Automation, containerisation and pipelinisation for bioinformatics

easier for you to run it, easier for others to reproduce it

FY2025 Current Trends in Bioinformatics Course

バイオインフォマティクス特講

Course presented November 10th, 2025.

Some general illustrations removed to be sure with copyright compliance

Charles Plessy, Senior Staff Scientist

Okinawa Institute of Science and Technology Graduate University

About me

Where I am from

- Graduated from Strasbourg, France.
- Made transgenic zebrafish for my Ph. D.
- Learned bioinformatics by myself on Linux computers.
- Moved to Japan in 2004.

Where did I work in Japan? What did I do?

RIKEN: 2004–2018

- Transcriptome sequencing
- Single cells
- Development of new technologies
- Compute cluster power user
- **Pipeline user**

OIST: 2018–2025

- Genome sequencing
- Development of new methods
- Bioinformatics user group founder
- Compte cluster power user (always)
- **Pipeline developer**

I would not be talking to you if there were no Free software

- A software is Free (libre, open source) when you have the rights to use, study, modify and redistribute it.
- I would not have reached the level I have today without Free Software.
- Free software often form communities based on a topic or a geographical area.
- Community is the key and AI can not replace it.

debian med

東京エリアDebian勉強会

Why this lecture?

Two ways to approach this lecture

→ Learn to run computations

- Pay attention to the details
- Try the toy examples in your computer
- Remember the keywords for AI search

→ Learn to understand bioinformaticians

- Get the big picture
- Remember what the challenges are
- Think how you can remove or prevent roadblocks upfront

Reproducibility crisis: 再現性の危機

Have you heard about it?

Unreproducible research wastes time, money, health of volunteers, and erodes trust in science.

- Behavioural causes: 疑わしい研究慣習 (questionable research practices)

- Technical causes: Did you know that your software decays with time?

Your "future self" thanks you

The work you will do for reproducibility is an investment for society and the future...

...but more often that you would think, it also pays off directly to you.

From a take-home messages a previous lecture (Takahiro Suzuki, 2024 / 10 / 07)

Catch-up cutting-edge technologies

New Technologies open novel Biology

Commoditization of NGS

From the Genome center to a Lab (or Outsourcing)
Insufficient number of bioinformaticians

How can we solve that problem in an ageing country?

Social/Medical implementation of Omics

Genomics medicines
Diagnosis
Personalized medicine

What you will learn today

Reproductible science is a duty, and it also makes you more efficient.

There are tools to make reproductible **bioinformatics**.

These **tools** also empower you to design bigger research projects that you complete by yourself.

nf-core is a community of **people** that builds pipelines for bioinformatics projects, and I will show you how to use them.

Containers (コンテナ)

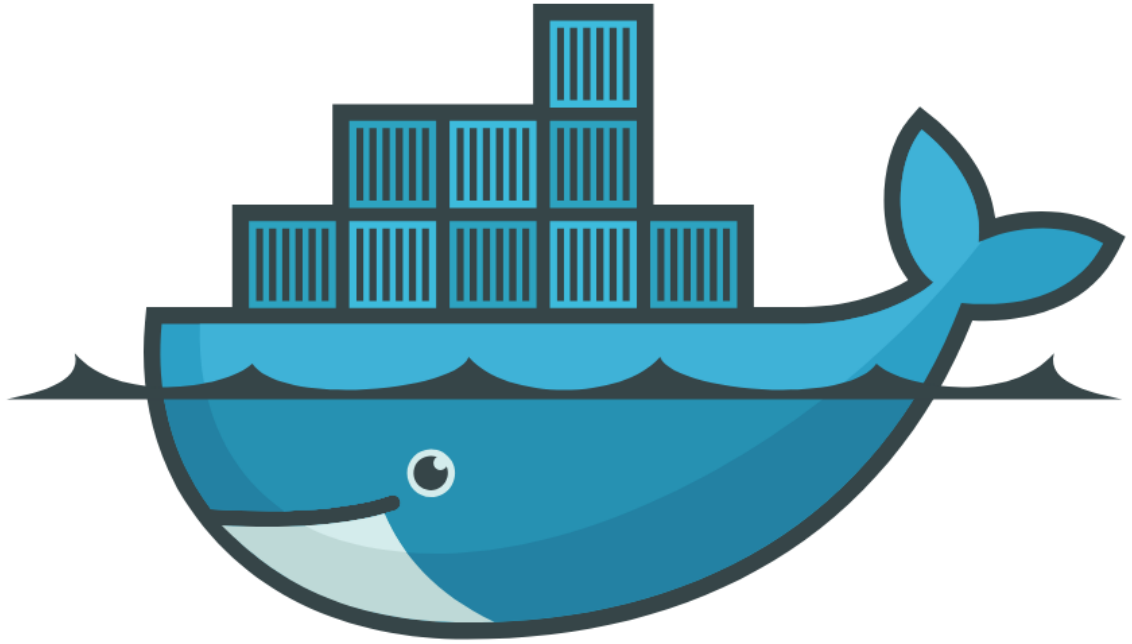
Containers

- Containers enabled globalization by standardizing size.
- Shippers handle containers, not individual product shapes.
- Locked end-to-end, they improve security.

Containers

- Containers ship software in a standard way.
- They run anywhere, regardless of host system (Java and Go aimed for this).
- They improve security by isolating the system, though exceptions are needed for real work.

Docker is an iconic container system (ドッカーコンテナ), widely used to standardize and ship software.



docker

One software, one container

- Just
  - Download ("pull")
  - Use

- No need to install or configure the container,

- but you still need to install Docker...

As you can imagine, reality is a little bit more complicated…

Industry and academia: different problems, different solutions

Docker

- By default, can not access your files
- Needs Administrator privileges
- Many containers are built on other containers
- Designed for desktops and cloud

Singularity

- By default, can access all your files
- Runs as ordinary user
- Most containers are custom-built
- Designed for HPC environments

Podman ??
Apptainer??

- Podman is a compatible Docker replacement.
- Apptainer is a fork of Singularity.
- I use both, but to keep things simpler, today I only said Docker / Singularity

podman

Regulatory agencies (EU, FDA, ...) want more guarantees about software provenance and security.

- If you become in charge of purchase, security or compliance, you will hear about the:

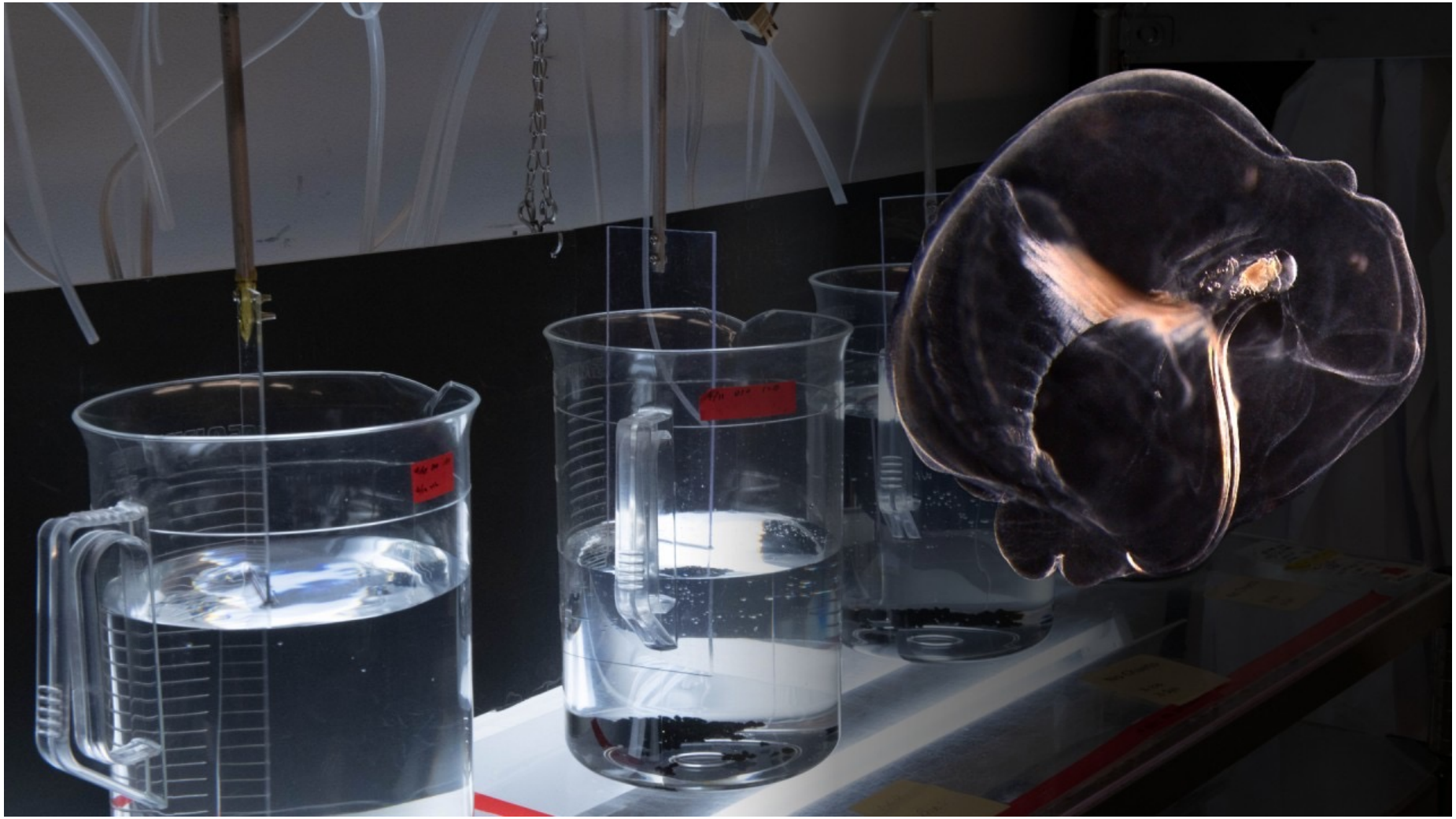
  Software Bill Of Materials
  (ソフトウェア部品表)

  It allows the trust of what is inside the software black box to pass through each step of the supply chain.

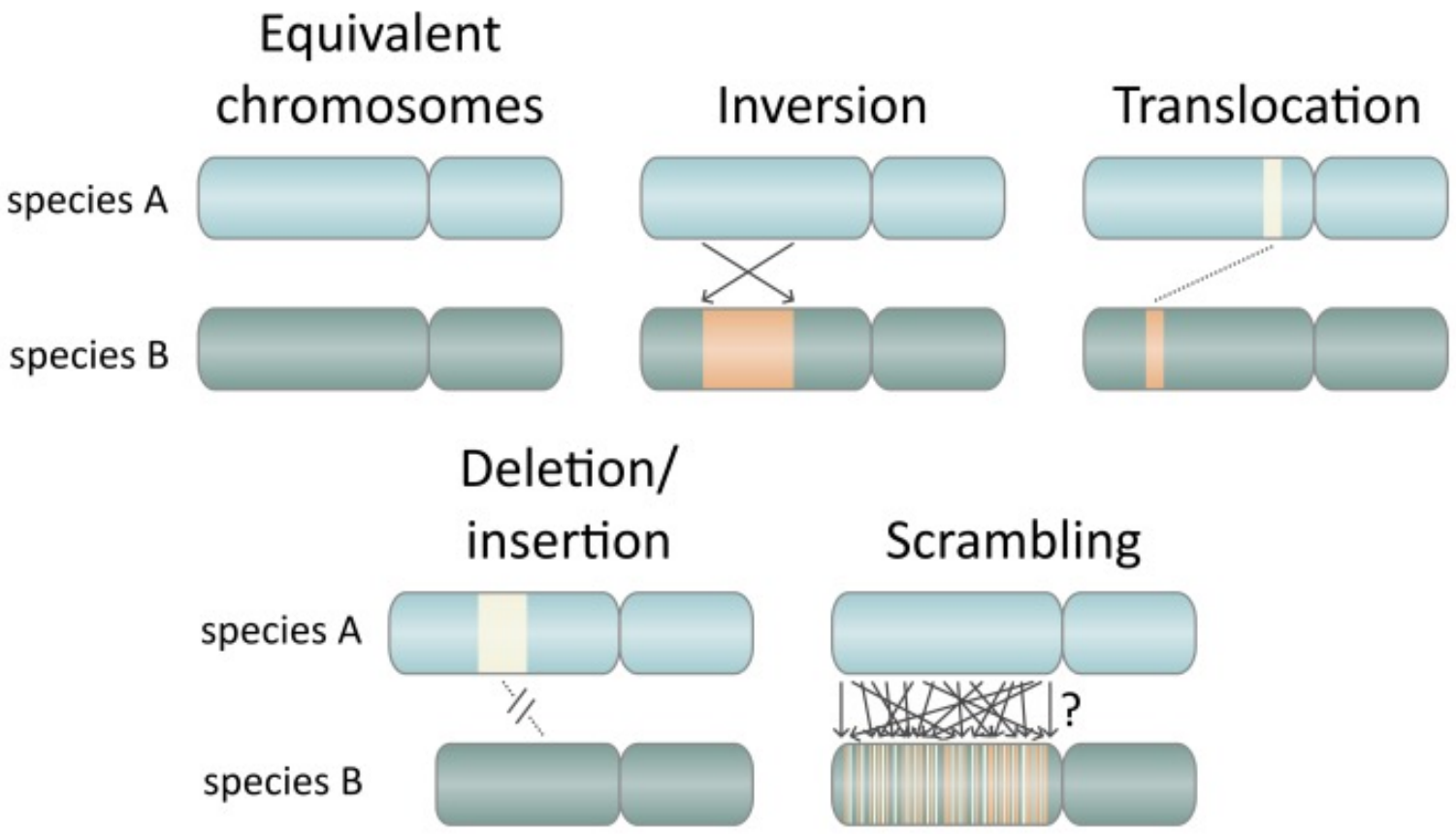
My own reproductible research

I study the genome of the zooplankton *Oikopleura dioica*

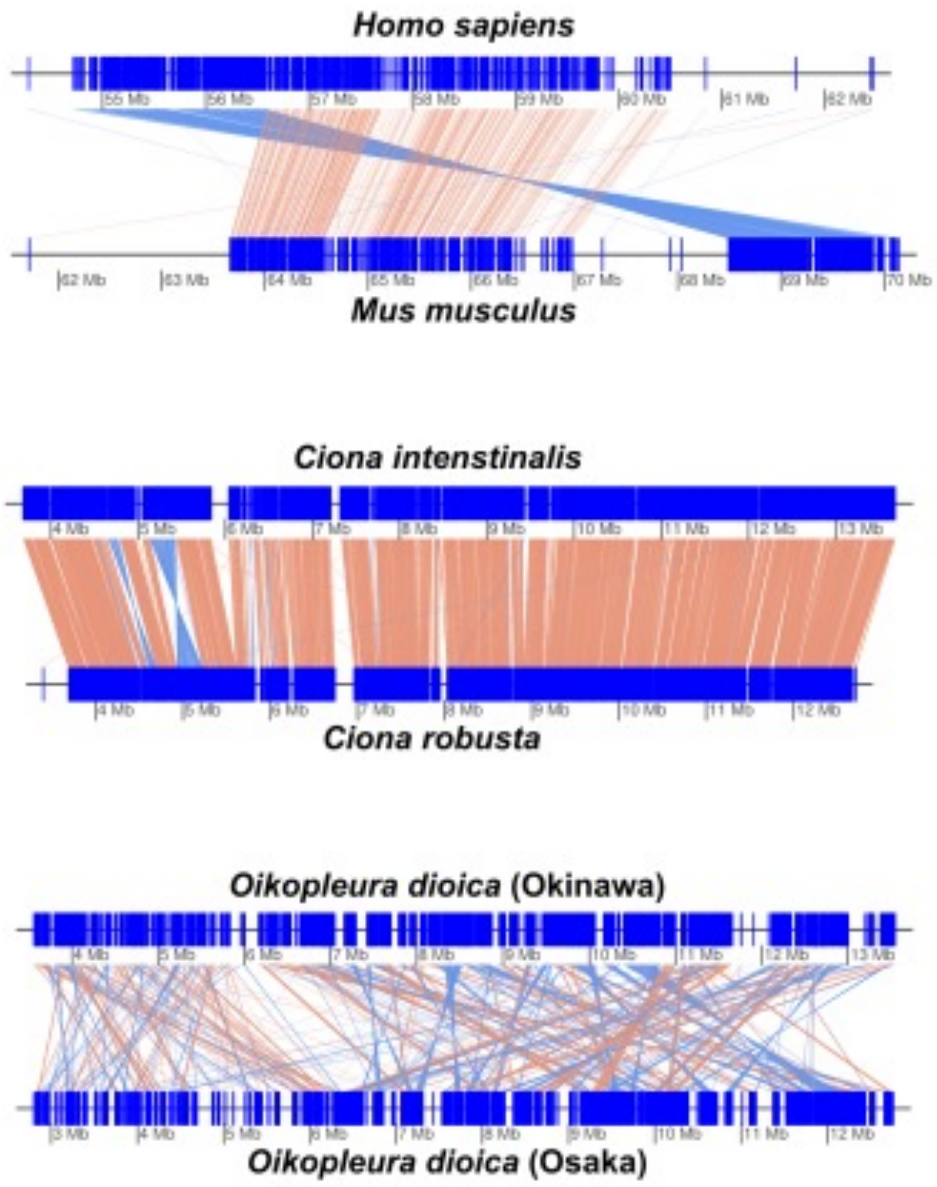
- Small genome
- Fast life cycle
- Unexpected cryptic species



We searched for structural variations, by aligning genomes together.



We discovered ***genome scrambling***



Credit for research digest: Adrian Skov, OIST press team

To help other researchers to align genomes the way I did, I created a nf-core pipeline.

nf-core

On this page

nf-core/pairgenomealign

Pairwise genome comparison pipeline using the LAST software to align a list of query genomes to a target genome, and plot the results

comparative-genomics | dot-plot | genomics | last | pairwise-alignment | synteny | whole-genome-alignment

Launch version 2.2.1

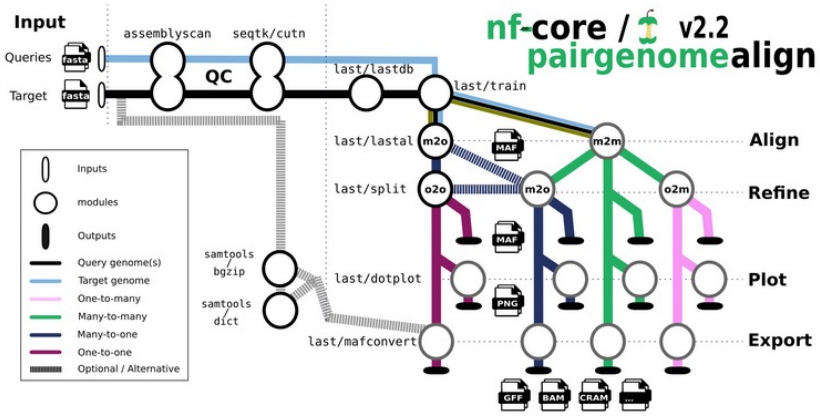
https://github.com/nf-core/pairgenomealign

Introduction

2.2.1

Introduction

nf-core/pairgenomealign is a bioinformatics pipeline that aligns one or more *query* genomes to a *target* genome, and plots pairwise representations.



The main steps of the pipeline are:

1. Genome QC (`assembly-scan`).
2. Genome indexing (`lastdb`).
3. Genome pairwise alignments (`lastal`).
4. Alignment plotting (`last-dotplot`).
5. Alignment export to various formats with `maf-convert` , plus `Samtools` for SAM/BAM/CRAM.

Software pipelines

Beniimo taruto pipeline video on:
https://www.youtube.com/watch?v=QvnBdIQZFUU

What are Nextflow pipelines (パイプライン)? Why do you need them?

- **Nextflow** downloads, installs and runs containers for you.
- To run a Nextflow pipeline you only need to download a single file (and install Java).
- On your HPC cluster, Java and a container system are usually pre-installed.
- **nf-core** is a user and bioinformatics pipeline community.

*With a **tool** and a **community**, you can do reproducible bioinformatics.*

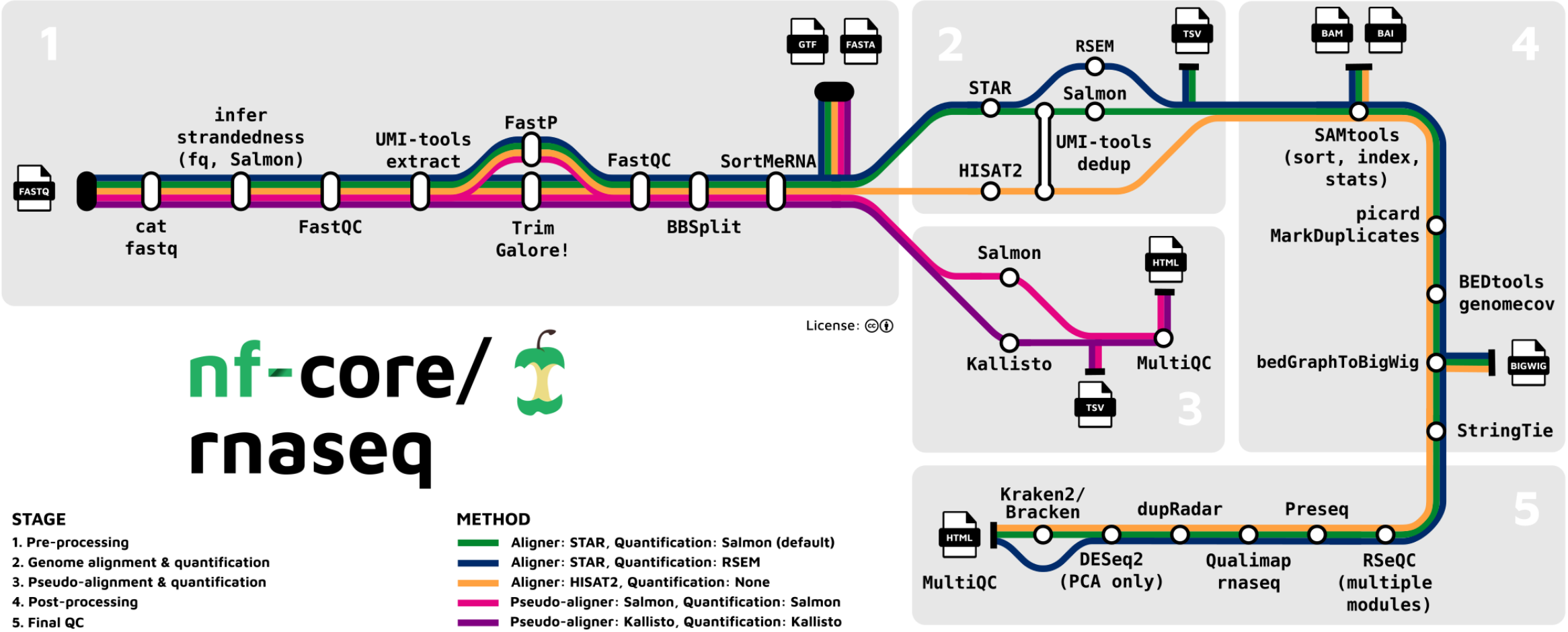
Let's run a simple pipeline:

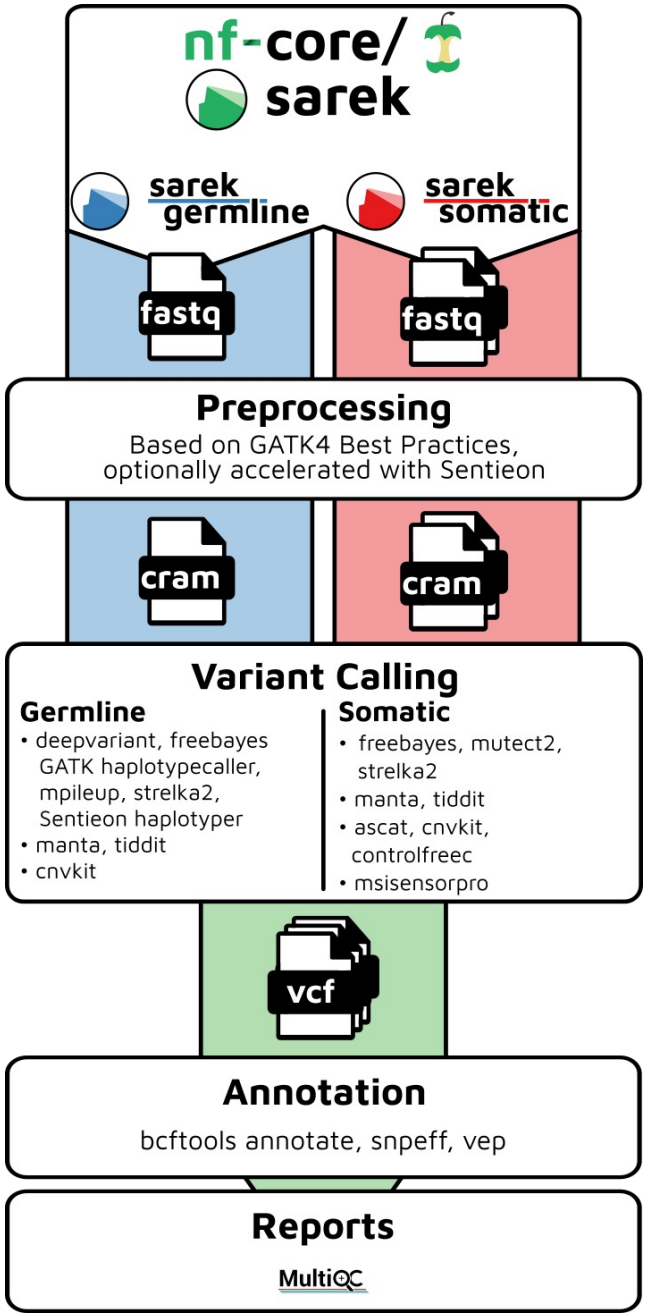
nf-core/demo is a simple nf-core style bioinformatics pipeline for workshops and demonstrations. It was created using the nf-core template and is designed to run quickly using small test data files.

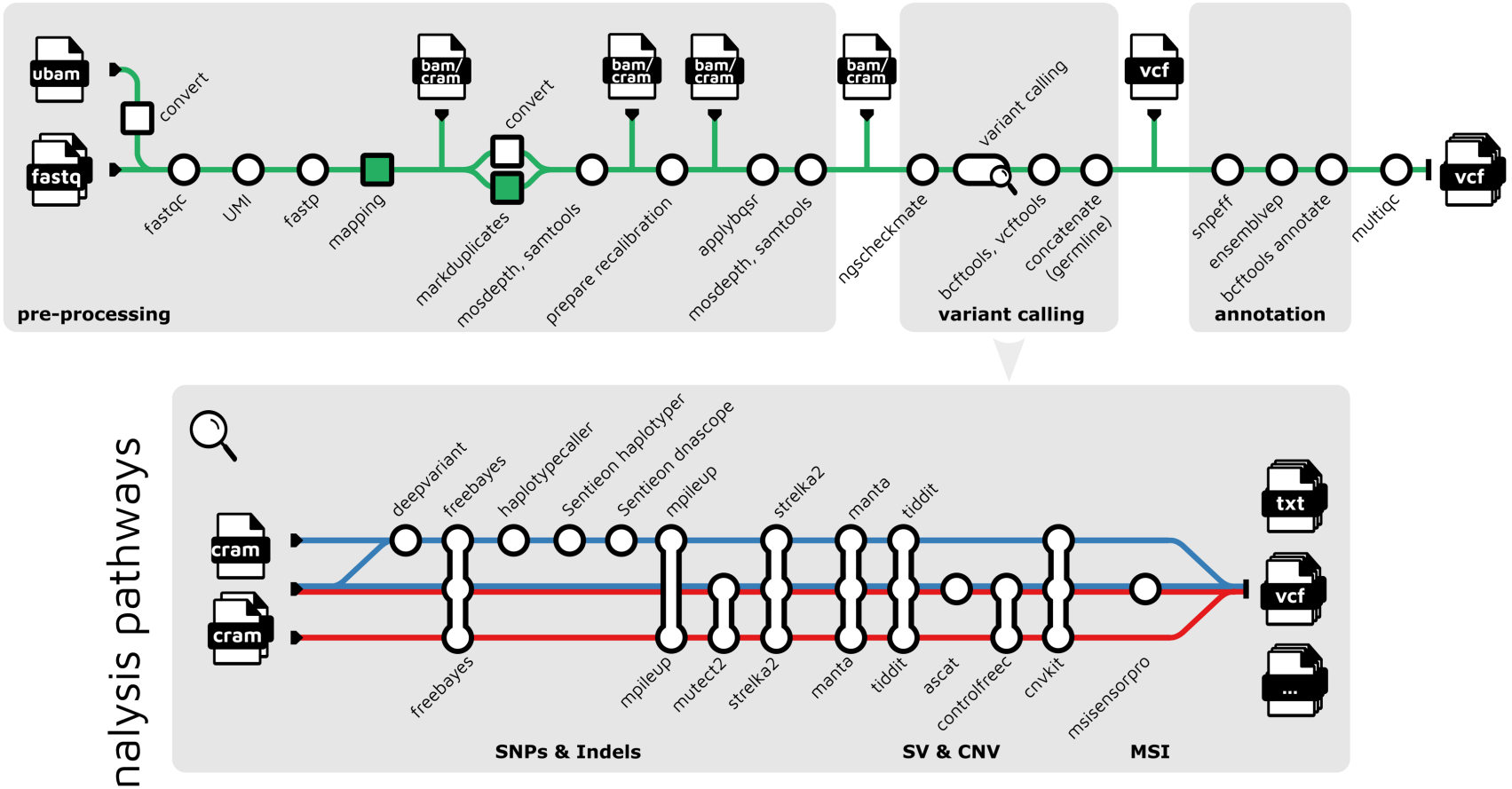
nf-core/
demo


```
$ nextflow run -profile test,singularity nf-core/demo --outdir results
```

Real pipelines are more complicated, but that gives you freedom!







nf-core/sarek Example analysis pathways

- Mandatory
- Optional
- Optionally Sentieon accelerated

- Core workflow
- Germline variant calling
- Tumor only variant calling
- Tumor-normal pair variant calling

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

The harders part is always to plug the data in the analysis. Pipelines help you to stay organised.

Full samplesheet

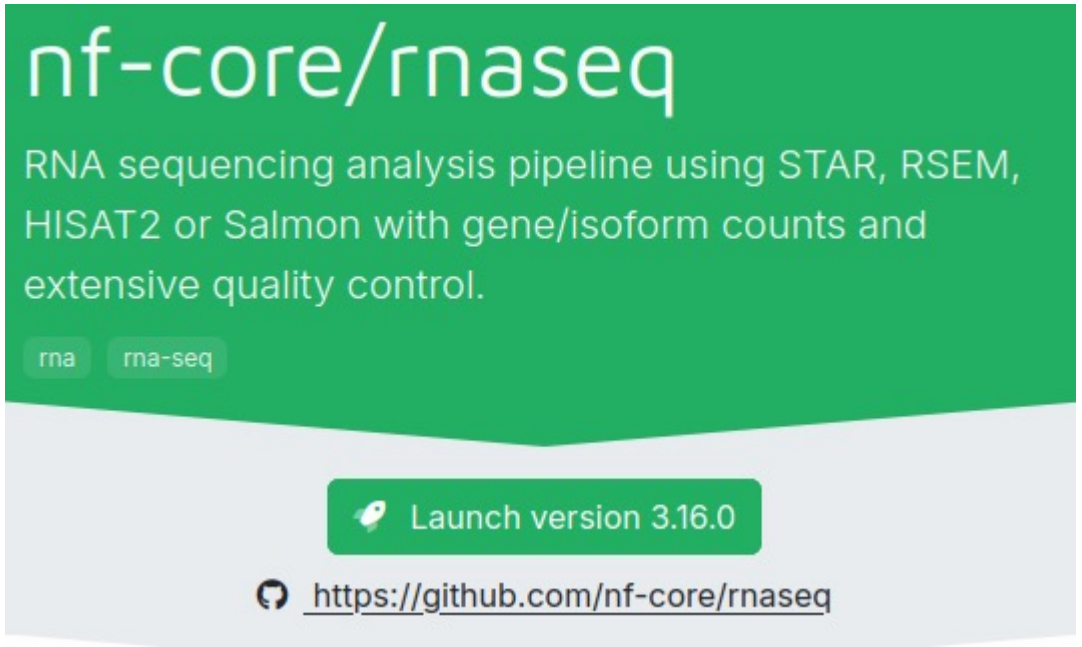
The pipeline will auto-detect whether a sample is single- or paired-end using the information provided in the samplesheet. The samplesheet can have as many columns as you desire, however, there is a strict requirement for the first 4 columns to match those defined in the table below.

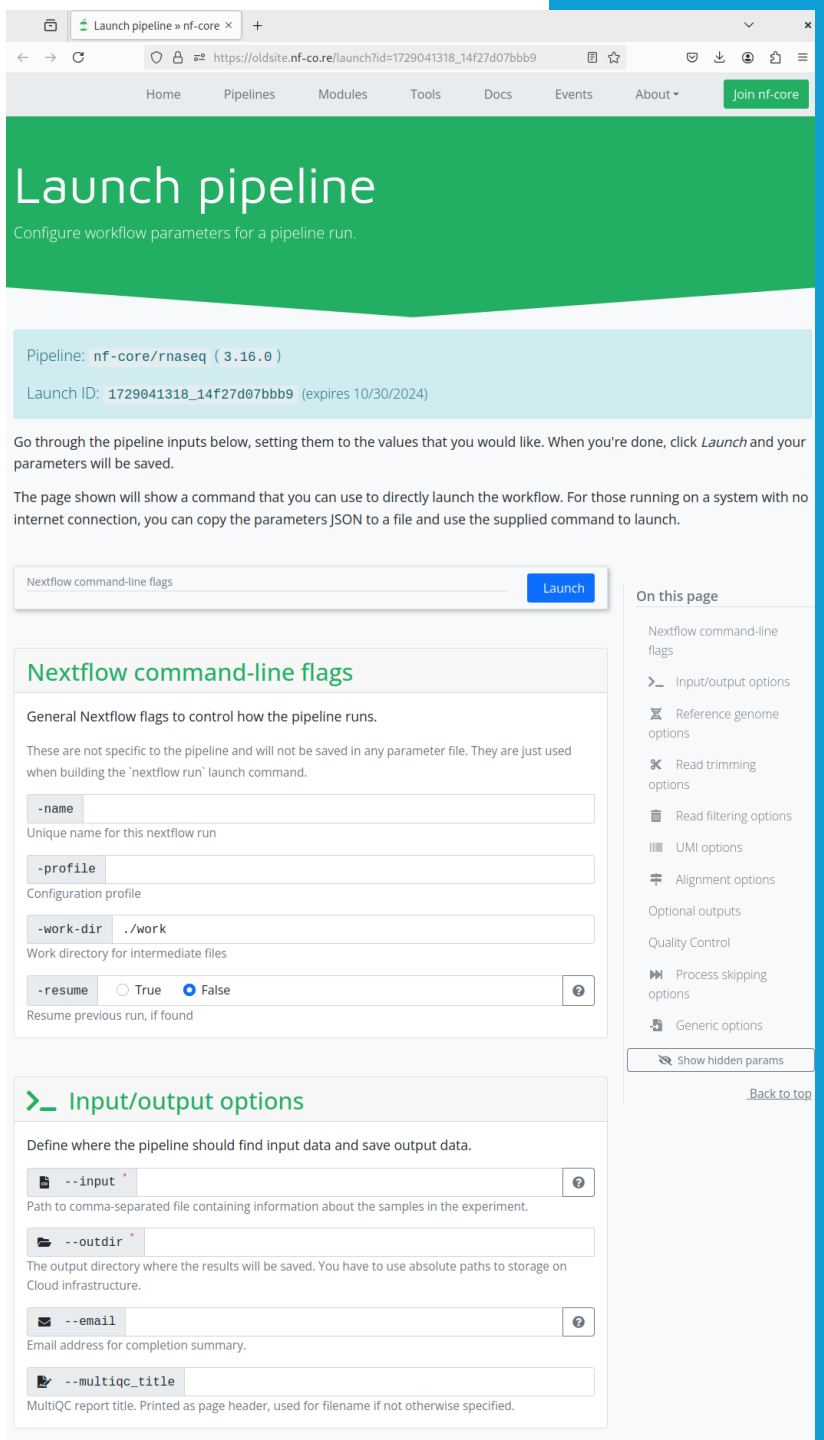
A final samplesheet file consisting of both single- and paired-end data may look something like the one below. This is for 6 samples, where `TREATMENT_REP3` has been sequenced twice.

samplesheet.csv

```csv
sample,fastq_1,fastq_2,strandedness
CONTROL_REP1,AEG588A1_S1_L002_R1_001.fastq.gz,AEG588A1_S1_L002_R2_001.fastq.gz,forward
CONTROL_REP2,AEG588A2_S2_L002_R1_001.fastq.gz,AEG588A2_S2_L002_R2_001.fastq.gz,forward
CONTROL_REP3,AEG588A3_S3_L002_R1_001.fastq.gz,AEG588A3_S3_L002_R2_001.fastq.gz,forward
TREATMENT_REP1,AEG588A4_S4_L003_R1_001.fastq.gz,,reverse
TREATMENT_REP2,AEG588A5_S5_L003_R1_001.fastq.gz,,reverse
TREATMENT_REP3,AEG588A6_S6_L003_R1_001.fastq.gz,,reverse
TREATMENT_REP3,AEG588A6_S6_L004_R1_001.fastq.gz,,reverse
```

nf-core pipelines are designed to be easy to start from the Web





The web interface tells you what to type on your terminal!

Or you can pay them to run your analysis in a cloud.

Launch parameters saved

Your workflow parameters are ready to go! Follow the instructions below for instructions on how to launch your pipeline:

If your system has an internet connection

The easiest way to launch this workflow is by using the `nf-core/tools` helper package.

Once installed (see documentation), simply run the following command and follow the prompts:

```
nf-core launch --id 1729041550_be6d0d15d58d
```

Launch using Seqera Platform

Clicking the button below will take you to the Seqera Platform launch page with all parameters set, ready for launch (requires a Seqera Platform account).

🚀 Seqera Platform ❯ Launch

Launching with no internet and without nf-core/tools

You can run this pipeline with just Nextflow installed by copying the JSON below to a file called `nf-params.json`:

```
{
    "input": "mySampleSheet.csv",
    "outdir": "myResults",
    "email": "charles@plessy.oist.jp"
}
```

Then, launch Nextflow with the following command:

```
nextflow run nf-core/rnaseq -r 3.16.0 -name YCU_course_2024 -profile oist -params-file nf-params.json
```

Sequera, the startup behind Nextflow, provides free AI support

(I wrote the prompt in English, translated it in Google, and used it to query Sequera AI)

Nextflow初心者です。HPCクラスターでnf-coreテストパイプラインを実行したいと考えています。管理者から、ノードあたり64CPU、512GBのSLURMキュー「compute」にアクセスできると言われました。JavaとSingularityはインストール済みです。Nextflowのインストールと設定、そしてnf-coreテストパイプラインの実行方法を教えてください。

Seqera AI Chat

seqera.io/ask-ai/chat-v2

seqera Platform Open Source Resources Solutions Company Seqera AI Pipelines Containers Log in Sign up

1. Nextflowのインストール

手順1: Javaの確認

まず、Javaがインストールされているか確認してください（Java 11+が必要）：

```
java -version
```

手順2: Nextflowのインストール

最もシンプルな方法は自己インストールパッケージを使用することです：

```
# Nextflowをダウンロード
curl -s https://get.nextflow.io | bash

# 実行可能にする
chmod +x nextflow

# パスが通った場所に移動
mkdir -p $HOME/.local/bin/
mv nextflow $HOME/.local/bin/

# PATHに追加（~/.bashrcに追加）
echo 'export PATH="$PATH:$HOME/.local/bin"' >> ~/.bashrc
source ~/.bashrc

# インストール確認
nextflow info
```

2. SLURM用のNextflow設定

HPCクラスター用の設定ファイル `nextflow.config` を作成します：

```
// nextflow.config
process {
    executor = 'slurm'
    queue = 'compute'

    // デフォルトリソース設定
    cpus = 4
    memory = '16 GB'
    time = '2h'

    // Singularityの設定
    container = 'default-container'
}

singularity {
    enabled = true
    autoMounts = true
    cacheDir = '$HOME/.singularity'
}

// nf-coreパイプライン用の推奨設定
params {
    max_cpus = 64
    max_memory = '512 GB'
    max_time = '240h'
}
```

3. 環境変数の設定

Singularityキャッシュディレクトリを設定します：

```
# ~/.bashrcに追加
export NXF_SINGULARITY_CACHEDIR="$HOME/.singularity"
export SINGULARITY_CACHEDIR="$HOME/.singularity"
export SINGULARITY_TMPDIR="/tmp"

# ディレクトリを作成
mkdir -p $HOME/.singularity

# 設定を反映
source ~/.bashrc
```

4. nf-core CLIツールのインストール（オプション）

nf-coreパイプラインを簡単に使用するため：

```
pip install nf-core
```

5. テストパイプラインの実行

簡単なテスト（推奨開始方法）

```
# Nextflowの基本テスト
nextflow run hello

# nf-coreテストパイプラインの実行
nextflow run nf-core/test-datasets -profile singularity
```

より本格的なnf-coreパイプラインのテスト

```
# RNA-seqパイプラインのテストプロファイル
nextflow run nf-core/rnaseq -profile test,singularity
```

Computation on human subject data

Example HPC clusters facilitating computation on restricted-access data

Managing a copy of personal data is a big responsibility.

For Nation-wide project you can run computations on secured servers instead of transporting the data.

ToMMoスーパーコンピュータ

https://sc.megabank.tohoku.ac.jp/

- The ToMMo supercomputer is divided into **three virtual units**:
  - **Unit A**: Public development (accessible via browser).
  - **Unit B**: Data distribution via **data visiting** (not remote download).
  - **Unit C**: Data preprocessing.

- It has SLURM, Singularity and Java installed → You can use nf-core pipelines!

NBDC data on スパコン SHIROKANE

https://gc.hgc.jp/lead/nbdc/

- The supercomputer Shirokane (Human Genome Center, https://www.at.hgc.jp/) is one of the external servers designated by DBCLS (https://dbcls.rois.ac.jp/about.html) that can handle human data in the NDBC Human Database (https://humandbs.dbcls.jp).

- It has Univa Grid Engine, Singularity and Java installed → You can use nf-core pipelines!

Many thanks to

Nick Luscombe and my colleagues in his reserch unit at OIST, who provided plenty of important feedback on earlier verions of this lecture.

Jordan Ramilowski for advices on the lecture content and format.

Takahiro Suzuki for sharing his slides with me last year.

Thank you for listening

And see you at MBSJ 2025 in December?

Glossary

- Reproducibility crisis: 再現性の危機
- Container (コンテナ)
- Job scheduler (ジョブ管理システム)
- Pipeline (パイプライン)
- Software Bill Of Materials (ソフトウェア部品表)

To go further

- https://ja.wikipedia.org/wiki/再現性の危機
- https://nf-co.re/
- https://www.debian.org/devel/debian-med/
- https://github.com/oist/BioinfoUgrp