Nextflow for bio-informatics

Luc Cornet
ENCCB workshop

BCCM

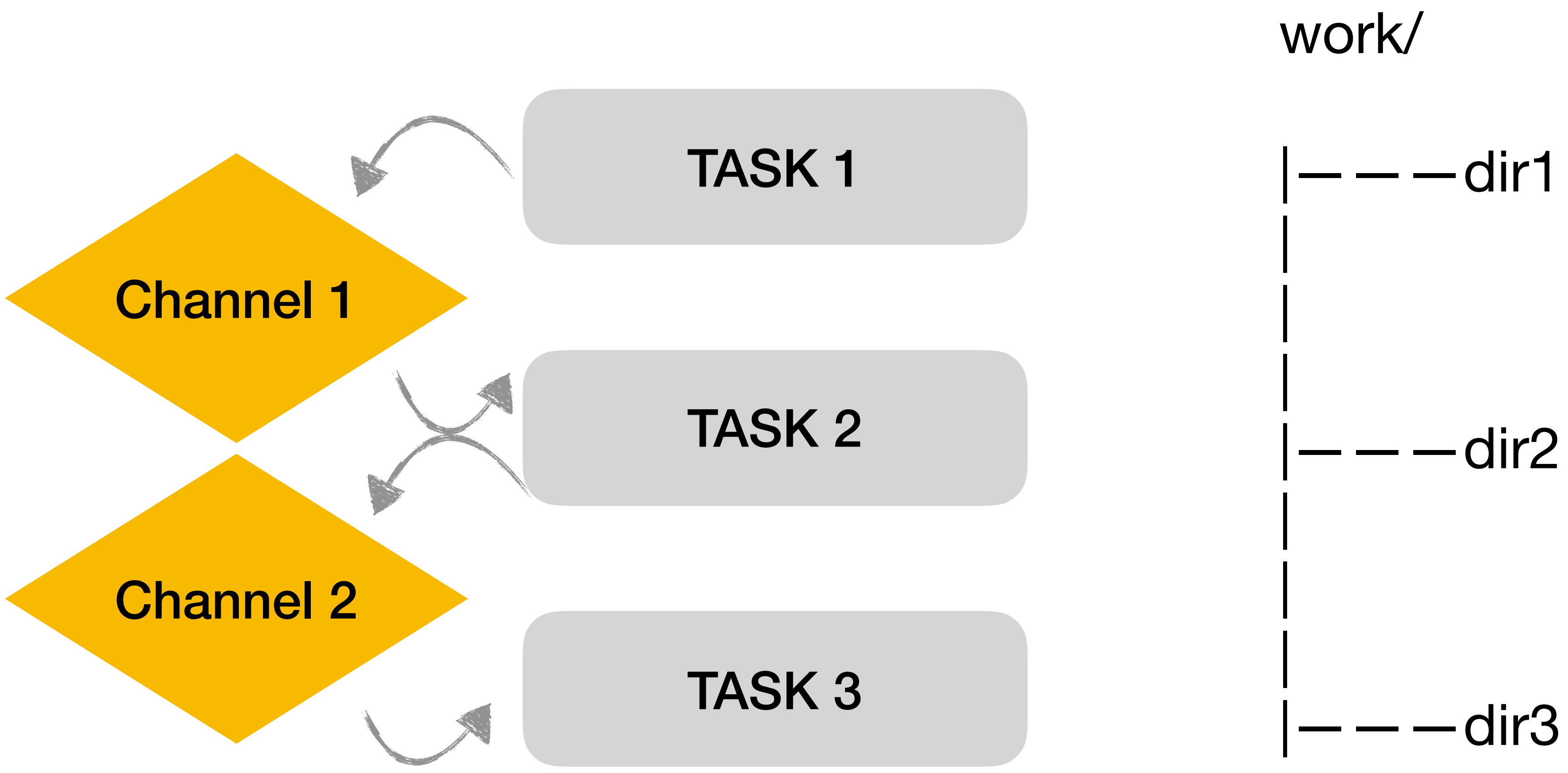
GEN-ERA

What is Nextflow?

- **Published in 2017** (Di Tommaso et al., 751 citations)
- **Scalable workflow system using container(s)**
- **Domain specific language (DSL)**

nextflow

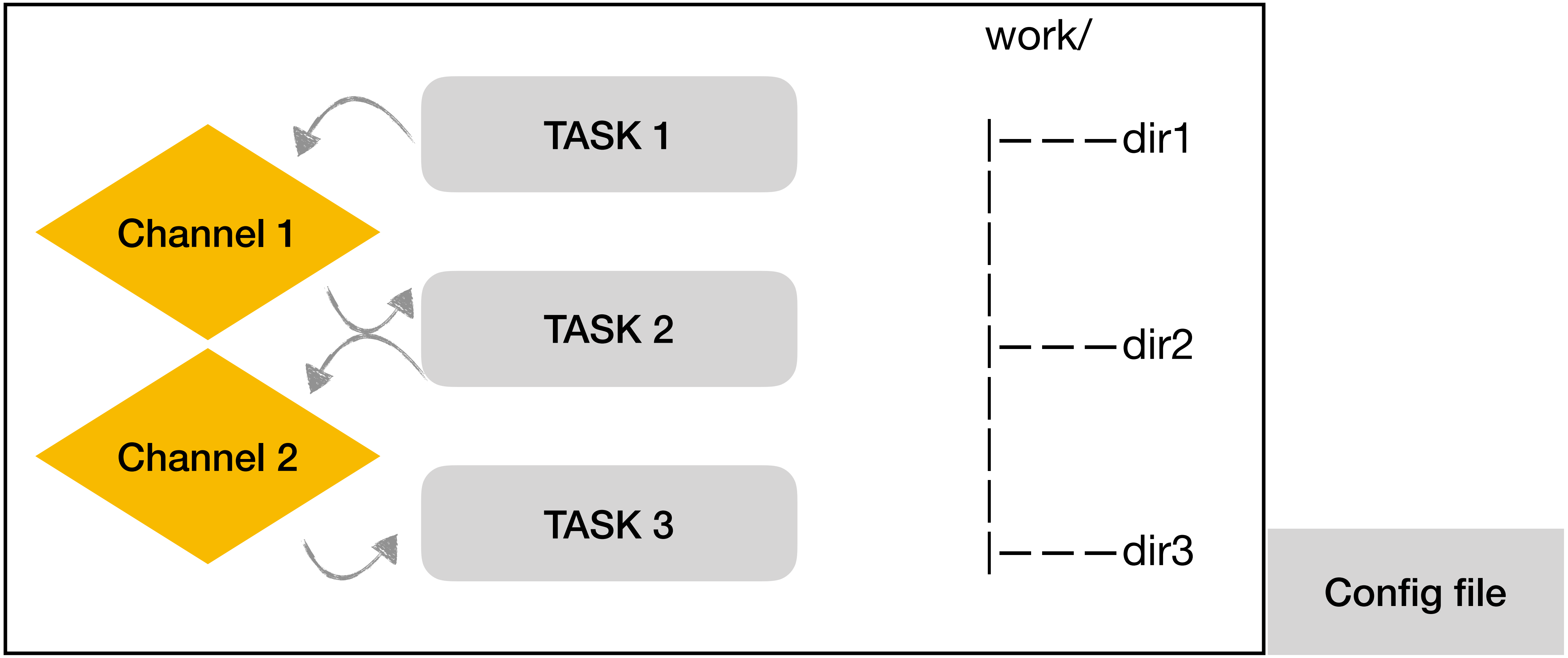
How it works?



A Nextflow scripts is composed of process, interacting through channels

nextflow

How it works?



Containers can be used as operating system.

nextflow

What is the GEN-ERA project?

- BCCM collections in the genomic era (GEN-ERA)
- Establishment of modern genomics practices in the collections
- Partners: IHEM, ITM, LMG, MUCL, UCL
- FAIR principles: Findable, Accessible, Interoperable, Reusable

Who am I?

Bioinformatician

Principal investigator of the project

BCCM/IHEM - Sciensano

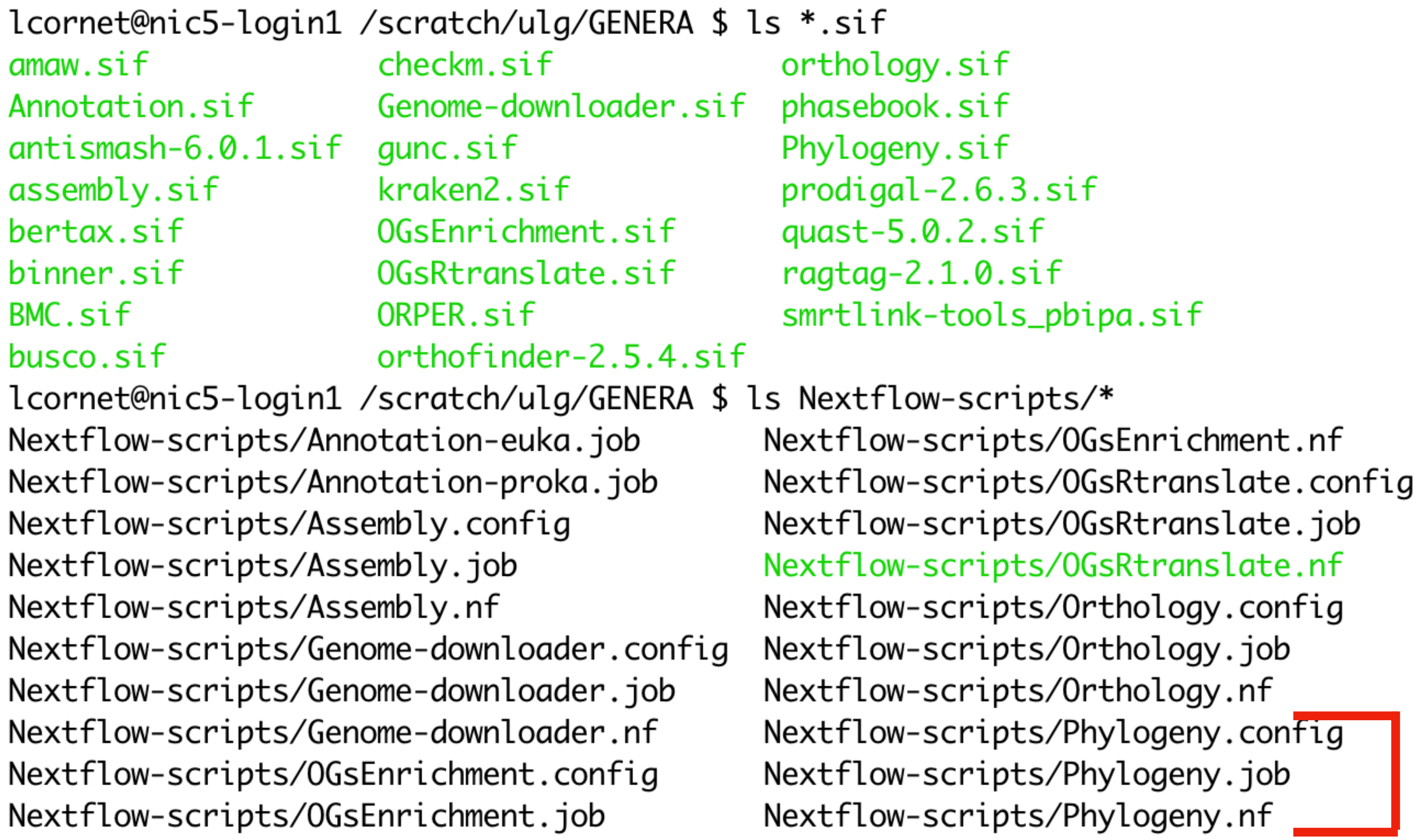
BCCM
BELGIAN
CO-ORDINATED
COLLECTIONS OF
MICRO-ORGANISMS

How is Nextflow used in the GEN-ERA project?

How to do bioinformatics with beginners?

- Nextflow scripts are used as programs
- Each of the scripts should be executed with a single command
- All tools pre-installed on Nic5 with singularity containers
- All databases are shared on Nic5
- Interface on Github

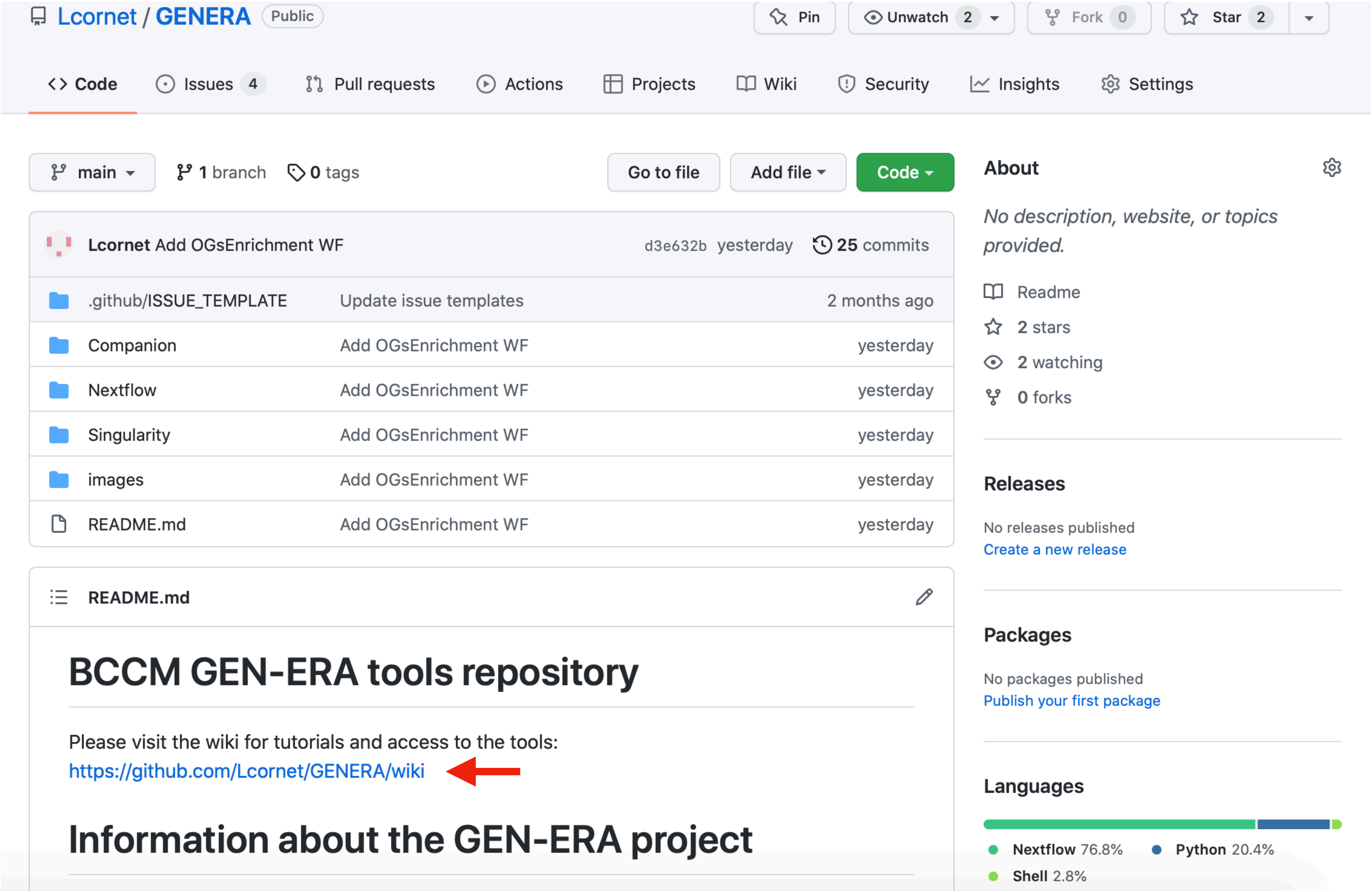
How is Nextflow used in the GEN-ERA project?



```
lcornet@nic5-login1 /scratch/ulg/GENERA $ ls *.sif
amaw.sif             checkm.sif             orthology.sif
Annotation.sif       Genome-downloader.sif  phasebook.sif
antismash-6.0.1.sif  gunc.sif               Phylogeny.sif
assembly.sif         kraken2.sif            prodigal-2.6.3.sif
bertax.sif           OGsEnrichment.sif      quast-5.0.2.sif
binner.sif           OGsRtranslate.sif      ragtag-2.1.0.sif
BMC.sif              ORPER.sif              smrtlink-tools_pbipa.sif
busco.sif            orthofinder-2.5.4.sif
lcornet@nic5-login1 /scratch/ulg/GENERA $ ls Nextflow-scripts/*
Nextflow-scripts/Annotation-euka.job       Nextflow-scripts/OGsEnrichment.nf
Nextflow-scripts/Annotation-proka.job      Nextflow-scripts/OGsRtranslate.config
Nextflow-scripts/Assembly.config           Nextflow-scripts/OGsRtranslate.job
Nextflow-scripts/Assembly.job              Nextflow-scripts/OGsRtranslate.nf
Nextflow-scripts/Assembly.nf               Nextflow-scripts/Orthology.config
Nextflow-scripts/Genome-downloader.config  Nextflow-scripts/Orthology.job
Nextflow-scripts/Genome-downloader.job     Nextflow-scripts/Orthology.nf
Nextflow-scripts/Genome-downloader.nf      Nextflow-scripts/Phylogeny.config
Nextflow-scripts/OGsEnrichment.config      Nextflow-scripts/Phylogeny.job
Nextflow-scripts/OGsEnrichment.job         Nextflow-scripts/Phylogeny.nf
```

Shared Folder

How is Nextflow used in the GEN-ERA project?



Lcornet Add OGsEnrichment WF d3e632b yesterday 25 commits

.github/ISSUE_TEMPLATE	Update issue templates	2 months ago
Companion	Add OGsEnrichment WF	yesterday
Nextflow	Add OGsEnrichment WF	yesterday
Singularity	Add OGsEnrichment WF	yesterday
images	Add OGsEnrichment WF	yesterday
README.md	Add OGsEnrichment WF	yesterday

README.md

BCCM GEN-ERA tools repository

Please visit the wiki for tutorials and access to the tools:
https://github.com/Lcornet/GENERA/wiki

Information about the GEN-ERA project

About

No description, website, or topics provided.

- Readme
- 2 stars
- 2 watching
- 0 forks

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- Nextflow 76.8%
- Python 20.4%
- Shell 2.8%

How is Nextflow used in the GEN-ERA project?

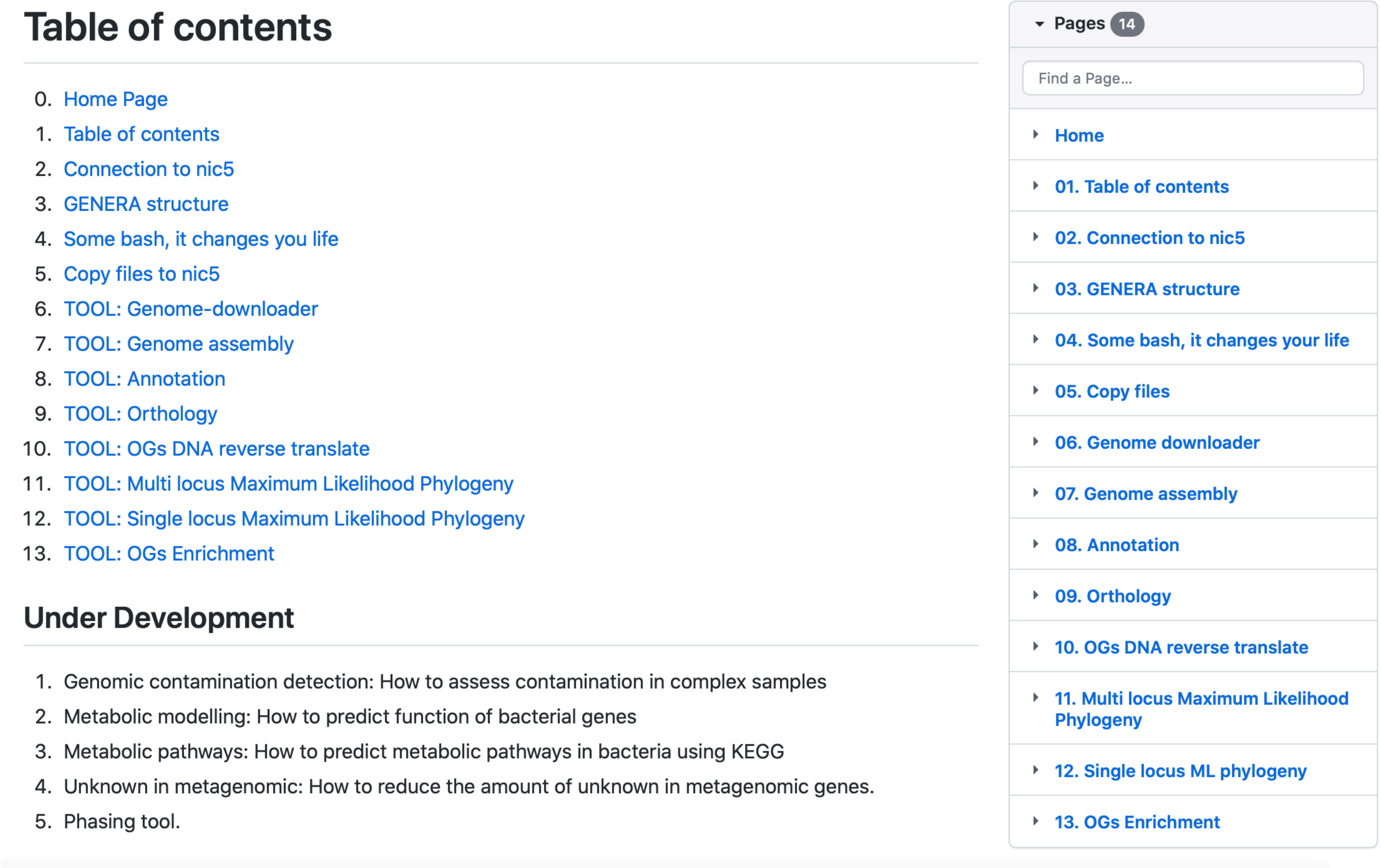


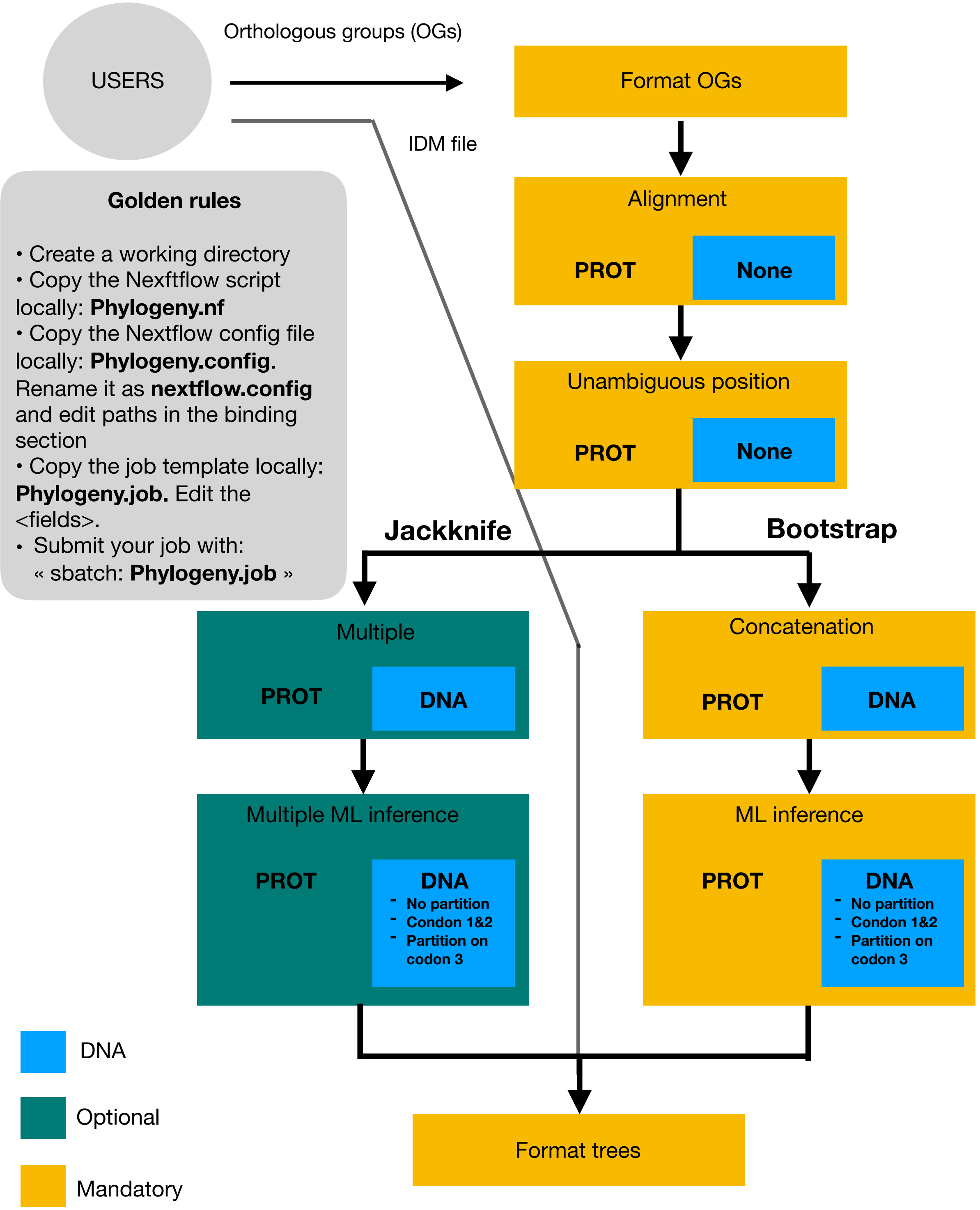
Table of contents

0. Home Page
1. Table of contents
2. Connection to nic5
3. GENERA structure
4. Some bash, it changes you life
5. Copy files to nic5
6. TOOL: Genome-downloader
7. TOOL: Genome assembly
8. TOOL: Annotation
9. TOOL: Orthology
10. TOOL: OGs DNA reverse translate
11. TOOL: Multi locus Maximum Likelihood Phylogeny
12. TOOL: Single locus Maximum Likelihood Phylogeny
13. TOOL: OGs Enrichment

Under Development

1. Genomic contamination detection: How to assess contamination in complex samples
2. Metabolic modelling: How to predict function of bacterial genes
3. Metabolic pathways: How to predict metabolic pathways in bacteria using KEGG
4. Unknown in metagenomic: How to reduce the amount of unknown in metagenomic genes.
5. Phasing tool.

How is Nextflow used in the GEN-ERA project?



USERS

Orthologous groups (OGs)

IDM file

Format OGs

Alignment

PROT

None

Unambiguous position

PROT

None

Jackknife

Bootstrap

Multiple

PROT

DNA

Multiple ML inference

PROT

DNA

- No partition
- Condon 1&2
- Partition on codon 3

Concatenation

PROT

DNA

ML inference

PROT

DNA

- No partition
- Condon 1&2
- Partition on codon 3

Format trees

DNA

Optional

Mandatory

Golden rules

- Create a working directory
- Copy the Nexftflow script locally: **Phylogeny.nf**
- Copy the Nextflow config file locally: **Phylogeny.config**. Rename it as **nextflow.config** and edit paths in the binding section
- Copy the job template locally: **Phylogeny.job.** Edit the <fields>.
- Submit your job with: « sbatch: **Phylogeny.job** »

FIELDS

Path to directory containing OGs (Prot must be .faa and DNA .fna) - mode DNA or prot - Jackknife activated or not - Jackknife number of replicates - Jackknife width.

https://github.com/Lcornet/GENERA/wiki/11.-Multi-locus-Maximum-Likelihood-Phylogeny

```
nextflow Phylogeny.nf --OG=OGs --IDM=file.idm --jackk=yes
```

Mandatory arguments:

--OG	Path to OG directory in fasta format (.faa for prot and .fna files for DNA)
--IDM	Path to IDM file

Optional arguments:

--mode	specify prot or DNA, default = prot
--jackk	activate jackknife, default = no
--rep	number of jackkninfe replicates
--width	Width of jackknife replicates
--cpu	number of cpus to use, default = 1

For the users:

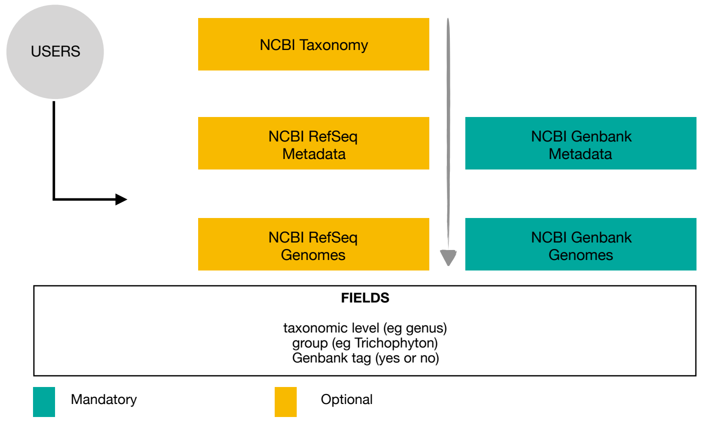
1. **Download of the suite of 3 files**
2. **Edit the paths**
3. **Launch the job**

How is Nextflow used in the GEN-ERA project?

How to do bioinformatics with beginners?

- Training to each collection (1/2 day)
- Interaction through GitHub for request-update-question-bug
- Video for each tool (Scheduled)

Genome-downloader



FIELDS

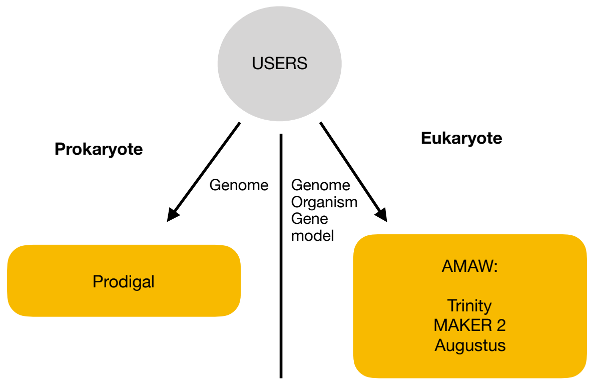
taxonomic level (eg genus)
group (eg Trichophyton)
Genbank tag (yes or no)

🟩 Mandatory 🟨 Optional

Golden rules

- Create a working directory
- Copy the Nexftflow script locally: **Genome-downloader.nf**
- Copy the Nextflow config file locally: **Genome-downloader.config**. Rename it as **nextflow.config** and edit paths in the binding section
- Copy the job template locally: **Genome-downloader.job**. Edit the <fields>.
- Submit your job with: « **sbatch Genome-downloader.job »**

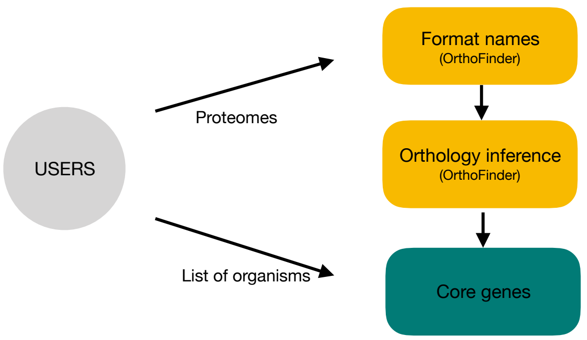
Genome-annotation



Golden rules

- Create a working directory
- Copy the job template l locally: **Annotation*.job.** Edit the <fields>.
- Submit your job with: « **sbatch Annotation.job** »

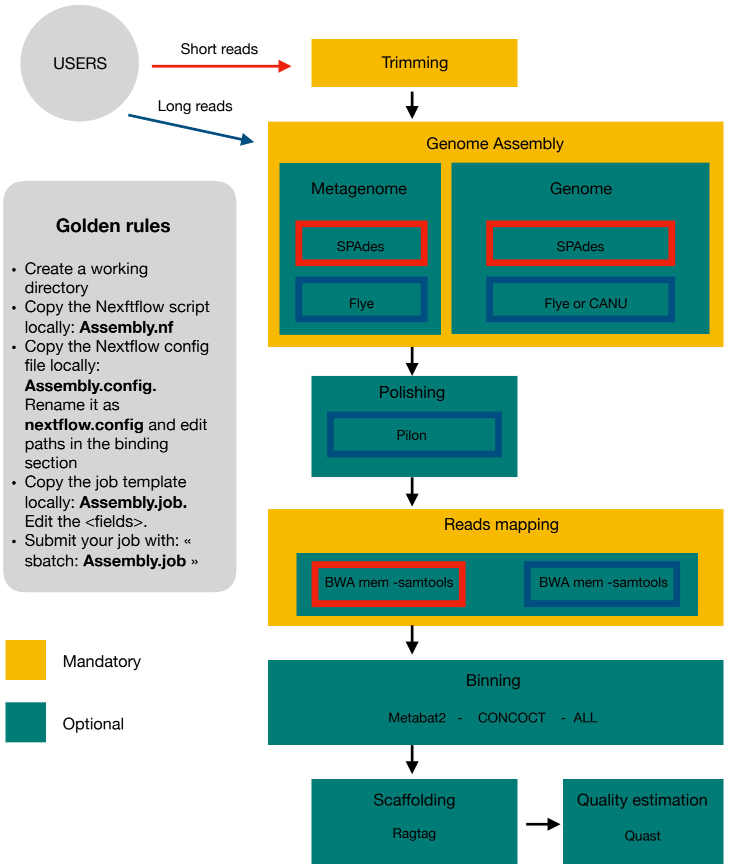
Orthology inference



Golden rules

- Create a working directory
- Copy the Nexftflow script locally: **Orthology.nf**
- Copy the Nextflow config file locally: **Orthology.config.** Rename it as **nextflow.config** and edit paths in the binding section
- Copy the job template locally: **Orthology.job**. Edit the <fields>.
- Submit your job with: « sbatch **Orthology.job** »

Genome-assembly



Golden rules

- Create a working directory
- Copy the Nexftflow script locally: **Assembly.nf**
- Copy the Nextflow config file locally: **Assembly.config.** Rename it as **nextflow.config** and edit paths in the binding section
- Copy the job template locally: **Assembly.job.** Edit the <fields>.
- Submit your job with: « sbatch: **Assembly.job** »

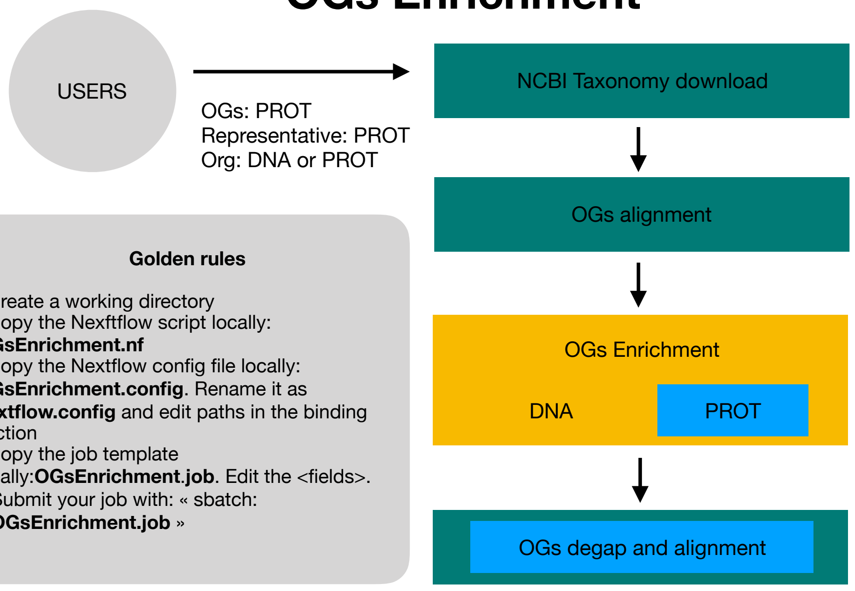
🟨 Mandatory

🟩 Optional

FIELDS

PATH to short reads R1 file - PATH to short reads R2 file - PATH to long reads (ONT/PB) - Metagenome mode activated or not - binning mode activated or not - ragtag scaffolding mode activated or not.

OGs Enrichment

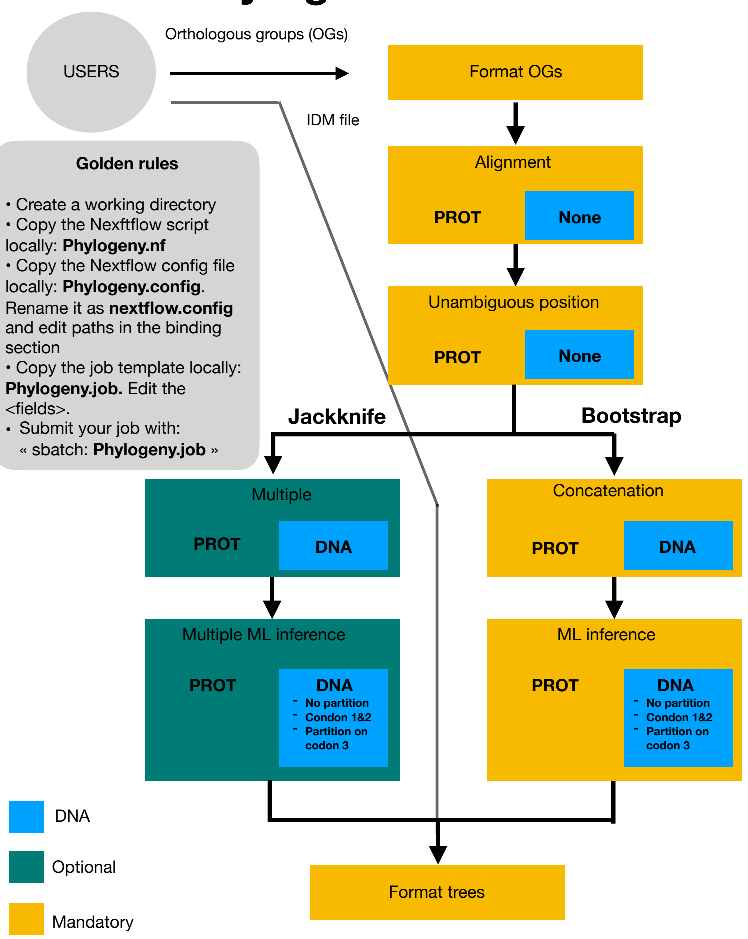


Golden rules

- Create a working directory
- Copy the Nexftflow script locally: **OGsEnrichment.nf**
- Copy the Nextflow config file locally: **OGsEnrichment.config**. Rename it as **nextflow.config** and edit paths in the binding section
- Copy the job template locally:**OGsEnrichment.job**. Edit the <fields>.
- Submit your job with: « sbatch: **OGsEnrichment.job** »

FIELDS:
Path to OGs (.faa); Path to representative proteomes (.faa); Path to org to add in fasta (.fna or .faa); specify the orgType DNA or Prot; alignement is optional; forty-two

Phylogenomics



Golden rules

- Create a working directory
- Copy the Nexftflow script locally: **Phylogeny.nf**
- Copy the Nextflow config file locally: **Phylogeny.config**. Rename it as **nextflow.config** and edit paths in the binding section
- Copy the job template locally: **Phylogeny.job.** Edit the <fields>.
- Submit your job with: « sbatch: **Phylogeny.job** »

🟦 DNA

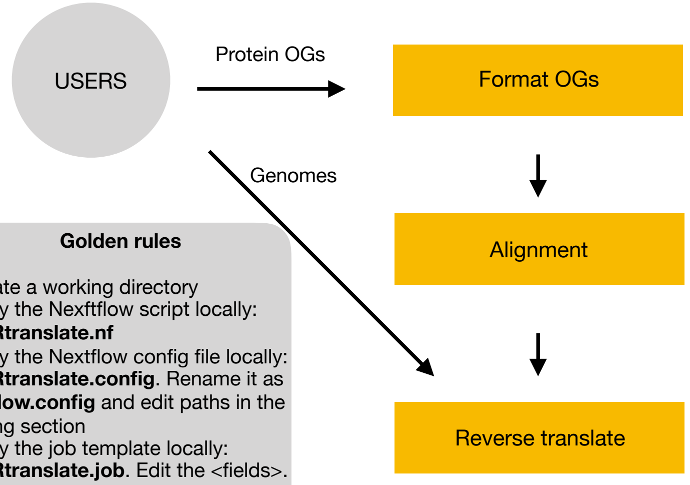
🟩 Optional

🟨 Mandatory

FIELDS

Path to directory containing OGs (Prot must be .faa and DNA .fna) - mode DNA or prot - Jackknife activated or not - Jackknife number of replicates - Jackknife width.

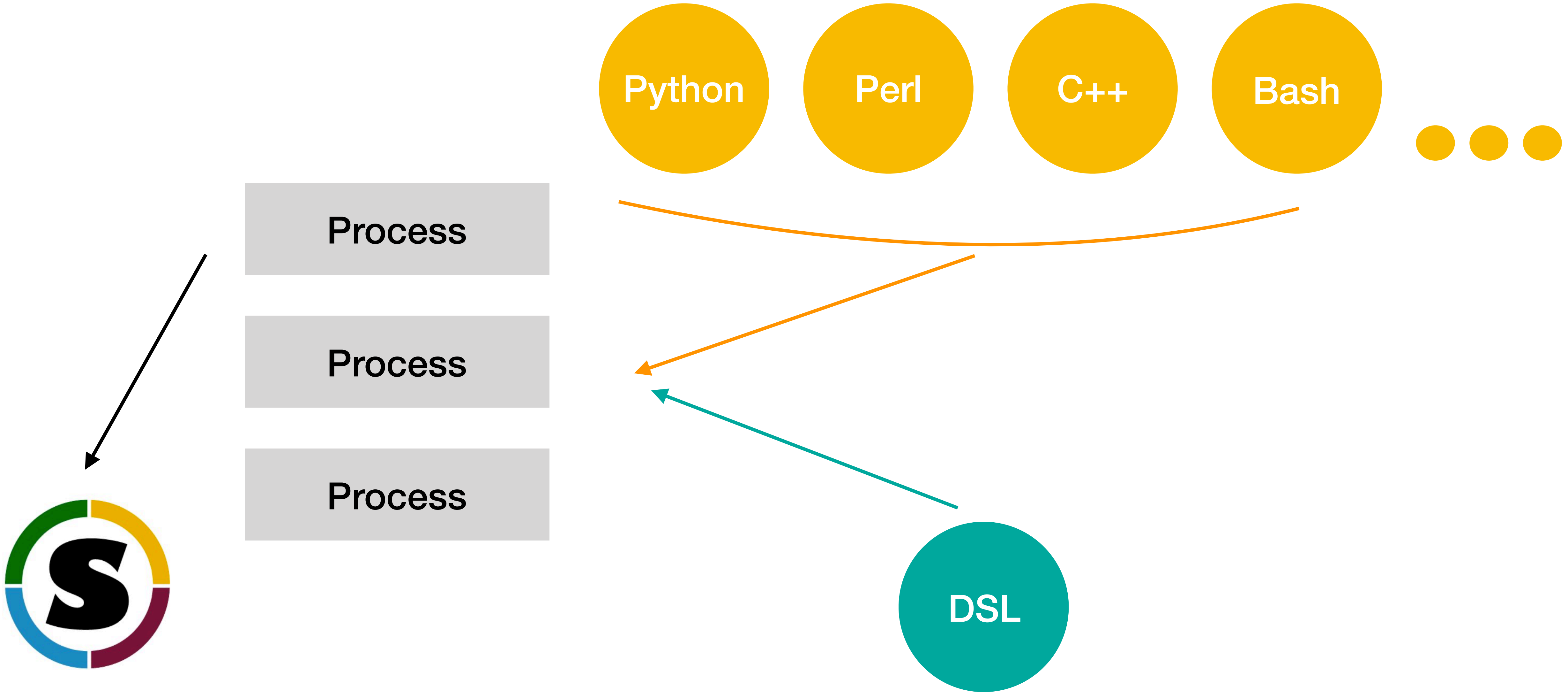
OGs Reverse translate



Golden rules

- Create a working directory
- Copy the Nexftflow script locally: **OGsRtranslate.nf**
- Copy the Nextflow config file locally: **OGsRtranslate.config**. Rename it as **nextflow.config** and edit paths in the binding section
- Copy the job template locally: **OGsRtranslate.job**. Edit the <fields>.
- Submit your job with: « sbatch: **OGsRtranslate.job** »

Very easy to develop



Python

Perl

C++

Bash

Process

Process

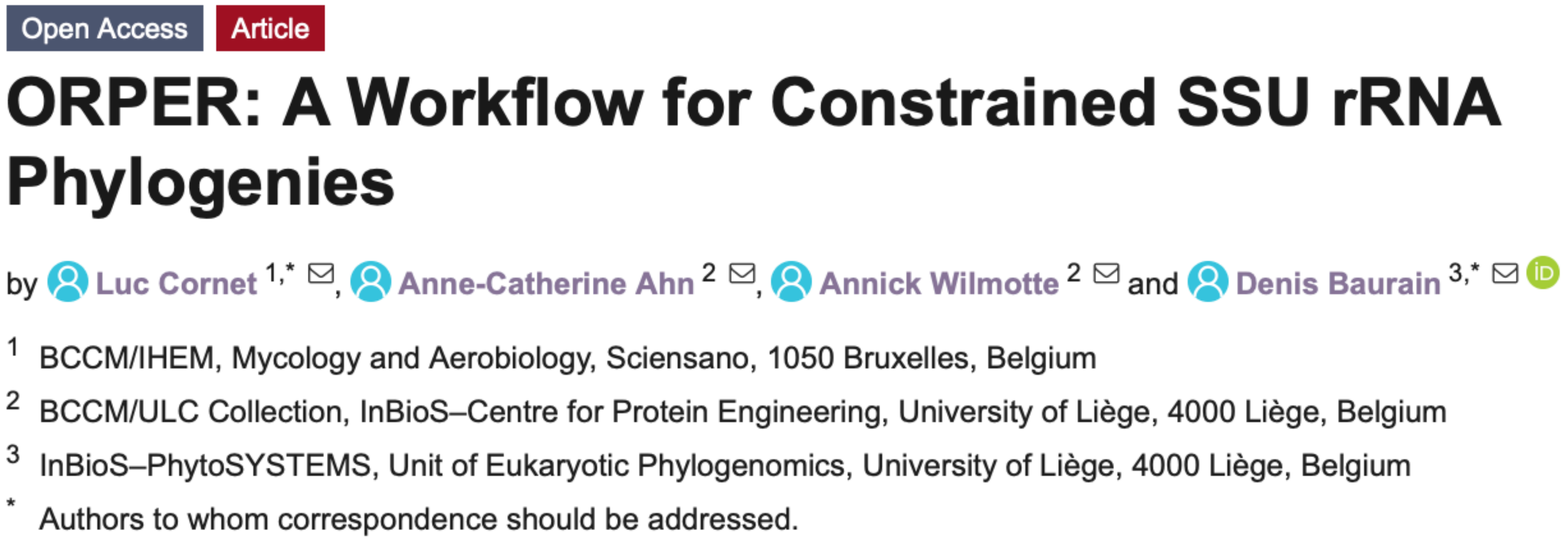
Process

DSL

nextflow

Others advantages

- Interoperable -> collaborations
- Reproducibility
- Possibility to publish the tools



Open Access Article

ORPER: A Workflow for Constrained SSU rRNA Phylogenies

by Luc Cornet[1,*], Anne-Catherine Ahn[2], Annick Wilmotte[2] and Denis Baurain[3,*]

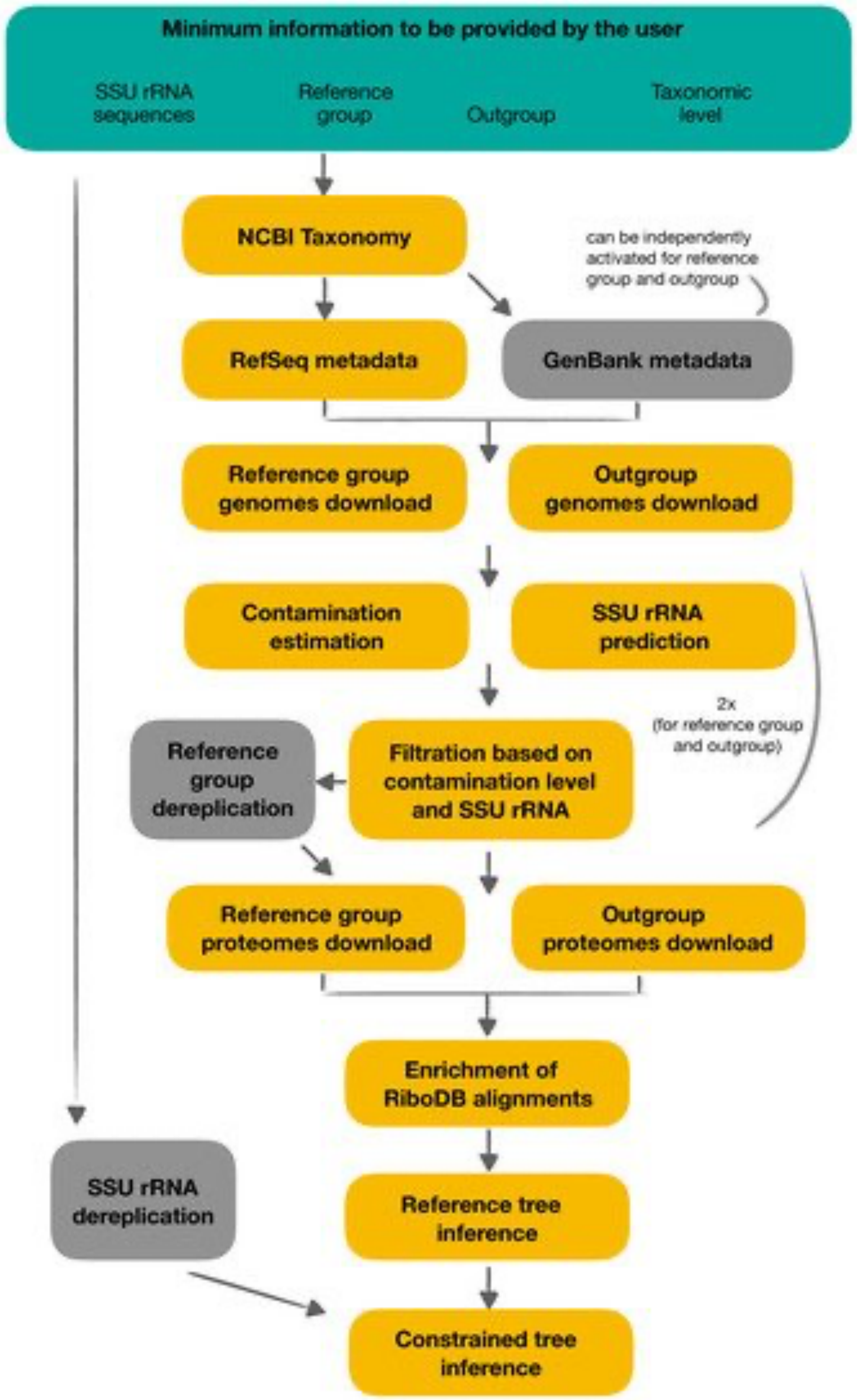
[1] BCCM/IHEM, Mycology and Aerobiology, Sciensano, 1050 Bruxelles, Belgium

[2] BCCM/ULC Collection, InBioS–Centre for Protein Engineering, University of Liège, 4000 Liège, Belgium

[3] InBioS–PhytoSYSTEMS, Unit of Eukaryotic Phylogenomics, University of Liège, 4000 Liège, Belgium

[*] Authors to whom correspondence should be addressed.

ORPER



Minimum information to be provided by the user

SSU rRNA sequences

Reference group

Outgroup

Taxonomic level

NCBI Taxonomy

can be independently activated for reference group and outgroup

RefSeq metadata

GenBank metadata

Reference group genomes download

Outgroup genomes download

Contamination estimation

SSU rRNA prediction

2x (for reference group and outgroup)

Reference group dereplication

Filtration based on contamination level and SSU rRNA

Reference group proteomes download

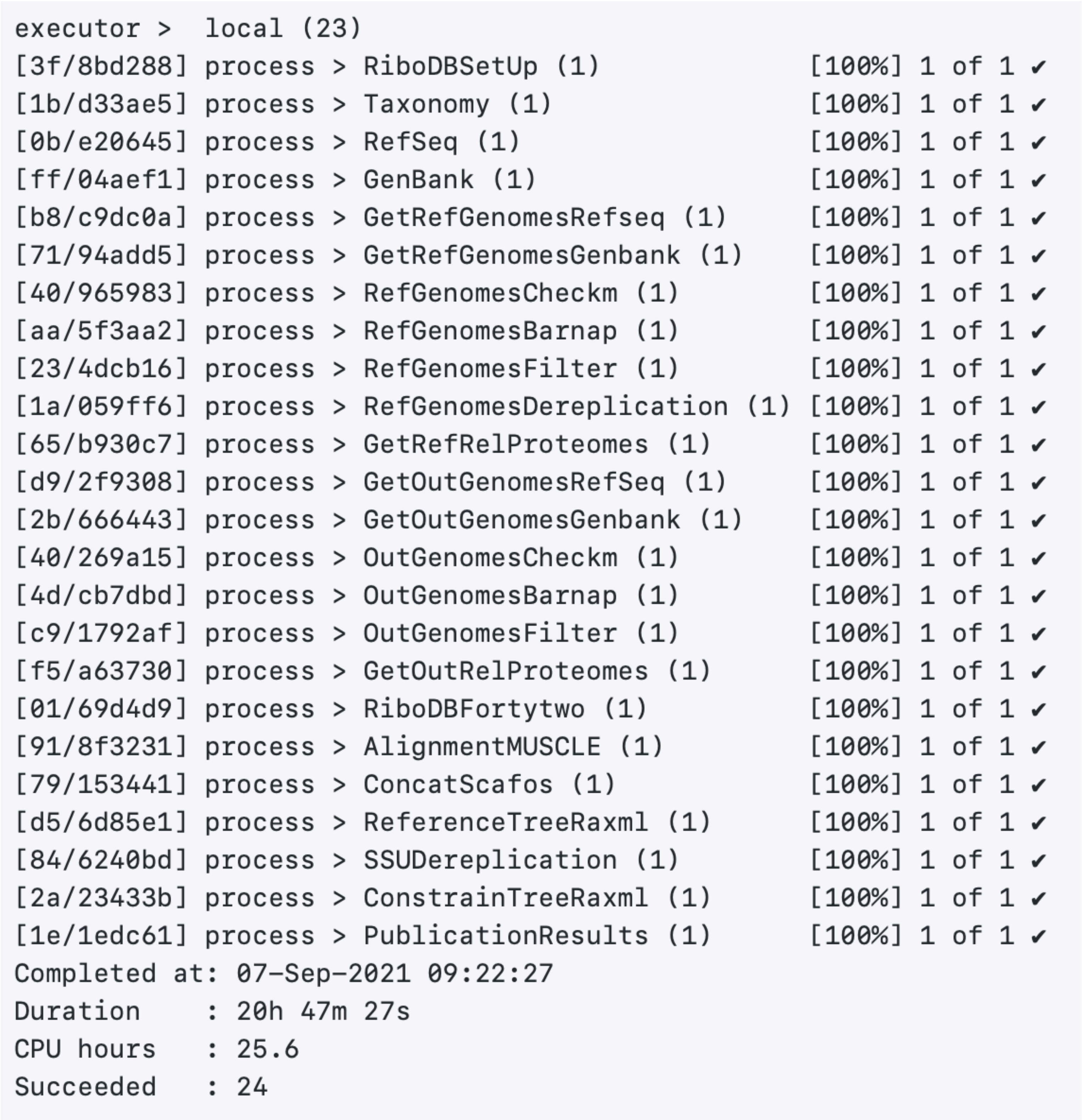
Outgroup proteomes download

Enrichment of RiboDB alignments

SSU rRNA dereplication

Reference tree inference

Constrained tree inference



```
executor >  local (23)
[3f/8bd288] process > RiboDBSetUp (1)               [100%] 1 of 1 ✔
[1b/d33ae5] process > Taxonomy (1)                  [100%] 1 of 1 ✔
[0b/e20645] process > RefSeq (1)                    [100%] 1 of 1 ✔
[ff/04aef1] process > GenBank (1)                   [100%] 1 of 1 ✔
[b8/c9dc0a] process > GetRefGenomesRefseq (1)       [100%] 1 of 1 ✔
[71/94add5] process > GetRefGenomesGenbank (1)      [100%] 1 of 1 ✔
[40/965983] process > RefGenomesCheckm (1)          [100%] 1 of 1 ✔
[aa/5f3aa2] process > RefGenomesBarnap (1)          [100%] 1 of 1 ✔
[23/4dcb16] process > RefGenomesFilter (1)          [100%] 1 of 1 ✔
[1a/059ff6] process > RefGenomesDereplication (1) [100%] 1 of 1 ✔
[65/b930c7] process > GetRefRelProteomes (1)        [100%] 1 of 1 ✔
[d9/2f9308] process > GetOutGenomesRefSeq (1)       [100%] 1 of 1 ✔
[2b/666443] process > GetOutGenomesGenbank (1)      [100%] 1 of 1 ✔
[40/269a15] process > OutGenomesCheckm (1)          [100%] 1 of 1 ✔
[4d/cb7dbd] process > OutGenomesBarnap (1)          [100%] 1 of 1 ✔
[c9/1792af] process > OutGenomesFilter (1)          [100%] 1 of 1 ✔
[f5/a63730] process > GetOutRelProteomes (1)        [100%] 1 of 1 ✔
[01/69d4d9] process > RiboDBFortytwo (1)            [100%] 1 of 1 ✔
[91/8f3231] process > AlignmentMUSCLE (1)           [100%] 1 of 1 ✔
[79/153441] process > ConcatScafos (1)              [100%] 1 of 1 ✔
[d5/6d85e1] process > ReferenceTreeRaxml (1)        [100%] 1 of 1 ✔
[84/6240bd] process > SSUDereplication (1)          [100%] 1 of 1 ✔
[2a/23433b] process > ConstrainTreeRaxml (1)        [100%] 1 of 1 ✔
[1e/1edc61] process > PublicationResults (1)        [100%] 1 of 1 ✔
Completed at: 07-Sep-2021 09:22:27
Duration     : 20h 47m 27s
CPU hours    : 25.6
Succeeded    : 24
```

Thank you

BCCM

GEN-ERA