack!



Some cool stats - 2023

WHO IS USING NEXTFLOW?



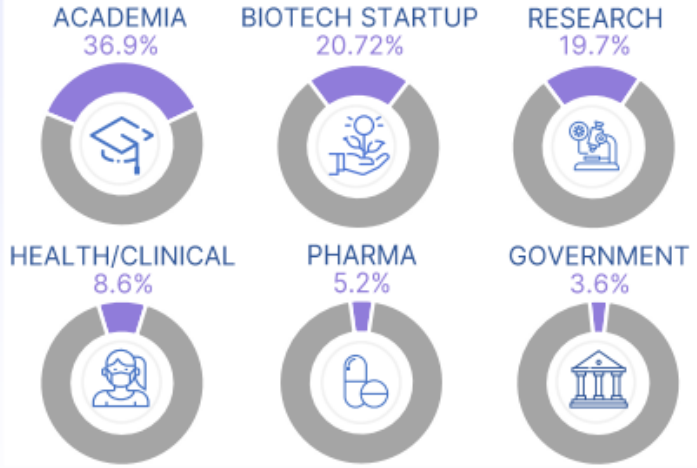
Bioinformaticians	66.5%
Principal Investigators	9.2%
Software Engineers	8.4%
Data Scientists	7.4%
HPC Sysadmin	1.6%
Others	7.0%

Most Nextflow users are **Bioinformaticians**

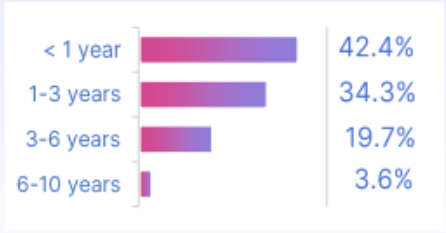
NEXTFLOW USERS ARE GLOBAL

31.1%	4.0%	2.0%	1.4%
14.7%	3.0%	1.6%	1.4%
10.2%	2.8%	1.6%	1.2%
8.0%	2.6%	1.6%	1.0%

USERS BY INDUSTRY

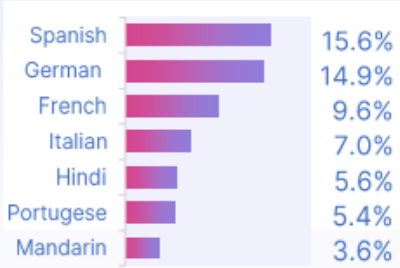


RAPID ADOPTION



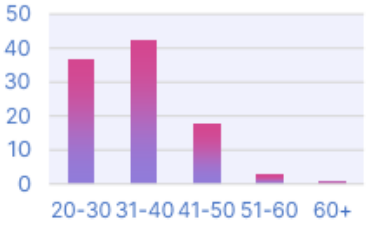
42% of Nextflow users are new in the last year

LANGUAGES SPOKEN



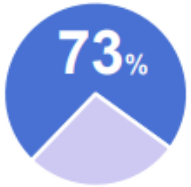
In addition to English

AGE OF NEXTFLOW USERS



79% of Nextflow users are between 20 and 40 years old

GENDER



Most Nextflow users are male

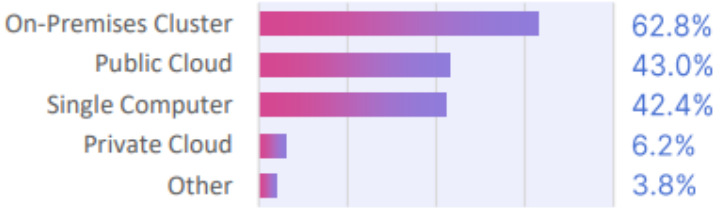
HAPPY USERS

99% user satisfaction rate

94% would recommend Nextflow to a colleague

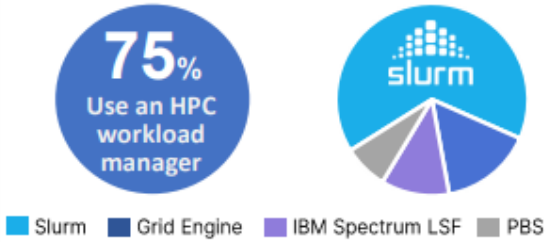
Some cool stats - 2023

WHERE DO YOU RUN NEXTFLOW?



On-premises HPC clusters remain the dominant runtime platform for Nextflow

SLURM RULES THE *HPC ROOST*

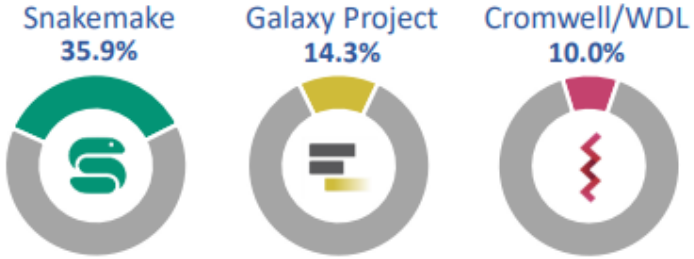


Slurm is used to manage **66%** of all on-premises Nextflow clusters

WORKLOADS SHIFT TO THE CLOUD

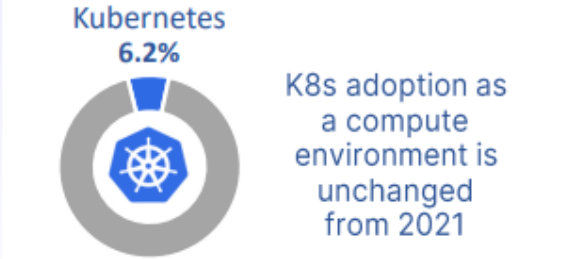
Among Nextflow users, the use of public cloud is up **20%** since 2022[1]

NEXTFLOW USERS ALSO RUN:



Other tools used by Nextflow users include CWL, Airflow, and Prefect

LIMITED KUBERNETES ADOPTION



Amazon EKS, Google GKE, Azure AKS, OpenShift, Rancher, etc.

PREFERRED CLOUDS

Among users planning a move to the cloud within 2 years, most plan to use AWS

Public cloud adoption is higher in the private sector where **80%** use public cloud[2]

[1] In March 2022, 35.7% of 384 survey respondents indicated they were using the public cloud. In March 2023, 502 respondents, 43% are using public cloud – a 20% increase
[2] Includes 170 of 401 respondents who self-identified as being in Biotech startups, Pharmaceutical companies, or healthcare/diagnostics/clinical care. Of these, 136 use public cloud.

Making your Bioinformatics analysis reproducible, scalable and portable is a trend with no turning back!

Use this great power wisely!

References

- Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017 Apr 11;35(4):316-319. doi: 10.1038/nbt.3820.
- Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, Garcia MU, Di Tommaso P, Nahnsen S. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020 Mar;38(3):276-278. doi: 10.1038/s41587-020-0439-x.
- Kim YM, Poline JB, Dumas G. Experimenting with reproducibility: a case study of robustness in bioinformatics. Gigascience. 2018 Jul 1;7(7):giy077. doi: 10.1093/gigascience/giy077.
- Mangul S, Mosqueiro T, Abdill RJ, Duong D, Mitchell K, Sarwal V, Hill B, Brito J, Littman RJ, Statz B, Lam AK, Dayama G, Grieneisen L, Martin LS, Flint J, Eskin E, Blekhman R. Challenges and recommendations to improve the installability and archival stability of omics computational tools. PLoS Biol. 2019 Jun 20;17(6):e3000333. doi: 10.1371/journal.pbio.3000333.


IT WORKED
ON MY COMPUTER
BUT I DON'T
HAVE YOUR COMPUTER
imgflip.com

Thank you!