

Big Data analytics mit Nextflow und nf-core



Alexander Peltzer

<https://nf-co.re>

Slides: <http://bit.ly/nf-core-tuebix>

Challenges: Big Data

- Data in computational (biology, physics, chemistry ...) is
 - big (PB scale)
 - diverse (e.g. sequencing, proteomics, ...)
 - erroneous (e.g. contains sequencing errors)

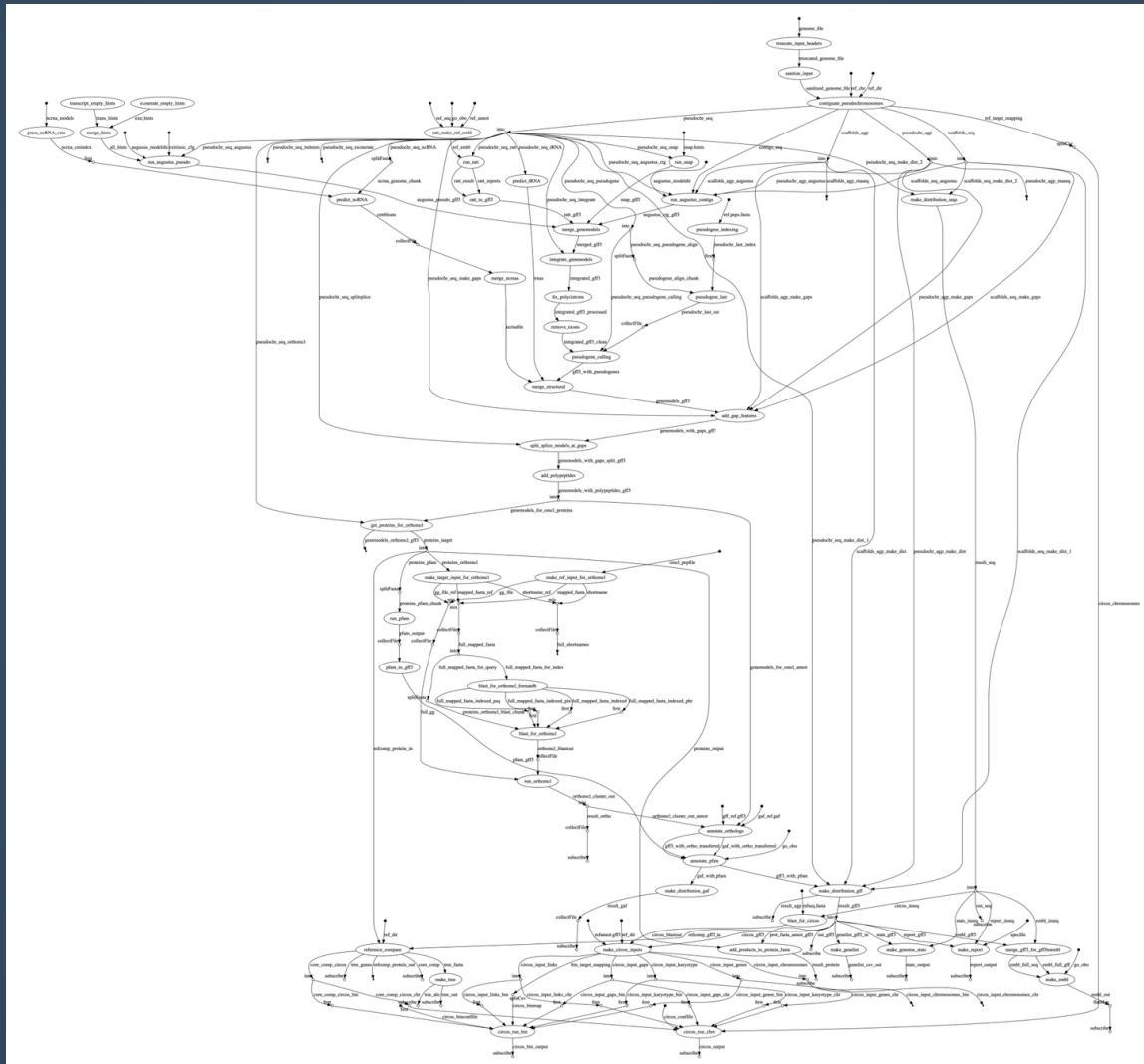
We need methods and tools to analyze such data!

Challenges: Software dependencies

Workflows / Pipelines consist of

- different tools
- dozens of individual methods

Complex dependency trees and
configuration requirements!





Challenges: Software dependencies

Workflows / Pipelines consist of

- different tools
- dozens of individual methods

"[...] of the tools selected for our comprehensive and systematic usability test, **49% were deemed "difficult to install," and 28% of the tools failed to be installed [...]."**

- *Mangul et al, biorxiv preprint, October 25 2018*

Challenges: Reproducibility

- Large-scale projects more common today
 - 1,000 Genomes Project
 - 100,000 Genomes Project UK
 - (EU 1,000,000 Genomes Project)
- Reproduce results with older data / integrate with newer data

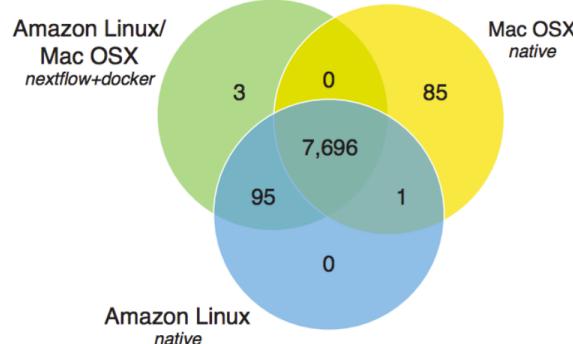
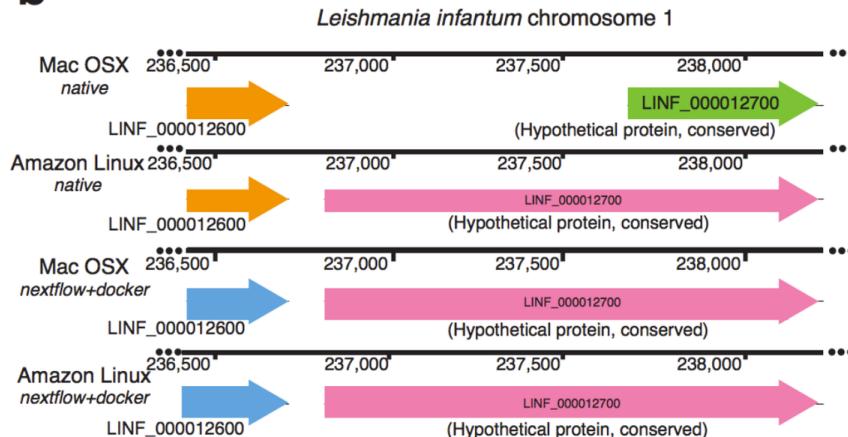
**Many paper results are not reproducible at all
or require a lot of effort !**

Challenges: Environmental stability

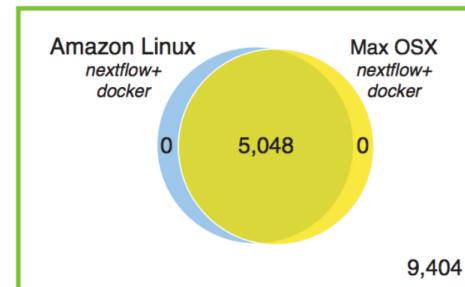
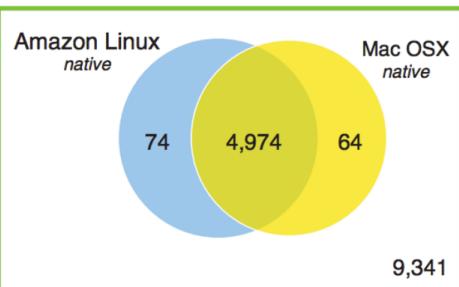
- Portability and stability of code between different OS should be ensured
- Are results different? Yes, they are ...

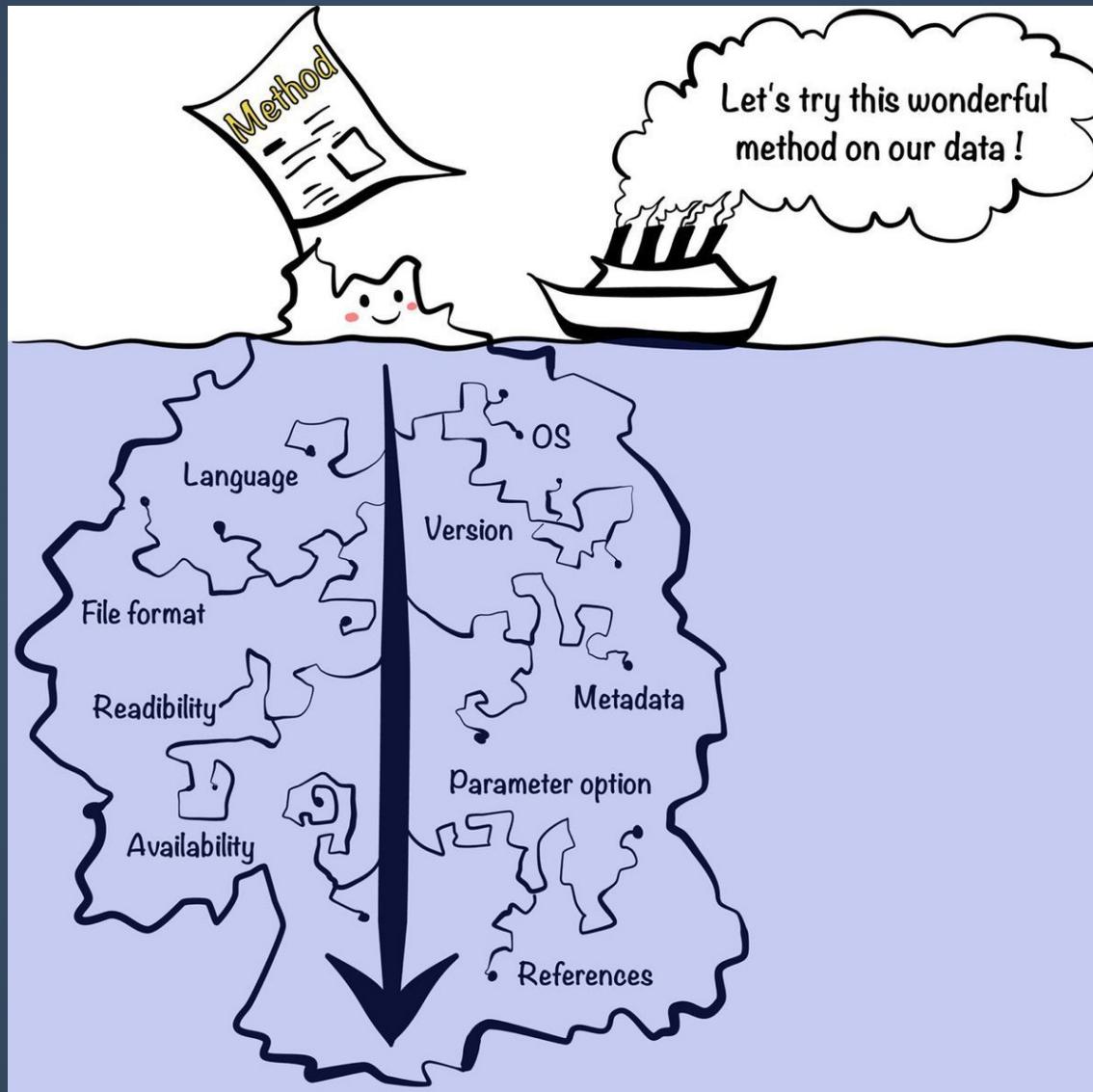


Challenges: Software dependencies

aGene annotation of *Leishmania infantum* with Companion**b****c**

Transcript quantification and differential expression with Kallisto and Sleuth





All of this is a requirement for:

FAIR Data Sharing

- **Findable**
- **Accessible**
- **Interoperable**
- **Reproducible**

Strong move towards FAIR-Sharing, FAIR-Processing in scientific disciplines!

Nextflow

- Custom DSL (domain-specific language) for
 - fast prototyping
 - enabling task composition
 - easy parallelization
- Self-contained: Containerize tasks (e.g. with Docker)
- Isolation of dependencies: Keep container - rerun analysis at any point!



Automated



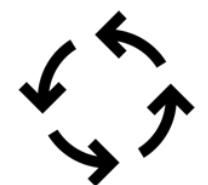
Parallelizable



Reliable



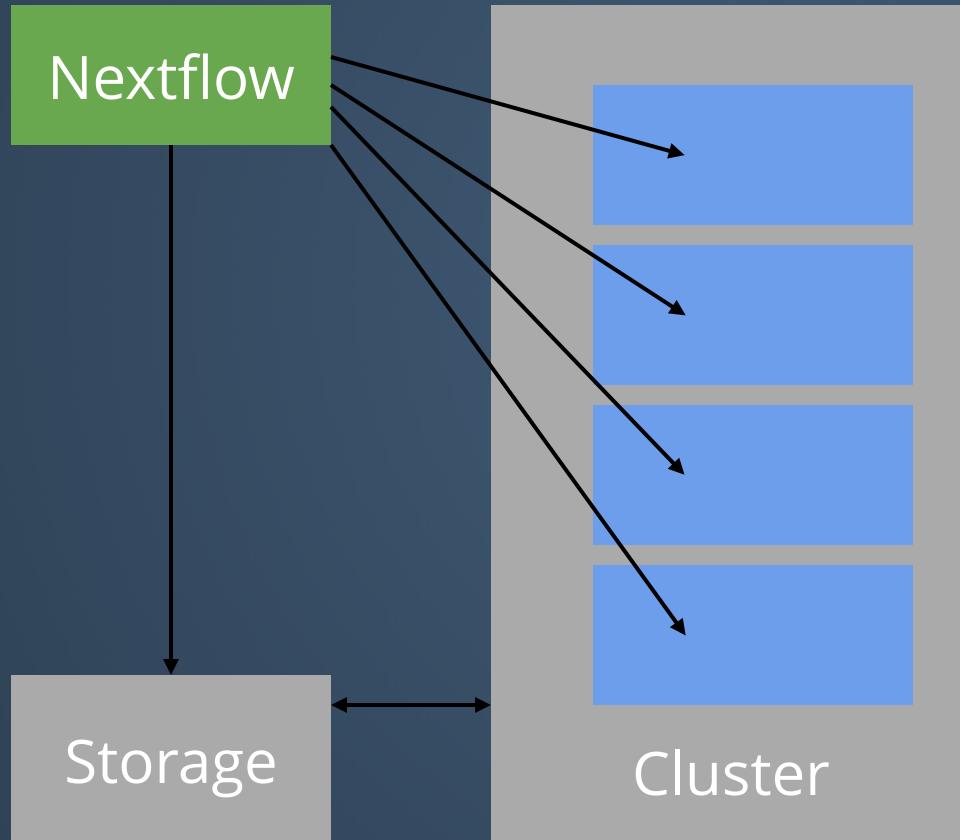
Easy for others to run



Reproducible results

nextflow

Nextflow: Centralised Orchestration



- Submit jobs to cluster nodes
- Store data on shared storage

Nextflow: Executor abstraction

```
#Run script locally
process.executor = 'local'

#Run script on PBS/Torque
process.executor = 'pbs'

#Run script on Kubernetes cluster
process.executor = 'k8s'

#Run script on AWS Batch
process.executor = 'awsbatch'
```

=> Improves code portability

nextflow

```
#!/usr/bin/env nextflow
input = Channel.fromFilePairs(params.reads)

process fastqc {
    input:
    file reads from input

    output:
    file "*_fastqc.{zip,html}" into results

    script:
    .....
    fastqc -q $reads
    .....

}
```

nextflow

```
#!/usr/bin/env nextflow
input = Channel.fromFilePairs(params.reads)

process fastqc {
    input:
    file reads from input

    output:
    file "*_fastqc.{zip,html}" into results

    script:
    """
    fastqc -q $reads
    """
}
```

```
nextflow run main.nf --reads "*.fastq.gz"
```



Singularity

BIOCONDA®



- Community effort to collect production ready analysis pipelines
- Save time in development, more testing, more updates
- <https://nf-co.re>



Phil Ewels



Alex Peltzer



Harshil Patel



Maxime Garcia



Sven Fillinger



Andreas Wilm

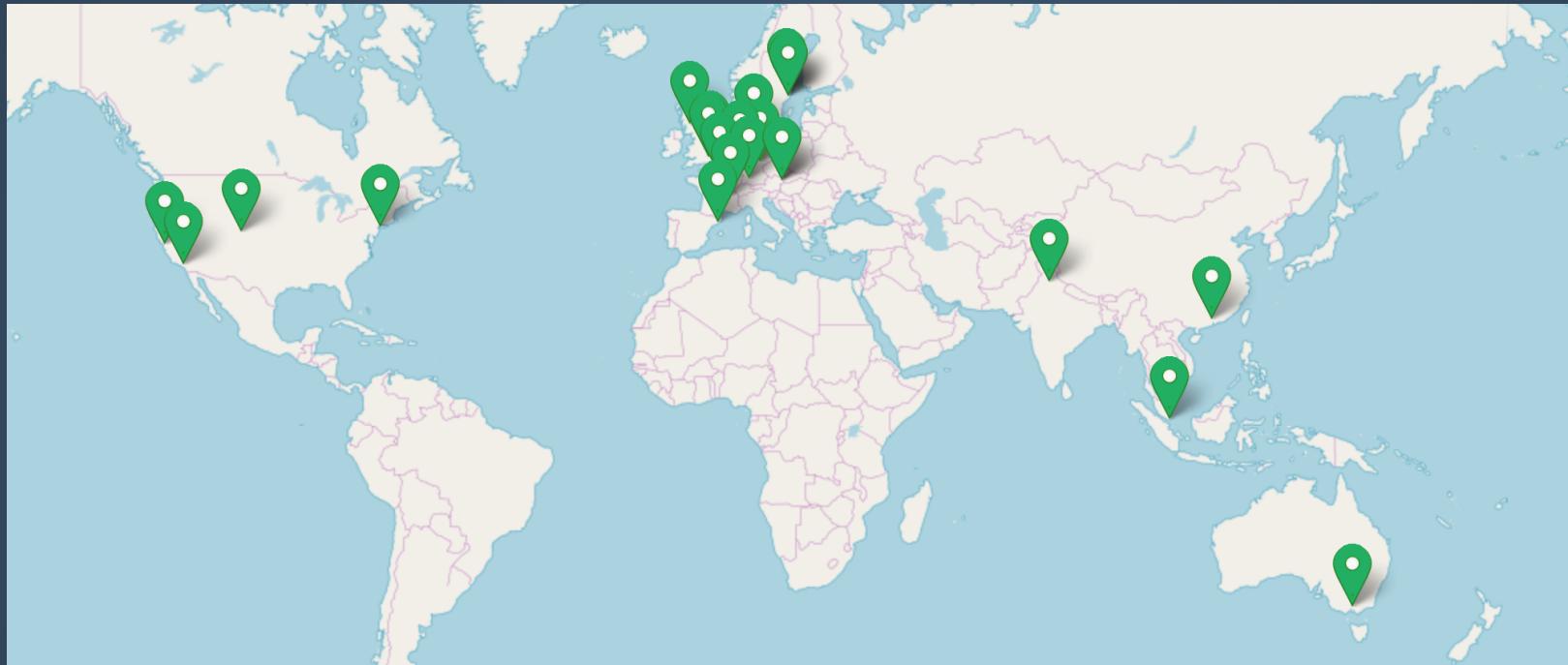
+ many others!



- Initially supported by SciLifeLab, QBiC and A*Star Genome Institute Singapore

SciLifeLab







nextflow



conda-forge

BIOCONDA®



Twitter

nf-core



Slack



Google
Group



GitHub

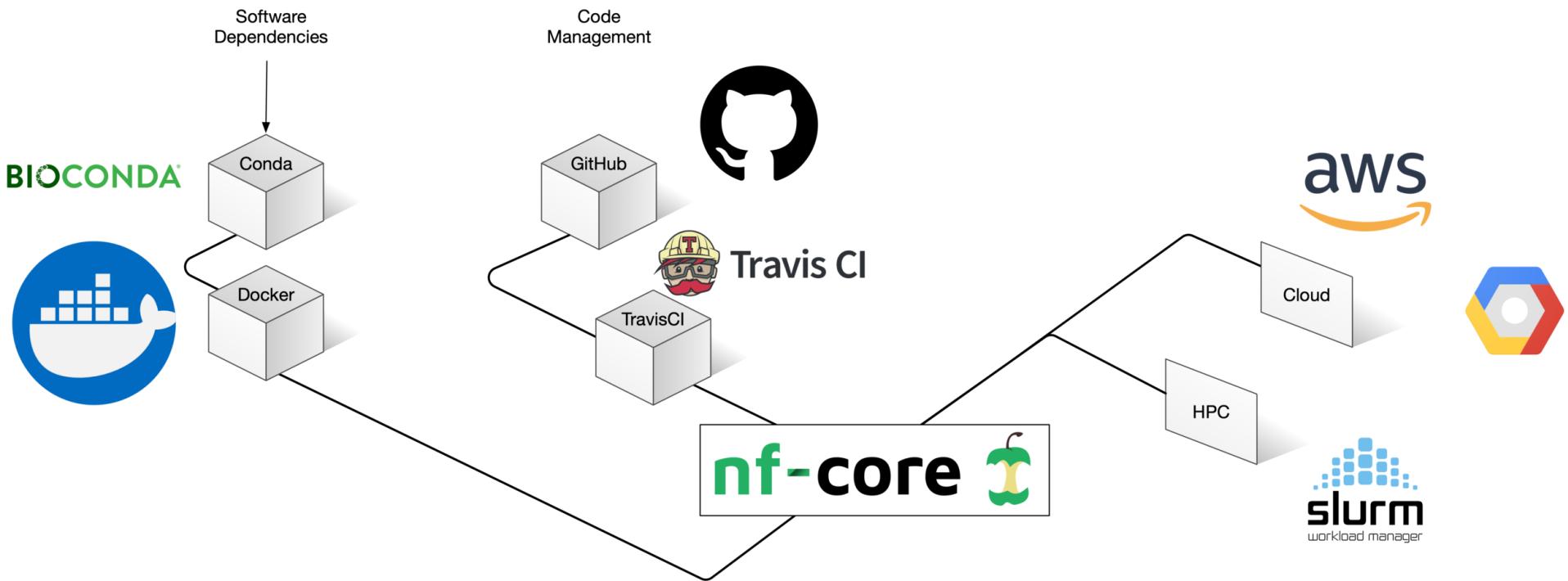


nf-co.re



All pipelines adhere to requirements

- Nextflow based
- MIT license
- Software bundled in Docker / Singularity
- Continuous integration testing (e.g. Travis CI)
- Stable release tags
- Common pipeline usage and structure
- Software bundled in bioconda





Need help?

- nf-core Tools: To get a skeleton for new pipelines
 - Synchronization of best-practices across pipelines!
 - Linting app: To check what conforms with nf-co.re
- Slack: To communicate with the community!



nf-core create: Create a pipeline from scratch

nf-core lint: Check pipeline conforms with best-practices

- 15 stable
- 15 in development

<https://nf-co.re/pipelines>

nf-core/rnafusion ✓ ⭐ 11
fusion fusion-genes gene-fusion rna rna-seq
RNA-seq analysis pipeline for detection gene-fusions
Version 1.0 Published 1 week ago

nf-core/mhcquant ✓ ⭐ 4
mass-spectrometry peptides
Identify and quantify peptides from mass spectrometry raw data
Version 1.2.4 Published 3 weeks ago

nf-core/hlatyping ✓ ⭐ 1
dna hla hla-typing immunology optotype personalized-medicine rna
Precision HLA typing from next-generation sequencing data
Version 1.1.3 Published 3 weeks ago

nf-core/methylseq ✓ ⭐ 21
bisulfite-sequencing dna-methylation methyl-seq
Methylation (Bisulfite-Sequencing) analysis pipeline using Bismark or bwa-meth + MethylDackel
Version 1.3 Published 3 weeks ago

nf-core/eager ✓ ⭐ 12
adna ancientdna pathogen-genomics population-genetics
A fully reproducible and state of the art ancient DNA analysis pipeline.
Version 2.0.5 Published 4 weeks ago

nf-core/ampliseq ✓ ⭐ 20
16s amplicon-sequencing docker singularity
16S rRNA amplicon sequencing analysis workflow using QIIME2
Version 1.0.0 Published 2 months ago

nf-core/rnaseq ✓ ⭐ 77
rna rna-seq
RNA sequencing analysis pipeline using STAR or HISAT2, with gene counts and quality control
Version 1.2 Published 2 months ago

nf-core/deepvariant ✓ ⭐ 13
deep-variant dna google variant-calling
Google's DeepVariant variant caller as a Nextflow pipeline
Version 1.0 Published 3 months ago

nf-core/ddamsproteomics ⚠️ ⭐ 2
Quant proteomics as practiced at Lehtiö lab for NF-core
No releases yet

nf-core/epitopeprediction ⚠️ ⭐ 2
A bioinformatics best-practice analysis pipeline for epitope prediction and annotation
No releases yet

nf-core/exoseq ⚠️ ⭐ 8
exome exome-sequencing genomics variant-calling
Exome Sequencing analysis pipeline (Work in progress)
No releases yet

nf-core/atacseq ⚠️ ⭐ 5
atac-seq chromatin-accessibility
UNDER CONSTRUCTION: ATAC-seq peak-calling and differential analysis pipeline.
No releases yet

nf-core/lncpipe ⚠️ ⭐ 9
differential-expression lncrna rna-seq-analysis transcriptome
UNDER DEVELOPMENT--- A Nextflow-based pipeline for comprehensive analyses of long non-coding RNAs from RNA-seq datasets

nf-core/mag ⚠️ ⭐ 6
annotation assembly binning bioinformatics metagenomics
Assembly and binning of metagenomes



Comes with interactive reports!

https://multiqc.info/examples/wgs/multiqc_report.html

Comes with proper documentation!

Pipeline overview

The pipeline is built using [Nextflow](#) and processes data using the following steps:

- [FastQC](#) - read quality control
- [TrimGalore](#) - adapter trimming
- [STAR](#) - alignment
- [RSeQC](#) - RNA quality control metrics
 - [BAM stat](#)
 - [Infer experiment](#)
 - [Junction saturation](#)
 - [RPKM saturation](#)
 - [Read duplication](#)
 - [Inner distance](#)
 - [Gene body coverage](#)
 - [Read distribution](#)
 - [Junction annotation](#)
- [dupRadar](#) - technical / biological read duplication
- [Preseq](#) - library complexity
- [featureCounts](#) - gene counts, biotype counts, rRNA estimation.
- [StringTie](#) - FPKMs for genes and transcripts
- [Sample_correlation](#) - create MDS plot and sample pairwise distance heatmap / dendrogram
- [MultiQC](#) - aggregate report, describing results of the whole pipeline

... and a lot more!

Documentation

Extensive documentation covering installation, usage and description of output files ensures that you won't be left in the dark.



CI Testing

Every time a change is made to the pipeline code, nf-core pipelines use continuous-integration testing to ensure that nothing has broken.



Travis CI

Stable Releases

nf-core pipelines use GitHub releases to tag stable versions of the code and software, making pipeline runs totally reproducible.



Docker

Software dependencies are always available in a bundled docker container, which Nextflow can automatically download from dockerhub.



Singularity

If you're not able to use Docker, built-in support for Singularity can solve your HPC container problems. These are built from the docker containers.



Bioconda

Where possible, pipelines come with a bioconda environment file, allowing you to set up a new environment for the pipeline in a single command.



Acknowledgements

NF-Core Team

Phil Ewels (SciLifeLab, Stockholm)

Maxime Garcia (SciLifeLab, Stockholm)

Harshil Patel (The Francis Crick Institute, London)

Sven Fillinger (QBiC/Tü)

Paolo di Tommaso (CRG, Barcelona)

Evan Floden (CRG, Barcelona)