

GenPat Platform

User manual

Bionformatic Unit

© IZSAM "G. Caporale"

Index

1. Welcome 6
1.1 GENPAT Platform 6
1.1.1 Introduction 6
1.1.2 Official samples management 6
1.1.3 Quick links 6
1.2 Quick start 7
1.2.1 User Guide 7
1.2.2 Upload procedure 7
1.2.3 Download procedure 7
1.2.4 User Profiles 7
1.3 Features 9
1.3.1 Main features 9
1.3.2 Table features 11
1.3.3 Data Import / Export through files 13
1.4 Use cases 14
1.4.1 Use cases 14
1.4.2 Useful videos on sample and table management 15
1.4.3 Useful videos on the AMR workflow 16
1.4.4 Useful videos on hybrid assembly 17
1.4.5 Useful videos on FastQ metrics 18
2. Official samples 19
2.1 Management of official samples 19
2.1.1 Rules for data access 19
2.1.2 Quality of data 20
2.1.3 IZS and LAB Profiles 21
2.1.4 Upload of official samples 21
2.1.5 Bioinformatics analyses 22
2.1.6 Download 22
2.1.7 Support 22
2.2 EFSA submissions 23
2.2.1 Submission process to EFSA (for Reference Laboratories) 23
2.2.2 Approval process (EFSA Country Officer) 23
2.2.3 Comparison between samples on WGS Portal and samples on the platform 24
3. Main objects 26
3.1 Samples 26
3.1.1 Cards of Samples 26

3.1.2 Table functionalities for samples 26
3.1.3 Public and private samples 27
3.2 Alias 28
3.2.1 Interface of Alias of samples 28
3.2.2 List of aliases per sample 29
3.3 Metadata 30
3.3.1 Codings according to EFSA standards 30
3.3.2 Pathogenic species 30
3.3.3 Host species 31
3.3.4 Pathogenic material 32
3.3.5 Sampling Point 32
3.4 Analysis results 33
3.4.1 Report error 33
4. Filter samples 34
4.1 Tags 34
4.1.1 Tag 34
4.1.2 Manage tags 35
4.2 Cart 37
4.2.1 Introduction 37
4.2.2 Manage Cart 38
4.2.3 Cart activation 40
5. Analyses 41
5.1 Introduction 41
5.1.1 Run analyses 41
5.1.2 Preference selection and saving 42
5.1.3 Check analysis 42
5.2 Run analysis steps 43
5.2.1 Step 1: Samples 43
5.2.2 Step 2: Tools 43
5.2.3 Step 3: Inputs 44
5.3 Check analyses 46
5.3.1 Analysis card 46
5.3.2 Import of analyses 47
5.3.3 Replay analysis 47
5.3.4 Keep analysis 47
5.4 Tool Descriptions 49
5.4.1 Introduction 49
5.4.2 Bioinformatics tools for non-Illumina Sequencing Technologies 56
5.4.3 Single samples 57

5.4.4 Multi sample 142
5.4.5 Pipelines 193
6. Reports 271
6.1 Reports 271
6.2 Checks 272
6.2.1 Tables of checks 272
6.2.2 Check of the assembly 275
6.2.3 Check of the reads 277
6.2.4 Check of the species 279
6.2.5 Check Taxa 280
6.2.6 Check Coverage 282
6.2.7 Check of cgMLST 283
6.2.8 Check of Kleborate 284
6.2.9 Quality checks for ML-based classification 285
7. Downloads 286
7.1 Download procedure 286
7.1.1 Preset 287
7.1.2 View downloads 287
7.1.3 wget for Windows users 288
8. Uploads 292
8.1 Introduction 292
8.1.1 Topics 292
8.2 Upload 293
8.2.1 Preparation and upload of metadata 293
8.2.2 Uploading sequences 294
8.2.3 File processing 295
8.3 Upload from NCBI 296
8.4 PT GenPat Upload 297
8.4.1 Upload Instructions 297
8.4.2 Metadata Files 297
8.4.3 Description of Required Metadata 298
8.4.4 Additional Notes 298
8.5 Upload check 299
8.5.1 Success 299
8.5.2 Failed 299
8.6 Warnings 300
8.7 Further information 301
8.7.1 Metadata glossary 301
8.7.2 Metadata and .zip rules 305

9. Dashboards 308
9.1 SPREAD, Spatiotemporal Pathogen Relationships and Epidemiological Analysis Dashboard 308
9.1.1 Description 308
9.1.2 Using SPREAD 308
9.1.3 Overview of General Functions 309
9.1.4 Components 309
9.2 Auspice 314
9.2.1 Quick guide 314
9.2.2 Main user interface 314
10. SOP 318
10.1 NGSManagerSOP 318
10.1.1 Versions 318
10.1.2 Validation and reproducibility 322
11. News 323
11.1 2026 323
11.1.1 September 2026 323
11.1.2 April 2026 326
11.1.3 January 2026 327
11.2 2025 328
11.2.1 December 2025 328
12. Support 329
12.1 Password management 329
12.1.1 Password change procedure 329
12.1.2 Forgotten Password 332
12.1.3 Password requirements 333
12.2 Help desk 334
12.3 Useful videos 335
12.3.1 NGS & Bioinformatics - ERFAN 335
12.3.2 CIS demo at Matrix WP6 335
12.4 Keyboard shortcuts 336
12.4.1 Shortcuts list 336
12.5 FAQ 337
12.5.1 Data management 337
12.5.2 Cart and tag 338
12.5.3 Interface 338
12.5.4 Data check and visualization 339
12.5.5 Analyses 339
12.5.6 Profile management 341

1. Welcome

1.1 GENPAT Platform

1.1.1 Introduction

The *National Reference Centre for Genomic Sequences of Pathogenic Microorganisms: database and bioinformatic analysis* (GENPAT) was formally established at the central headquarters of the Istituto Zooprofilattico Sperimentale dell'Abruzzo e del Molise (IZSAM) by Decree of the Italian Ministry of Health on May 30, 2017 (G.U.R.I. n°. 196, of August 23, 2017).

The mission and objectives of the Centre are:

- to develop a national platform for the collection and storage of genomic sequences of pathogenic microorganisms, for the execution of bioinformatics analyses and for the archiving and sharing of the results;
- to establish a structured and permanent network of reference officers, within the individual Istituti Zooprofilattici Sperimentali, in order to ensure a better coordination of the activities to be carried out at the national level;
- to provide techical and scientific assistance to the Italian Ministry of Health and to the competent authorities;
- to manage the organization of training courses, within its area of competence, for the staff of the National Health Service and of other competent authorities;
- to encourage research activities within the specific area of expertise;
- to implement any other useful activity related to its specific area of competence, including coordination and collaboration with other administrations and organisms in the sector.

Responsible: Cesare Cammà - c.camma@izs.it

1.1.2 Official samples management

In agreement with the respective National Reference Laboratories, starting from Monday, June 3rd, 2024, authorized users are able to view minimal metadata and genomic characterization results of all *L. monocytogenes* and *Campylobacter* strains isolated within the national territory, in the context of official controls and other official activities as defined by Regulation (EU) 2017/625 and Legislative Decree No. 27 of February 2, 2021.

Further information can be found in the dedicated section Management of official samples.

1.1.3 Quick links

- Quick start
- Management of official samples
- Main Features
- Useful videos

1.2 Quick start

1.2.1 User Guide

The video below shows the main activities that can be performed on the platform, including:

- sample table management;
- use of filters, sample selection and operations on them;
- launching analyses on selected samples;
- surveillance of the status of launched analyses and viewing the results.

Main activities

1.2.2 Upload procedure

The 2 following video show the procedures to upload samples on the platform. In the first video, a compressed `.zip` file is uploaded, containing the `.fasta` files with their corresponding metadata; the second video shows the import ofsequences directly from **NCBI**.

Upload of sequence files

Upload of sequence file

Upload from NCBI

Upload from NCBI

1.2.3 Download procedure

It is possible to download the samples directly using either a sample collection from the cart or from an active tag. The following video provides a brief demonstration of one of the available download procedures; further information, along with additional video tutorials is available in the wiki section concerning the download procedure.

Download procedure

1.2.4 User Profiles

Each user can be associated with more than one profile and, therefore, users can belong to different **groups**.

Users will be able to view samples or perform operations related to the group they belong to, and only within the limits of what is permitted for users of that group. From the user's perspective, groups can therefore act as a general filtering role.

The following video shows how to change group, highlighting how this affects the main navigation menu and the samples that can be viewed.

> **Note:** For privacy and data security reasons, the video was created using the Cohesive demo application; therefore, the interface may show some minor differences.

Switching the user group

1.3 Features

1.3.1 Main features

Flexibility of the database schema

The system features extremely flexible database management schemas; no software development is required to perform operations such as creating new tables or new fields in modules.

The following video shows how, in addition to defining additional attributes, it is possible to create a completely new custom database from scratch, through the use of:

- *classes*, i.e. database tables;
- class *attributes*, i.e. the table columns (available for every type, including lists with fixed values, single or multiple values, external keys, files or formulas) and their corresponding layout;
- *domains*, i.e. relation types, even with any specific additional attributes;
- *lookup types*, i.e. lists with fixed values that can be associated to a class attribute;
- *reports*, i.e. the possibility of specializing a class in "subclasses" (also at multiple levels), by adding specific attributes and domains and sharing the attributs of superclass;
- *views* based on filters or SQL queries;
- user interface behaviours that are specific for each class: data validation, automatisms during data update, contextual menus, online help, etc.

Note: For privacy and data security reasons, some of the videos in this page were created using the Cohesive demo application; therefore, the interface may show some minor differences.

Flexibility of databases

General modification

This feature allows you to change the value of one or more attributes of a selected set of cards from the current class. The function can be activated from the contextual menu of the class by enabling the multiselection.

General modification

Data Import / Export

The platform supports importing and exporting data in `.csv`, `.xls`, `.xlsx` formats, as well as in relational database form.

The system is based on the configuration of specific templates to ensure the management of complex operations in a simplified manner. For a detailed description and video guides on this topic, please consult the dedicated Data Import/Export section of this Wiki.

Tables

The platform features an integrated table of samples within the interface: it lists the platform entries and allows the interaction with them. Tables support customisation tools and the use of filters; moreover, it provides various functions to export, download, print or store the entries in the Cart. Data concerning an entry of interest can be accessed and cunsulted by simply expanding its card. A detailed guide, including a video, is available in the section dedicated to tables.

Tags

Tags are a platform feature that allows the user to group a selection of samples under the same label. When a tag is active, only the samples belonging to that tag will be displayed in the sample table; in this way, tags can act as filters.

Further information about tags is available both on the dedicated page and in the Tag management subsection along with demonstration videos on how to create, activate, edit or delate a tag.

Cart

The cart is a feature that allows the temporary storage of selected samples and can be used to create new tags or to launch analyses. The analysis launch, in fact, can be performed using an active tag or the samples available in the cart. The cart also features the **Activate cart** mode which allows you to operate on the platform by filtering all samples and launched analyses, displaying only what is related to the samples available in the cart.

In the introduction of the Wiki page dedicated to the cart you can find information and videos about the main functions of the cart. For a more in-depth discussion, please refer instead to the Manage Cart page of this Wiki.

Wiki integration on Run analysis

In the analysis launch system of the platform, for each available and selected analysis/pipeline, you can find a link to the related documentation, in addition to a brief description.

Wiki Integration

Repeat Analyses

Two quick actions are available on the platform to allow the user to request, for a new set of samples, the launch of an analysis that has previously been successfully completed. The two functions make it possible to keep the parameters used previously, in order to repeat the analysis on a new dataset, provided either by an active tag or by a selection of samples in the cart.

Additional information and a video tutorial are available on the dedicated page.

Dashboards

The direct visualization of analysis results within the platform is enabled through the use of several dashboards and functions, including an integrated and extended version of GrapeTree, **SPREAD** (further informaton is available in the dedicated section of this wiki).

The following video shows the development over time of the SARS-CoV-2 epidemic in the Abruzzo region (Italy), using *SPREAD*. The analysis shown was carried out using the metadata of samples: the dates - grouped by year/month of sampling - are used to build the tree while the corresponding GeoJSON data are used for placement on the map.

SPREAD timelapse

1.3.2 Table features

The **Main objects** on the platform, namely Samples, Alias code, Metadata and Analysis results, are managed through tabular display systems integrated into the platform.

Users can interact with the *entries* in the table of each main element of the platform and customize it based on one's own needs and preferences. The tables also support the use of basic and advanced filters, as well as export, download and print options.

> **Note:** Please, consider that **Main objects** tables share the same management and similar functionalities, while still considering the presence of some specific options available in the contextual menu of each element's table.

Table management and customization

The following video shows how a table can be managed and modified by changing the order and size of the columns, selecting which columns to display and adjusting the ascending or descending order for each column. Changes made to a table are temporary unless saved using the designated button; they can be reset by removing the saved preferences, as shown in the video.

Demo on table customization

> **Note:** Grid preferences are saved, based on the selected group at the time of saving. Consequently, if a user is assigned to multiple groups, it is possible to have different preference configurations for each group.

The columns of the tables on the platform automatically adjust their width according to the available space; for some tables, it is possible to disable this automatic behavior using a toggle located at the top right, allowing the width of the columns to be adjusted based on the content or personal preferences. If the total table width exceeds the available horizontal space, you can navigate through the columns using horizontal scrolling.

Change the table view

Basic filters

From the icon with the funnel, basic filters can be applied to the columns, either individually or in combination, to search the entries displayed in the table or to narrow down the search field.

Basic filters

Entry detail

By double-clicking on an item's row (or, as an alternative, clicking the **[+]** toogle of an entry) users can expand an item's card and access additional resources and links.

Demo on opening an entry detail

Advanced filters

Tables also provides the use of advanced filters that can be applied to the list of entries. The icon actually allows to either choose an attribute through the *Add filter* option or apply filters that were previously saved.

Advanced filters allow searching by one or more attributes (which can be represented by text strings, dates or metadata), using one or more search terms or boolean operators for each attribute. Additionally, advanced filters can be modified, copied, saved and applied by default to the table.

Use of advanced filters

Context menu

In addition to the filters, the platform's tables provide the user with a context menu that allows customization of display options and enabling/disabling the multi-selection of entries. Additional options provided in the context menu, such as import and export systems, may be available only for the *Sample* table or may differ among the main element tables.

Use of the context menu

> **Important note:** the functions available in the context menu **require selecting the samples of interest** on the page, just like when adding samples to the cart. Multiple selection, enabled by default, can be enabled or disabled using the corresponding option in the context menu or via the keyboard shortcut `Ctrl+Shift+m`.
>
> Any attempt to export a file without first specifying the objects to be exported will result in a failure or in the export of an empty file.

Hierarchy tree

In the context menu of tables containing information about the **hierarchy** or the **parent** element of entries (*e.g.* in the metadata), it is possible to find a `See hierarchy tree` icon whose activation provides an interactive tree visualization of the table's elements, as shown in the video below.

Hierarchy tree

> **Note:** if the table lists more than 2000 items (rows), the `Expand view` function is automatically disabled.

Export the Result Code (RIS_CD)

Tables of quality Checks list all the analyses which have been run and imported, identifying each of them with a specific unique code, the "**RIS_CD**".

A **Result Code (RIS_CD)** (*e.g.*, "210712-10003548-4TY_cgMLST-chewbbaca") identifies an analysis run on a specific date, together with the corresponding results. **The first 6 digits of the prefix represent the execution date** of the analysis (in the "YYMMDD" format); the final part consists of the name of the analysis (*e.g.*: "4TY_cgMLST-chewbbaca").

In the various tables of *Checks* it is possible to export the selected codes as CSV or to copy them; this allows to use RIS_CD codes to select samples to run an analysis or to download the results.

> **Note:** the selection by **RIS_CD** codes allows to download or run analyses even on older results, thus giving access to **all analyses imported** into the platform. Running an analysis on the results of an older analysis, therefore requires the use of **RIS_CD** code.

The short video below shows the workflow to follow in order to define the inputs of an analysis (or the files to download) using the RIS_CD code.

Export of RIS_CD

1.3.3 Data Import / Export through files

The system offers the possibility to import and export data through sources available in the most common formats:

- `CSV`, `XLS` and `XLSX` files (which can also be exported);
- relational databases.

In order to manage the quite complex operations in a simplified way, the system is based on the configuration of specific *templates*. Each template will provide diversified access permissions and will include all information useful for the operation automation: operation type, attributes mapping, sync mode (unique key and deleting mode), etc.

During import, it is possible to operate in *merge* mode, which allows you to: - edit existing strings (identified by a unique key); - insert new strings; - remove those no longer present through (logical deletion or by updating the application status).

A complete error handling is provided, displayed in the user interface when working in interactive mode and sent by e-mail. Once the errors have been corrected, the file can be reloaded without modification, since the strings already imported will be ignored.

> **Note:** For privacy and data security reasons, some of the videos in this page were created using the Cohesive demo application; therefore, the interface may show some minor differences.

Template 1
The import and export functionalities can be accessed from the contextual menu of the CMDBuild classes, for which an import and/or export template has been configured.

Template 2

1.4 Use cases

1.4.1 Use cases

This section collects demo videos for different workflows and use cases.

The available pages are: - AMR workflow - Hybrid assembly - Management of samples and tables - FastQ metrics

1.4.2 Useful videos on sample and table management

Activation and deactivation of columns

Activate and deactivate columns

1.4.3 Useful videos on the AMR workflow

Run an AMR analysis

Run AMR analysis

Checking results and metrics of AMR analyses

Check AMR metrics

AMR report

AMR report

1.4.4 Useful videos on hybrid assembly

Run a hybrid assembly

Run hybrid assembly

1.4.5 Useful videos on FastQ metrics

Checking ONT quality metrics

Check ONT quality metrics

Checking FastQ metrics

Check FastQ metrics

2. Official samples

2.1 Management of official samples

It is hereby communicated that, in agreement with the respective National Reference Laboratories, starting from Monday, June 3rd, 2024, a new method for accessing the genomes of *L. monocytogenes*,*Campylobacter* and, from 2025, *Brucella* strains, uploaded to the platform of the National Reference Centre for Genomic Sequences of Pathogenic Microorganisms is in operation: the database and bioinformatics analyses (**GENPAT**) are available at https://genpat.izs.it.

This new system will allow, authorized users, to view the minimal metadata and the genomic characterization results of all *L. monocytogenes*, *Campylobacter* and *Brucella* strains, isolated within the national territory, in the context of official controls and other official activities as defined by Regulation (EU) 2017/625 and Legislative Decree No. 27 of February 2, 2021.

It will also be possible to export allelic profiles and assembled sequences (*de novo assembly*) for potential use within the activities and purposes of the Servizio Sanitario Nazionale (National Health Service).

2.1.1 Rules for data access

The GENPAT platform allows users with an "IZS Listeria" profile to view minimal metadata and genomic characterization results of the following sequences:

- Strain sequences of *L. monocytogenes* isolated since 01/01/2021, within the national territory, in the context of official controls and other official activities as defined by Regulation (EU) 2017/625 and Legislative Decree (D. Lgs.) No. 27 of February 2, 2021;
- Sequence quality values compliant with the accredited SOP `IZS TE B2.1.9 SOP032` ;
- Allelic call quality values compliant with the accredited SOP `IZS TE B912.1 SOP001` .

Furthermore, the platform allows users with an "IZS Campylobacter" profile to view minimal metadata and genomic characterization results of the following sequences:

- Strain sequences of *Campylobacter* isolated since 01/01/2021, within the national territory, in the context of official controls and other official activities as defined by Regulation (EU) 2017/625 and Legislative Decree (D. Lgs.) No. 27 of February 2, 2021;
- Sequence quality values compliant with the accredited SOP `IZS TE B2.1.9 SOP032` ;
- Allelic call quality values compliant with the accredited SOP `IZS TE B3.1.3 SOP071` .

Furthermore, the platform allows users with an "IZS Brucella" profile to view minimal metadata and genomic characterization results of the following sequences:

- Strain sequences of *Brucella* isolated since 01/01/2021, within the national territory, in the context of official controls and other official activities as defined by Regulation (EU) 2017/625 and Legislative Decree (D. Lgs.) No. 27 of February 2, 2021;
- Sequence quality values compliant with the accredited SOP `IZS TE B2.1.9 SOP032` ;
- Allelic call quality values compliant with the accredited SOP `IZS TE B3.1.3 SOP076` .

2.1.2 Quality of data

For each sample entered into the platform, the pipeline NgsManager reports quality data (*Quality Check* section) in the table of Samples and the quality values are verified in accordance with the above-mentioned SOPs. Main values are summarized below.

NGS (`IZS TE B2.1.9 SOP032` **)**

Parameter	Acceptable value
Number of total reads *Campylobacter sp.*	≥ 320.000
Number of total reads *L. monocytogenes*	≥ 580.000
Q-score	≥ 30

***L. monocytogenes* (** `IZS TE B912.1 SOP001` **)**

Parameter	Acceptable value
Genome lenght of *L. monocytogenes*	± 5%
Sequence Type and CC	100%
Number of cgMLST identified loci	≥ 95% (1660)

***Campylobacter* (** `IZS TE B3.1.3 SOP071` **)**

Parameter	Acceptable value
Genome lenght of *C. jejuni*	± 5%
Genome lenght of *C. coli*	± 5%
Sequence Type e CC	100%
Number of cgMLST identified loci *C. jejuni*	≥ 95% (644)
Number of wgMLST identified loci *C. jejuni*	≥ 30% (838)
Number of cgMLST identified loci *C. coli*	≥ 95% (501)
Number of wgMLST identified loci *C. coli*	≥ 30% (743)

***Brucella* (** `IZS TE B3.1.3 SOP076` **)**

Parameter	Acceptable value
Number of cgMLST identified loci *B.melitensis*	≥ 95% (2455)
Number of wgMLST identified loci *Other Brucella species*	≥ 95% (1676)
SNP analysis: horizontal coverage	≥ 97%
SNP analysis: vertical coverage	≥ 30

Furthermore, the computed species is identified based on genomic data (`Computed species` field in the *Samples* table). This value must be consistent with the value of the declared species (field `Species`).

The automated pipeline checks the quality values and the consistency between the declared "species" and "computed species"; then, depending on the correctness of these values, the pipeline sets, in the `Check` column of the "Samples" table, the result as `OK` if all values are satisfied, `Fail` or `Warning` with a specific indication of the values that are not met.

2.1.3 IZS and LAB Profiles

Each user associated with an organization (`ORG`) and authorized to manage official samples is assigned two profiles on the GENPAT platform:

- **IZS**: for official samples management;
- **LAB**: for the managemnet of samples belonging to the organization, both official and non-official.

The following table summarizes

	IZS profile	**LAB profile**
Access to samples	No organization is associated for sample filters. This allows access to all official samples from Italy (as previously defined in the section "Rules for Data Access"), collected from 2021 onwards and with the field `Check=OK` .	`ORG` has been set as the associated organization for sample filters. This allows access to all samples submitted by `ORG` , regardless of their type or data quality.
Access to metadata	The user has a limited access to the metadata of official samples from all over Italy. Furthermore, he has a restricted access to the level of detail within the metadata. It is possible, for example, to know the municipality of the sampling point but not the exact location.	The user has full access to metadata (and to the level of detail) of the samples submitted by `ORG` .
Download	It allows downloading the allelic profiles and the *de novo* assembly of the official Italian samples.	It allows downloading any analysis result available on GENPAT for samples belonging to ORG `ORG` .
Upload	`ORG` has been set as the associated organization for uploads. This allows to send sequences of samples belonging to `ORG` .	`ORG` has been set as the associated organization for upoads.
Analyses	It allows to run comparison analyses among the official Italian samples.	It allows to launch any analysis available on GENPAT for samples belonging to `ORG` .

N.B.: As an **IZS** profile, in the *Samples with anomalies* table, it is possible to view the samples sent by `ORG` that have a result other than `OK` in the `Check` column. These samples cannot be used to run analyses. It should be noted that anomalous samples belonging to other organizations are not visible in the *Samples with anomalies* table.

2.1.4 Upload of official samples

NGS sequence data (raw data)

For isolates of *L. monocytogenes* and *Campylobacter* collected from official sampling, the communication regarding the sending of the strain, or the NGS sequence data (raw data), must be carried out through the SEAP application, in the same way as it has been done up to now.

An automatic procedure ensures that the sequence data and metadata sent to SEAP are available on GENPAT platform.

SEAP will soon be integrated into the GENPAT platform. Appropriate communication will be provided as soon as this integration is completed.

Assembly data

In addition, for official laboratories that wish to upload *de novo* assembled files directly to the GENPAT platform, a dedicated upload page is available.

N.B.: This function is available subject to verification of the bioinformatics pipeline used. For this purpose, please contact the GENPAT CRN.

For details about mandatory and optional metadata, please refer to the table in the Metadata glossary.

2.1.5 Bioinformatics analyses

Automatic analyses

All samples submitted to the GENPAT platform are automatically processed through the NgsManager pipeline. In particular, for the production of *de novo* assemblies and allelic profiles, this pipeline is compatible with the OneHealth analytical pipeline provided by EFSA.

Manual analyses

The analyses available on the official samples are:

- Grapetree: it builds tree graphs using the "Minimum Spanning Tree" (MSTree) and the "Neighbor Joining" (NJ) algorithms based on the allelic profiles of the provided samples, as identified in the cgMLST analysis.
- ReporTree: similarly to *Grapetree*, it creates distance graphs among microorganisms using the allelic profiles from the cgMLST analysis. The tree is generated using the "Minimum Spanning Tree" (MSTreeV2) algorithm..
- ReporTree (SNP): this analysis is available only for *Brucella* and it creates distance graphs among microorganisms based on a SNP alignment produced with the Snippy tool, derived from mappings performed on the same genomic reference (GCF_000740415.1). The tree is build using the Minimum Spanning Tree (MSTreeV2) algorithm.
- Surveillance *Listeria monocytogenes*: currently available only for *Listeria monocytogenes*.

2.1.6 Download

With the IZS profile, it is possible to download the allelic profiles and the *de novo* assembly. To download the files, it is mandatory to provide a note specifying the reason and the intended use of these data.

2.1.7 Support

For any support request, please use the **Help desk** functions accessible through the Wiki contextual menu, located at the top right of the platform. Here you can find the detailed page with an explanatory video.

2.2 EFSA submissions

Authorized profiles can use the platform to manage the **selection** and **approval** of samples to be sent to **EFSA**.

The process takes place in several phases and involves two main groups:

- Who sends to EFSA (Reference laboratories);
- Who approves (EFSA Country Officer).

2.2.1 Submission process to EFSA (for Reference Laboratories)

1. Approval

If you have the necessary permissions, you will find the **Send for approval** option in the context menu of the **Samples** page.

Using this option, you can submit a selection of samples for approval to your **EFSA Country Officer**, who can approve or reject the submission to EFSA.

Sample submissions with the status **Under approval**, **Approved for submission**, and/or **Not approved** are listed in the respective sections, grouped under **EFSA Submissions** in the side navigation menu.

Only sample submissions with the status **Approved for submission** can actually be sent to EFSA.

Demo video on sending for approval.

2. Submit to EFSA

Within the **EFSA Submission** section of the side navigation menu, the **2 - Approved for submission** page allows you to select the sample to send and then click on **Send to EFSA**.

A modal will open asking you to enter an **authorization token** provided by EFSA.

After submission, a card is generated allowing you to monitor the status of the submission and view any reports related to the sending.

The completed submissions are collected within the **4 - Submitted to EFSA** page, always under the **EFSA Submission** section in the left navigation menu.

Submission procedure to EFSA.

2.2.2 Approval process (EFSA Country Officer)

A user with the **EFSA Country Officer** profile will receive a notification indicating the presence of samples to be approved for submission to EFSA.

By following the link in the notification, the user will access the **1 - To be approved** page, under **EFSA Submission** in the left navigation menu.

In this page, the user will find the list of samples awaiting approval; by selecting them and clicking on **Edit submission status**, he will be able to **approve** or **not approve** the request.

The following short video shows how to change the status of a submission and view the status of approvals and submissions on the platform.

Approval and sample status display.

Sample details

Within the views available in the side menu, under **EFSA Submission**, summary information on the samples is provided, such as species, submission process status and time-related information.

For more details on samples and associated metadata, please consult the **Samples** section available under **Main objects** in the side navigation menu.

On this page, the user can find a complete list of all samples (submitted to EFSA, awaiting approval, not approved, or approved) along with all related metadata; from this page, it is also possible to apply basic or advanced filters to view specific details or perform targeted exports. For detailed guides on these features, please refer to the Table features section of the wiki.

In the **Samples** section there is a predefined advanced filter **Sent to EFSA** that filters the table by showing all samples sent to EFSA. This filter is accessible by clicking the funnel icon located immediately above the table.

Submitted to EFSA filter
Filter view

2.2.3 Comparison between samples on WGS Portal and samples on the platform

This section explains how to compare samples available on the **WGS Portal** (web interface of EFSA's One-Health WGS System) with those registered on the **platform**, highlighting the different submission types managed by the two systems.

> **Note:** always apply the **Show privates** filter on WGS Portal before starting the comparison: without this filter you would also see public samples and the discrepancy would be even greater.

General differences

After enabling **Show privates**, it is normal to find a discrepancy between the number of samples shown on WGS Portal and those visible on our platform. The gap depends on the origin of the data and the flows by which samples were uploaded over time.

Possible discriminants

The portal may show samples with code `MIG-` and `EFSA-`:

- `MIG-` samples are records imported by EFSA into the **WGS Portal** system before the feeding flow was activated. These are not present in the **platform**;
- samples with the `EFSA-` prefix are those inserted into the **WGS Portal** after activation of the feeding flow. Among these there are different sequence types depending on the feeding type, as reported below.

Currently we recognize four ways samples arrived at EFSA:

1. EFSA migration

- Manual inserts performed by EFSA at the first activation of WGS Portal.
- Typical identifier: `MIG-`.
- Available only on WGS Portal (field *Code* = `MIG-***`).

2. Pre-platform submissions

- Manual submissions by data providers before the flow managed through our platform was introduced.
- Typical identifier: `EFSA-`.
- Visible on WGS Portal; not always present or aligned in GenPat.

3. EFSA Submission via platform

- Official and more recent flow, managed through our platform.
- Typical identifier: `EFSA-`.
- The full list is available on the platform within the **Sent to EFSA** page; corresponding samples are also present on WGS Portal.

4. Direct EFSA entries

- EFSA uploads samples directly to WGS Portal, without going through GenPat.
- Typical identifier: `EFSA-`.
- On WGS Portal the **Submitter** field is set to `EFSA`.

Summary table

Submission type	Description	Code	Data origin	Where to find them
EFSA migration	Records uploaded by EFSA at WGS Portal activation	`MIG-`	EFSA → WGS Portal	WGS Portal (*Code* = *MIG-*)
Pre-platform	Manual submissions by data providers prior to GenPat	`EFSA-`	Data provider → EFSA	WGS Portal
Direct EFSA entries	Samples uploaded by EFSA on WGS Portal	`EFSA-`	EFSA → WGS Portal	WGS Portal (*Submitter* = *EFSA*)
EFSA Submission via GenPat	Flow managed by the GenPat platform	`EFSA-`	Data provider → Platform → EFSA	GenPat (**Sent to EFSA**) + WGS Portal

Important: samples sent by Italian *data providers* to EFSA have an identifier that starts with the `EFSA-` prefix.

In summary

- Always check the **Submitter** field on WGS Portal to understand who uploaded the sample.
- Samples with `MIG-` come from the initial EFSA migration and are not present on the platform.
- To compare data correctly, consider the introduction date of the different flows and the submission mode associated with each code.

3. Main objects

3.1 Samples

The "**Samples**" page is one of the system's main areas and is available in the "**Main objects**" section of the "**Navigation menu**". It shows a table with the available samples on the platform along with their associated metadata which are displayed according to the user's profile and any active filters, such as tag or cart.

> **Tip** In the section dedicated to table functionalities you can find a useful guidance on how to interact with this page.

3.1.1 Cards of Samples

By double-clicking on a row in the table (or by clicking on the [+] icon if available) you can access the **detailed card** of each sample: - the top section presents the **main metadata** along with information derived from the analyses conducted on the sample; - the bottom section presents the list of the **relations** between the sample and other databases, eventual aliases and the list of **tags** it is associated with.

Some sections present a quick access link ( ) indicated by an icon placed next to the section title.

The video below briefly shows which actions you can execute on samples.

> **Note:** For privacy and data security reasons, the video was created using the Cohesive demo application. Therefore, the interface may show some minor differences.

Features of the sample table

3.1.2 Table functionalities for samples

As with the other platform main objects displayed in a tabular format, (Alias, Metadata and Analysis results), the top ribbon allows to access different options with **Samples** as well, such as:

- **filter** tools,
- **export** tools,
- direct operations on samples.

The general functioning is described in the dedicated section of the table management; more specifically, the table of Samples features certain specificities, such as the use of the **Cart** or some particular items available based on the user's profile within the context menu .

Cart

The Cart is a function available for samples only and it represents the preparatory step to interact with them. Through the cart you can:

- run analysis,
- download,
- add samples to an existing tag or create new ones.

One or more **selected samples** can be added to the Cart using the corresponding button . The arrow on the right side of the button allows you to choose among the following options:

- **Add to cart**,
- **Replace cart**,
- **Remove from cart**.

> **Important note:** Adding samples to the cart always requires the **preliminary selection** of the samples of interest. Multi selection is active by default but it can be turned on/off from the context menu option or using its key binding `Ctrl+Shift+m`. It is not possibile to add samples to the cart if none of them have been selected.

A complete guide to the cart and to all its functions is available at the Cart Wiki page page.

3.1.3 Public and private samples

Two types of data are available within the table of Samples: *Public samples* and *Private samples*.

- **Public samples** correspond to sequences deposited in NCBI databases (e.g., *Assembly*, *Nucleotide*, *SRA*, etc.) and can be imported into the platform through the procedure of Upload da NCBI.
- **Private samples** are sequences not available in public databases, uploaded to the platform either through automated procedures following sequencing or via Manual upload methods.

To make things easier, two preconfigured filters are available and allow you to display **only public samples** or **only private samples**. It is also possible to set one of the two filters as the default, so it will be active every time you access the platform, based on your preferences. The following video shows the use of filters.

Filters for public and private samples
In the section of this wiki dedicated to table functionalities you can find further information about the use of the basic filters and of the advanced filters.

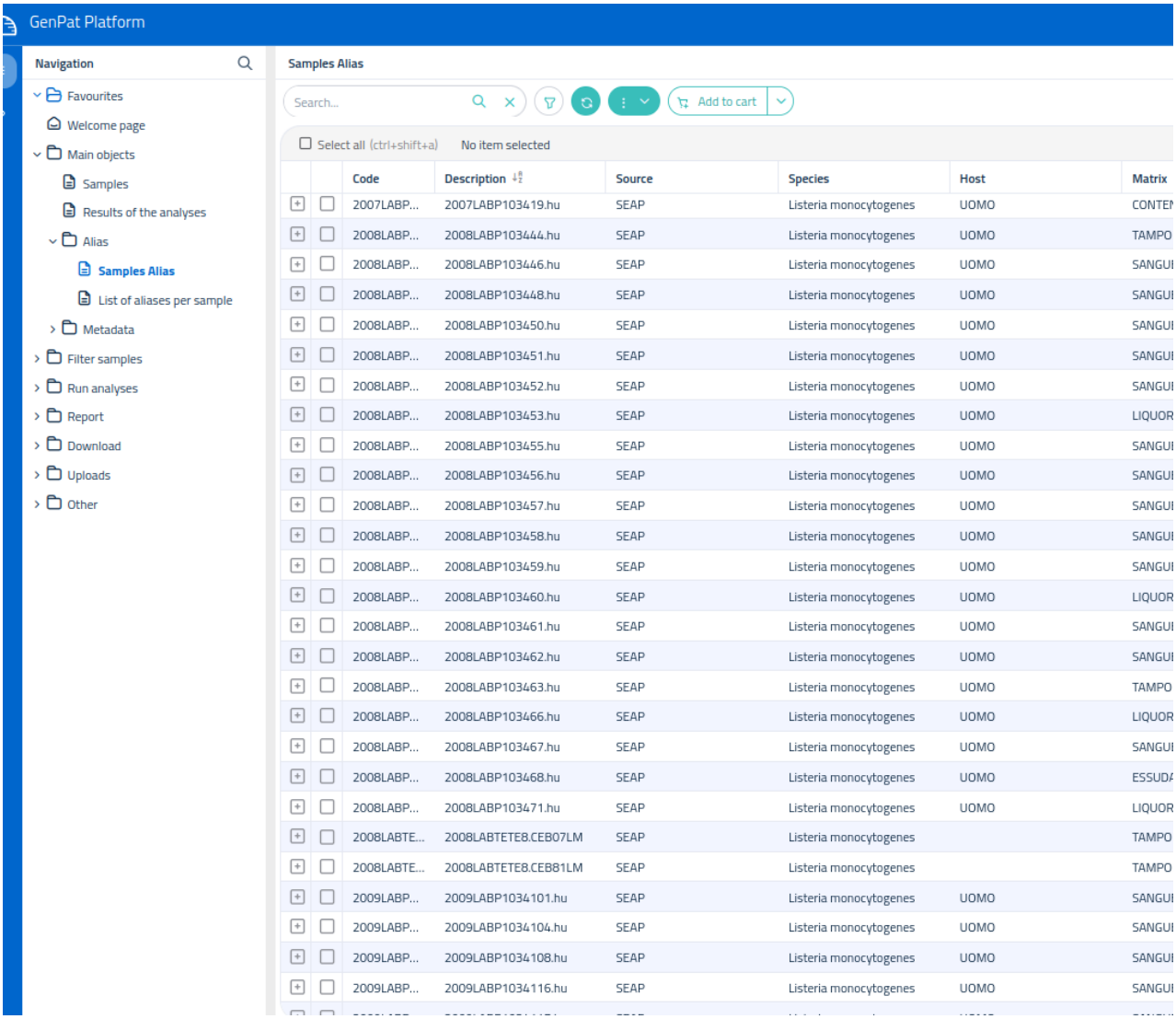
3.2 Alias

In the **Main Objects** section of the **Navigation** menu is also available the group of **Alias** functionalities, concerning the relations and correspondences among codes referring to the same samples, such as SEAP and SILAB codes or the original codes of external samples uploaded to the platform (e.g., from ENA or NCBI databases).

3.2.1 Interface of Alias of samples

In the **Alias of samples** page, a table displays the sample codes in the platform along with the corresponding original codes before their import (**Identification** column).

As with the **Samples** page, the *entries* for the aliases are also organized using a table system; therefore, the user has access, for the Alias, to the same functionalities and interaction methods available for samples.

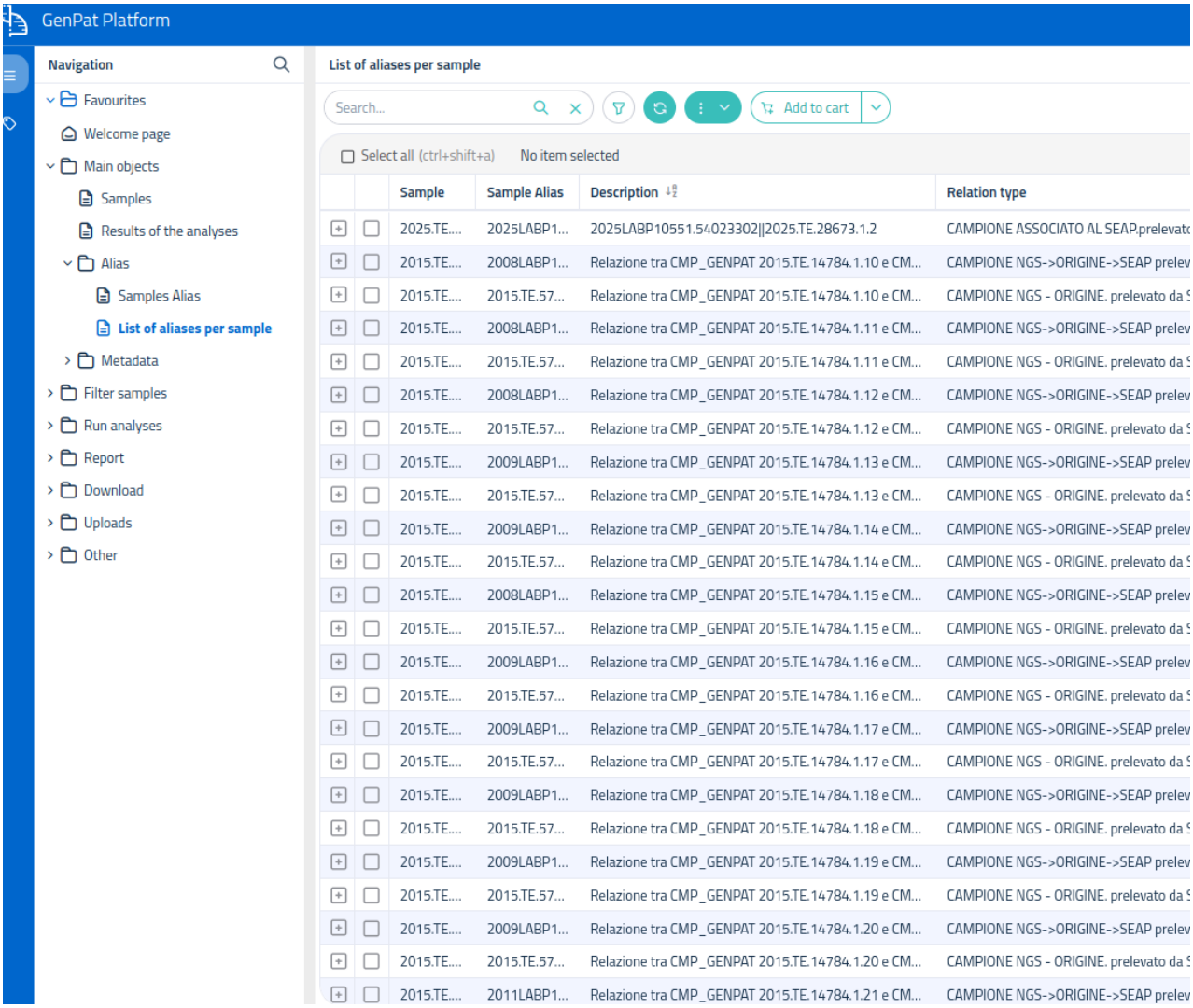


Concerning Alias table, the columns you can view are fewer than those available in the *Samples* table, as they include only the sample codes (Code and original Identification) and other metadata associated with the sample.

3.2.2 List of aliases per sample

This page includes, within the table, the codes of the platform's *entries* and their corresponding relations in the database with all their Alias codes, together with a relation description field. The **List of aliases per sample** is designed so that users can associate the platform code with the original code (and *vice-versa*), they can track back to the origin of the sample and to its import into the platform from the external database.

In this case as well, all table features are available for the user.



3.3 Metadata

In the "Main objects" section of the side navigation menu, the set of features related to **Metadata** (Wikipedia link) is available.

Metadata on the platform are divided into the following categories:

- **Species**: list of organism species, together with their corresponding identification codes. The category is further divided into:
  - **Pathogenic species**;
  - **Host Species**
- **Pathogenic materials**: type of origin of the sample (culture, genetic material, or sequence file) with the corresponding identification codes.
- **Sampling Point**: list of places where the sample was originally collected (names of entities, companies, hospitals and/or latitude and longitude coordinates), together with their corresponding identification codes.

Metadata are essential for uploading a new sample to the platform. Please, refer to the **Upload** section for more details.

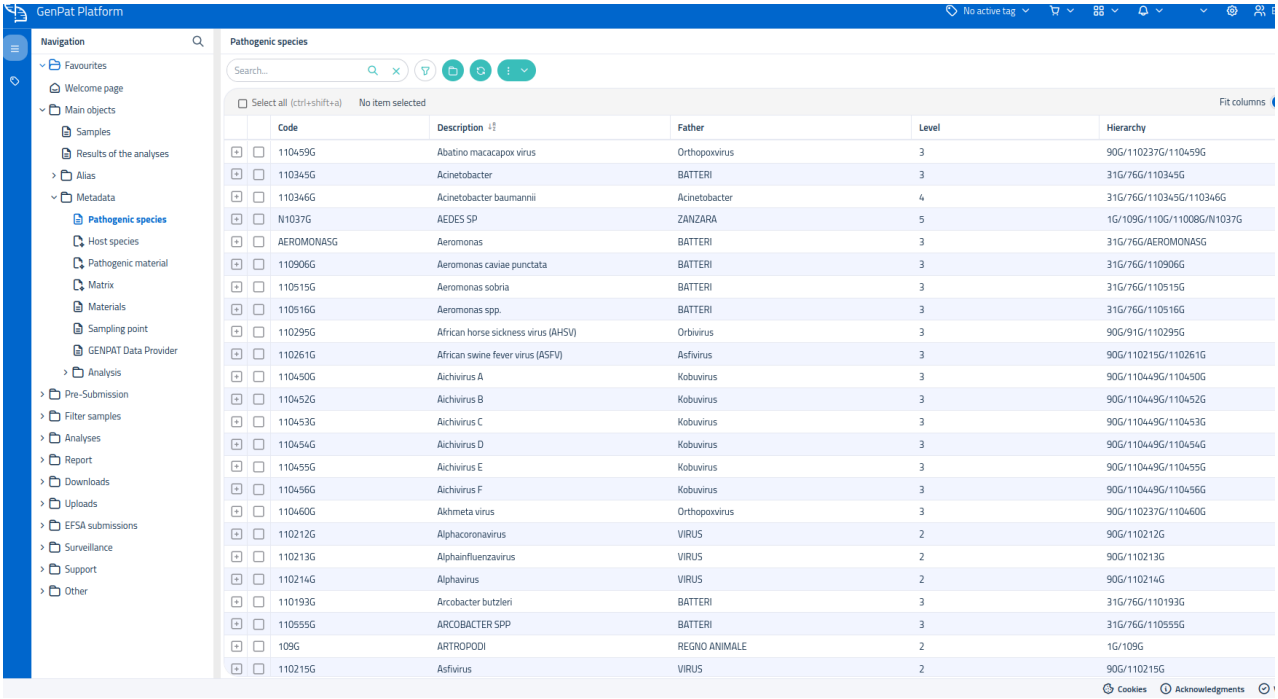
The metadata have hierarchical trees that can be accessed through the contextual menu available at the level of individual categories.

3.3.1 Codings according to EFSA standards

On the platform, to ensure uniform coding for external users, sample metadata comply with the specifications EFSA FOODEX2.

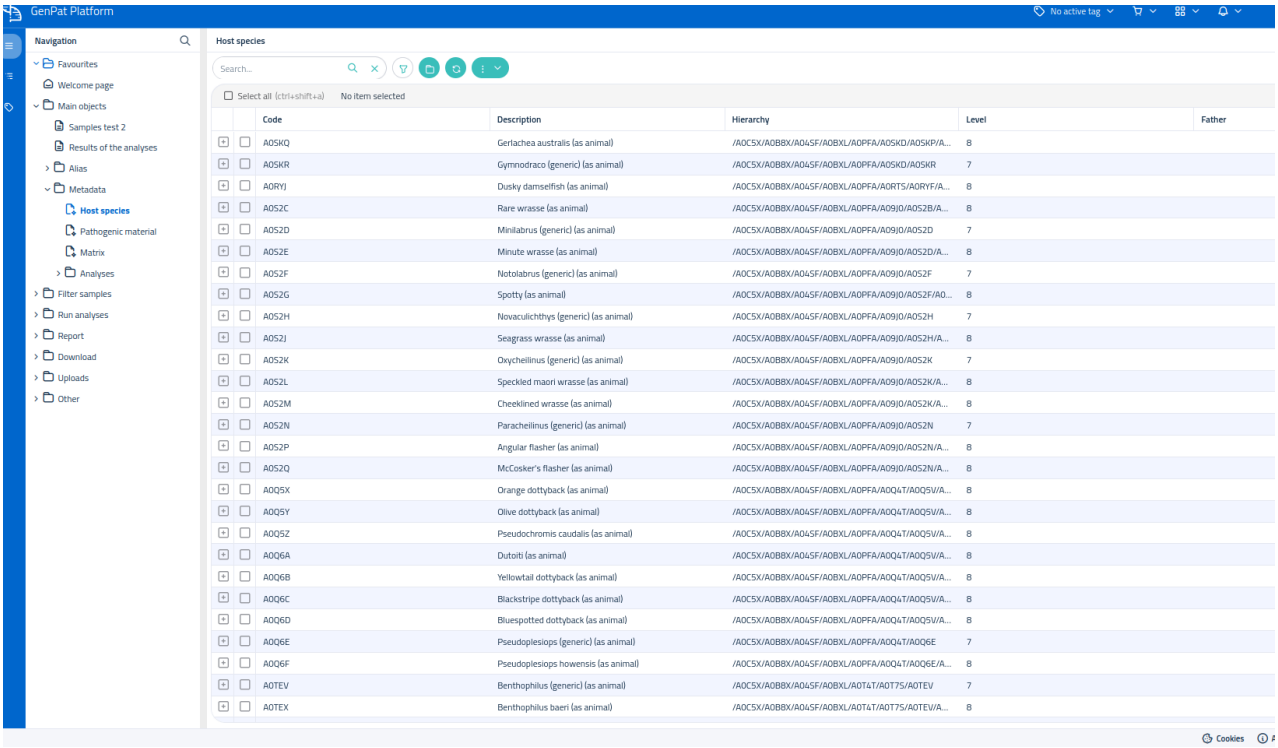
3.3.2 Pathogenic species

The *Pathogenic species* table lists the pathogen species and the corresponding codes used to identify them in the platform. The available operations are the same as those in the samples table.



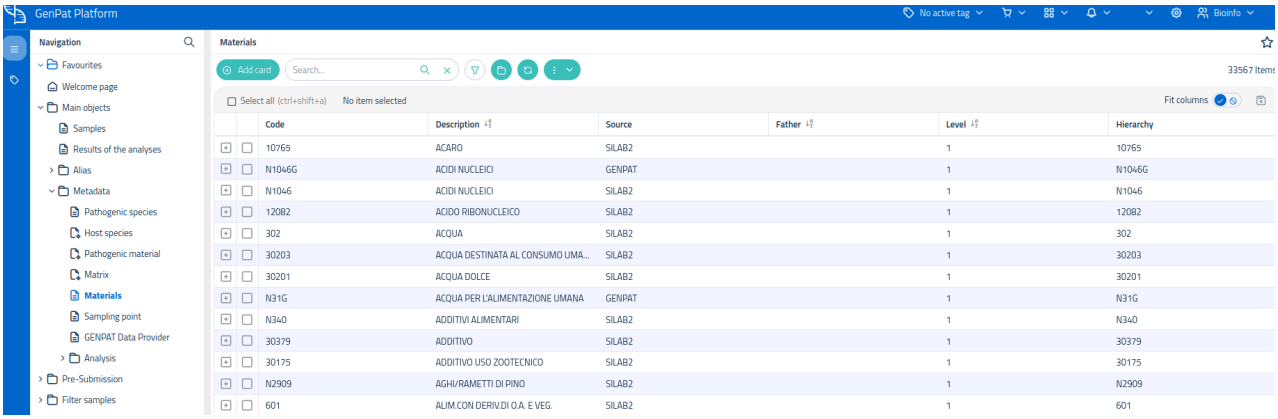
3.3.3 Host species

The *Host Species* table lists the species of host organisms and their corresponding codes. The list of "host" organisms refers to those whose reference genomes have been added to the platform. The available operations are the same as those in the samples table.



3.3.4 Pathogenic material

In the *Pathogenic material* table the type of origin of the sample is reported, wheher it is a culture, genetic material or a sequence file, along with the corresponding codes. The available operations are the same as in the samples table.



3.3.5 Sampling Point

The *Sampling point* table lists the codes of the locations where the initial collection or retrieval of the material, from which a sample was isolated, took place. The "sampling point" includes the province, municipality, any address and the facility or organization. The basic operations available on this table are the same as those possible on thesample. However, for certain authorized profiles, it is also possible to independently add sampling points from external sources which, once selected, will become part of the platform's sampling points.

Add sampling point

The external sources currently available on the platform are:

- Companies registered in the National Database (BDN);
- Recognized or registered Italian establishments.

To perform a search in both external sources, it is necessary to have a code (**company code** for BDN and **Approval Number** for the other organisms). The functions for adding a sampling point are available, for authorized users, during the preparation and upload of metadata. Please, find here below a demonstration video

Upload Preparation page

3.4 Analysis results

You can access this page from the **Main objects** of the side navigation menu of the platform. It lists, in a tabular view, all the results of the bioinformatics analyses in the platform that have already been performed on the individual samples and that are available, as input, for the potential launch of new analyses. The bioinformatics analyses are marked by the *source BIOINFONAS*.

Note: being able to check the presence of analysis results on a sample in the platform is essential to know whether all the files required for downstream analyses are available; indeed, these analyses may in turn require, as input, the results obtained from previous analyses. Please find here below two links to the analysis launch documentation.

- How to run analyses
- Introduction to the tools

The available interaction options with the table and its functions are the same as for table of samples.

3.4.1 Report error

If the same analysis has been run multiple times on the same sample, the system will automatically use the most recent imported results as input for downstream analyses, whenever required.

If the results of an analysis are not good for usage, it is possible to mark that analysis as an error-carrying run, in order to prevent that the system uses it as input for further analyses. Regarding this, in the page context menu you can find the **Report error** option that allows you to mark the selected items in the table as errors.

Important: when marking an analysis as an error, the system will automatically mark as errors all other analyses that used as input the results of the marked one. **Example:** if an assembly obtained from 2AS_denovo_shovill is reported as incorrect, the system will automatically invalidate any typing analyses (MLST, cgMLST, etc.) that were calculated from that assembly.

The following video briefly shows how to **report an error**:

Report an error

4. Filter samples

4.1 Tags

4.1.1 Tag

What is a tag

Tags are a platform feature that allows grouping samples into collections. Creating a tag is useful for samples related to specific projects or outbreaks. If samples belong to a tag, as long as it remains active, only the samples associated with it can be displayed in the platform tables.

> For example, if samples related to a hypothetical outbreak XXX have been analyzed, it is possible to create a tag "outbreak_XXX" to associate the samples. This allows quick access to the samples every time it is necessary to run an analysis, avoiding the process of searching and selecting each sample before the analysis.
>
> A tag corresponding to a collection of samples can therefore be activated to view samples or results related to them, but also to launch a new analysis or pipeline on all samples belonging to the tag, specifying only the desired tag instead of all samples as analysis input.

The short video below shows tag activation. For in-depth information about tags, refer to the Manage Tags page.

Activate a tag

Tag menu

The tag menu is accessible from the sidebar by clicking on the "paperclip"-shaped icon, located below the navigation icon. The menu works like a filesystem, with a tree structure that allows files to be displayed hierarchically. By right-clicking on the various items, you can operate on individual files as well as on folders of files, including: copying a tag name, adding it to favorites, activating it, renaming it, managing it, or removing it.

The *functions* allowed by the tag menu are:

- Add a **New tag** using the `+` button at the top right. This option allows the user to name the tag, choose the parent tag folder, and decide whether to allow its visibility to their group;
- Add a **New folder** using the `+` button at the top right. This option works exactly like adding a new tag;
- Choose the folder in which to save the tag of interest;
- Add or remove the tag from favorites;
- Move a tag of interest from one folder to another by dragging and dropping it;
- Decide who is allowed to view it: if you are *the user who created the new tag*, you can choose whether to make it visible to your group or not.

The following video shows the functions available in the tag menu.

Tag menu and its functionalities

4.1.2 Manage tags

This guide illustrates, through a series of video tutorials, the operations available for tags.

The **Manage tags** page can be accessed either from the side navigation menu by clicking on the **Filter Samples** section, or by selecting the **Activate tag** option from the dropdown menu that appears when clicking *No active tag* in the top bar. On this page, you can find information about the active tag (if any); you can also deactivate the current tag or activate an existing one.

The following video briefly shows the contents of the page: in the **Manage tags** section you will find 4 expandable elements ("cards"); each card is dedicated to one of the main operations available for tags: **Activate tag**, **Create tag**, **Edit tag** and **Remove tag**.

Manage tags

Activate tag

The first card provides access to tag activation: by clicking on "Select a tag to activate", it will be possible to choose and select a tag from those available, listed in a popup window. Once the tag is selected, it will be activated by clicking the "Activate tag" button.

Activate tag

Most used, Last used, Favorites

The **Most used**, **Last used** and **Favorites** cards allow you to quickly select tags to activate according to your preferences. The **Most used** and **Last used** cards are automatically populated with a list of **5 tags** you use most and a list of the last **5 tags** you activated, respectively.

The **Favorites** list, on the other hand, depends on direct user input and has no limit on the number of selectable tags. However, it's worth remembering that adding too many would be like having none at all! The following video provides a brief demonstration of how to add and remove tags from the favorites list.

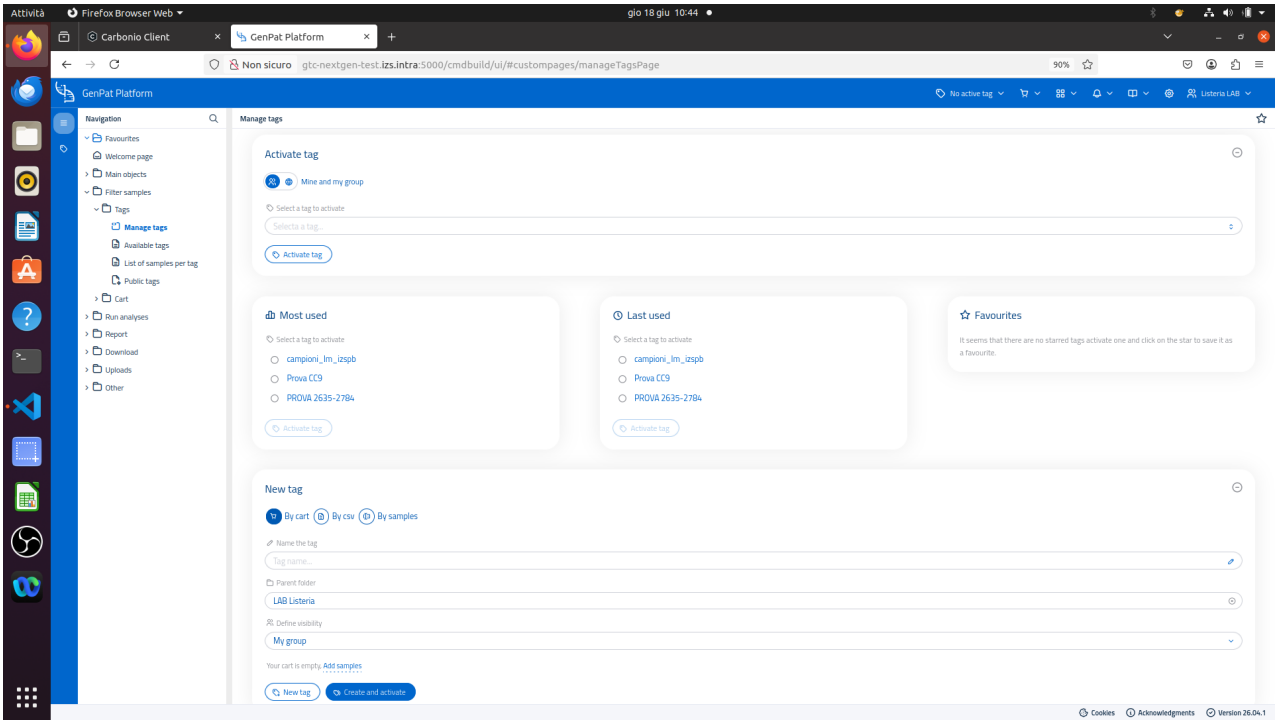
Manage Favorites

Create tag

The **Create tag** card offers 3 different ways to create a new tag: you can use a collection of samples from the cart, a CSV file, or sample codes (in the latter case, sample codes are written or pasted as a comma-separated list directly in the "Sample codes" field). Selecting one of the options will update the card, which will show the specific fields for the chosen option.

Create tag

> For all options and methods for creating a tag, the **Visibility** option is available. This feature allows you to specify whether you want to save the tag only for personal view and use, or make it available to the user's group as well.



Edit tag samples

The **Edit tag** card allows you to modify an existing tag by adding or removing samples: just like **Create tag**, here too you can choose between cart, CSV or list of codes. You can also modify the visibility of the selected tag and remove all its contents without deleting the tag.

Edit tag

Remove tag

The last card (**Remove tag**) is meant to allow deletion of a selected tag. As with the **Activate tag** section, clicking on the search field or the "Select tag" button (on the right) allows you to select the desired tag; to delete it, click the red "Remove tag" button.

Remove tag

4.2 Cart

4.2.1 Introduction

The cart is a highly flexible function, integrated into the platform. Its operation is similar to the typical shopping cart on e-commerce sites but it includes some specific features.

The cart is intended to **temporarily** store a selection of samples which can be stacked or removed in order to be used for different purposes: filtering the platform, running the analyses, execute download operations, viewing reports, creating tags.

For any further detailed information about the cart function, please refer to the dedicated Manage Tags page. For analysis, download and report generation, please refer to the corresponding sections of the documentation as well. This page describes how to add samples to the cart and how activate it in order to filter the platform.

Adding samples to Cart

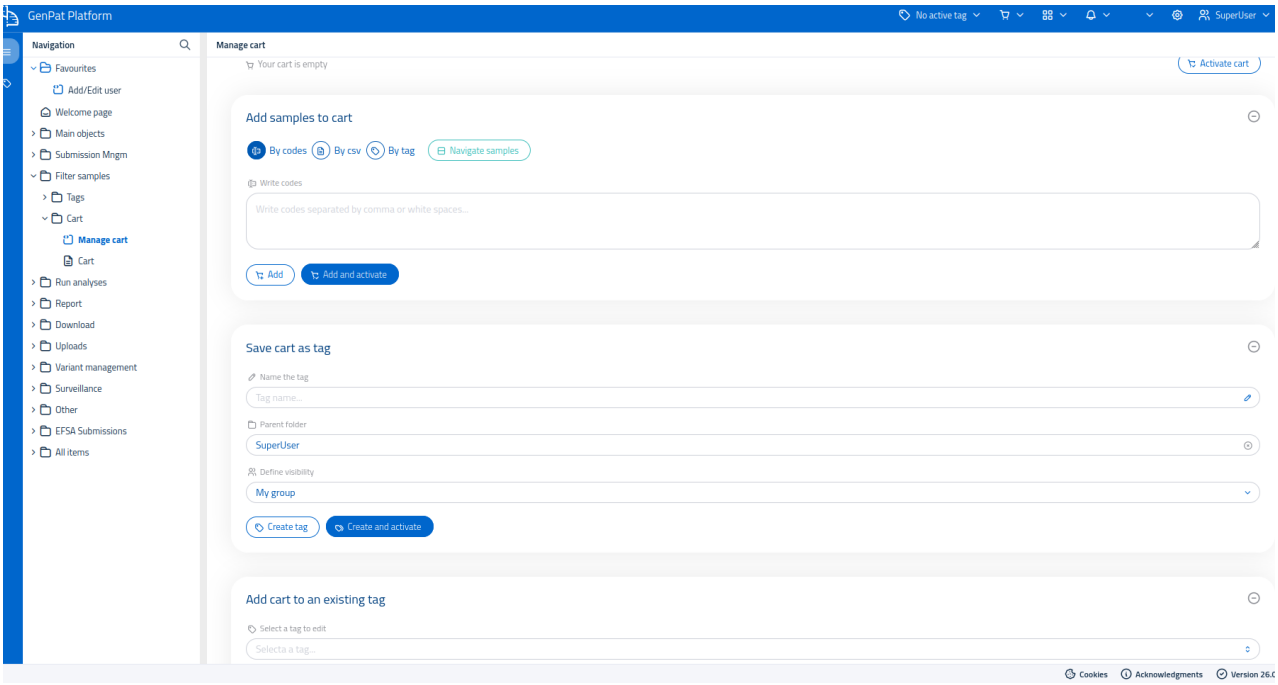
The following video briefly shows how to add **entries** to the cart.

Add samples to cart

4.2.2 Manage Cart

The cart function in this platform works similarly to the cart of an e-commerce site; it can be accessed from the navigation menu on the left (**Filter samples**), where both cart access and cart management functions are available. The cart can also be accessed from the icon on the right side of the top bar, which acts as a shortcut to the **Manage cart** page.

The **Manage cart** page offers two quick action buttons at the top right corner: **Add samples** (to manually add samples from the samples' Table), and **Empty cart**.



The cart's three main functions are grouped into three expandable *cards*: - **Add samples to cart**, - **Save cart as tag** - **Add cart to an existing tag**.

Add samples to cart

From this card it's possible to edit the cart's contents by adding samples in three different ways: from a saved *tag*, from a *csv* file or from *codes* (*i.e.* directly typing or pasting sample codes in the "*Write codes*" field). The selection of one of these modes will update the card, which will show the options and fields available for the selected mode.

The following video briefly displays the steps of this phase.

Add samples to cart

Save cart as tag

This card will allow you to create a new tag directly from the samples contained in the cart. For any further information about tags, please visit the dedicated section or dive deeper into the new features available for tags in the Manage tags page.

The following short video displays how to save a cart as tag.

Save cart as a tag

Add cart to an existing tag

The last expandable card allows the user to add samples currently in the cart to an existing tag. To know more about tags, please visit the dedicated section or check the tags functionalities in the Manage tags page.

The following video shows how to add samples in the cart to a specified tag.

Add cart to an existing tag

4.2.3 Cart activation

The Cart is used for the *temporary* storage of samples; moreover, by being "activated", it also acts as a filter function.

The **Manage cart** page includes an activation button (at the top right); once clicked, the samples in the cart will be used as filtering terms across the platform.

The **Activate cart** function allows you to filter the platform using a selection of samples, enabling operations on them (such as running analyses, requesting downloads or viewing reports) without the need to activate a tag.

Cart activation (as an alternative to tag activation) is required to run an analysis from the cart. The cart will be automatically activated if selected as the source for sample collection for the analysis.

> Cart activation overrides any active tag and its deactivation brings the platform back to the previous status (i.e.**if a tag was active before cart activation, than it will be reactivated**).

The filtering action performed by the cart has two fundamental differences compared to that of the tag:

- The cart is **transient**: it is not saved and can be easily and instantly modified by adding and removing samples;
- When a cart is active, the whole **platform** is filtered based on the samples in the cart; this means that, in addition to sample lists, the contents of **View analyses**, **View downloads** and **View imports** are also filtered, making it easier and faster to consult the records of operations and analyses performed on specific entries, without the need to search or filter them from the entire dataset.

The following video shows how to activate and deactivate the cart and how these actions impact the platform.

Cart activation

5. Analyses

5.1 Introduction

The page for launching analyses can be accessed from the side navigation menu (**Analysis > Run analysis**), through the shortcut on the top-right function bar (button) or directly from the **Welcome page**.

5.1.1 Run analyses

The system's core functionalities include a simplified workflow to run analyses on desired samples, whose interface is shared with the download procedure.

The *run analysis* procedure is guided and organized in 3 phases: step **1** (**Samples**) requires to specify a source from which to retrieve the list of samples to be analyzed or those associated with the files needed for the analysis; step **2** (**Tools**) leads to the analysis selection screen; step **3** (**Inputs**) is intended to allow the choice of the source of the files that will be used as input.

Detailed information about the 3 phases of the launch analysis system are available at the following links:

1. Selection of samples (**Samples**);
2. Choice of the analysis type (**Tools**);
3. Selection of the input (**Inputs**).

The following video shows all the phases of the analysis launch procedure.

> **Note:** For privacy and data security reasons, the video was created using the Cohesive demo application; therefore, the interface may show some minor differences.

Phases of the run analysis procedure

> **Note 1:** launching an analysis from the **cart** will automatically trigger its activation. For further information on cart activation, please refer to the dedicated section activate the cart.
>
> **Note 2:** for launching an analysis, it is also possible to use one or more **RIS_CD** codes via the dedicated button at the top-right corner. Using the RIS_CD code is useful to select univocally the results of a specific run, providing easy access to older analysis results as well.

Further information on the use of RIS_CD for launching analyses and for Download functions is available in the "Export of the Result Code (RIS_CD)" section of this wiki.

5.1.2 Preference selection and saving

When selecting tools, you can manage your preferences with a dedicated system that includes:

- the ability to mark tools as favorite, on the base of your *role*;
- a *Favorites* category that shows all the tools you've starred.

Preferences selection and saving

5.1.3 Check analysis

Once the analysis is completed, it is possible to view its card in "**Check analysis**", as shown in the previous video. The page *Check analyses* lists the runs launched by the user along with useful information such as the type of analysis, the tool used and the processus status. From here, it will also be possible to replay the analysis or import the analysis if it was succesfully completed. For more information, please refer to the dedicated page.

5.2 Run analysis steps

The analysis execution system is logically divided into three phases:

1. **Samples** (sample selection);
2. **Tools** (selection of the analysis type and bioinformatics tools);
3. **Inputs** (input selection).

A quick guide to launching an analysis is available on the introduction to "Run analyses" Wiki page which also includes a video demonstrating the entire procedure.

> **Note:** For privacy and data security reasons, some of the videos in this page were created using the Cohesive demo application; therefore, the interface may show some minor differences.

This wiki section presents the 3 steps: *Samples*, *Tools* and *Inputs*.

5.2.1 Step 1: Samples

In order to proceed with the selection of the analysis, it is first necessary to provide the system with a collection of sample codes. This step is necessary to prepare the launch of an analysis on files associated with each of the specified samples (for example, *raw reads* of the samples in the case of trimming or fasta assemblies in the case of cgMLST with chewBBACA). The selected analysis will be executed for each specified code.

The user can choose either the **cart** or a **tag** to select the codes. A tag is a static collection, containing sample codes, defined and saved by the user. The cart allows a transient collection of codes, defined by the user, which can be enabled, as a filtering function, through the cart activation. If the user chooses to provide the code collection via the cart, the cart is automatically activated, acquiring priority over any active tag. Deactivating the cart restores the platform to its prior state before its activation.

When the cart is empty or no tag is active, it is possible to click the appropriated link to select the tag (*Activate tag* button) or to populate the cart (*Manage cart* button) from the designated pages.

A functionality is also available for quickly populating the cart using a list of codes or a CSV file.

The short video below shows the execution of the code selection phase via tag.

Code selection via tag

5.2.2 Step 2: Tools

The *Tools* step provides the user with an interface for selecting the analysis (or the pipeline) to be performed on the previously selected samples and, subsequently, the bioinformatics tool to be used for carrying out the analysis. The types of analysis available on the platform are displayed as boxes showing their name and typology. However, they can also be filtered by category, viewed in a tree layout or searched using the dedicated search field.

After selecting the analysis, the user is presented with a confirmation screen that provides useful information about the chosen analysis and the selected bioinformatics program. On this page, the user can select the bioinformatics tool (also referred to as the *method*) from the drop-down menu in the *Select method* section, if two or more bioinformatics tools - that can perform the same analysis - are available on the platform.

The short video below shows how to select the analysis and the tool.

Selection of analysis and tool

Preferences selection and saving

When selecting tools, you can manage your preferences with a dedicated system that includes:

- the ability to mark tools as favorite, on the base of your *role*;
- a *Favorites* category that shows all the tools you've starred.

Preferences selection and saving

5.2.3 Step 3: Inputs

The final step required to complete the run analysis process is the definition of the inputs. In the first step (*Samples*), the codes of the samples of interest were provided, but it is still necessary to specify which file, among those associated with each of these samples, should be used for the selected analysis; for example, an analysis such as "trimming" uses the raw reads of samples as input while clustering with Grapetree requires the tabular files generated by cgMLST. In the first case, the FASTQ files of the reads, from the samples entered in step 1, will be selected whereas, in the second case, the corresponding allele profile files generated by chewBBACA will be selected.

> This step therefore requires that the necessary files to launch an analysis be previously generated and imported, (this means that, for example, it would not be possible to run GrapeTree on samples for which chewBBACA has not been run beforehand).

The fields of the interface in this final step are already pre-filled with the standard input for the selected bioinformatics tool; therefore, user intervention will only be necessary if more than one suitable input is available. From the same interface, user will be able to specify certain options, if available, and, for some analyses, to access the advanced lauch mode.

The short video below shows how to select inputs for an analysis run.

Input selection

Inputs: advanced mode

Samples in the platform can be identified by two strings: - `.TE.` (sequenced samples); - `.EXT.` (samples manually imported or imported from an external database).

Some bioinformatics tools in the platform need to receive a specific instruction, in order to use files from both types as input.

To allow this instruction to be provided, the third and final phase of the analysis launch system provides a switch to move from the "**Basic mode**" for the input selection to the "**Advanced mode**" which enables the input selection through checkboxes.

The following short video shows the use of the *Advanced mode* for input selection on one of the tools for which it is available: the CFSAN pipeline.

Advanced mode for input selection

Selection of multiple references

Some analyses (such as mapping performed with **Ivar**, or the pipelines that include it, such as Mapping of segmented virus) provide the possibility of selecting more than one reference file. The reference selection is performed by opening the pop-up window of the reference genomes, through which it is possible to choose any number of files by ticking the corresponding checkboxes. Repeating this process will not overwrite the previous selected references: they will simply be added to the list.

The list of selected references can be cleared (**Clear** button); moreover, the references in the list can be removed individually through the icon.

> **Note:** the order of references in the list affects the final result of the multi-reference mapping. Please, place the main reference genomes first.

The following video shows how to interact with the options for selecting multiple references.

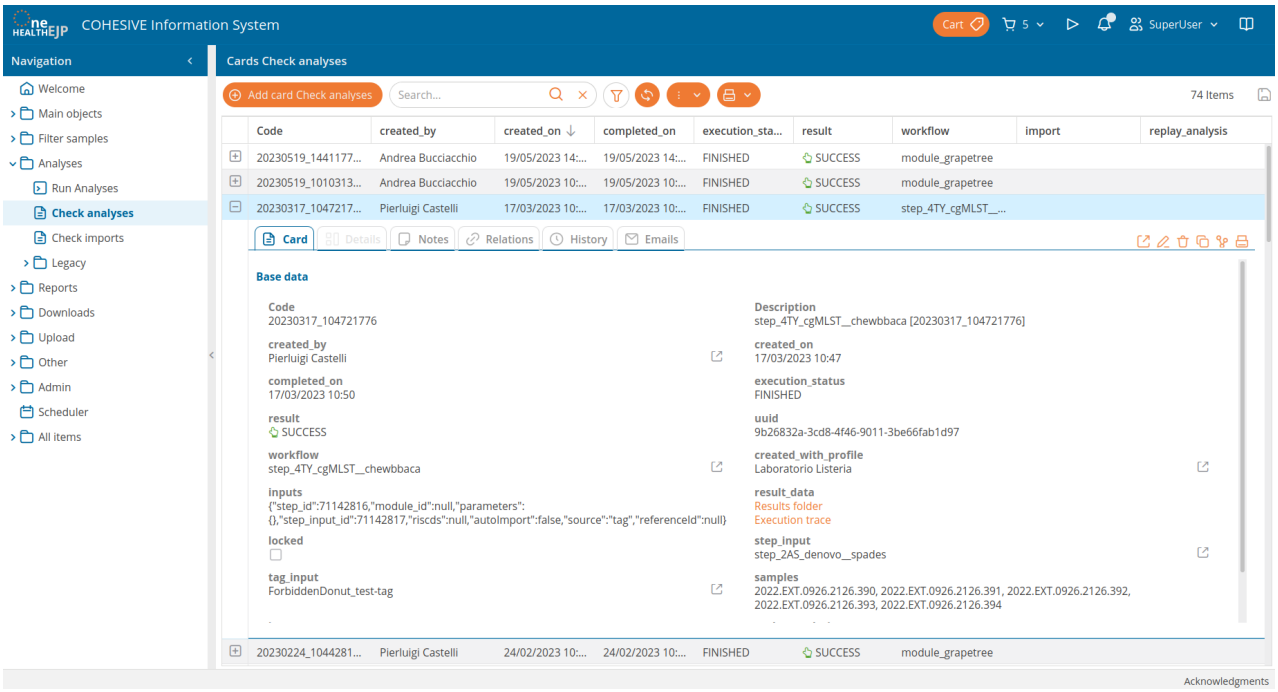
Multiple reference selection options

5.3 Check analyses

Launched analyses are listed on the **Launched pipelines** page, accessible from the **Analyses** section in the side navigation menu. Each analysis is associated with an expandable card which can be opened using the "+" button located to the left side of the analysis code.

5.3.1 Analysis card

The analysis card lists information about the executed run. If the process has successfully completed, you can access technical information about the process, the used inputs and the functions "Replay analysis" and "Import" of the results. Additionally, in the card you will also find a link to the **Result folder**, within the *Result data* section.



Result folder

An analysis results can be viewed through a system that allows you to browse and search the generated files, without leaving the platform. Some download tools are available to allow you multiple selection and search, as well as the option to download either a single generated file or a folder containing multiple files. The following video briefly shows the various operations made available by these functions.

Results folder

Note The maximum size for downloading multiple files, through `All`, `Search` and `Selected` options, is currently **20GB**. The size indicator, inside the button, turns red when allowed limit is exceeded. The download of a single resource, however, has no limit.

5.3.2 Import of analyses

Once an analysis has been successfully completed, the generated files are stored in a temporary folder, where they will remain for 45 days. During this period, users will be able to access the results folder, as previously mentioned, from the *Result data* section of the analysis card; alternatively, the results can be downloaded using the download function of the platform that allows manual download and provides the `wget` command useful for the downloading from the terminal.

After 45 days, the temporary folder will be removed from the servers. **To ensure that the analysis results are saved permanently, they must be imported**.

> In most cases, **launching an analysis that needs, as input, the files produced by another analysis, requires that these files have been previously imported**.

To import results, simply click the **Import** button, located at the botton left of the analysis card, as briefly shown in the video below. The system will notify the user once the import is complete.

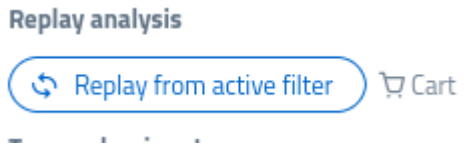
Import of analysis

> Please, be aware that certain analyses may not be eligible for import.

The list and progress of imports are available on the **View imports** page, in the **Analyses** section, of the side navigation menu.

5.3.3 Replay analysis

At the bottom of a completed analysis card, it is possible to access the **Replay analysis** option. This feature allows you to repeat analyses that have already been successfully performed, maintaining the same previous parameters and options. The new set of samples can be provided with the active cart or a active tag.



The following video shows how to replay an analysis.

> **Note:** For privacy and data security reasons, the video was created using the Cohesive demo application; therefore, the interface may show some minor differences.

Replaying an analysis

5.3.4 Keep analysis

Within the analysis card, there's the **Keep analysis** option, which allows you to choose whether to retain indefinitely a non-importable analysis, preventing its related data from being removed from the system once the 45-day time limit has expired.

Keep analysis

5.4 Tool Descriptions

5.4.1 Introduction

The analyses of bacterial or viral isolates available on the platform are organized into categories. The names of these groups follow a nomenclature designed to describe type of analysis and its execution level.

Nomenclature system

Prefixes

The categories of an analysis and its execution level are summarized by a short code, used as a prefix to the analysis name. Below is a summary table of these prefixes:

1PP pre-processing analysis
2AS assembly tools
2MG metagenomic analysis
3TX taxonomic classification (*taxa*)
4TY *in silico* typing
4AN genome annotation

The first element of the prefix is a number indicating the usual execution level: the **pre-processing** is normally performed before any other analysis and is therefore assigned level **1**; **taxonomic analyses** require files from assembly as input and are therefore placed at level **3**, after pre-processing and assembly.

Besides analyses that can be launched by users, there is an additional class, called **0SQ**, which identifies **S**equence **Q**uality checks automatically run on new reads added to the plaform.

Names of the analyses

The name of the analysis follows the code used as a prefix and describes the type of data manipulation that will be performed by the bioinformatic tools available for this purpose.

The *trimming* analysis, for example, is classified as a pre-processing analysis; therefore, its full name is **1PP_trimming**. Similarly, running a *de novo* assembly will require the use of the **2AS_denovo** analysis which will retain the same name, regardless of the fact that it will later be possible to choose among multiple bioinformatic tools.

In the sections below, all analyses belonging to each group available in the platform's run analysis feature are listed in dedicated tables, along with a brief description and links to their respective pages.

Suffixes

Many of the analyses available on the platform can be performed using multiple alternative software tools, each with its own characteristics and differences compared to the others. These programs are referred to as bioinformatic "tools" or "methods".

The name of an analysis is then completed using, as suffix, 2 underscore characters, ("_"), followed by the name of the selected tool; for example, "2AS_denovo__spades" and "2AS_denovo__unicycler" both perform *de novo* assembly; however, the former uses the software "spades" while the latter uses "unicycler".

Analyses that can be performed by multiple software tools will allow the selection of the method to be used via a drop-down menu, during the second step ("Tools") of the Run Analysis procedure.

Available tools for each analysis are listed in the respective wiki pages dedicated to the individual analyses.

NOTE: in most cases, long reads generated by Nanopore and Iontorrent systems and by Illumina long reads devices must be processed differently compared to short reads. Some bioinformatic available tools are therefore specific to perform a given analysis **exclusively** on long reads; there may also be entire analyses dedicated solely to the processing of such files.

The Bioinformatics tools for long reads page allows you to consult the lists of these analyses.

Single sample analysis

Prefix	Analysis name	Description	Tools
1PP	trimming	removal of low quality residues from raw reads	trimmomatic fastp chopper
	hostdepl	depletion of host sequences by mapping against host genome	Bowtie minimap2
	filtering	retain reads matching organism of interest	Bowtie minimap2
	downsampling	reduction of excessive coverage	BBnorm
	generated	FASTA to FASTQ conversion with dummy qualities	fasta2fastq
2AS	denovo	de novo genome assembly	SPAdes unicycler Shovill flye PlasmidSPAdes
	mapping	reference-based mapping	Bowtie ivar Snippy Medaka
	hybrid	hybrid short/long read assembly	unicycler
	indexing	Bowtie2 indexing	Bowtie2
2MG	denovo	metagenomic assembly	metaSPAdes
3TX	class	taxonomic classification	kraken kraken2 ConFindr Centrifuge kmerfinder
	species	closest species identification	blast abricate mash
4TY	MLST	Multi-Locus Sequence Typing	mlst
	cgMLST	core genome MLST	chewBBACA
	flaA	Campylobacter flaA typing	flaA
	lineage	SARS-CoV-2 lineage West Nile Virus lineage	Pangolin westnile
	wgMLST	whole genome MLST	chewBBACA
	plasmid	plasmid typing	MOB-suite
	ML	ML-based source attribution	Source Classifier
	serotype	serotype prediction	seqsero / dengue
4AN	genes	genome annotation	Prokka
	AMR	prediction of antibiotic resistance genes	abricate blast staramr ResFinder

filtering

Multi sample analysis

Analysis type	Analysis name / Tool	Description
Gene-by-gene based clustering	Grapetree	MST and NJ tree construction
	Reportree	MST construction and geoJSON generation
	Augur	Nextstrain phylogenetic analysis
	Surveillance	manual ReporTree execution
Pangenome extraction	Panaroo	accessory gene presence/absence matrix
	Snippy-core	core SNP identification
SNP-based clustering	CFSAN	reference-based SNP phylogeny
	kSNP3	reference-free SNP phylogeny
	VCF2MST	MST construction from VCF

Pipelines

In addition to the launch of individual analyses, the platform also delivers automatic pipelines, consisting of analyses already linked together. Pipelines are designed to make the execution of frequently used workflows simpler and faster.

In the table below, the individual analyses that make up each pipeline are listed ("Analysis" column; for information on the individual software used for the analyses in the pipelines, please refer to the corresponding Wiki pages).

Pipeline Name	**Description**	**Analysis**
Covid Emergency	fast assembly of SARS-CoV2 samples and lineage assignment	2AS_mapping + 4TY_lineage
Depletion & *de novo*	depletion of host reads from trimmed reads and subsequent *de novo* assembly	1PP_hostdepl + 2AS_denovo
Genome Draft	mapping and genome annotation. Mapping is performed using both Bowtie and Snippy.	2AS_mapping + 4AN_genes
Enterotoxin *S. aureus* finder	*de novo* assembly followed by BLAST to identify the presence of the gene encoding the enterotoxin in the genome	2AS_denovo + 4AN_AMR
NgsManager	macro-pipeline which, depending on sample type, executes the modules for *SARS-CoV2*, *Bacteria* or *Virus* samples	pipelines "Raw Read Processing", "Covid Emergency", "Bacterial WGS", "Bacterial typing", "Genome Draft"
Raw Read Processing	quality check of reads, trimming and virus/bacteria classification	0SQ_rawreads + 1PP_trimming + 3TX_class
Filtering and *de novo*	it removes non-target organisms' reads from the raw reads and, subsequently, starts the *de novo* assembly	1PP_filtering + 2AS_denovo
Typing on Bacteria	horizontal and vertical coverage calculation; species calculation; gene annotation; identification of virulence and antibiotic-resistance genes; typing	2AS_mapping + 3TX_species + 4AN_genes + 4AN_AMR + 4TY_wgMLST + 4TY_cgMLST + 4TY_MLST + 4TY_flaA
WNV - lineage calculation and mapping	lineage calculation for *West Nile Virus* samples and mapping against the reference of the calculated lineage	4TY_lineage + 2AS_mapping
Mapping of Segmented Virus	it performs the mapping of genome segments, in segmented virus, using multiple references	2AS_mapping
Plasmids (AMR)	pipeline for plasmid typing with MOB-Suite	4TY_plasmid + 4AN_AMR
nf-core/ampliseq	community nf-core pipeline for amplicon sequencing and denoising. It supports Illumina paired-end reads, Illumina single-end reads, PacBio o IonTorrent	nf-core/ampliseq
IAEA AIV workflow	automatically determines avian influenza virus reference sequence for each segment and reconstructs corresponding consensus sequence	To be added
IAEA LSDV workflow	identification of LSDV reference and reconstruction of consensus sequence	To be added
IAEA WGSBAC Pipeline for *Salmonella*	implementation of full WGSBAC Pipeline for *Salmonella* samples	WGSBAC GitLab
IAEA WGSBAC Pipeline for *Brucella*	implementation of full WGSBAC Pipeline for *Brucella* samples	WGSBAC GitLab

Pipelines for Quality Check (QC):

Pipeline Name	Description
QC FastQC	quality check pipeline to run the FastQC software individually on raw or trimmed reads
QC Nanoplot	quality check pipeline to run the Nanoplot software individually on raw reads from Nanopore devices
QC quast	quality check pipeline to run the Quast software individually on the assembly files or the hybrid assembly files

5.4.2 Bioinformatics tools for non-Illumina Sequencing Technologies

In this page we list the bioinformatic tools available for reads produced with sequencing technologies different from Illumina short reads. Each of the tools listed below is available in the corresponding analysis and must be used **exclusively** with the specified sequencing technologies. Similarly, such reads can be processed **only** with the appropriate tools and not with those used for Illumina short reads.

Analyses for reads from Oxford Nanopore Technologies (ONT)

Analysis	Tool	Notes
1PP_filtering__minimap2	minimap2	only for reads from Nanopore devices
1PP_hostdepl__minimap2	minimap2	only for reads from Nanopore devices
1PP_trimming__chopper	chopper	only for reads from Nanopore devices
2AS_denovo__flye	Flye	only for reads from Nanopore devices
2AS_hybrid__unicycler	Unicycler	used for hybrid assembly of Illumina short reads and Nanopore reads
2AS_mapping__medaka	Medaka	only for reads from Nanopore devices
3TX_class__centrifuge	Centrifuge	only for reads from Nanopore devices
module_nanopore.md		preprocessing and typing pipeline for reads from Nanopore devices

Analyses for reads from Thermofisher IonTorrent

Analysis	Tool	Notes
1PP_trimming__fastp	fastp	only for reads from IonTorrent devices
1PP_trimming__trimmomatic	trimmomatic	for reads from IonTorrent and paired-end short reads
2AS_denovo__plasmidspades	plasmidSPAdes	only for reads from IonTorrent devices (assembly of plasmid sequences)
2AS_denovo__shovill	shovill	for reads from IonTorrent and paired-end short reads
2AS_denovo__spades	SPAdes	for reads from IonTorrent and paired-end short reads
2AS_mapping__bowtie	bowtie	for reads from IonTorrent and paired-end short reads
2AS_mapping__ivar	iVar	for reads from IonTorrent and paired-end short reads
2AS_mapping__snippy	snippy	for reads from IonTorrent and paired-end short reads
3TX_class__confindr	ConFindr	for reads from IonTorrent and paired-end short reads
3TX_class__kraken	kraken	for reads from IonTorrent and paired-end short reads
3TX_class__kraken2	kraken2	for reads from IonTorrent and paired-end short reads
3TX_species__mash	Mash	for reads from IonTorrent and paired-end short reads
4AN_AMR__resfinder	resfinder	for reads from IonTorrent and paired-end short reads

5.4.3 Single samples

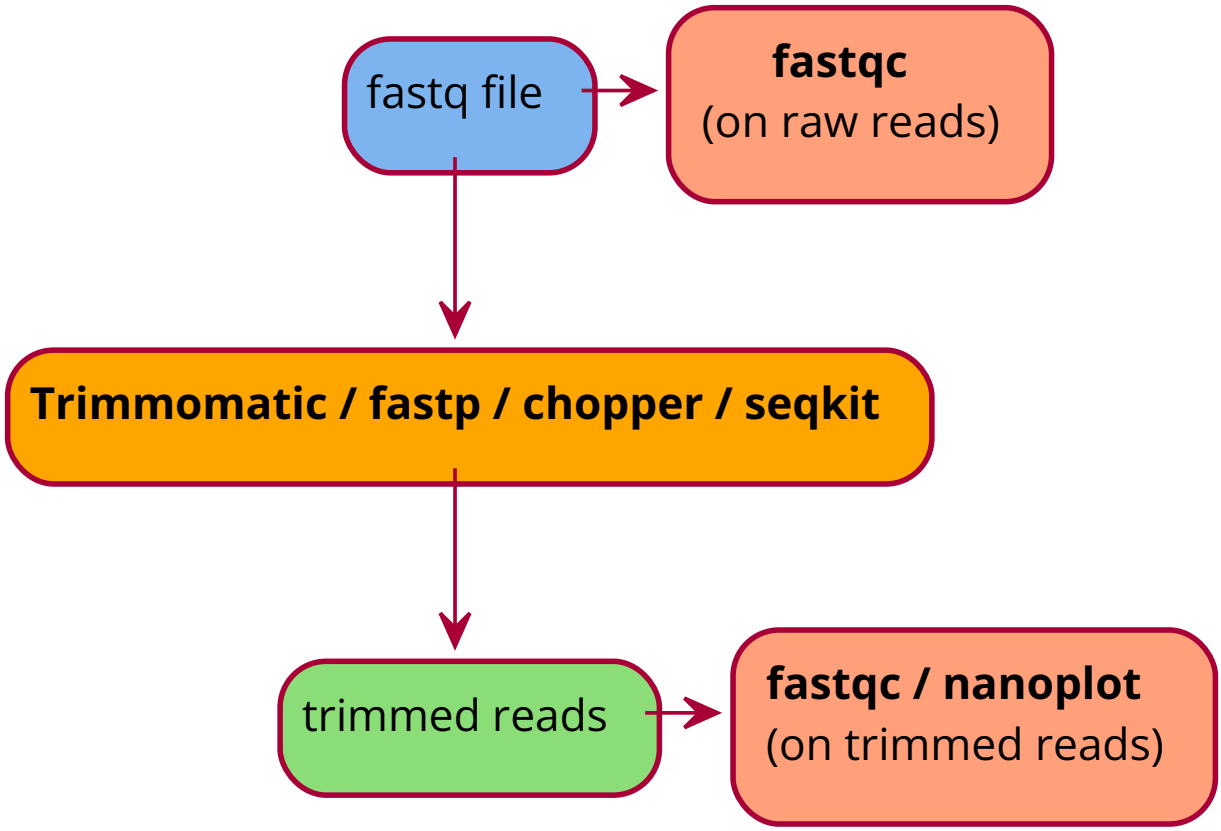
1 Pre-processing

1PP_trimming

Introduction

The trimming is intended to remove low-quality residues from the reads obtained by the sequencer.

The available tools for trimming in GenPat are "trimmomatic" and "fastp". The **1PP_trimming** analysis consists of running one of the trimming tools and **fastQC**. The latter computes quality statistics for the reads and is run both on the raw reads, before trimming, and on the newly trimmed reads.



Tool
Input
Result
Quality Check tool

Run 1PP_trimming analysis

Once the analysis **1PP_trimming** has been selected in the dedicated run analyses page, it will be possible to choose the bioinformatics software to use among those available for the analysis. The available tools in this case are:

- **trimmomatic** - Read trimming tool for Illumina NGS data

  Trimmomatic GitHub page: https://github.com/usadellab/Trimmomatic

- **fastp** - an all-in-one preprocessing for FastQ files

  Fastp GitHub page: https://github.com/OpenGene/fastp

- **chopper** - filtering and trimming for fastq files from long read sequencing such as PacBio or ONT

  Chopper GitHub page: https://github.com/wdecoster/chopper

Note 1: for *long read* trimming, please use **chopper** exclusively; for *short read* trimming, please use **trimmomatic** or **fastp**.

Note 2: Chopper parameter guide is available on the Chopper GitHub official page.

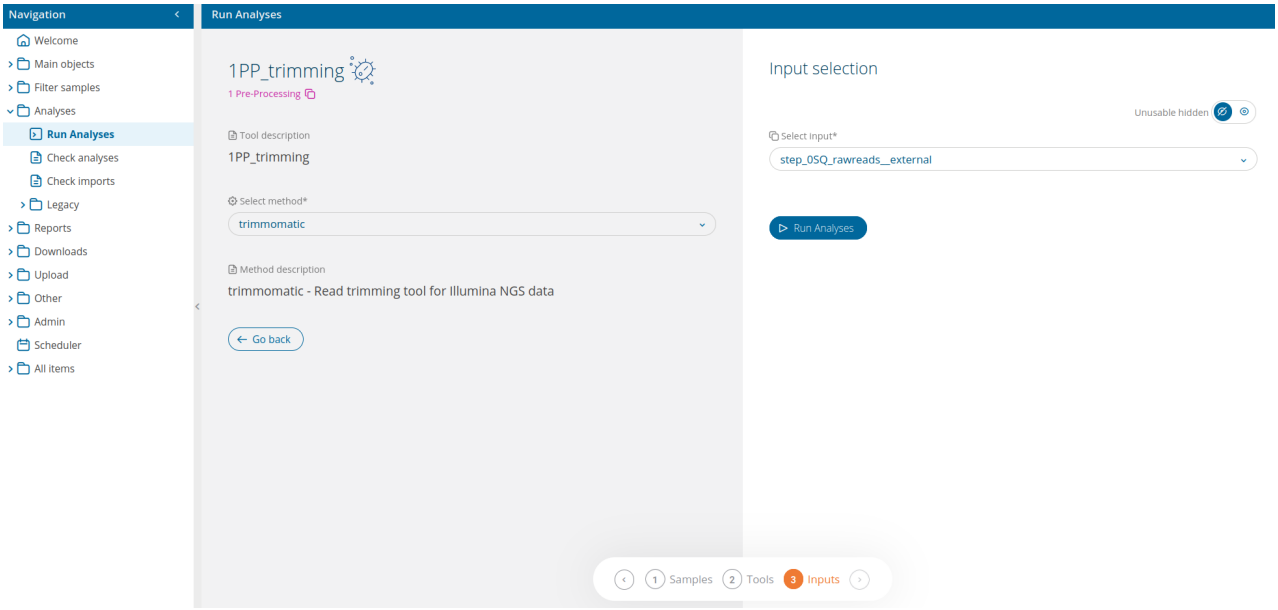
- **seqkit** - a cross-platform and ultrafast toolkit for FASTA/Q file manipulation

  Seqkit GitHub page: https://bioinf.shenwei.me/seqkit/usage/#grep

Note 3: this tool removes all reads that contain the adapter sequence specified in "*Custom adapter sequence*", using the tool's grep function.

Note 4: For Illumina reads only. For this analysis, either the *0SQ_rawreads* or *1PP_trimming* analysis can be used as input.

The input is specified in the last section: since trimming is performed on the fastQ files obtained from the sequencer, the available options will be: "step_0SQ_rawreads__fastq" for fastQ files from internal reads (code 20XX.TE.XXXX.X.X) and "step_0SQ_rawreads__external" for imported fastQ files (code 20XX.EXT.XXXX.X.X). The input selection interface of *1PP_trimming* provides the advanced input selection mode as this analysis can be performed on both internal source files and imported files.



Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

Result folder

> To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are 3 subfolders:

- **meta**: (*metadata*), where the log and configuration files of the executed process are saved;
- **qc**: (*quality check*) where check quality files are saved. It also contains the two sub-folders *meta* and *result*. For this analysis, quality check is performed with **fastqc**. QC calculation is not performed when running the analyses with Seqkit;
- **result**: where the files containing the results produced by the analysis are saved.

The tables below show the lists of files present in the folders, based on the tool used, along with some useful information. Note that the file name will include the name of the program chosen for the analyses.

Trimmomatic

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_R1_trimmomatic.fastq	trimmed read 1 (R1)	folder: result
DSXXXXXXXX-DTXXXXXX_ID_R2_trimmomatic.fastq	trimmed read 2 (R2)	folder: result
DSXXXXXXXX-DTXXXXXX_ID_unpaired_trimmomatic.fastq	unpaired trimmed reads	folder: result
DSXXXXXXXX-DTXXXXXX_ID_R1_trimmomatic_fastqc.html	quality of reads R1	folder: qc > result
DSXXXXXXXX-DTXXXXXX_ID_R1_trimmomatic_fastqc.zip	quality of R1 (zip file)	folder: qc > result
DSXXXXXXXX-DTXXXXXX_ID_R2_trimmomatic_fastqc.html	quality of reads R2	folder: qc > result
DSXXXXXXXX-DTXXXXXX_ID_R2_trimmomatic_fastqc.zip	quality of R2 (zip file)	folder: qc > result
DSXXXXXXXX-DTXXXXXX_ID_unpaired_trimmomatic_fastqc.html	quality of unpaired reads	folder: qc > result
DSXXXXXXXX-DTXXXXXX_ID_unpaired_trimmomatic_fastqc.zip	quality of unpaired reads (zip file)	folder: qc > result

fastp

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_fastp_R1.fastq.gz	trimmed read 1 (R1)	folder: result
DSXXXXXXXX-DTXXXXXX_ID_fastp_R2.fastq.gz	trimmed read 2 (R2)	folder: result
DSXXXXXXXX-DTXXXXXX_ID_fastp_unpaired.fastq.gz	unpaired trimmed reads	folder: result

chopper

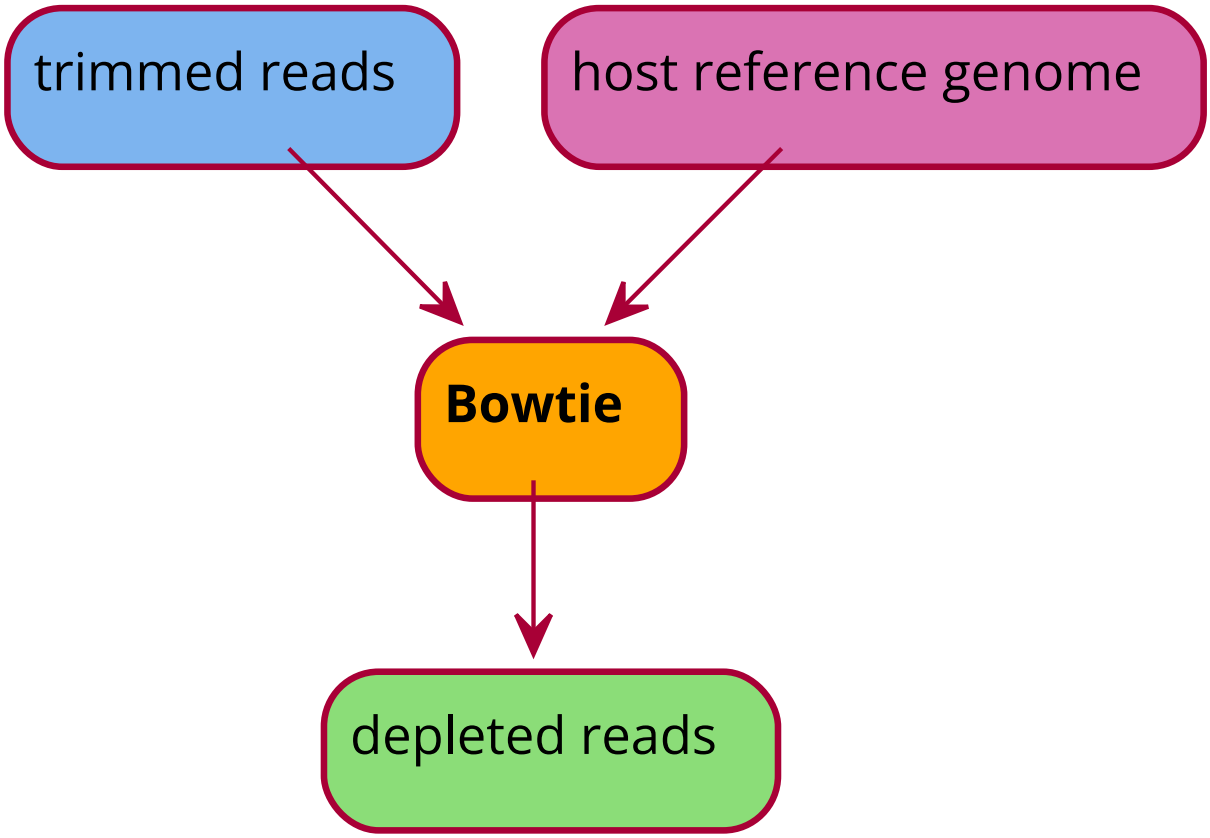
File	Description	Location
filtered_reads.fastq.gz	trimmed long reads	folder: result

1PP_hostdepl

Introduction

The *1PP_hostdepl* analysis maps the reads against the genome of the selected host organism, allowing the removal of host-contaminating reads from the FASTQ files (host sequence *depletion*).

The output files consist of new FASTQ files with cleaned reads.



Tool
Input
Result
Reference

Run 1PP_hostdepl analysis

Once the analysis **1PP_hostdepl** has been selected in the dedicated run analyses page, it will be possible to choose the bioinformatics software ("method") to use from those available for the analysis. The available tools for this analysis are:

- For all reads:
- **Bowtie** - An ultrafast, memory-efficient short read aligner.

Bowtie GitHub page: https://github.com/BenLangmead/bowtie

Only for long reads from nanopore-based devices:

- **minimap2** - a versatile sequence alignment program that aligns DNA or mRNA sequences against a large reference database.

Minimap2 GitHub page: https://github.com/lh3/minimap2

In the section dedicated to the input selection, it is necessary to specify both the input reads and a reference, as an additional parameter. Host depletion is performed by mapping the reads from the FASTQ files against a reference genome - the host genome - whose contaminating sequences must be identified and removed from the reads. In the definition of the *Host* parameter, it will therefore be possible to select any of the reference genomes for species uploaded into GenPat.

Note: If the desired host is not among those available, please send a request by e-mail to the bioinformatics department (bioinformatica@izs.it), to have the reference genome added to GenPat.

The input selection interface of this analysis provides the advanced input selection mode, so as to be able to simultaneously select FASTQ files that have been previously processed by different methods.

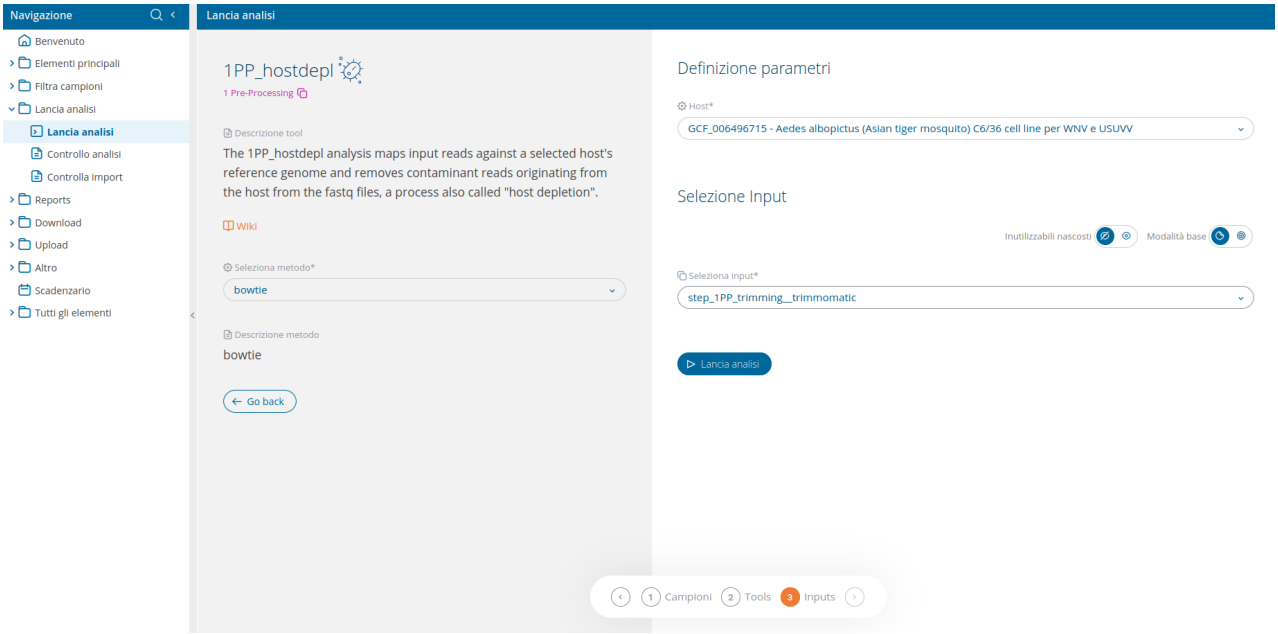
Inputs for Bowtie:

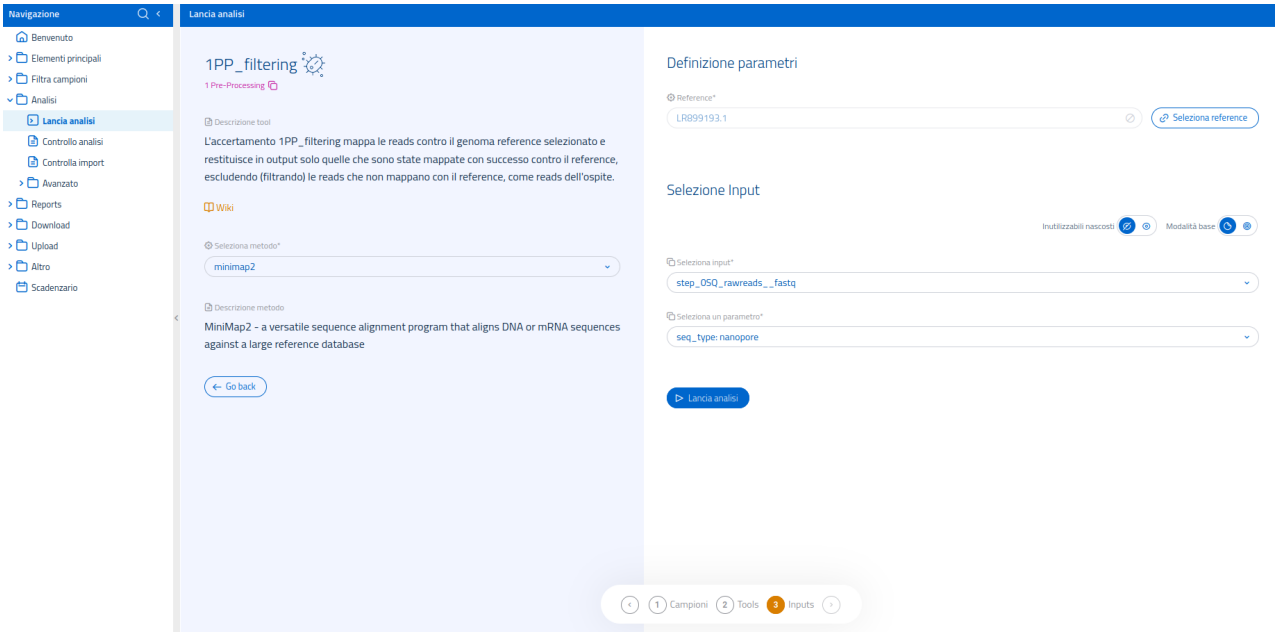
- step_1PP_trimming
- step_1PP_hostdepl
- step_1PP_downsampling

Inputs for minimap2:

- step_0SQ_rawreads

If the selected software is **minimap2**, inputs can only come from **step_0SQ_rawreads** of **single** long reads produced by nanopore-based devices. For further information about the **minimap2** software, please refer to the minimap2 official GitHub page.





Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

Result folder

> To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are two subfolders:

- **meta** (*metadata*), where the log and configuration files of the executed process are saved;
- **result**, where the files containing the results produced by the analysis are saved.

The tables below show the list of the files generated by the available tools for 1PP_hostdepl, along with some useful information.

Bowtie:

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_R1_vdhost_HOSTID.fastq	depleted read 1 (R1)	Folder: result
DSXXXXXXXX-DTXXXXXX_ID_R2_vdhost_HOSTID.fastq	depleted read 2 (R2)	Folder: result

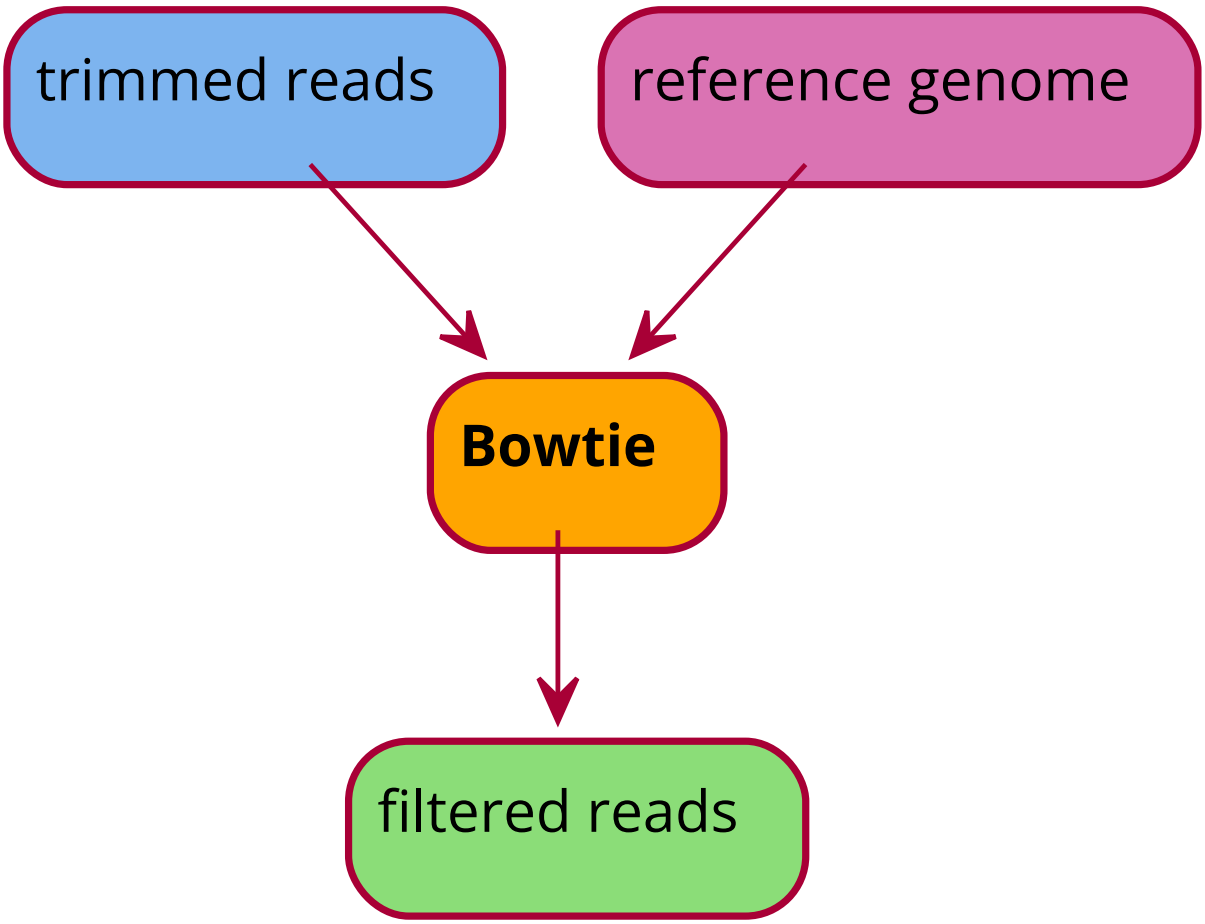
minimap2:

File	Description	Location
DSXXXXXX-DTXXXXXX_ID_minimap2_HOSTID.fastq.gz	depleted fastq of single read from nanopore device (compressed file)	Folder: result

1PP_filtering

Introduction

The *1PP_filtering* analysis maps the reads against the selected reference genome and outputs only those that have been successfully mapped to the refernece, excluding ("filtering out") the others.



Tool
Input
Result
Reference

Run 1PP_filtering analysis

Once the analysis **1PP_filtering** has been selected in the dedicated run analyses page, it will be possible to choose the bioinformatics software ("method") to use among those available for the analysis. The available tools for this analysis are:

- For all reads:
- **Bowtie** - An ultrafast, memory-efficient short read aligner.

Bowtie GitHub page: https://github.com/BenLangmead/bowtie

- Only for long reads from nanopore-based devices:
- **minimap2** - a versatile sequence alignment program that aligns DNA or mRNA sequences against a large reference database.

Minimap2 GitHub page: https://github.com/lh3/minimap2

In the section dedicated to input selection, it will be necessary to specify both the input reads and a reference genome, chosen from the available ones, which will be used for mapping.

The input selection interface of this analysis provides the advanced input selection mode, so as to be able to simultaneously select FASTQ files that have been previously processed by different methods.

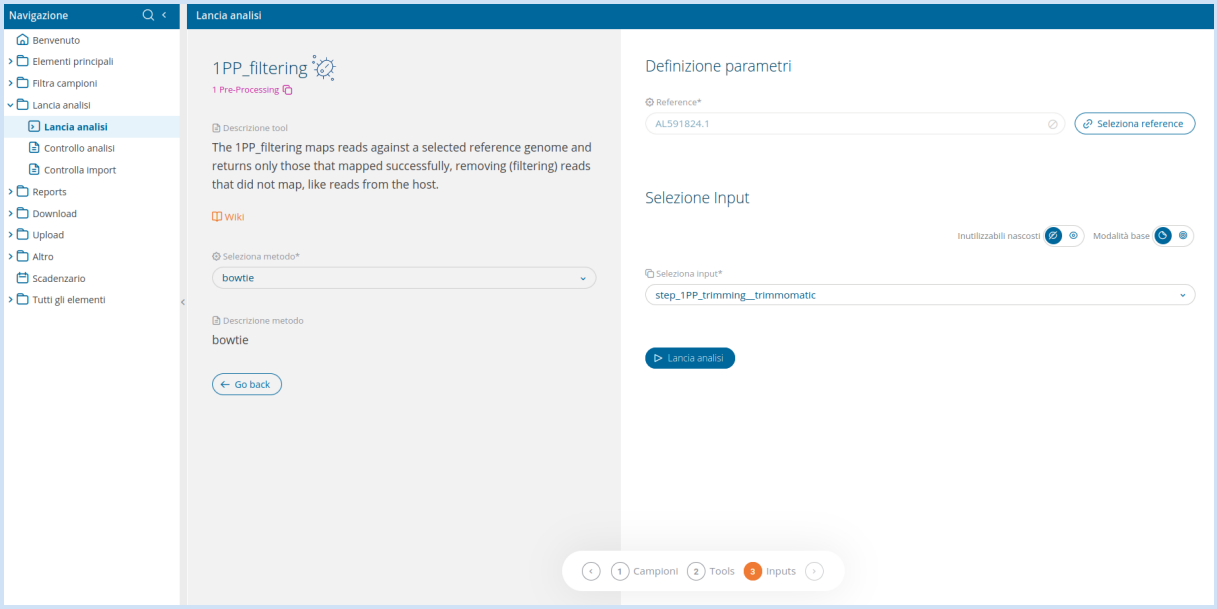
Inputs for Bowtie:

- step_1PP_trimming
- step_1PP_hostdepl
- step_1PP_downsampling
- step_1PP_filtering

Inputs for minimap2:

- step_0SQ_rawreads

If the selected software is **minimap2**, inputs can only come from **step_0SQ_rawreads** of **single** long reads produced by nanopore-based devices. For further information about the **minimap2** software, please refer to the minimap2 official GitHub page. -->



Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

Result folder

To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are two subfolders:

- **meta** (*metadata*), where the log and configuration files of the executed process are saved;
- **result**, where the files containing the results produced by the analysis are saved.

The tables below show the list of the files generated by the available tools for 1PP_filtering, along with some useful information.

Bowtie:

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_R1_bowtie_REFID.fastq.gz	filtered read 1 (R1) (compressed file)	folder: result
DSXXXXXXXX-DTXXXXXX_ID_R2_bowtie_REFID.fastq.gz	filtered read 2 (R2) (compressed file)	folder: result

minimap2:

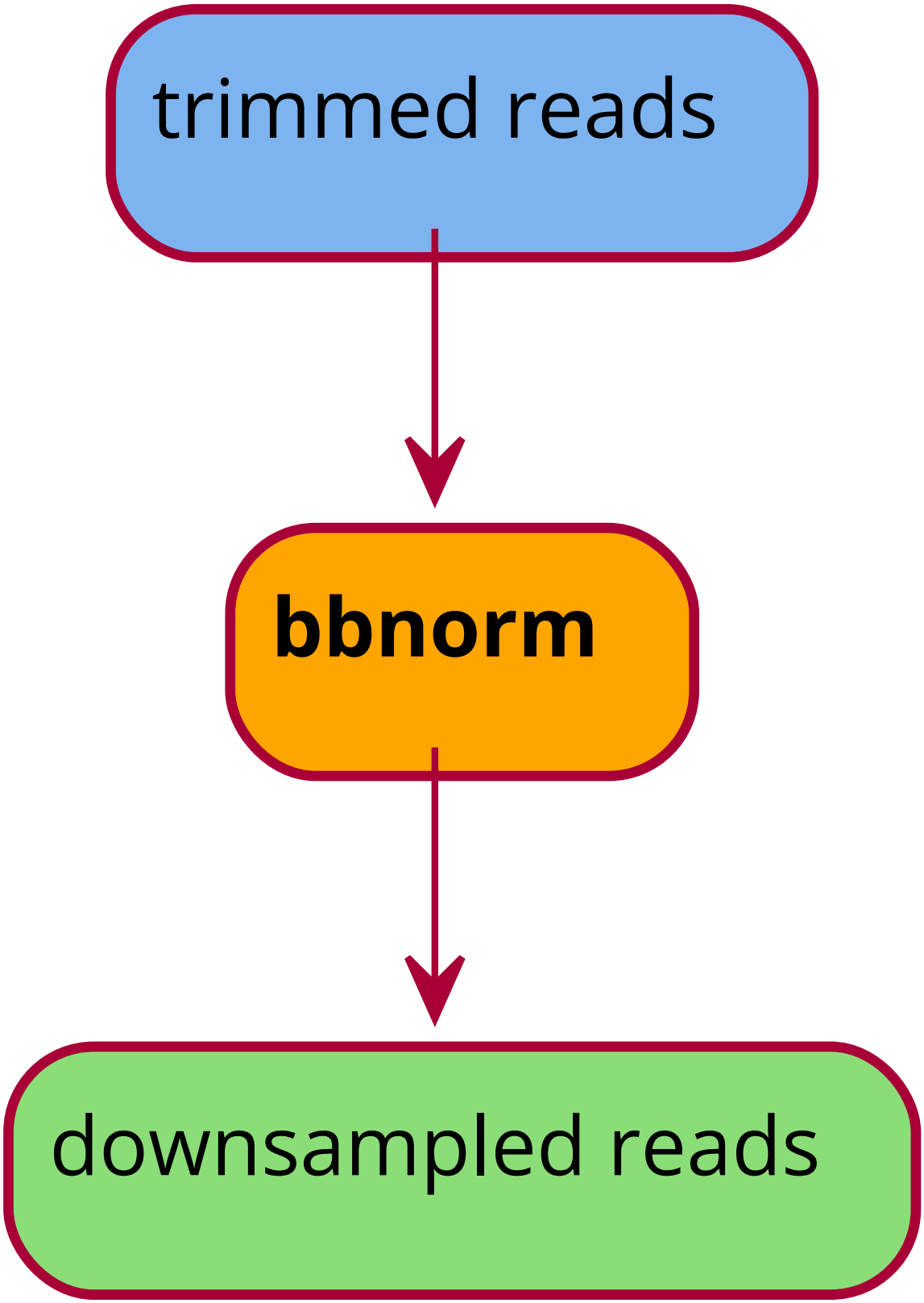
File	Description	Location
DSXXXXXX-DTXXXXXX_ID_minimap2_REFID.fastq.gz	filtered fastq of nanopore's single read (compressed file)	result directory

1PP_downsampling

Introduction

Downsampling can be defined as a process of reducing read depth (vertical coverage), at specific positions or within certain regions of the genome, with the aim of decreasing the number of uninformative reads and improving the quality of the assembly and downstream analyses. BBnorm normalizes read depth using k-mer hashing.

Sequencing protocols can often lead to the accumulation of reads in a specific region of the genome. This excess of information, in these cases, can slow down downstream analyses, which are forced to process a larger volume of data, even though much of it is not informative. In these cases, downsampling allows avoiding prolonged computation times by discarding excess read pairs until a specified threshold is reached, that is, the desired vertical coverage.



Tool

Run 1PP_downsampling analysis

Once the analysis **1PP_downsampling** has been selected in the dedicated run analyses page, it will be possible to choose the bioinformatics software ("method") to use among those available for the analysis. The available tool in this case is:

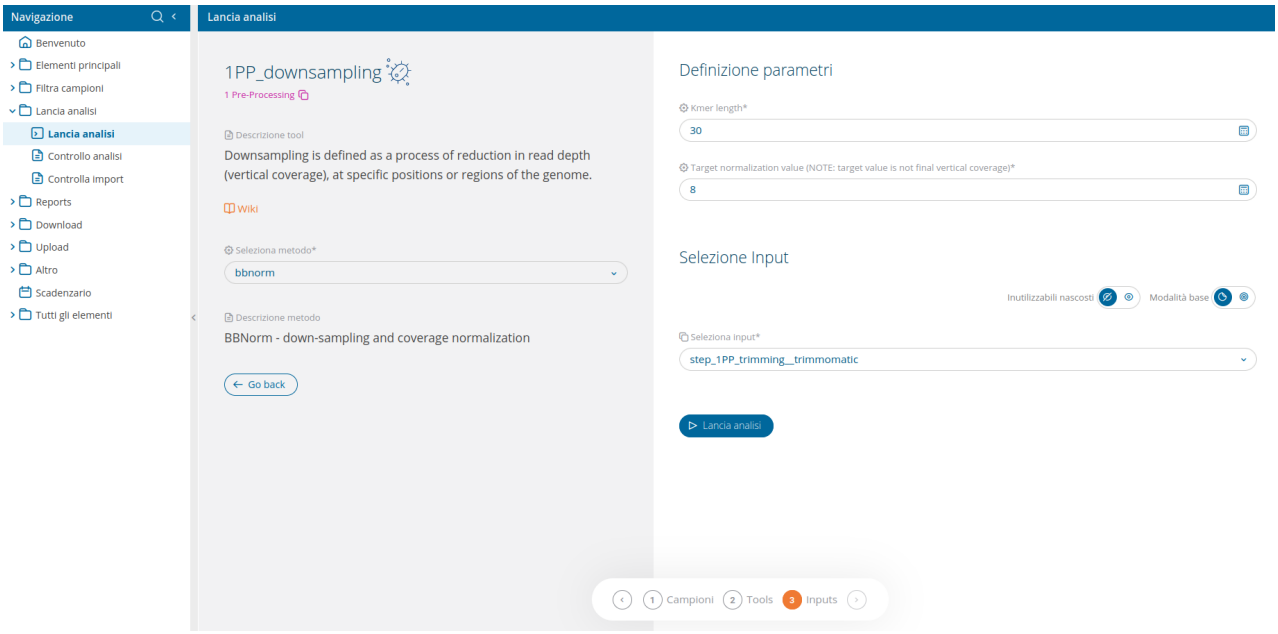
- **BBnorm** - Kmer-based error-correction and normalization tool.

> **BBnorm user manual:** https://manpages.debian.org/unstable/bbmap/bbnorm.sh.1.en.html

In the parameter definition screen, it will be necessary to specify both the input reads and some additional parameters, used to determine the final vertical coverage desired for the genomes of the samples.

The input selection interface provides the advanced input selection mode. In fact, downsampling may be performed following upstream analyses that are different from one another:

- step_1PP_trimming
- step_1PP_hostdepl
- step_1PP_downsampling



Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

Parameters

> An overview of the parameters of BBnorm is available at the following link: https://manpages.debian.org/unstable/bbmap/bbnorm.sh.1.en.html.

The two parameters "kmer lenght" and "Target" depend on:

- the species of the sample (and, consequently, the genome lenght);
- for "Target", it depends on the kmer lenght to be used, as specified in the "kmer lenght".

Kmer lenght has a default value of 31. It supports any kmer lenght but a shorter lenght than 32 is recommended for improved tool efficiency.

Target indicates the normalization of the **k-mers** used by BBnorm to perform *downsampling*.

> **Warning:** the "Target normalization value" parameter **does not** correspond to the final vertical coverage obtained by running the tool; rather, according to the BBnorm parameter guide, it controls the vertical depth of the k-mers used by the tool to normalize the reads.

Due to the way the software works, it is therefore necessary to experimentally determine the value of the second parameter, in relation to the first ("k-mer lenght") and the species.

Two examples for *Listeria monocytogenes* are listed below:

- species *L. monocytogenes*, kmer length = 30, target = 31 --> vertical coverage = 40X
- species *L. monocytogenes*, kmer length = 30, target = 8 --> vertical coverage = 10X

Result folder

> To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are two subfolders:

- **meta** (*metadata*), where the log and configuration files of the executed process are saved;
- **result**, where the files containing the results produced by the analysis are saved.

The table below shows the list of the main files of interest present in the folders, along with some useful information.

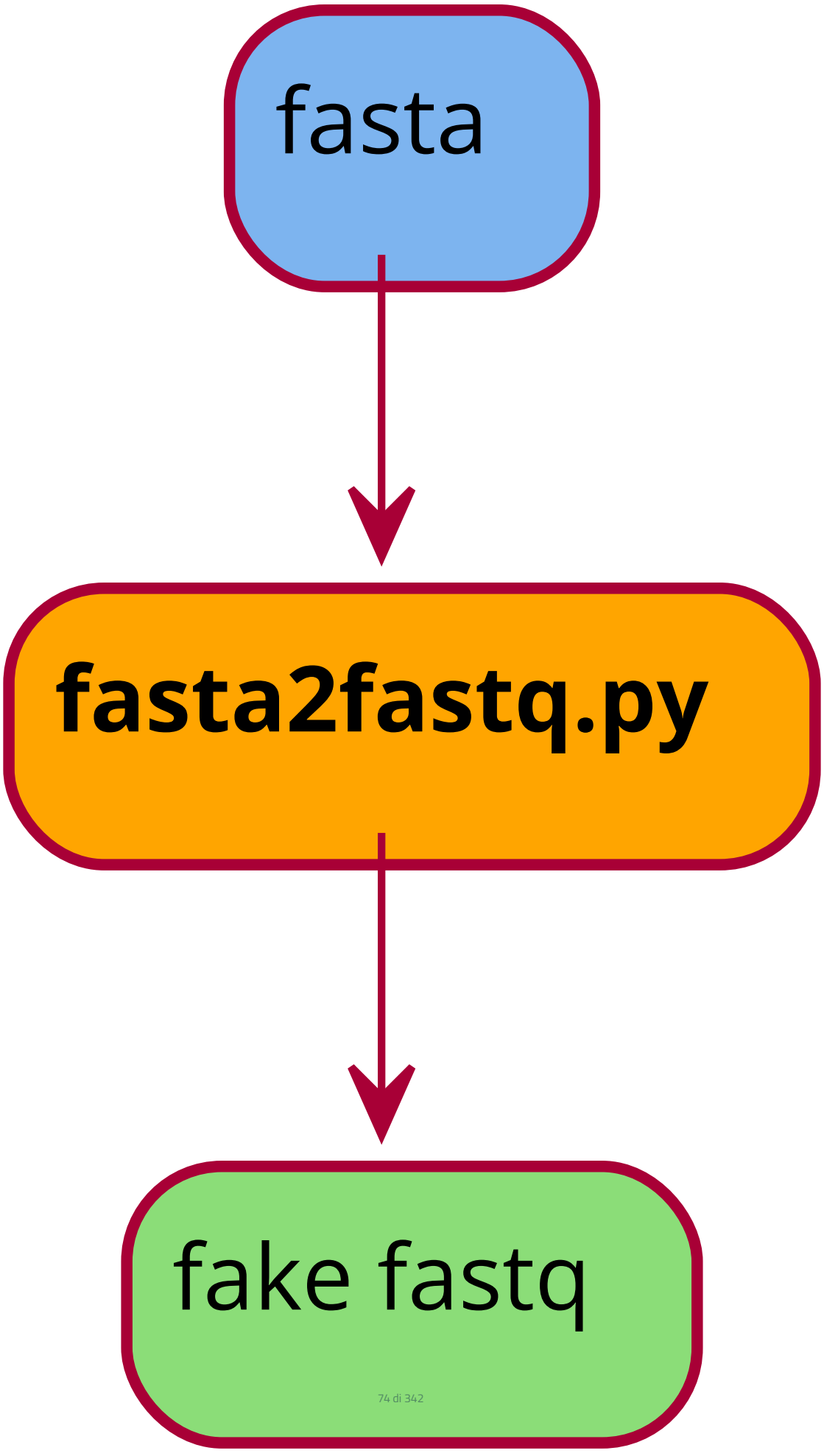
File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_bbnorm_kXX_tX_R1.fastq.gz	downsampled read 1 (R1)	folder: result
DSXXXXXXXX-DTXXXXXX_ID_bbnorm_kXX_tX_R2.fastq.gz	downsampled read 2 (R2)	folder: result

1PP_generated

Introduction

The *1PP_generated* builds a FASTQ file from a FASTA file. To achieve this result, a script specifically created for GenPat, (**fasta2fastq.py**), is used. It generates lines with dummy quality scores and modifies the headers so that they comply with the FASTQ format. The software generates files for both antiparallel DNA strands (R1 and R2).

Warning: Since the generated quality lines are dummy, the purpose of this analysis is solely to allow the use of sequences available only as FASTA files within analyses whose programs normally accept only FASTQ format as input. The output files, therefore, must not be used as a substitute for actual FASTQ sequences in workflows where the evaluation of read quality matters.


fasta
fasta2fastq.py
fake fastq

Run 1PP_generated analysis

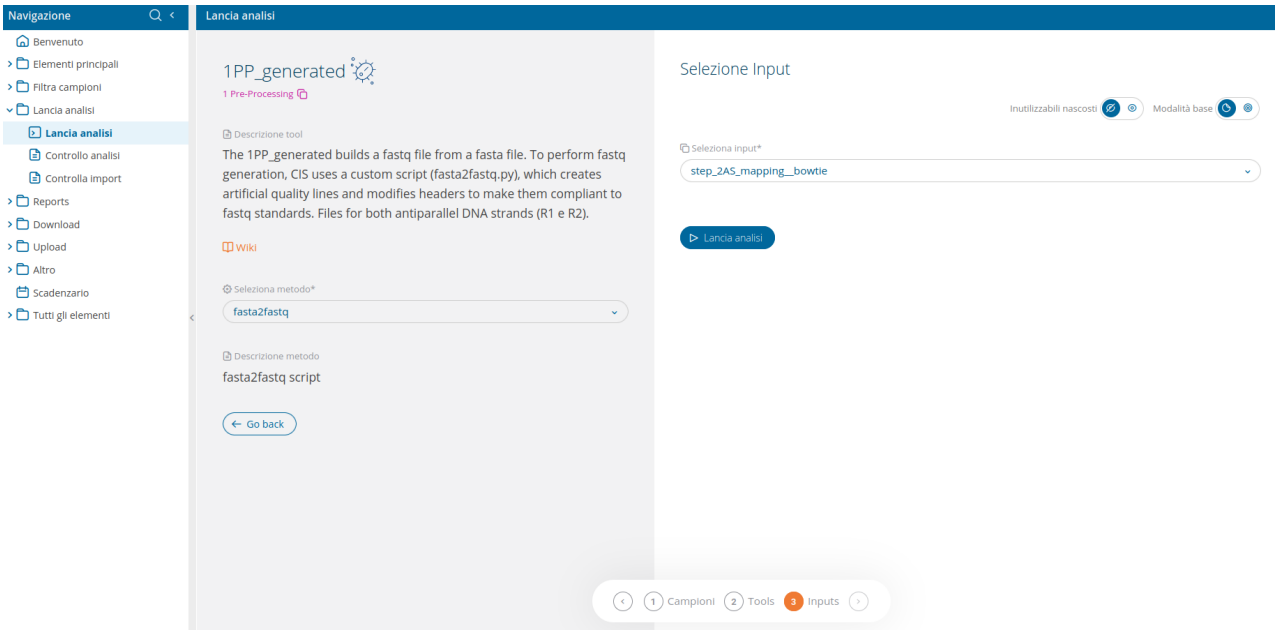
Once the analysis **1PP_generated** has been selected from the run analyses interface, the user will be able to select which bioinformatic tool ("method") to use. The only available tool for this analysis is **fasta2fastq**.

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.

Available inputs for 1PP_generated are:

- step_2AS_mapping
- step_2AS_denovo
- step_2AS_import

It will be necessary to specify the input sequences - usually originating from *de novo* assembly or from mapping - and the reference genome used for mapping, in case it is selected as input.



Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

Result folder

> To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are two subfolders:

- **meta** (*metadata*), where the log and configuration files of the executed process are saved;
- **result**, where the files containing the results produced by the analysis are saved.

The table below lists available fasta2fastq.py output files.

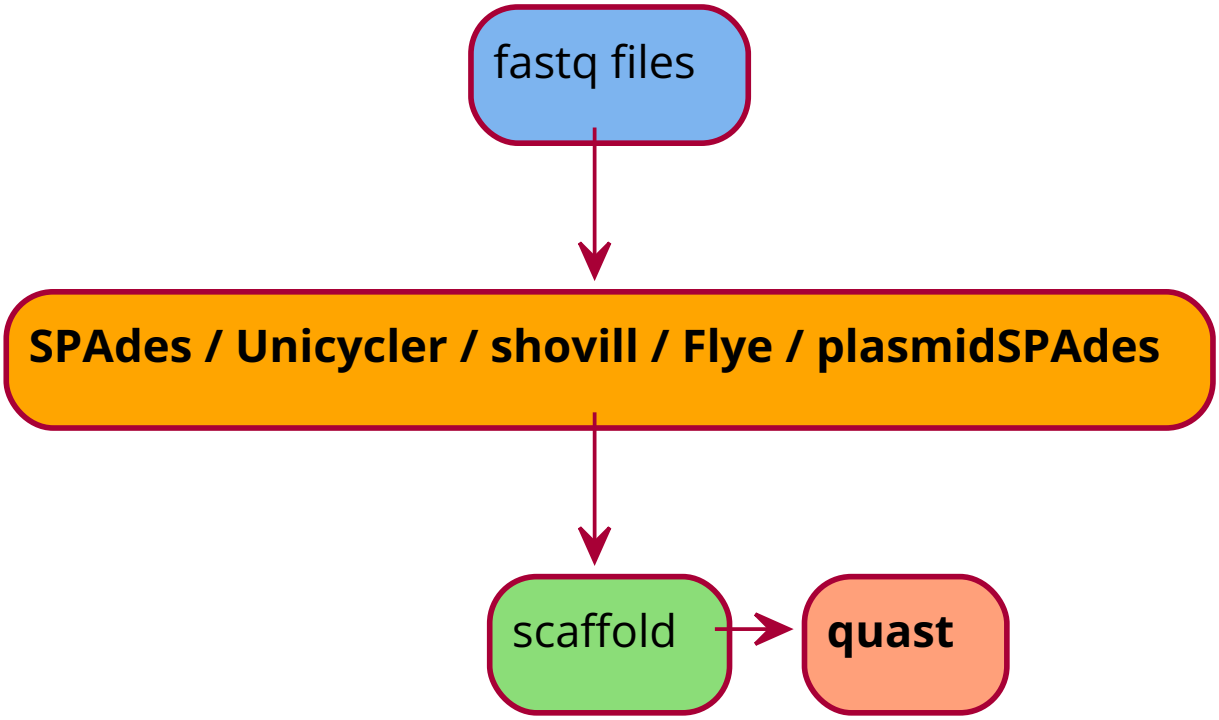
File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_fasta2fastq_R1.fastq.gz	dummy strand fastq + (compressed file: .gz)	folder: results
DSXXXXXXXX-DTXXXXXX_ID_fasta2fastq_R2.fastq.gz	dummy strand fastq - (compressed file: .gz)	folder: results

2 Assembly

2AS_denovo

Introduction

2AS_denovo analysis performs the *de novo* assembly of sequences, producing the genome's scaffolds.



Tool
Input
Result
Quality Check tool

Run 2AS_denovo analysis

Once the analysis **2AS_denovo** has been selected in the dedicated run analyses page, it will be possible to choose the bioinformatics software ("method") to use, among those available for the analysis. The available tools in this case are:

- **SPAdes** - St. Petersburg genome assembler.

  SPAdes GitHub page: https://github.com/ablab/spades

- **Unicycler** - hybrid assembly pipeline for bacterial genomes.

  Unicycler GitHub page: https://github.com/rrwick/Unicycler

- **shovill** - a pipeline for assembly of bacterial isolate genomes from Illumina paired-end reads.

  Shovill GitHub page: https://github.com/tseemann/shovill

- **Flye** - *de novo* assembler for single-molecule sequencing reads

  Flye GitHub page: https://github.com/fenderglass/Flye

- **plasmidSPAdes** - St. Petersburg genome assembler for bacterial plasmid assembly.

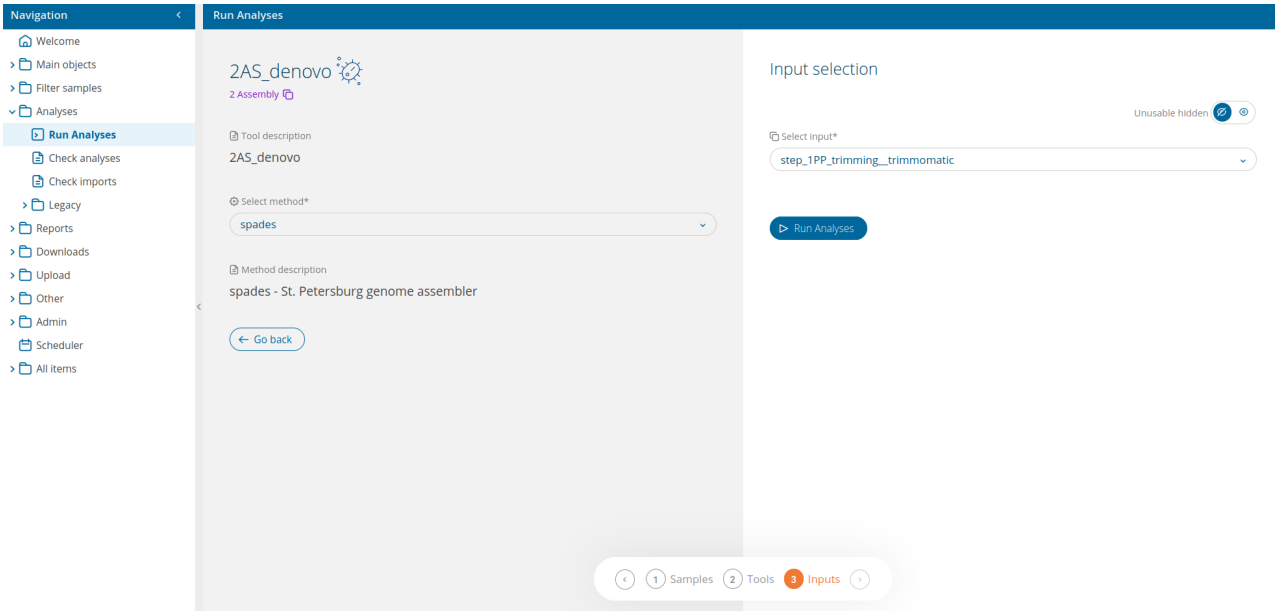
  PlasmidSPAdes GitHub page: https://github.com/ablab/spades

Note: the tool **Flye** is exclusively dedicated to the assembly of long reads generated by Oxford Nanopore devices; the tool **plasmidSPAdes** is an iteration of **SPAdes** designed for the use with plasmids.

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.

Accepted inputs can originate from:

- step_1PP_trimming
- step_1PP_hostdepl
- step_1PP_downsampling
- step_1PP_generated
- step_3TX_class
- step_4AN_AMR



Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

Result folder

> To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are 3 subfolders:

- **meta**: (*metadata*), where the log and configuration files of the executed process are saved;
- **qc**: (*quality check*) where check quality files are saved. It also contains the two sub-folders *meta* and *result*. For this analysis, quality check is performed with **quast**;
- **result**: where the files containing the results produced by the analysis are saved.

The tables below show the lists of files present in the folders, along with some useful information.

SPAdes

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_spades_contigs.fasta	file of contigs from the *de novo* assembly	folder: result
DSXXXXXXXX-DTXXXXXX_ID_spades_scaffolds.fasta	file of the scaffolds	folder: result
DSXXXXXXXX-DTXXXXXX_ID_spades_scaffolds_L200.fasta	file containing scaffolds longer than 200bp	folder: result
DSXXXXXXXX-DTXXXXXX_ID_quast.csv	file of quality statistics of the assembly	folder: qc > result

Unicycler

File	Description	Location
001_best_spades_graph.gfa	*de novo* assembly in .gfa format - cleaning step 1	folder: result
002_overlaps_removed.gfa	*de novo* assembly in .gfa format - cleaning step 2	folder: result
003_bridges_applied.gfa	*de novo* assembly in .gfa format - cleaning step 3	folder: result
004_final_clean.gfa	*de novo* assembly in .gfa format - cleaning step 4	folder: result
005_polished.gfa	*de novo* assembly in .gfa format - cleaning step 5	folder: result
DSXXXXXXXX-DTXXXXXX_ID_unicycler_assembly.fasta	file of contigs from the *de novo* assembly	folder: result
DSXXXXXXXX-DTXXXXXX_ID_unicycler_scaffolds_L200.fasta	file containing scaffolds longer than 200bp	folder: result
assembly.gfa	*de novo* assembly in .gfa format	folder: result
DSXXXXXXXX-DTXXXXXX_ID_quast.csv	quality statistics file of the assembly	folder: qc > result

shovill

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_shovill.fasta	file of contigs from the *de novo* assembly	folder: result
DSXXXXXXXX-DTXXXXXX_ID_quast.csv	quality statistics file of the assembly	folder: qc > result

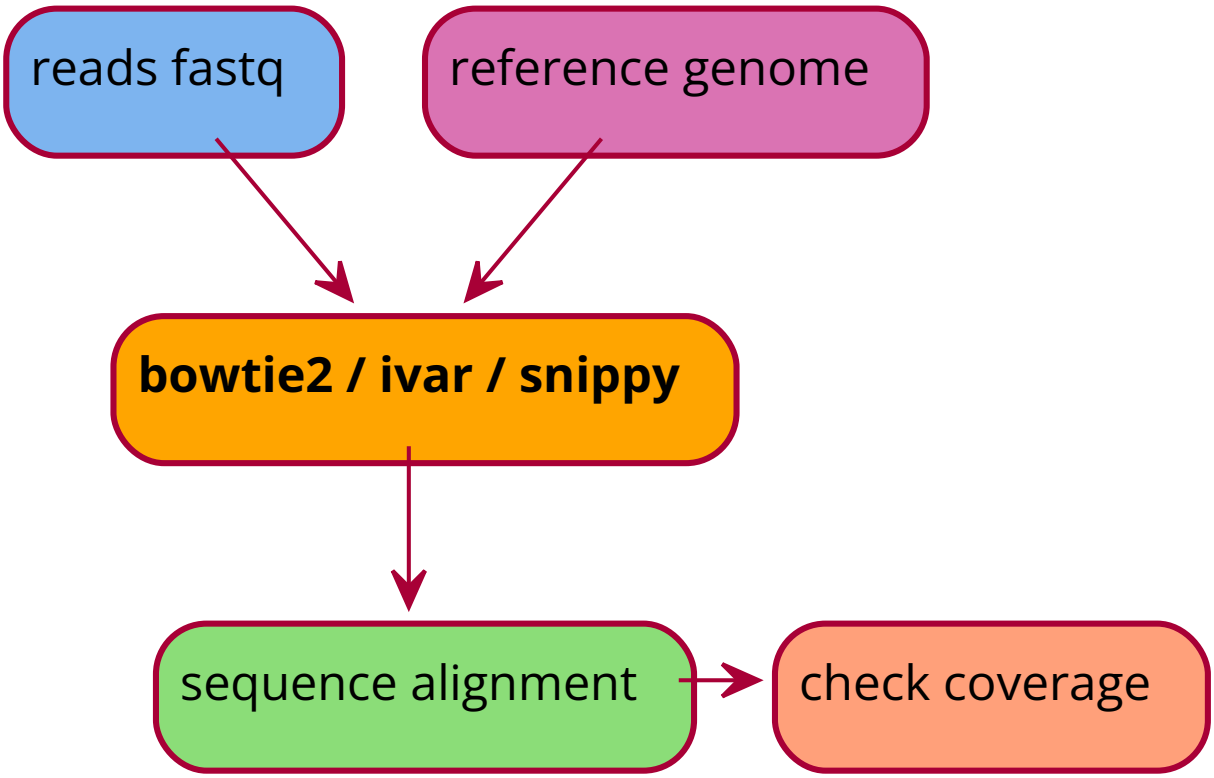
Flye

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_flye.fasta	file of contigs from the *de novo* assembly	folder: result
DSXXXXXXXX-DTXXXXXX_ID_quast.csv	quality statistics file of the assembly	folder: qc > result

2AS_mapping

Introduction

2AS_mapping analysis performs the **reference-based assembly** (*i.e.* **mapping**) of sequences.



Tool
Input
Result
Reference
Quality check

Run 2AS_mapping analysis

Once the analysis **2AS_mapping** has been selected in the dedicated run analyses page, it will be possible to choose the bioinformatics software ("method") to use, among those available for the analysis. The available tools in this case are:

- **bowtie2** - an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences.

  Bowtie2 GitHub page: https://github.com/BenLangmead/bowtie2

- **iVar** - computational package for viral amplicon-based sequencing

  iVar GitHub page: https://github.com/andersen-lab/ivar

- **Snippy** - Rapid haploid variant calling and core genome alignment

  Snippy GitHub page: https://github.com/tseemann/snippy

Note: When using the *snippy* tool, it is essential that the selected sample has the 4AN_genes analysis associated with it (Prokka).

- **medaka** - a tool to create consensus sequences and variant calls, from Nanopore sequencing data.

  Medaka GitHub page: https://github.com/nanoporetech/medaka

Note: medaka is an open-source tool developed by the Oxford Nanopore Technologies Ltd., to be used exclusively on sequencing data generated by Nanopore devices.

In the section dedicated to the *parameter definition*, it will be necessary to select a reference for the mapping, from the pop-up table accessible by clicking the "Select" button. The pop-up lists sequences of two types, both of which can be used as reference:

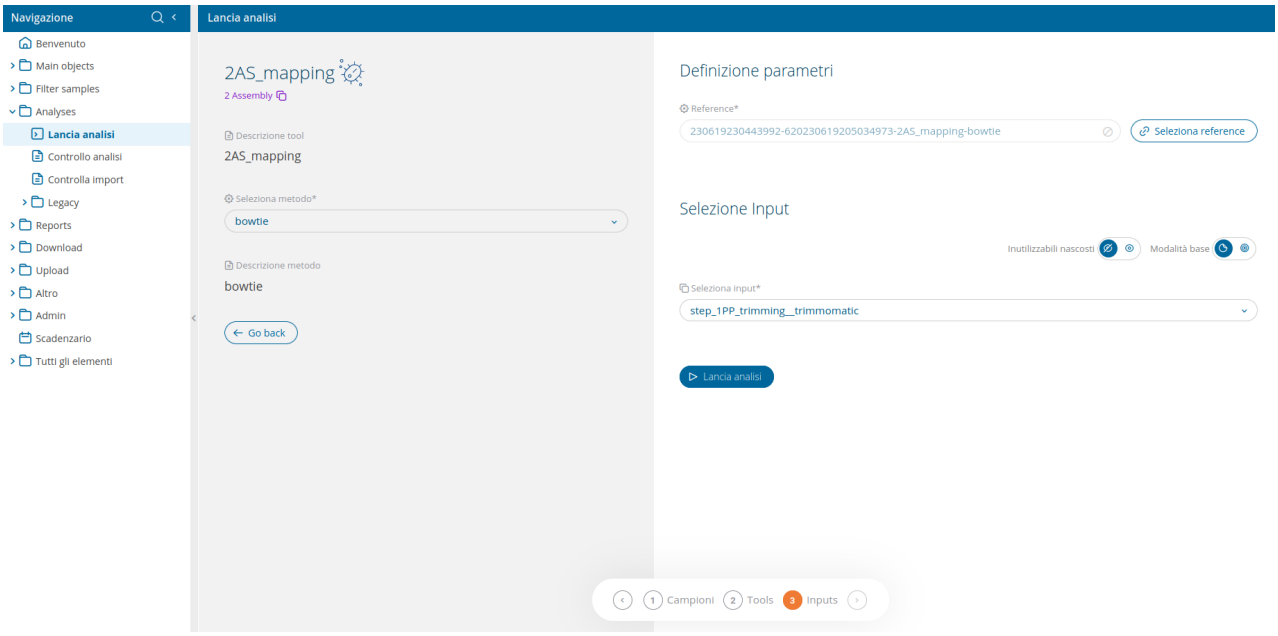
1. **fasta file reference**: files imported from external databases - such as NCBI - (listed in the GenPat reference table: https://genpat.izs.it/cmdbuild/ui/#classes/T_REFERENCE/cards);
2. the consensus sequence or the *de novo* assembly of **a sample available in the platform** (internal database).

With the "Ivar" method, it is possible to select more than one reference (please, refer to the "Multiple references" section below).

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.

The accepted inputs are the same as those that can also be used for the 2AS_denovo analysis:

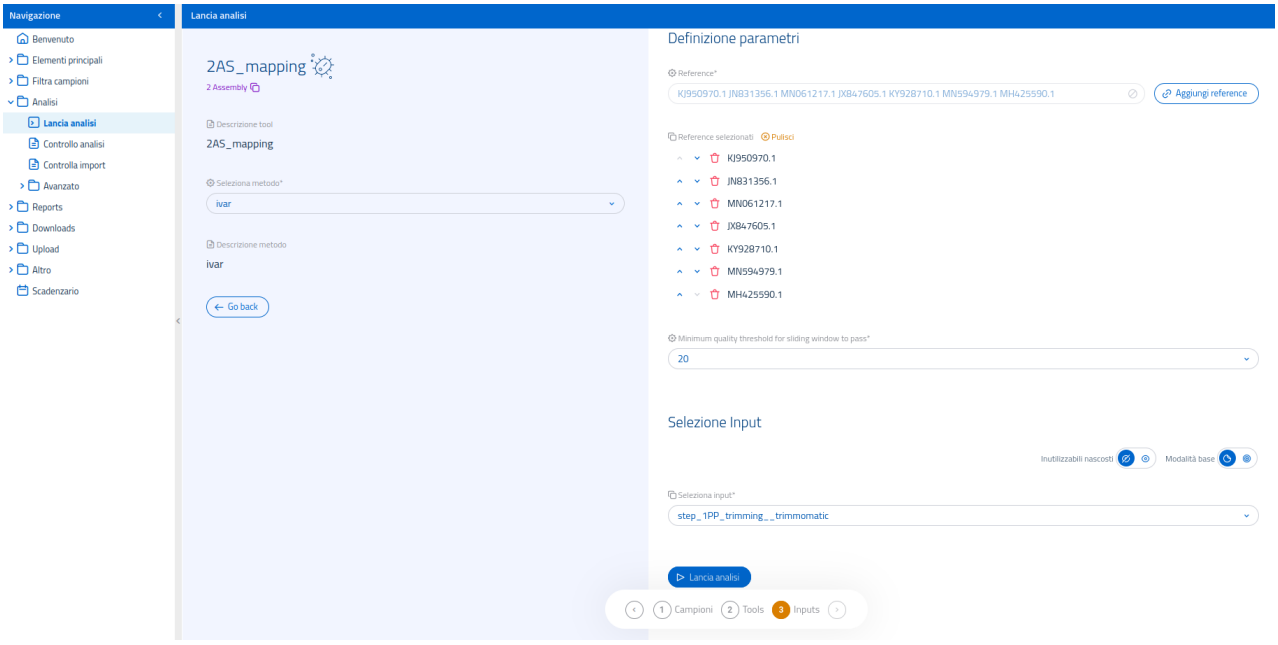
- step_1PP_trimming
- step_1PP_hostdepl
- step_1PP_downsampling
- step_1PP_generated
- step_3TX_class
- step_4AN_AMR



Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

Multiple references

If *2AS_mapping* is performed with **Ivar**, it is possible to select more than one reference, as shown in the image below. A guide to the multiple reference selection system is available in the dedicated section of the wiki on running analyses.



Result folder

To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are 2 subfolders:

- **meta**: (*metadata*), where the log and configuration files of the executed process are saved;
- **result**: where the files containing the results produced by the analysis are saved.

The tables below list the files produced by the available tools for 2AS_mapping.

Bowtie2

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_bowtie_REFID.fasta	consensus file	folder: result
DSXXXXXXXX-DTXXXXXX_ID_bowtie_REFID.sorted.bam	alignment file in .bam (Binary Alignment Map) format	folder: result
DSXXXXXXXX-DTXXXXXX_ID_bowtie_REFID.sorted.bam.bai	.bai (BAM file index) file	folder: result
DSXXXXXXXX-DTXXXXXX_ID_bowtie_REFID.var.flt.vcf	file of identified variants in .vcf format	folder: result
DSXXXXXXXX-DTXXXXXX_ID_bowtie_REFID_coverage_plot.png	coverage distribution graph	folder: result

Ivar and Snippy

Please note that using **ivar** includes running the tools *snippy*, *samtools* and finally *ivar*, whereas **2AS_mapping_snippy** runs exclusively the software it is named after.

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_ivar_REFID.fasta	consensus sequence obtained with ivar (**note:2AS-mapping_snippy does not produce this file**)	folder: result
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.aligned.fa	reference version with `-` in positions with `sequencing depth = 0` (number of reads aligning at that position = 0) and `N` for `depth` between 0 and the minimum number of reads considered in the siet coverage (**no variants are listed in this file**)	folder: result
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.bam	alignment file in .bam (Binary Alignment Map) format	folder: result
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.bam.bai	.bai (BAM file index) file	folder: result
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.bed	.bed (Browser Extensible Data) file	folder: result
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.consensus.fa	reference genome with representation of all variants	folder: result
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.consensus.subs.fa	reference genome with representation of substitution variants only	folder: result
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.txt	summary of the snippy run	folder: result
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.tab	variant table in .tsv format	folder: result
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.csv	variant table in .csv format	folder: result
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.filt.vcf	variants filtered by Freebayes	folder: result
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.raw.vcf	variants not filtered by Freebayes	folder: result
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.subs.vcf	table of substitution-only variants in .vcf format	folder: result
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.gff	variants in GFF3 format	folder: result
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.html	.html version of the .tab table of variants	folder: result
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.vcf	output file from Snippy (identified variants) in .vcf format (summary header + table)	folder: result
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.vcf.gz	snippy compressed .vcf output file	folder: result
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.vcf.gz.csi	bcftools index for the .vcf.gz file	folder: result
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID_coverage_plot.png	coverage distribution graph	folder: result

For further information on the files generated by *snippy*, their formats and their contents, please refer to the Snippy manual.

medaka

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_medaka_REFID.fasta	fasta file of the alignment	folder: result

Data visualization

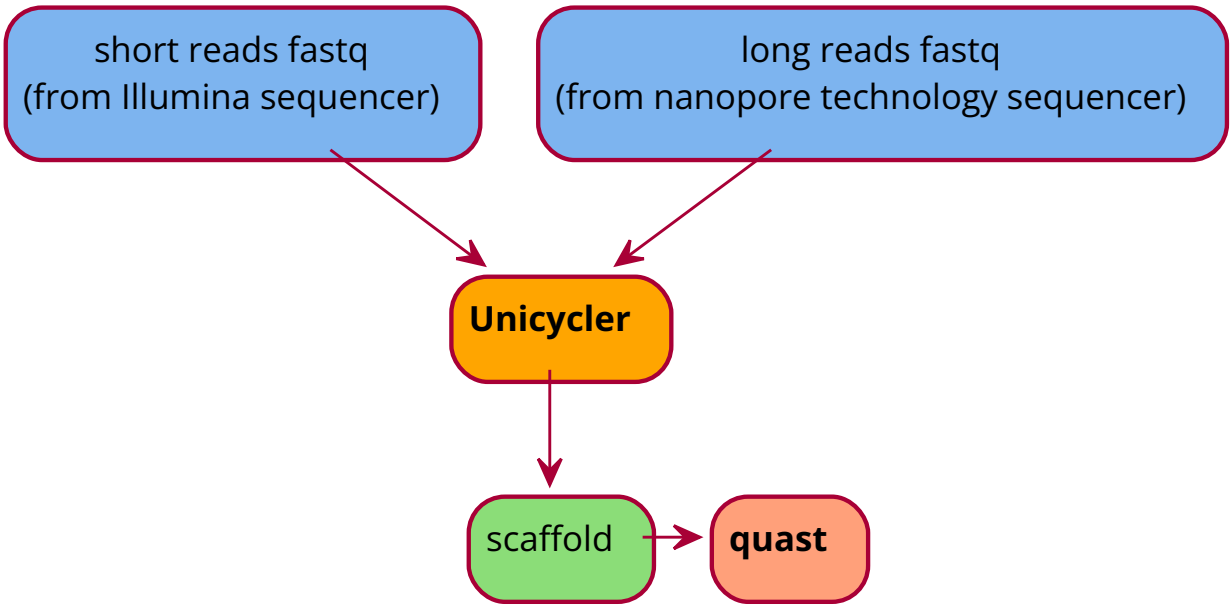
To evaluate the quality of the generated sequences, read alignments against the reference can be visualized using a dedicated software (*i.e.* Tablet, BioEdit and uGene) capable of reading .bam and .bam.bai files:

- Tablet (GNU/Linux, macOS, Windows): https://ics.hutton.ac.uk/tablet/download-tablet/;
- BioEdit (Windows): https://thalljiscience.github.io/;
- uGene (GNU/Linux, macOS, Windows): http://ugene.net/ugene/.

2AS_hybrid

Introduction

2AS_hybrid analysis performs the hybrid assembly of short and long reads. Hybrid assembly grants the coverage of repetitive DNA regions and place them in the correct positions, thanks to third-generation sequencing technologies that produce long reads; at the same time, hybrid assembly increases accuracy, thanks to the use of second-generation sequencing technologies that produce short reads.



Tool
Input
Result
Quality Check tool

Run 2AS_hybrid analysis

Once the analysis **2AS_hybrid** has been selected in the dedicated run analyses page, it will be possible to choose the bioinformatics software ("method") to use, among those available for the analysis. The only tool that can be used for this analysis is:

- **Unicycler** - hybrid assembly pipeline for bacterial genomes.

Unicycler GitHub Page: https://github.com/rrwick/Unicycler

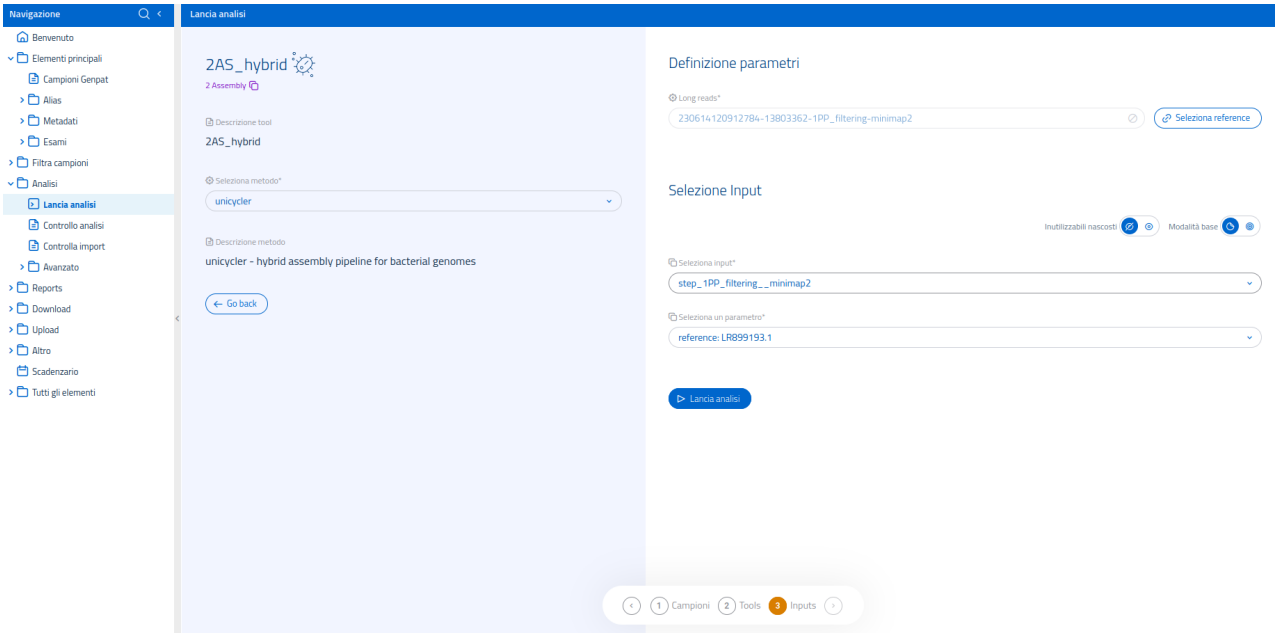
2AS_hybrid requires 2 inputs (please, refer to the image below):

1. *Long reads* from nanopore-based sequencers ("Long reads" field is mandatory);
2. *Short reads* from Illumina sequencers ("Input selection" field is mandatory). Possible short read inputs can be obtained from pre-processing analyses of the reads, such as:

   step_1PP_trimming
   step_1PP_filtering
   step_1PP_hostdepl
   step_1PP_downsampling.

> **Note:** the execution of the hybrid assembly will fail if long reads are used, as the second input, instead of Illumina short reads.

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.



Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

Result folder

> To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are 3 subfolders:

- **meta**: (*metadata*), where the log and configuration files of the executed process are saved;
- **qc**: (*quality check*) where check quality files are saved. It also contains the two sub-folders *meta* and *result*. For this analysis, quality check is performed with **quast**;
- **result**: where the files containing the results produced by the analysis are saved.

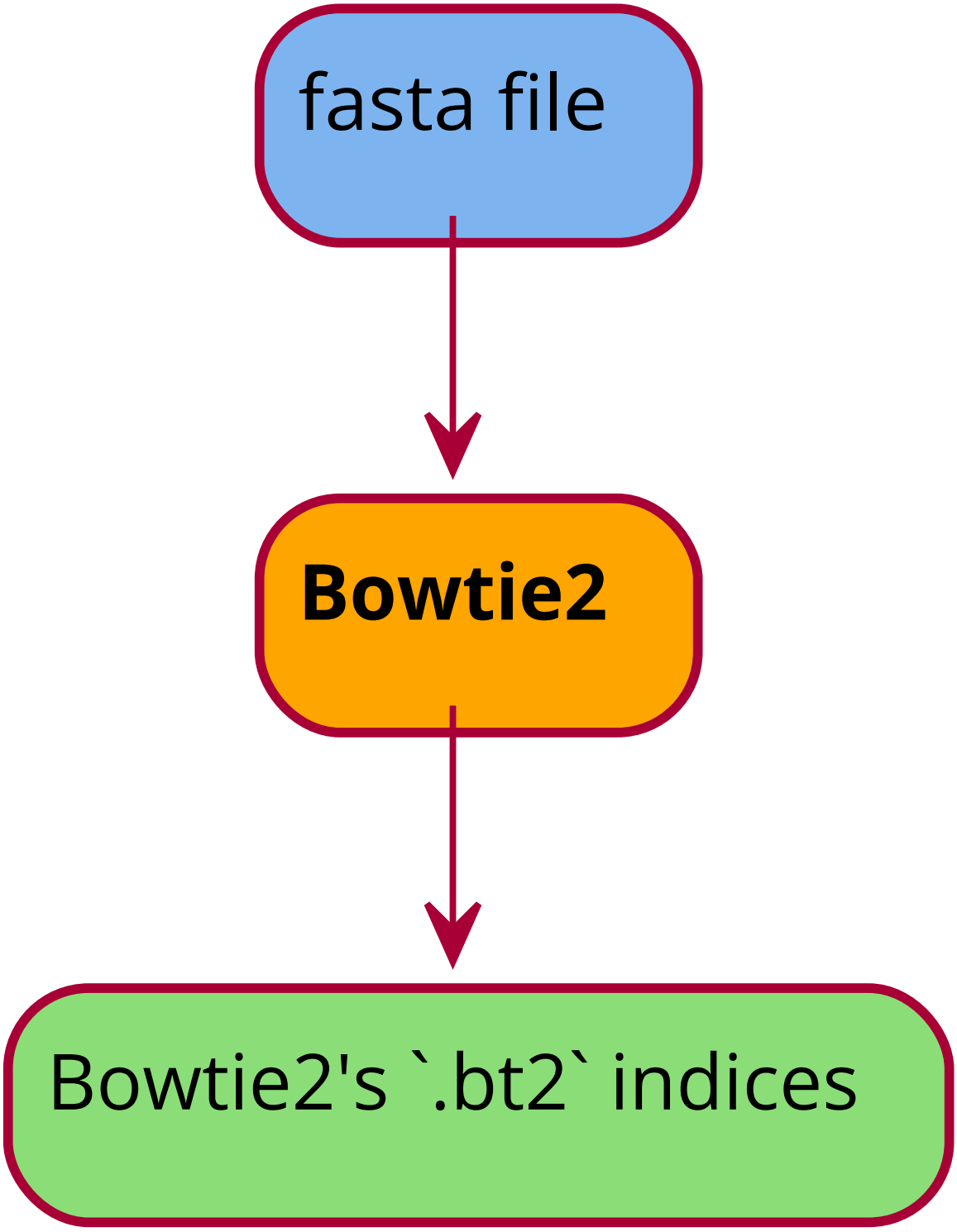
The table below contains a list of files present in the folders, along with some useful information.

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_unicycler.gfa	assembly in .gfa format	folder: result
DSXXXXXXXX-DTXXXXXX_ID_unicycler_assembly.fasta	hybrid assembly file in fasta format	folder: result
DSXXXXXXXX-DTXXXXXX_ID_quast.csv	quality statistics file of the assembly	folder: qc > result

2AS_indexing

Introduction

The *2AS_indexing* analysis indexes a FASTA file using Bowtie2 to generate to generate `.bt2` index files. The indexed reference genomes can be selected as host genomes for depletion in the host sequence depletion analysis.



Tool

Input

Result

Run 2AS_indexing analysis

The only bioinformatic tool performing indexing is Bowtie2; users can select it after chosing **2AS_indexing** analysis in the run analyses interface.

The input selection interface provides the advanced input selection mode.

The input for the indexing process is a FASTA file; consequently, all major analyses that produce a FASTA genome file can be used:

- step_2AS_import
- step_2AS_mapping
- step_2AS_denovo

Once the software is launched, the system generates a link to the Check analysis section and it will send the user a notification both when the analysis is successfully started and when the operation is completed.

> **Note:** this analysis is used to produce the files required for depletion of reads from the host organism. Only organisms for which this analysis has been **launched and imported**, either **manually or automatically after an upload from NCBI**, will be available as host genomes for read depletion in the host sequence depletion analysis.

Result folder

> To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are two subfolders:

- **meta** (*metadata*), where the log and configuration files of the executed process are saved;
- **result**, where the files containing the results produced by the analysis are saved.

The table below lists the files available in the result folder.

Bowtie:

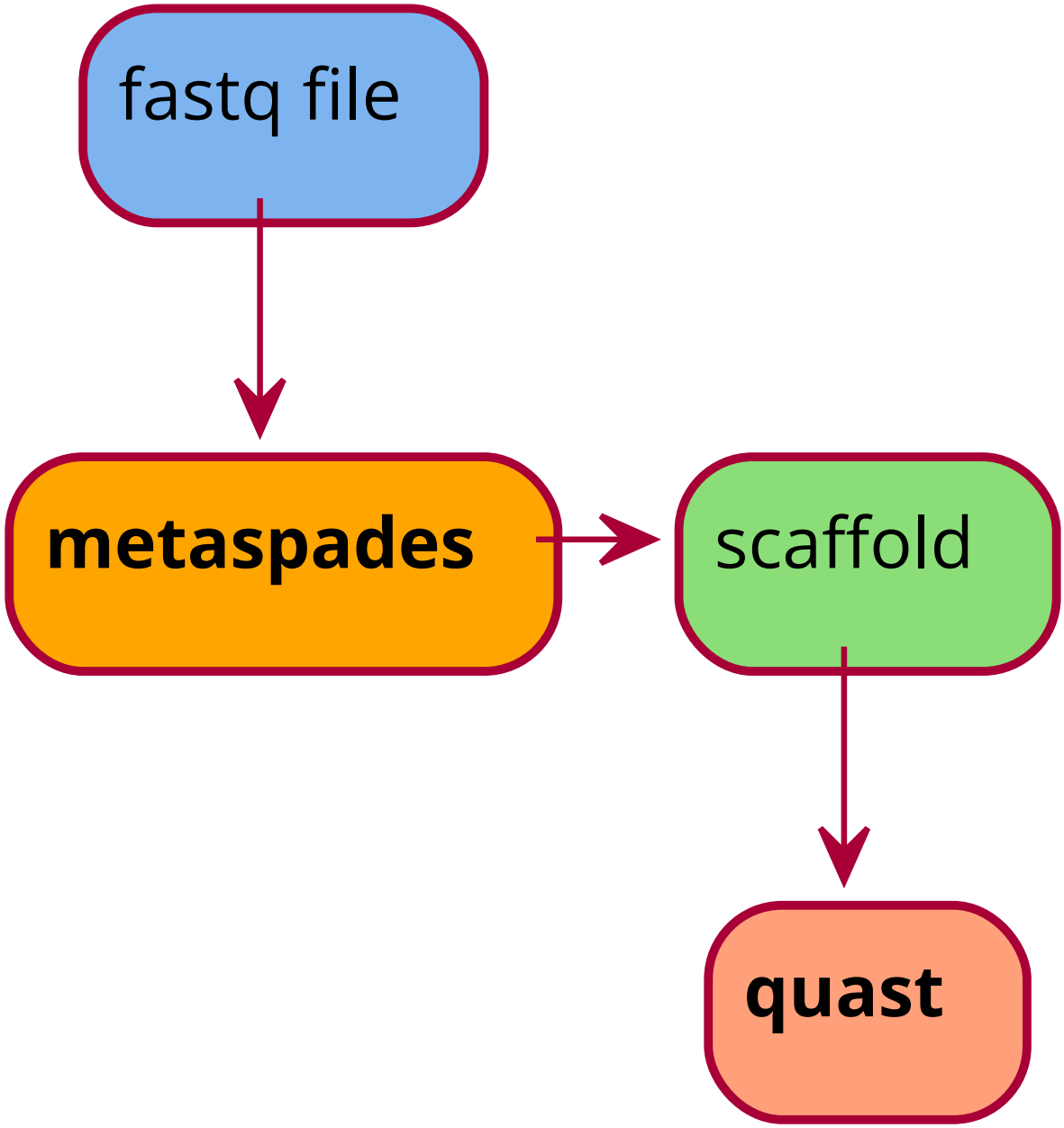
File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID.bt2	bowtie2 index files (`.bt2`)	folder: result
DSXXXXXXXX-DTXXXXXX_ID.fasta	the original fasta file on which indexing was performed	folder: result

2 Metagenomics

2MG_denovo

Introduction

2MG_denovo performs *de novo* assembly for metagenomic samples.



Tool
Input
Result
Quality Check tool

Run 2MG_denovo analysis

Once the analysis **2MG_denovo** has been selected in the dedicated run analyses page, it will be possible to choose the bioinformatics software to use, among those available for the analysis. The only tool available for this analysis is:

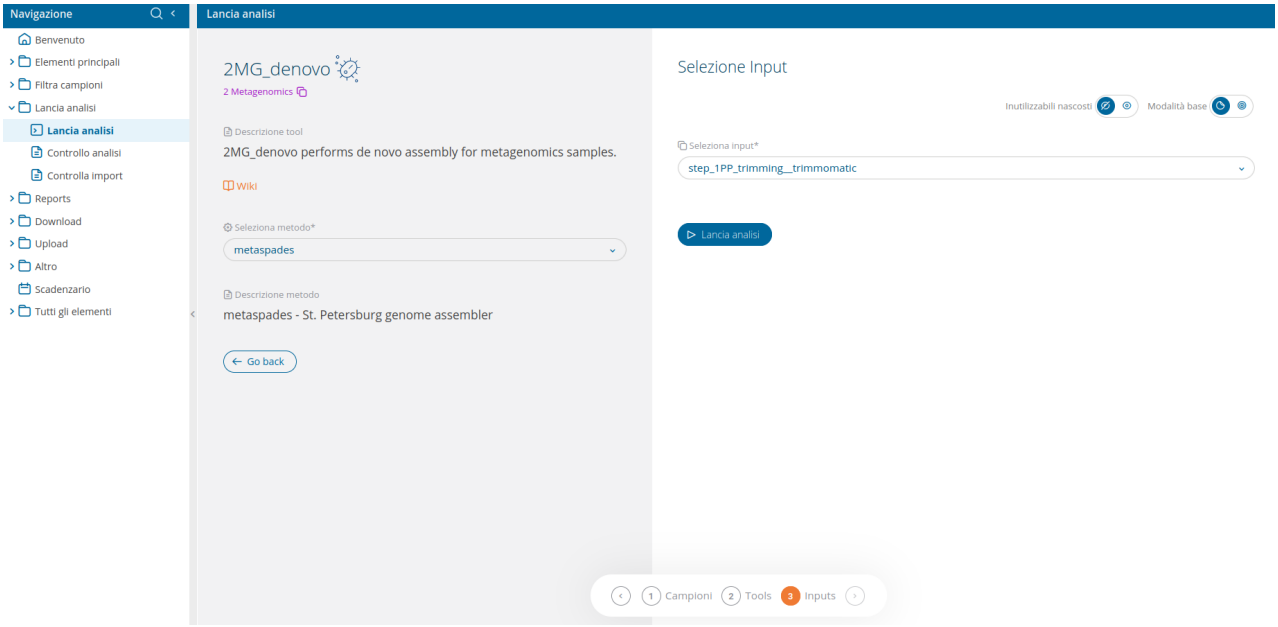
- **metaSPAdes** - St. Petersburg genome assembler

> **Pagina GitHub di metaSPAdes:** https://github.com/ablab/spades

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.

Accepted inputs can originate from:

- step_1PP_generated
- step_1PP_trimming
- step_1PP_hostdepl
- step_1PP_downsampling
- step_3TX_class
- step_4AN_AMR



Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

Result folder

> To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are 3 subfolders:

- **meta**: (*metadata*), where the log and configuration files of the executed process are saved;
- **qc**: (*quality check*) where check quality files are saved. It also contains the two sub-folders *meta* and *result*. For this analysis, quality check is performed with **quast**;
- **result**: where the files containing the results produced by the analysis are saved.

The table below lists the files generated by *metaSPAdes*.

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_spades_contigs.fasta	file of contigs from the *de novo* assembly	folder: result
DSXXXXXXXX-DTXXXXXX_ID_spades_scaffolds.fasta	file of the scaffolds	folder: result
DSXXXXXXXX-DTXXXXXX_ID_spades_scaffolds_L200.fasta	file containing scaffolds longer than 200bp	folder: result
DSXXXXXXXX-DTXXXXXX_ID_quast.csv	file of quality statistics of the assembly	folder: qc > result

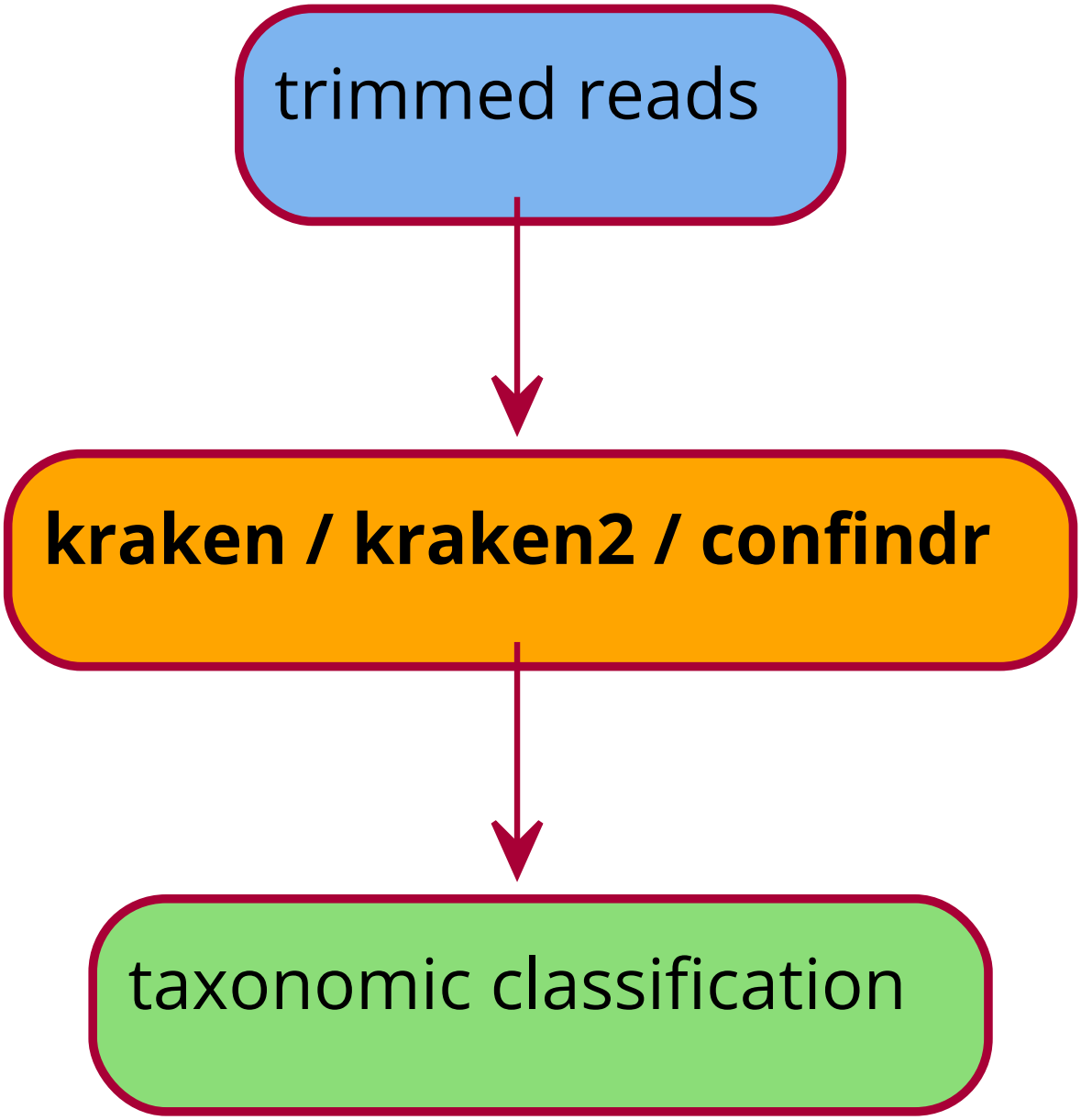
3 Taxonomy

3TX_class

Introduction

The *3TX_class* analysis performs taxonomic classification of the organisms of origin of the reads and includes a contamination checks.

3TX_class_kraken therefore produces the results related to contamination, which can be consulted in `Report > Detail > Checks/Metrics > Check of Taxa`.



Tool

Input

Result

Run 3TX_class analysis

Once the analysis **3TX_class** has been selected in the dedicated run analyses page, it will be possible to choose the bioinformatics software ("method") to use, among those available for the analysis. The available tools for this analysis are listed below.

For all microorganisms:

- **Kraken** - sequence taxonomic classification system
- **Kraken 2** - The second version of the Kraken sequence taxonomic classification system

> **Kraken GitHub page:** https://github.com/DerrickWood/kraken
>
> **Kraken 2 GitHub page:** https://github.com/DerrickWood/kraken2

Only for species belonging to *Escherichia*, *Salmonella* and *Listeria*:

- **ConFindr** - designed to find bacterial intra-species contamination in raw data generated by Illumina

  > **ConFindr GitHub page:** https://github.com/OLC-Bioinformatics/ConFindr/tree/main

Only for single long reads from **metagenomic** samples sequenced with Nanopore technology:

- **Centrifuge** - Classifier for metagenomic sequences

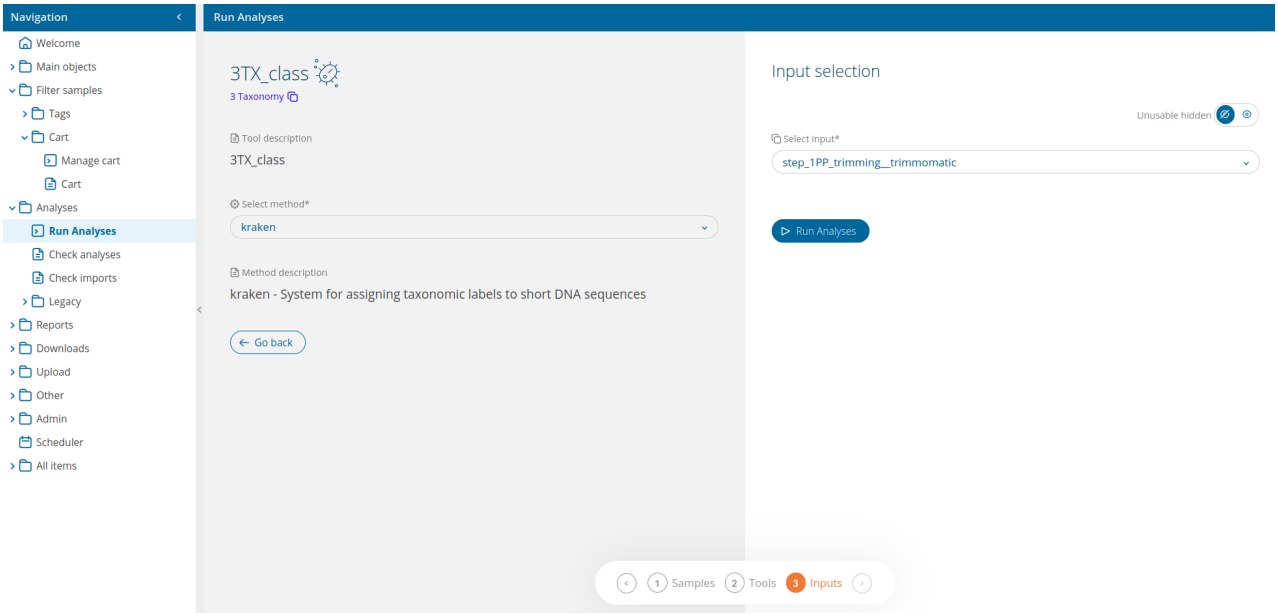
  > **Centrifuge GitHub page:** https://github.com/DaehwanKimLab/centrifuge

> **Note:** Kraken 2 differs from the first version. Since they are parallel software, both are available in the analysis launch system. For further information on how Kraken and Kraken 2 work and their differences, please refer to their official manual at the following link: https://github.com/DerrickWood/kraken2/blob/master/docs/MANUAL.markdown.

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.

Accepted inputs for *3TX_class* can originate from:

- step_1PP_trimming
- step_1PP_hostdepl
- step_1PP_downsampling



Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

Result folder

> To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are 2 subfolders:

- **meta**: (*metadata*), where the log and configuration files of the executed process are saved;
- **result**: where the files containing the results produced by the analysis are saved.

The tables below list the files generated by the available tools for *3TX_class*. For further details on the output files generated by the two versions of *Kraken*, please refer to kraken official manual and the help sheet of Kraken 2.

Kraken

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_kraken.tsv	kraken taxonomic classification file	folder: result

Kraken 2

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_bracken_genus.tsv	estimated abundance of reads identified by genus	folder: result
DSXXXXXXXX-DTXXXXXX_ID_bracken_genus_report.tsv	taxonomic classification report file at the genus level	folder: result
DSXXXXXXXX-DTXXXXXX_ID_bracken_species.tsv	estimated abundance of reads identified by species	folder: result
DSXXXXXXXX-DTXXXXXX_ID_bracken_species_report.tsv	taxonomic classification report file at the species level	folder: result
DSXXXXXXXX-DTXXXXXX_ID_kraken.tsv	taxonomic classification report file at the strain level	folder: result
DSXXXXXXXX-DTXXXXXX_ID_kraken.txt.gz	compressed text file containing sequences classified by Kraken 2	folder: result

ConFindr

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_contamination.csv	.csv table of contaminations identified in the sample by ConFindr	folder: result
DSXXXXXXXX-DTXXXXXX_ID_rmlst.csv	.csv table of alleles for each gene in the sample	folder: result
DSXXXXXXXX-DTXXXXXX_ID_confindr_report.csv	overall .csv table of samples and contaminations identified by ConFindr	folder: result

Centrifuge

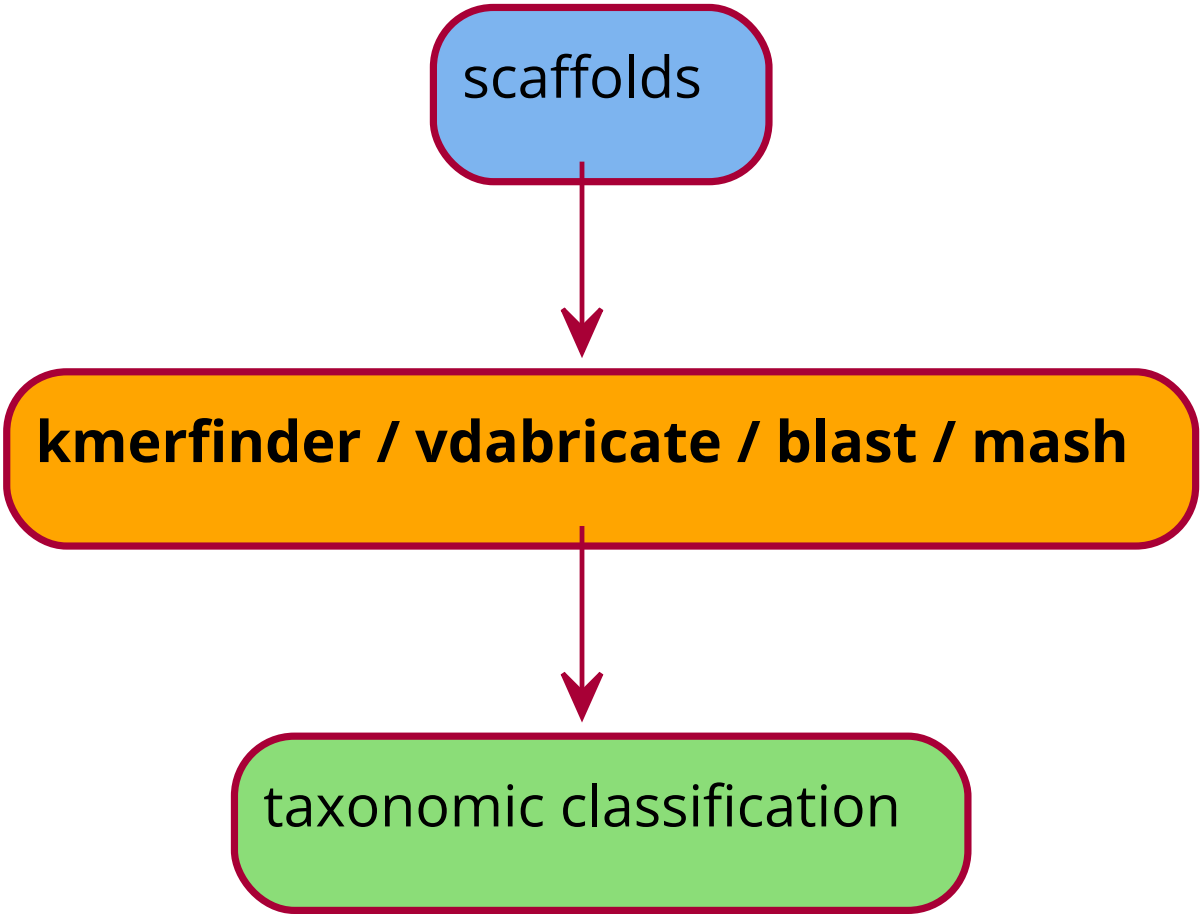
File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_centrifuge.output	tabular output file from Centrifuge containing the *taxa* ID, its score and the second-best score (open with a text editor or spreadsheet)	folder: result
DSXXXXXXXX-DTXXXXXX_ID_centrifuge.report	tabular report from Centrifuge containing the list of organisms detected in the sample and their statistics (open with a text editor or spreadsheet)	folder: result

For further information on Centrifuge outputs and software features, please refer to the Centrifuge GitHub official page.

3TX_species

Introduction

3TX_species assigns the closest bacterial or viral species to the sequence under analysis. If the tool used is **vdabricate**, the analysis identifies the best viral reference to be used for downstream analyses. **Kmerfinder** is the tool used for the best reference identification and determines the **assigned species**.



Tool
Input
Result

Run 3TX_species analysis

Once the analysis **3TX_species** has been selected in the dedicated run analyses page, it will be possible to choose the bioinformatics software ("method") to use, among those available for the analysis. The available tools for this analysis are:

For bacterial species assignment:

- **kmerfinder** - Prediction of bacterial species using a fast K-mer algorithm

  KmerFinder GitHub page: https://github.com/kcri-tz/kmerfinder.

For viral species assignment:

- **vdabricate** - Mass screening of contigs for viral species of a predefinite database using Abricate.

The *vdabricate* tool can be used to identify the best reference genome for a virus of interest, as determined by the sequence matches found in the viral sequence database (DB_Virus).

For all species:

BLAST - the Basic Local Alignment Search Tool identifies regions of similarity between biological sequences.

BLAST GitHub page: https://github.com/ncbi/blast_plus_docs.

- **Mash** - fast genome and metagenome distance estimation using MinHash.

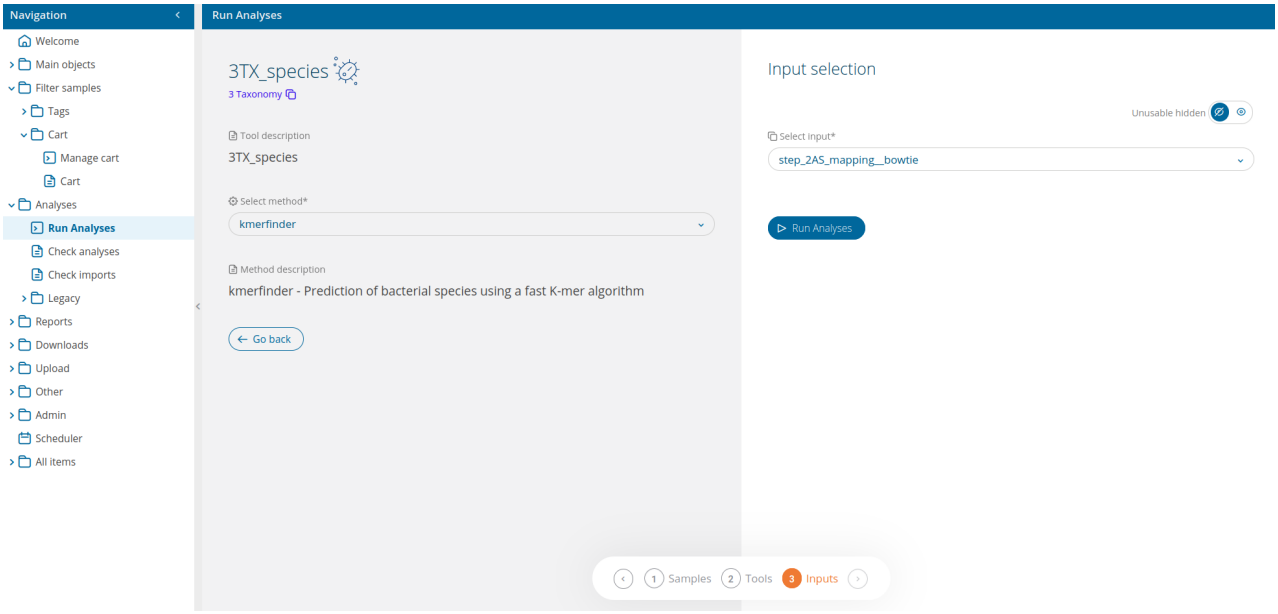
  Mash GitHub page: https://github.com/marbl/Mash .

For *3TX_species* it is necessary to specify the input sequences - whether derived from *de novo* assembly or from mapping - and the reference genome used for mapping, if the latter is selected as input.

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.

Accepted inputs for *3TX_species* can originate from:

- step_2AS_mapping
- step_2AS_denovo
- step_2MG_denovo



Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

Result folder

> To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are 2 subfolders:

- **meta**: (*metadata*), where the log and configuration files of the executed process are saved;
- **result**: where the files containing the results produced by the analysis are saved.

The tables below list the files generated by the available tools for *3TX_species*.

Kmerfinder

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_kmerfinder_bacterial.tsv	file containing the assigned bacterial species	folder: results
DSXXXXXXXX-DTXXXXXX_ID_kmerfinder_viral.tsv	file containing the assigned viral species	folder: results

vdabricate

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_vdabricate_calls.txt	identified matches in DB_Virus	folder: result

BLAST

File	Description	Location
results.csv	summary table of result files, with fields separated by ";"	folder: results
DSXXXXXXXX-DTXXXXXX_ID_blast_nt.asn.gz	sequence data in .asn (Abstract Syntax Notation) compressed format	folder: result
DSXXXXXXXX-DTXXXXXX_ID_blast_nt.csv	.csv table of sequence alignment results	folder: result
DSXXXXXXXX-DTXXXXXX_ID_blast_nt.txt	sequence alignment results in the standard BLASTn output	folder: result
DSXXXXXXXX-DTXXXXXX_ID_blast_nt_max_target_seqs_5.csv	.csv table of alignment results, filtered to the 5 closest sequences	folder: result

Mash

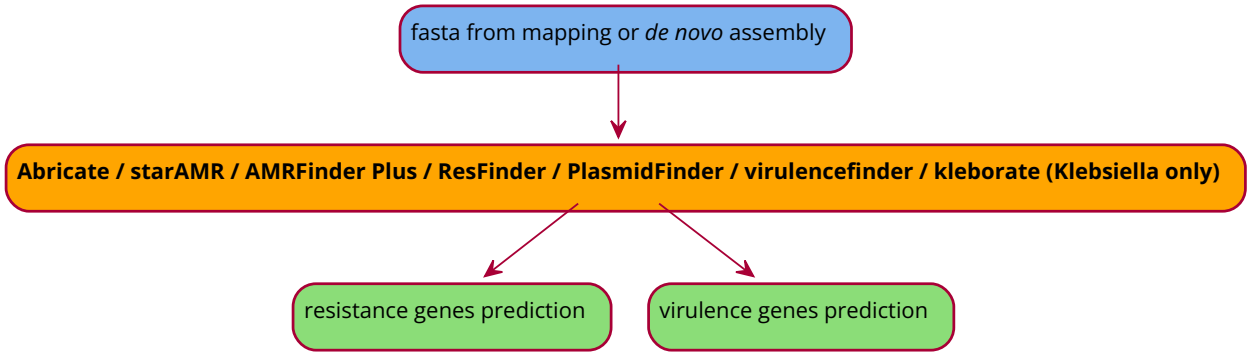
File	Description	Location
DS9573634-DT230426121427891_2019.TE.13476.1.1_mash_screen.tsv	intermediate table containing the sequence identity percentages, between the sample genome and the tested sequences	folder: result
DS9573634-DT230426121427891_2019.TE.13476.1.1_mash_screening_results.tsv	.tsv file containing the species final assignment	folder: result

4 Genome annotation

4AN_AMR

Introduction

4AN_AMR (Anti-Microbial Resistance) predicts the presence of genes related to antimicrobial resistance and virulence genes, in the reconstructed sequence of the microorganism of interest.



Tool
Input
Result

Run 4AN_AMR analysis

Once the analysis **4AN_AMR** has been selected in the dedicated run analyses page, it will be possible to choose the bioinformatics software ("method") to use, among those available for the analysis. The available tools in this case are:

For all microorganisms:

- **ABRicate** - Mass screening of contigs for antimicrobial resistance or virulence genes.

  ABRicate GitHub page: https://github.com/tseemann/abricate

- **AMRFinderPlus** - AMRFinderPlus identifies acquired antimicrobial resistance genes in bacterial protein and/or assembled nucleotide sequences as well as known resistance-associated point mutations for several taxa.

  AMRFinderPlus GitHub page: https://github.com/ncbi/amr/wiki/Running-AMRFinderPlus#usage

- **ResFinder** - identification of acquired antimicrobial resistance genes.

  ResFinder GitHub page: https://github.com/cadms/resfinder

- **PlasmidFinder** - gene identification for the identification and typing of plasmids.

  PlasmidFinder GitHub page: https://github.com/genomicepidemiology/plasmidfinder

For *Klebsiella pneumoniae* (but also for *K. oxytoca* and *Escherichia genus*):

- **Kleborate** - a tool to screen genome assemblies of Klebsiella pneumoniae and the Klebsiella pneumoniae species complex (KpSC).

  Kleborate GitHub page: https://kleborate.readthedocs.io/en/latest/Usage.html

For *Escherichia coli* only:

- **patho_typing** - identification of acquired antimicrobial resistance genes.

  Patho_typing GitHub page: https://github.com/B-UMMI/patho_typing

For *Staphylococcus aureus* only:

- **BLAST** - the Basic Local Alignment Search Tool finds regions of similarity between biological sequences.

  BLAST GitHub page: https://github.com/ncbi/blast_plus_docs

For *Campylobacter* only:

- **staramr** - (*AMR) compiles a summary report of detected antimicrobial resistance genes.

  staramr GitHub page: https://github.com/phac-nml/staramr

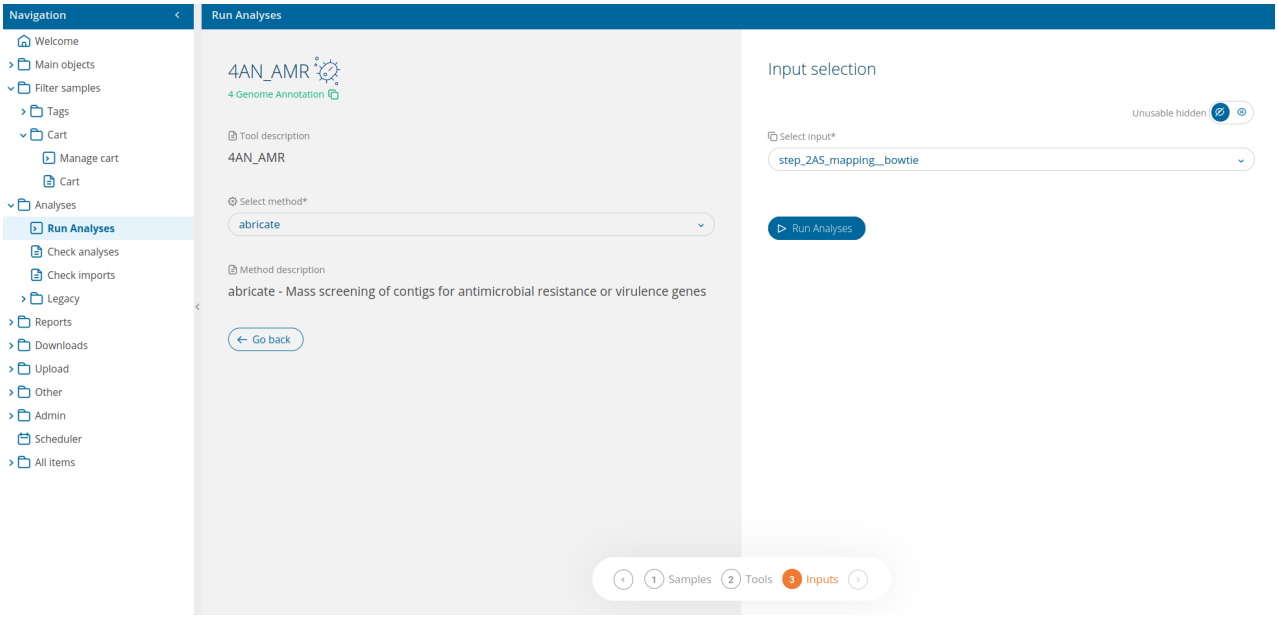
For *4AN_AMR* it is necessary to specify the input sequences - whether derived from *de novo* assembly or from mapping - and the reference genome used for mapping, if the latter is selected as input.

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.

Accepted inputs can originate from:

- step_2AS_mapping
- step_2AS_denovo
- step_2MG_denovo
- step_2AS_import

Note: in addition to the tools for the Anti-Microbial Resistance investigation, the available tools also include "**Filtering**" which **does not run AMR**: it filters AMR results (performed with one of the other available programs) on the base of the specified coverage and identity parameters. The usable inputs are *step_3TX_species__vdabricate* and *step_4AN_AMR__abricate*.



Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

Result folder

> To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are 2 subfolders:

- **meta**: (*metadata*), where the log and configuration files of the executed process are saved;
- **result**: where the files containing the results produced by the analysis are saved.

The tables below show the lists of files present in the folders, along with some useful information.

ABRicate

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_abricate_calls.txt	file with gene alignment results, coverage and database queries	folder: results
DSXXXXXXXX-DTXXXXXX_ID_abricate.summary	ABRicate summary file containing genes and their coverage	folder: results
DSXXXXXXXX-DTXXXXXX_ID_output_abricate_AMR.csv	antibiotic resistance genes summary file	folder: results
DSXXXXXXXX-DTXXXXXX_ID_output_abricate_VF.csv	virulence genes summary file	folder: results

For further details regarding the files generated by ABRicate, please refer to the official manual of ABRicate.

BLAST

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_blast_out.tsv	.tsv table containing BLAST results	folder: results

staramr

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_staramr_detailed_summary.tsv	.tsv table of results organized by data type (resistance gene/plasmid/MLST scheme...)	folder: results
DSXXXXXXXX-DTXXXXXX_ID_staramr_mlst.tsv	results of the 7-loci MLST scheme for species-specific genes + Sequence Type (ST)	folder: results
DSXXXXXXXX-DTXXXXXX_ID_staramr_plasmidfinder.tsv	.tsv results with sequences belonging to identified plasmids	folder: results
DSXXXXXXXX-DTXXXXXX_ID_staramr_pointfinder.tsv	AMR results in .tsv format with Pointfinder position	folder: results
DSXXXXXXXX-DTXXXXXX_ID_staramr_resfinder.tsv	resistance prediction results and resistance gene alignment	folder: results
DSXXXXXXXX-DTXXXXXX_ID_staramr_results.xlsx	AMR results in spreadsheet format	folder: results
DSXXXXXXXX-DTXXXXXX_ID_staramr_summary.tsv	AMR results in .tsv format	folder: results

ResFinder

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_resfinder_ResFinder_Hit_in_genome_seq.fsa	fasta file with the sequence of the resistance gene(s) identified in the sample genome with extended header	folder: results
DSXXXXXXXX-DTXXXXXX_ID_resfinder_ResFinder_Resistance_gene_seq.fsa	fasta file with the sequence of the resistance gene(s) identified in the sample genome with reduced header	folder: results
DSXXXXXXXX-DTXXXXXX_ID_resfinder_ResFinder_results.txt	text file with summary table and sequence identity	folder: results
DSXXXXXXXX-DTXXXXXX_ID_resfinder_ResFinder_results_tab.txt	.tsv file with the list of identified resistance genes (it lists only the genes that are present)	folder: results
DSXXXXXXXX-DTXXXXXX_ID_resfinder_ResFinder_results_table.txt	text file with the list of resistance genes (it specifies both the genes that are present and those that were not detected)	folder: results
DSXXXXXXXX-DTXXXXXX_ID_resfinder_pheno_table.txt	text file with results and phenotypes table	folder: results
DSXXXXXXXX-DTXXXXXX_ID_resfinder_std_format_under_development.json		folder: results

AMRFinderPlus

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_amrfinderplus.tsv	.tsv table with information on the identified resistance genes	folder: result

kleborate

File	Description	Location
DSXXXXXXXX-DTXXXXXX_kleborate.tsv	.tsv table containing all typing results	folder: result

Patho_typing

File	Description	Location
DSXXXXXXXX-DTXXXXXX_patho_typing_rematchModule_report.txt	.tsv table containing information on genes identified by ReMatCh	folder: result
DSXXXXXXXX-DTXXXXXX_patho_typing_report.txt	.tsv table containing the possible pathotypes identified	folder: result
DSXXXXXXXX-DTXXXXXX_patho_typing_extended_report.txt	.tsv table containing information on the genes identified for pathotype attribution	folder: result

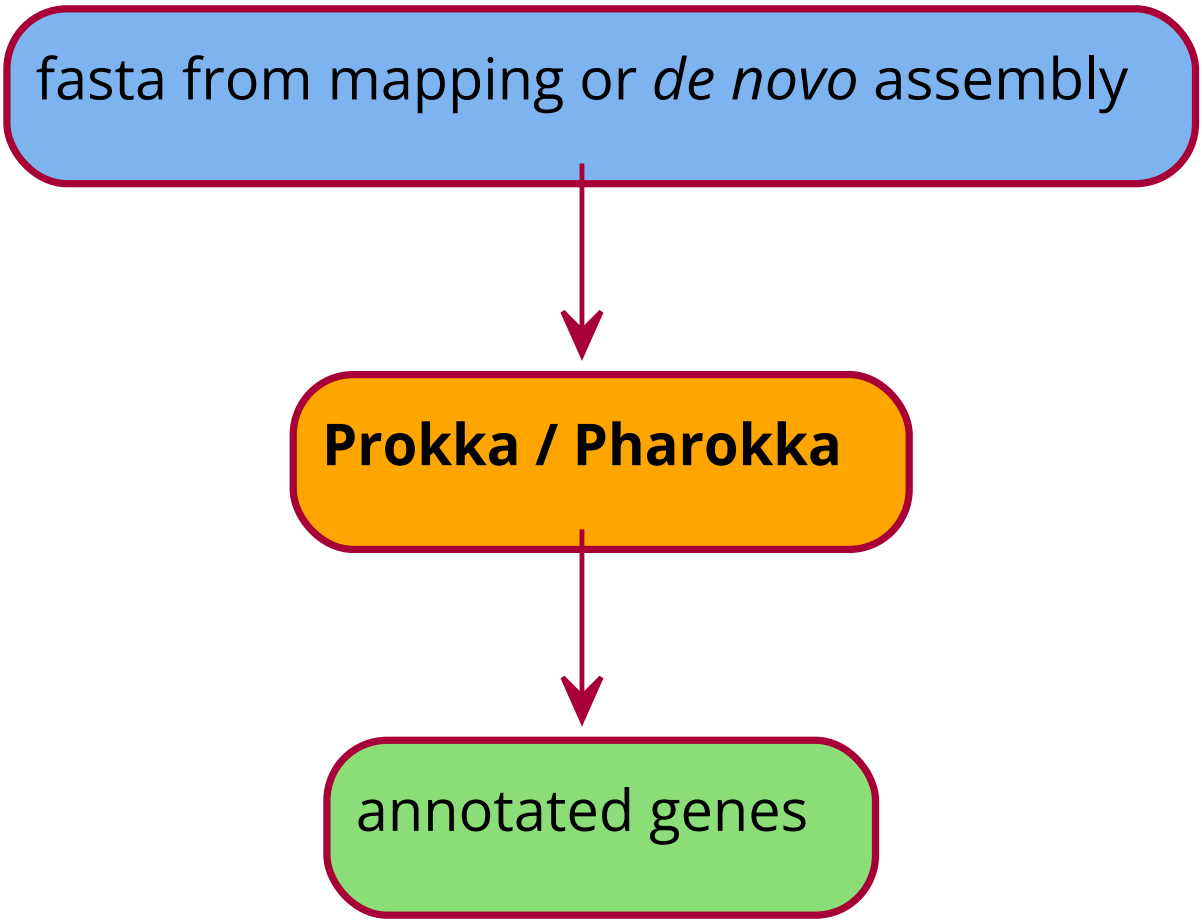
plasmidfinder

File	Description	Location
DSXXXXXXXX-DTXXXXXX_plasmidfinder_results.tsv	.tsv table containig information on the identified plasmids	folder: result
DSXXXXXXXX-DTXXXXXX_plasmidfinder_kma_*.res/aln/fsa	kma results for the identified plasmids	folder: result

4AN_genes

Introduction

4AN_genes performs the functional genome annotation, which identifies the possible encoded proteins and the ORFs.



Tool
Input
Result

Run 4AN_genes analysis

Once the analysis **4AN_genes** has been selected in the dedicated run analyses page, it will be possible to choose the bioinformatics software ("method") to use, among those available for the analysis. The available tools in this case are:

- **Prokka** - Tool to annotate genomes of bacteria, archaea and viruses.

  Prokka GitHub page: https://github.com/tseemann/prokka

- **Pharokka** - Annotation tool for bacteriophage genomes and metagenomes

  Pharokka GitHub page: https://github.com/gbouras13/pharokka

Prokka tool
Prokka is generally used to annotate genomes of bacteria, archaea and viruses. Its parameters include:

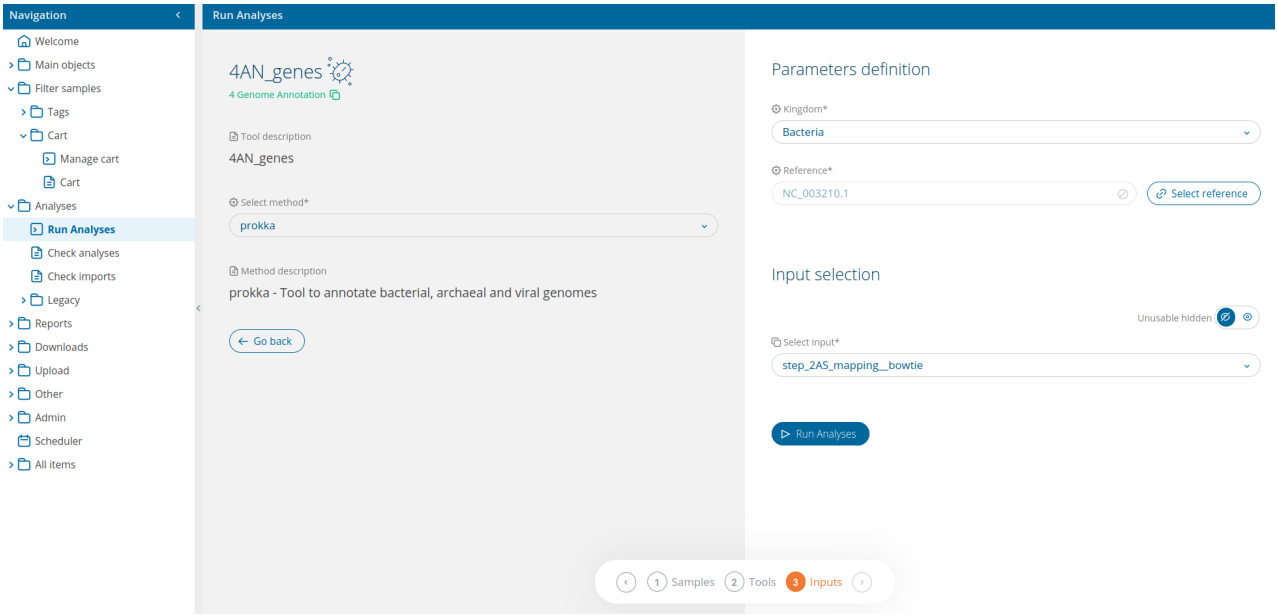
- selecting a **kingdom** among:
  - viruses;
  - bacteria;
  - archaea;
  - and, in addition, mitochondria;
- selecting a **reference** genome;
- selecting the **input** analysis.

For Prokka, the advanced input selection mode is also available.

Accepted inputs can originate from:

- step_2AS_mapping
- step_2AS_denovo
- step_2AS_hybrid
- step_2AS_import

It will be necessary to specify the input sequences - whether derived from *de novo* assembly or from mapping - and the reference genome used for mapping, if the latter is selected as input.



Pharokka tool

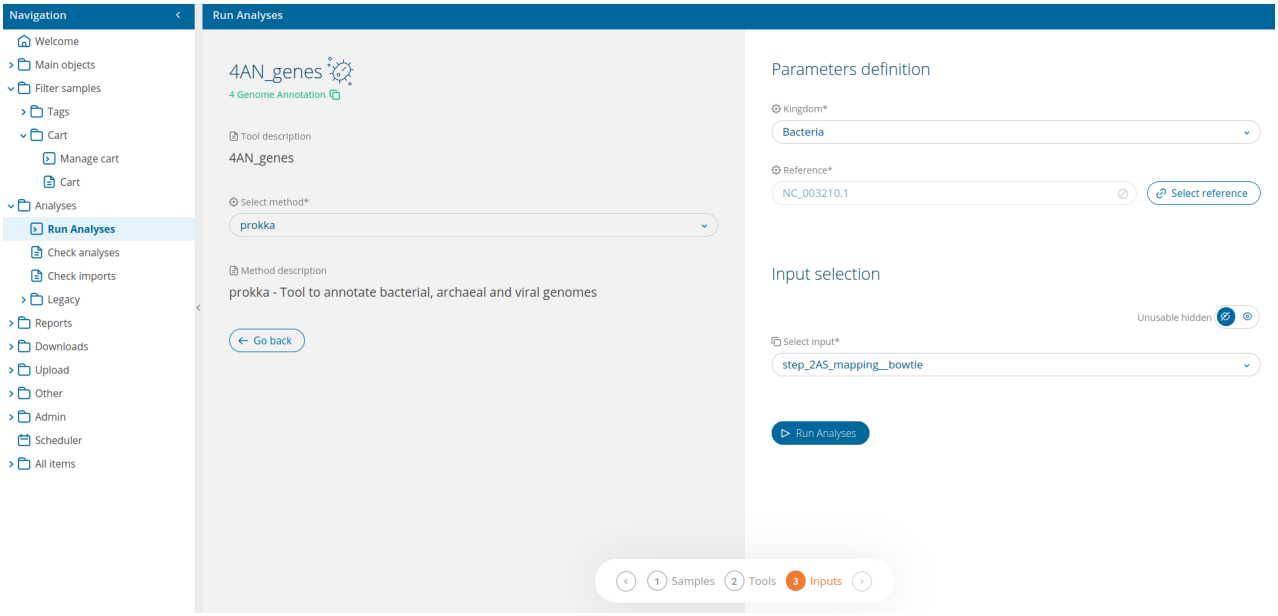
Pharokka is a tool developed **specifically and exclusively** for the annotation of bacteriophage genomes. The required parameters are only 2:

- the software to use for **CDS prediction** (the available tools are **Phanotate** and **Prodigal**; Phanotate is the *default* tool, as it is specific for prediction on the smaller phage genomes);
- the selection of the **input** analysis.

The advanced input selection mode is also available for Pharokka.

Accepted inputs can originate from:

- step_2AS_mapping
- step_2AS_denovo
- step_2AS_hybrid



In this case as well, it will be necessary to specify the reference genome used for mapping, if the latter is selected as input for the analysis.

Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

Result folder

> To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are 2 subfolders:

- **meta**: (*metadata*), where the log and configuration files of the executed process are saved;
- **result**: where the files containing the results produced by the analysis are saved.

The tables below list the files contained in the folders, along with some useful information.

Prokka

The following table lists the file generated by Prokka.

File	Description	Location
log_errore_controlli_esami.log	run's warning and error log	folder: results
metadata_samples.tsv	summary table in .tsv format of sample metadata	folder: results
results.csv	summary table separated by semicolon (";") containing sample IDs and information	folder: results
DSXXXXXXXX-DTXXXXXX_ID_prokka_REFID_result.err	report of inconsistencies and errors encountered during the run (text file)	folder: results
DSXXXXXXXX-DTXXXXXX_ID_prokka_REFID_result.faa	amino acid sequences from translation of identified coding genes (.faa format - fasta amino acid)	folder: results
DSXXXXXXXX-DTXXXXXX_ID_prokka_REFID_result.ffn	nucleotide sequences of identified coding genes (.fnn format - fasta nucleotide)	folder: results
DSXXXXXXXX-DTXXXXXX_ID_prokka_REFID_result.fna	nucleotide sequences of identified coding genes (.fna format)	folder: results
DSXXXXXXXX-DTXXXXXX_ID_prokka_REFID_result.fsa	sequences in .fsa format (fragment analysis data file)	folder: results
DSXXXXXXXX-DTXXXXXX_ID_prokka_REFID_result.gbk	output file in GenBank format	folder: results
DSXXXXXXXX-DTXXXXXX_ID_prokka_REFID_result.gff	output file in .gff format (General Feature Format)	folder: results
DSXXXXXXXX-DTXXXXXX_ID_prokka_REFID_result.log	run log of Prokka	folder: results
DSXXXXXXXX-DTXXXXXX_ID_prokka_REFID_result.sqn	file for GenBank submission in Sequin format	folder: results
DSXXXXXXXX-DTXXXXXX_ID_prokka_REFID_result.tbl	text file with information on sequence and *loci*	folder: results
DSXXXXXXXX-DTXXXXXX_ID_prokka_REFID_result.tsv	.tsv list of *loci* and proteins from mapped coding genes	folder: results
DSXXXXXXXX-DTXXXXXX_ID_prokka_REFID_result.txt	statistics on identified CDS	folder: results
proteins.faa	sequences of identified proteins in .faa format	folder: results

For more details on files generated by Prokka, please refer to Prokka official manual.

Pharokka

The following table lists the files generated by Pharokka.

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_pharokka_REFID.gbk	GenBank file, in .gbk format, with sequence features	folder: results
DSXXXXXXXX-DTXXXXXX_ID_pharokka_REFID.gff	GenBank file, in .gff format, with sequence features	folder: results
DSXXXXXXXX-DTXXXXXX_ID_pharokka_REFID_cds_final_merged_output.tsv	.tsv tabular file containing the genome annotation	folder: results

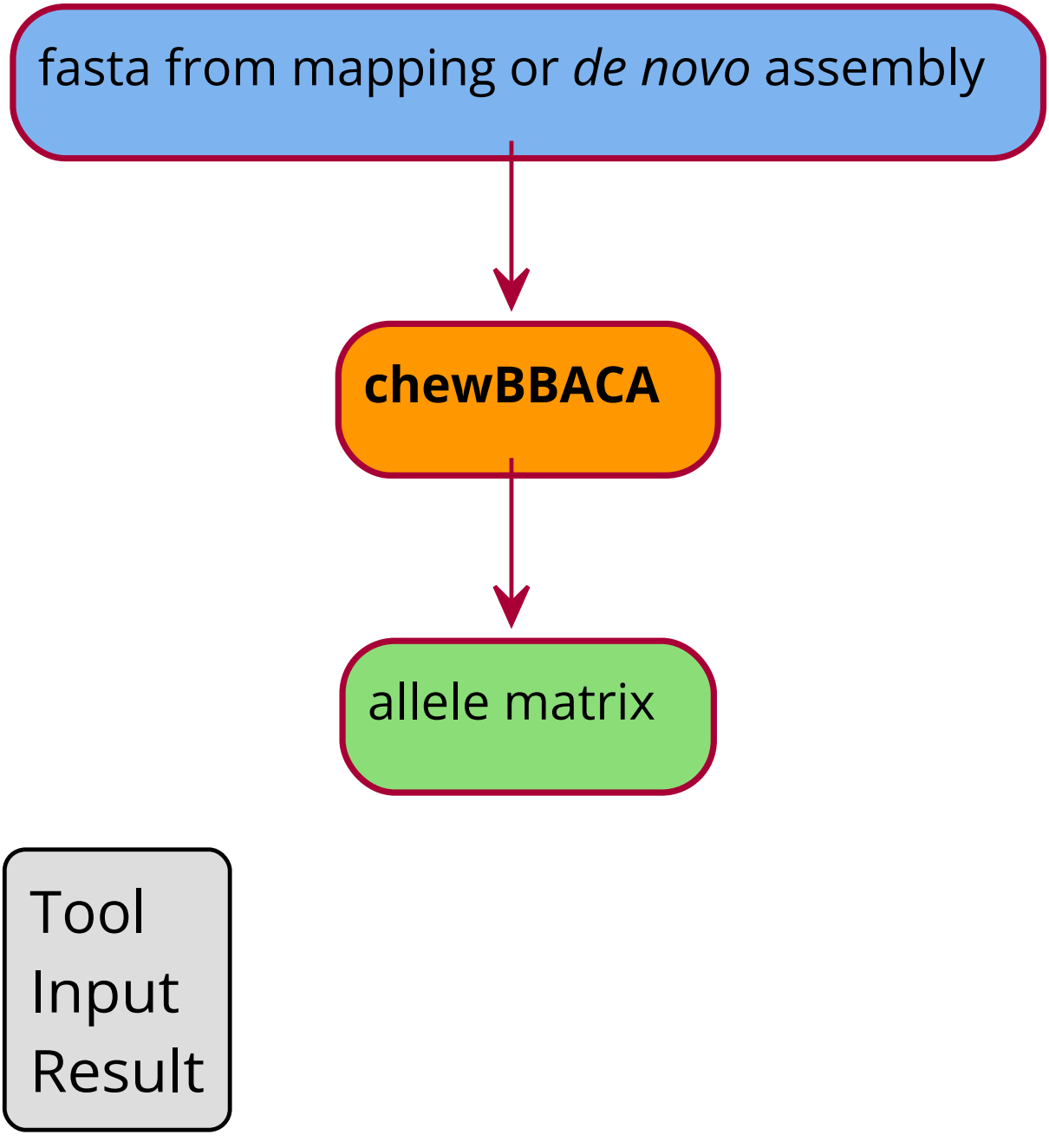
For more details on files generated by Pharokka, please refer to Pharokka official manual.

4 in silico typing

4TY_cgMLST

Introduction

4TY_cgMLST performs the "core genome Multi-Locus Sequence Typing" (cgMLST), a bacterial isolate characterization protocol, which allows identification of clones in microbial population.



Run 4TY_cgMLST analysis

Once the analysis **4TY_cgMLST** has been selected in the dedicated run analyses page, it will be possible to choose the bioinformatics software ("method") to use, among those available for the analysis. The available tool in this case is **chewBBACA** - BSR-Based Allele Calling Algorithm.

chewBBACA GitHub page: https://github.com/B-UMMI/chewBBACA

In the table below, the available schemes for the supported microorganisms are listed.

Species	Scheme name	Number of *loci*	Date of last update
Mycobacterium bovis	`m_bovis_2891_230720`	2891	20/07/2023
Listeria monocytogenes	`l_mono_chewie_1748_220623`	1748	23/06/2022
Campylobacter jejuni	`c_jejuni_678_180728`	678	28/07/2018
Campylobacter coli	`c_coli_528_220722`	528	22/07/2022
Staphylococcus aureus	`s_aureus_ridom_1861_210617`	1861	17/06/2021
Brucella melitensis	`b_melitensis_bme_2584_211118`	2584	18/11/2021
Klebsiella pneumoniae	`k_pneumoniae_pasteur_629_250617`	629	17/06/2025
Klebsiella pneumoniae	`k_pneumoniae_pasteur_2759_250617`	2759	17/06/2025
Escherichia coli	`e_coli_chewie_2360_210531`	2360	31/05/2021
Salmonella enterica	`s_enterica_chewie_3255_210531`	3255	31/05/2021
Brucella spp.	`b_wide_1764_220709`	1764	09/07/2022
Bacillus anthracis	`b_anthracis_pubmlst_3803_231214`	3803	14/12/2023
Acinetobacter baumannii	`a_baumannii_ridom_2390_250612`	2390	12/06/2025
Pseudomonas aeruginosa	`p_aeruginosa_ridom_3867_250612`	3867	12/06/2025
Serratia marcescens	`s_marcescens_ridom_2692_250926`	2692	26/09/2025

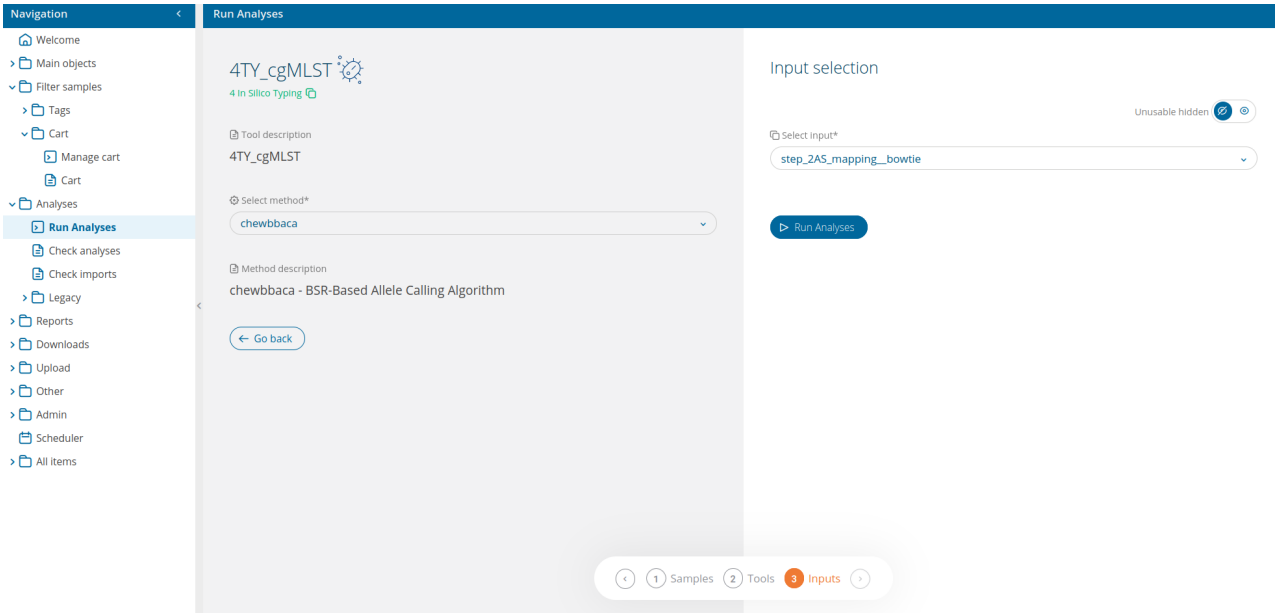
Note: Running *4TY_cgMLST* on a microorganism for which no cgMLST scheme is available, will result in the run falling.

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.

Accepted inputs can originate from:

- step_2MG_denovo
- step_2AS_denovo
- step_2AS_mapping
- step_2AS_import

It will be necessary to specify the input sequences - whether derived from *de novo* assembly or from mapping - and the reference genome used for mapping, if the latter is selected as input.



Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

Result folder

> To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are 3 subfolders:

- **meta**: (*metadata*), where the log and configuration files of the executed process are saved;
- **qc**: (*quality check*), where quality check files are saved;
- **result**: where the files containing the results produced by the analysis are saved.

> For allele calling with chewBBACA, the results will be provided using *crc32 encoding* (link to crc32 on Wikipedia).

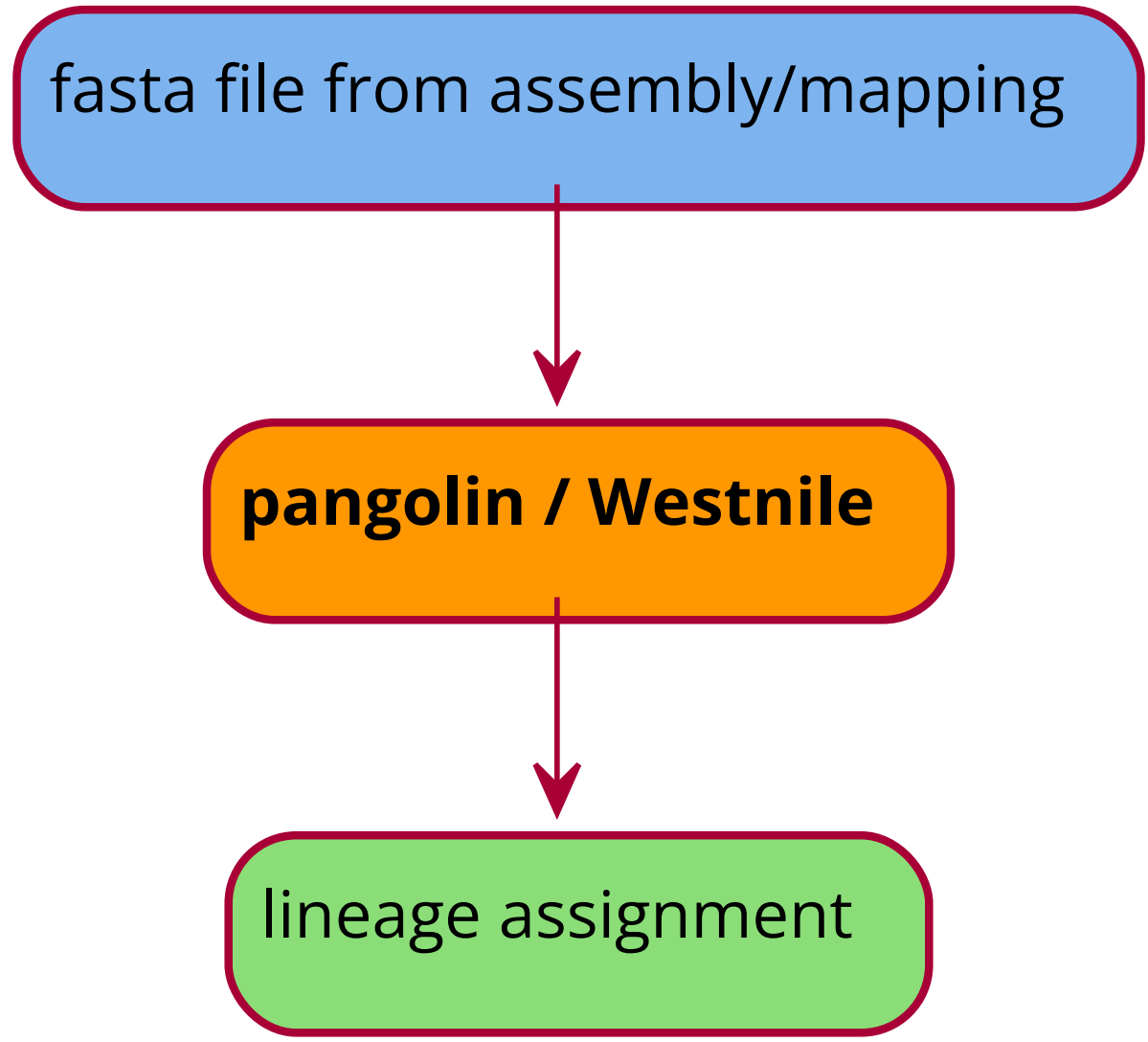
File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_chewbbaca_new_alleles.txt	sequences of newly-identified alleles	folder: result
DSXXXXXXXX-DTXXXXXX_ID_chewbbaca_results_contigsInfo.tsv	info about the contig mapped on each *locus*	folder: result
DSXXXXXXXX-DTXXXXXX_ID_chewbbaca_results_statistics.tsv	statistics on *loci* encoded as EXC, INF, LNF, PLOT, NIPH, ALM, ASM	folder: result
DSXXXXXXXX-ID_import_chewbbaca_check.csv	quality check with info on calledPerc, calledNum, annotated, new, notFound, discarded	qc > result directory
DSXXXXXXXX-DTXXXXXX_ID_chewbbaca_results_crc32.tsv	chewBBACA allelic calls with *crc32 encoding*, in .tsv format	cartella "result"

For more information on *locus* encoding and on chewBBACA output files, please refer to chewBBACA official guide.

4TY_lineage

Introduction

4TY_lineage assigns the lineage for SARS-CoV2 and West Nile Virus and is exclusive to samples of those viruses.



Tool
Input
Result

Run 4TY_lineage analysis

Once the analysis **4TY_lineage** has been selected in the dedicated run analyses page, it will be possible to choose the bioinformatics software ("method") to use, among those available for the analysis. The available tools in this case are:

SARS-CoV2 lineage assignment:

- **Pangolin** - Phylogenetic Assignment of Named Global Outbreak LINeages.

  Pangolin GitHub page: https://github.com/cov-lineages/pangolin.

West Nile Virus lineage assignment:

- **Westnile**

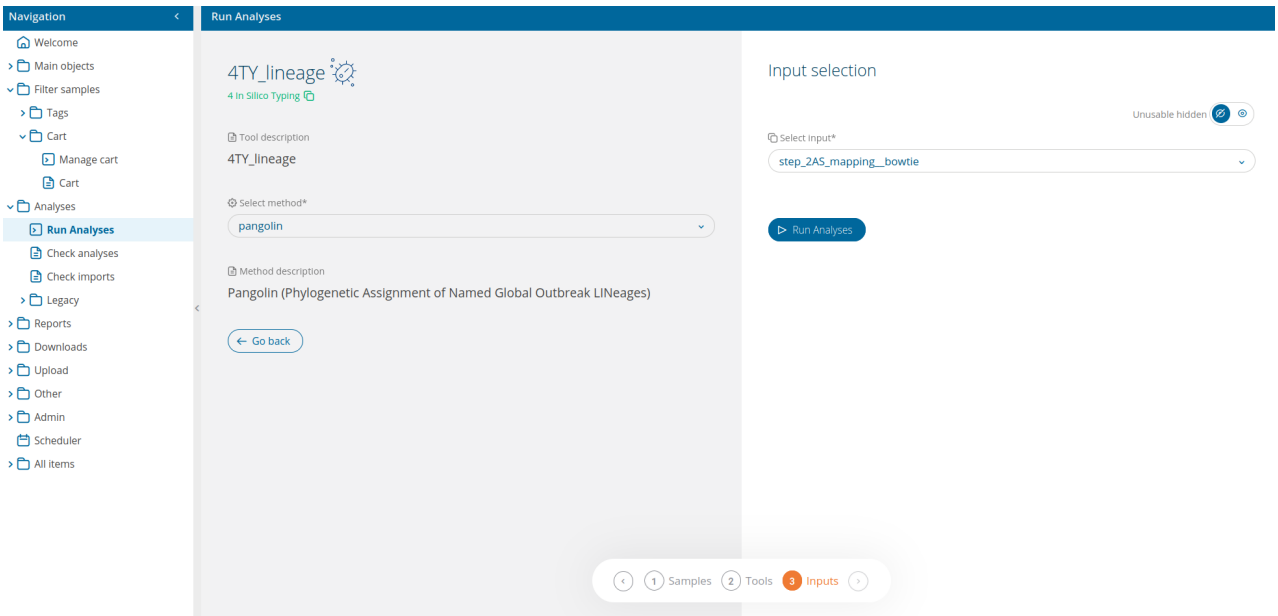
The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.

Accepted inputs can originate from:

- step_2MG_denovo
- step_2AS_denovo
- step_2AS_mapping
- step_2AS_import
- step_1PP_hostdepl
- step_1PP_downsampling
- step_1PP_trimming
- step_1PP_filtering

For **Westnile**, the only available input is the fasta from *2AS_mapping*.

t will be necessary to specify the input sequences - whether derived from *de novo* assembly or from mapping - and the reference genome used for mapping, if the latter is selected as input.



Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

Result folder

> To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are 2 subfolders:

- **meta**: (*metadata*), where the log and configuration files of the executed process are saved;
- **result**: where the files containing the results produced by the analysis are saved.

Pangolin

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_pangolin_lineage_report.csv	.csv file with lineage assignment	folder: resul

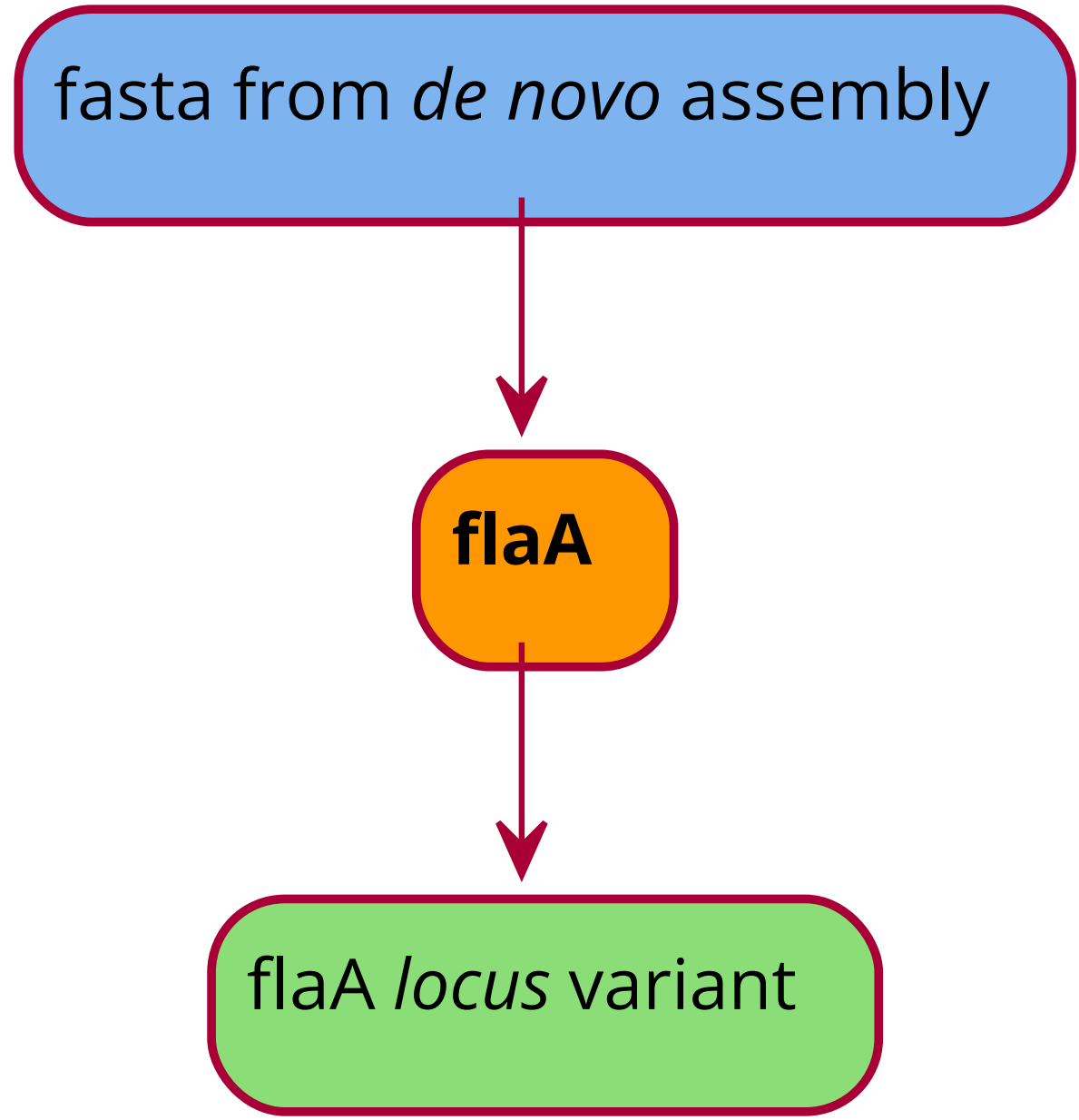
Westnile

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_westnile_lineage.csv	.csv file with lineage assignment	folder: result

4TY_flaA

Introduction

4TY_flaA determines, *in silico*, the flaA *locus* variant of *Campylobacter*.



Tool
Input
Result

Run 4TY_flaA analysis

Once the analysis **4TY_flaA** has been selected in the dedicated run analyses page, it will be possible to choose the bioinformatics software ("method") to use, among those available for the analysis. The available tool in this case is:

- **flaA** - Multilocus sequence typing, with *mlst* software, specific to the flaA profile.

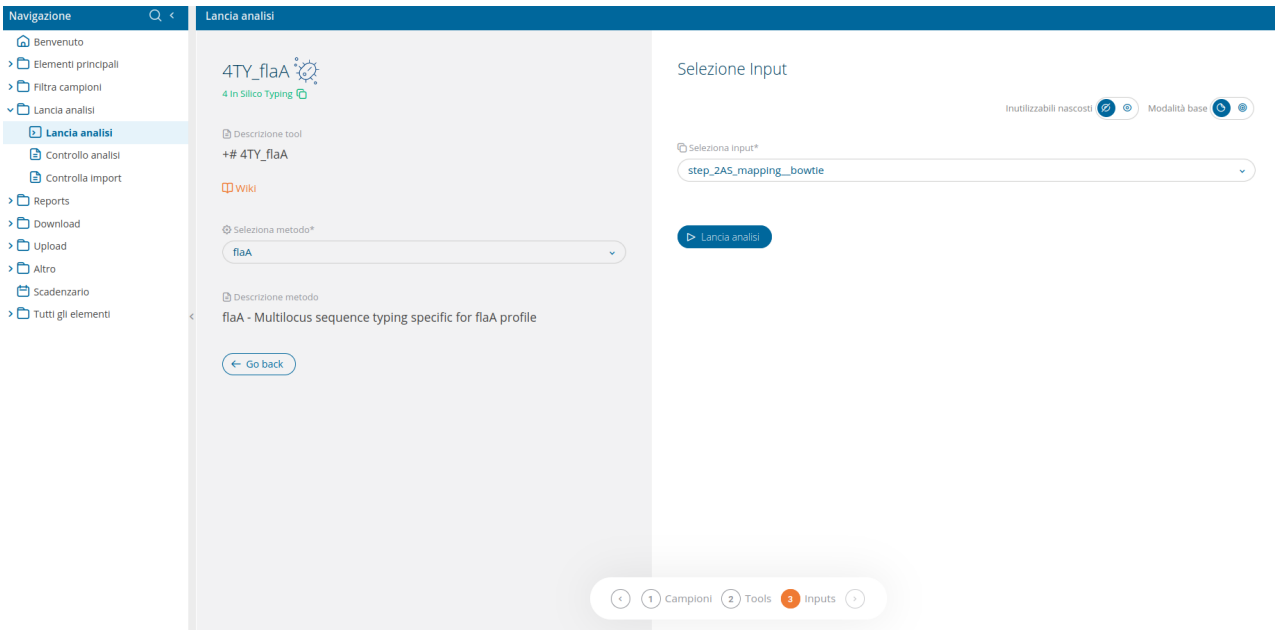
Pagina GitHub di mlst: https://github.com/tseemann/mlst.

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.

Accepted inputs can originate from:

- step_2MG_denovo
- step_2AS_denovo
- step_2AS_mapping
- step_2AS_import

It will be necessary to specify the input sequences - whether derived from *de novo* assembly or from mapping - and the reference genome used for mapping, if the latter is selected as input.



Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

Result folder

To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are 2 subfolders:

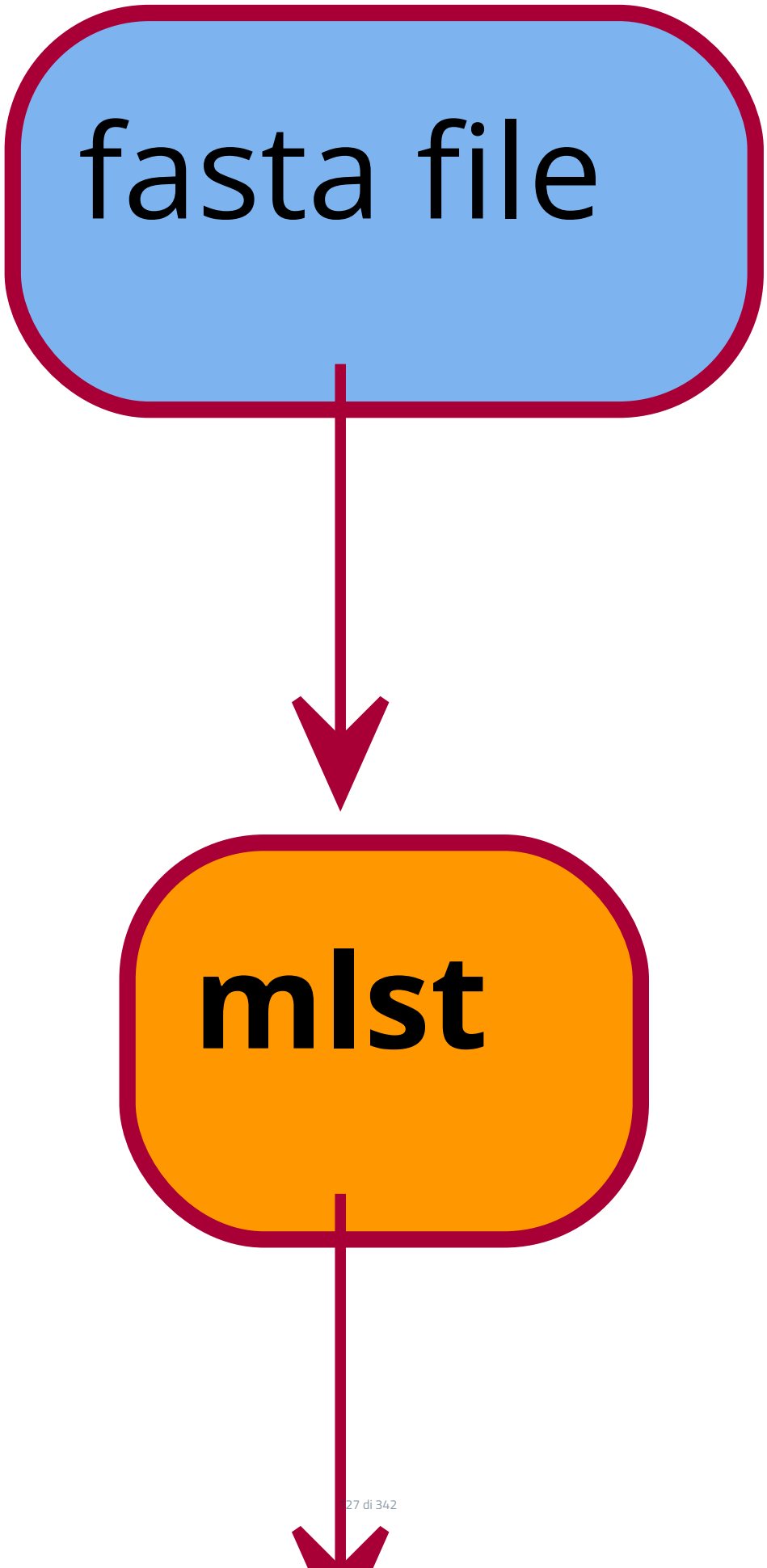
- **meta**: (*metadata*), where the log and configuration files of the executed process are saved;
- **result**: where the files containing the results produced by the analysis are saved.

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_flaA.tsv	allelic variant of the flaA locus	folder: result

4TY_MLST

Introduction

4TY_MLST performs the *in silico* Multi Locus Sequence Typing, assigning both Sequence Type (ST) and Clonal Complex (CC) where it is possible.


fasta file
mlst

Run 4TY_MLST analysis

Once the analysis **4TY_MLST** has been selected in the dedicated run analyses page, it will be possible to choose the bioinformatics software ("method") to use, among those available for the analysis. The available tool in this case is:

- **mlst** - *In vitro* Multi-Locus Sequence Typing

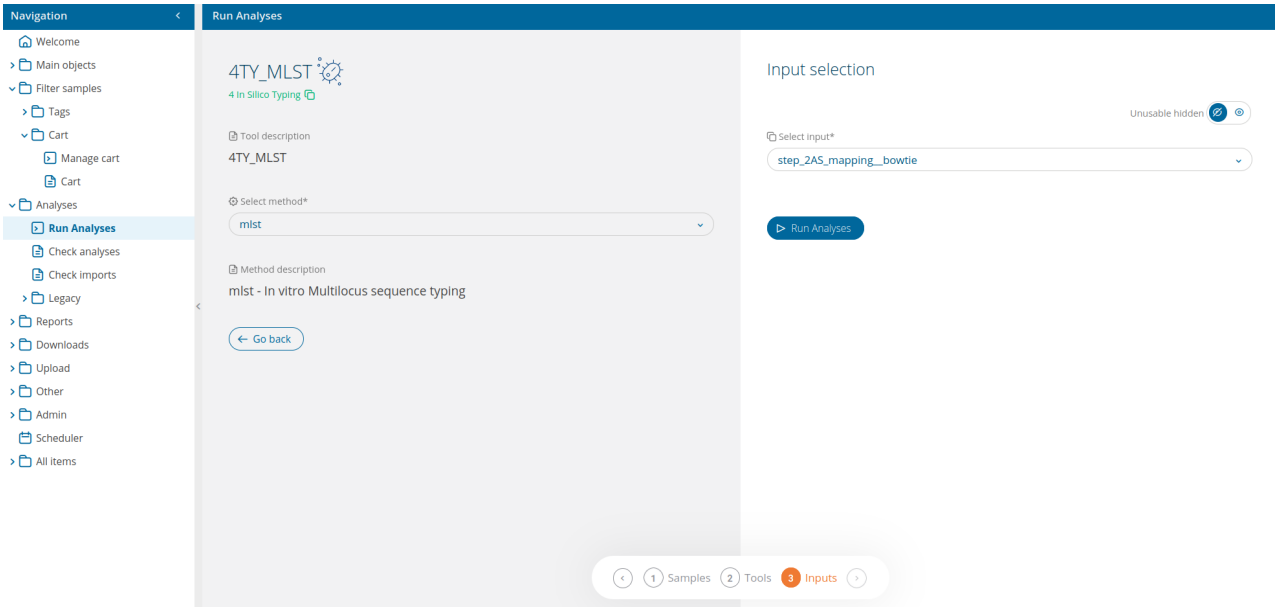
> **mlst GitHub page:** https://github.com/tseemann/mlst

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.

Accepted inputs can originate from:

- step_2MG_denovo
- step_2AS_denovo
- step_2AS_mapping
- step_2AS_import

It will be necessary to specify the input sequences - whether derived from *de novo* assembly or from mapping - and the reference genome used for mapping, if the latter is selected as input.



Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

Result folder

> To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are 2 subfolders:

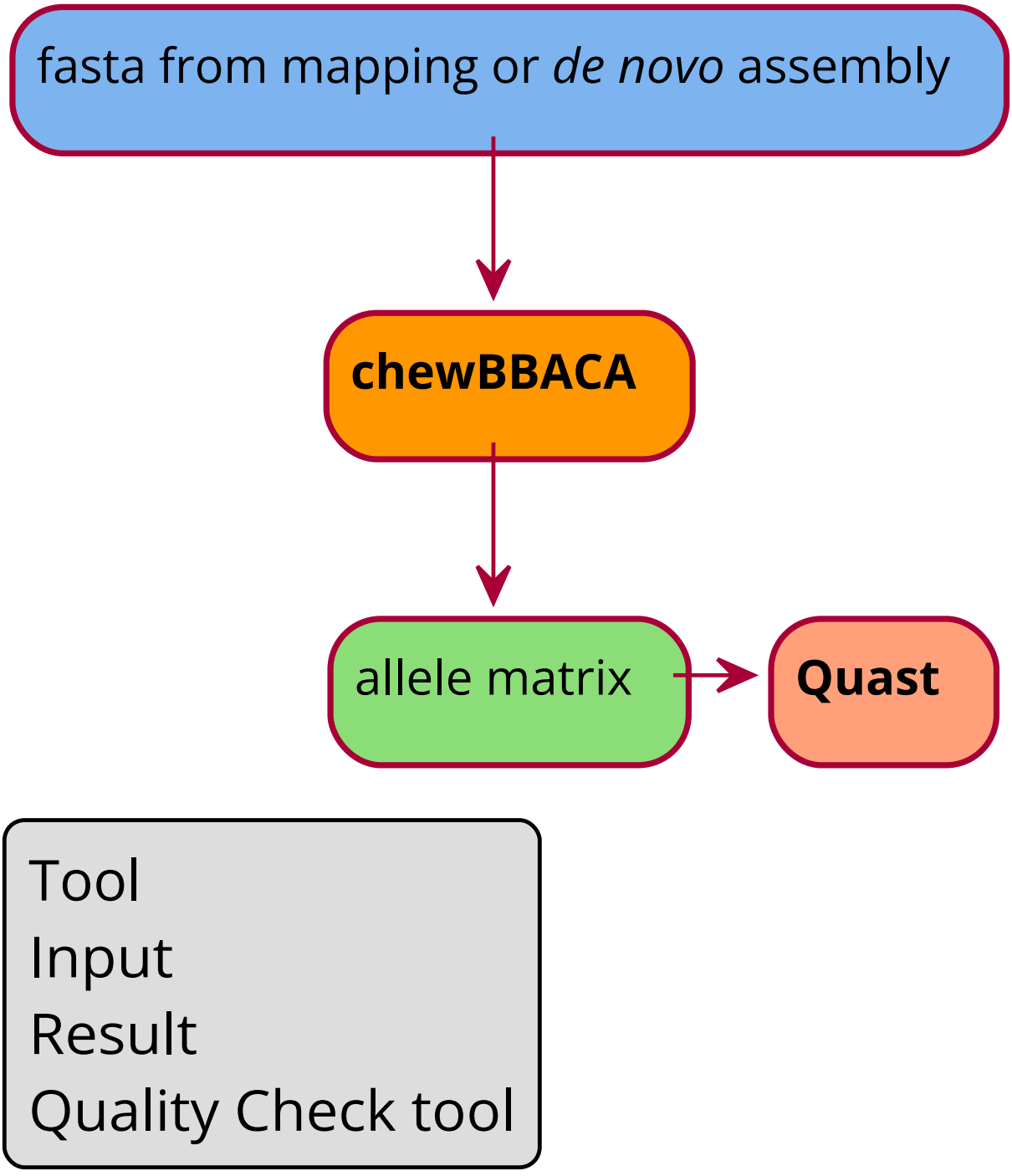
- **meta**: (*metadata*), where the log and configuration files of the executed process are saved;
- **result**: where the files containing the results produced by the analysis are saved.

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_mlst.tsv	result of *loci* calls for MLST	folder: result
DSXXXXXXXX-DTXXXXXX_ID_mlst_cc.csv	result with CC assignment	folder: result

4TY_wgMLST

Introduction

4TY_wgMLST performs *in silico* "whole genome Multi-Locus Sequence Typing" (wgMLST) for the bacterial species *Campylobacter jejuni* and *Campylobacter coli*. Unlike the cgMLST (core genome MLST) protocol, characterization is performed on the *whole genome*.



Run 4TY_wgMLST analysis

Once the analysis **4TY_wgMLST** has been selected in the dedicated run analyses page, it will be possible to choose the bioinformatics software ("method") to use, among those available for the analysis. The only method available for this analysis is **chewBBACA** - BSR-Based Allele Calling Algorithm.

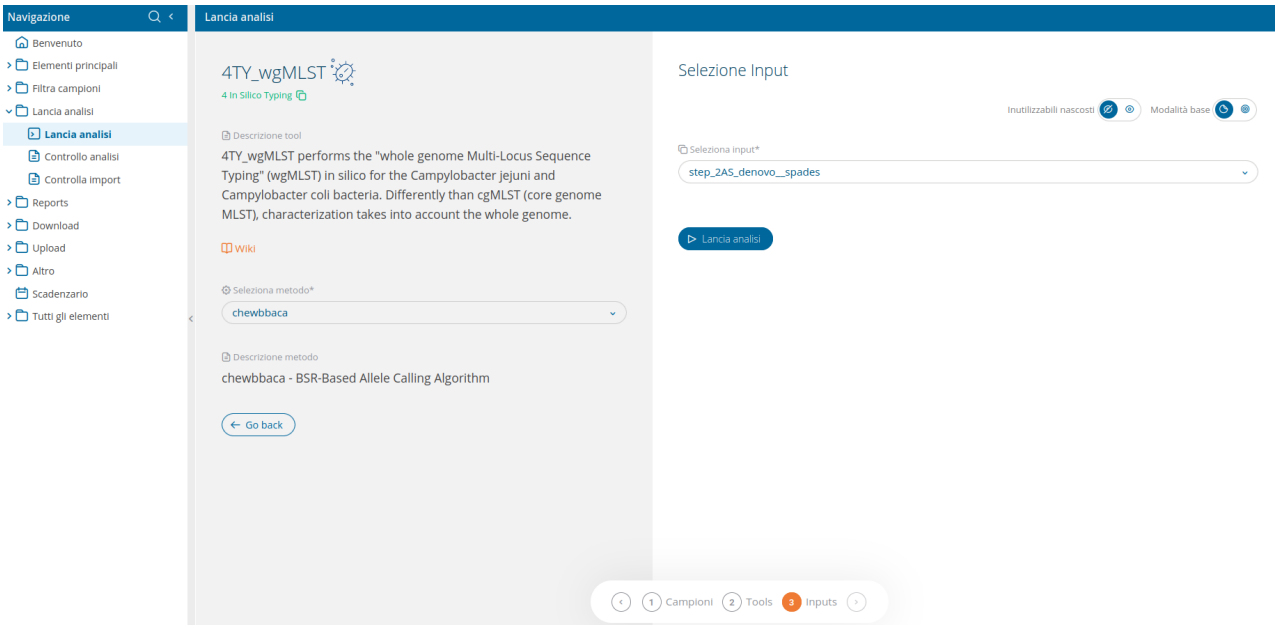
> **Note:** wgMLST schemas are available only for *Campylobacter jejuni* and *Campylobacter coli*. Running the analysis on another microorganism, for which no wgMLST schema is available, will cause the run to fail.

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.

Accepted inputs can originate from:

- step_2MG_denovo
- step_2AS_denovo
- step_2AS_mapping
- step_2AS_import

It will be necessary to specify the input sequences - whether derived from *de novo* assembly or from mapping - and the reference genome used for mapping, if the latter is selected as input.



Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

Result folder

> To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are 3 subfolders:

- **meta**: (*metadata*), where the log and configuration files of the executed process are saved;
- **qc**: (*quality check*), where quality check files are saved. It also contains the two subfolders *meta* and *result*. For this analysis, quality check is performed with **Quast**;
- **result**: where the files containing the results produced by the analysis are saved.

For allele calling with chewBBACA, results are provided using three different schemas:

- **IZS schema:** each allele is identified by a progressive numeric ID. ID assignment considers all *loci*, therefore numbering does **NOT** restart from 1 for each new *locus*.
- **MD5 schema:** each allele is identified by a 16-character alphanumeric code (MD5), obtained by applying a hash function to the allele sequence; this makes it interchangeable with other methods.
- **CRC32 schema:** control based on Cyclic Redundancy Check (CRC). Each allele is assigned a numeric check value determined by the remainder of the polynomial division of the data block. This value is used in the schema as the allele identifier.

Please note that the example file "DSXXXXXXXX-DTXXXXXX_ID_chewbbaca_results_ccoli_curated_210722.csv" shown in the table below can have a suffix for *C. coli* or *C. jejuni*, depending on the analyzed species.

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_chewbbaca_new_alleles.txt	sequence of newly-identified alleles	folder: result
DSXXXXXXXX-DTXXXXXX_ID_chewbbaca_results_contigsInfo.tsv	info about the contig mapped on each *locus*	folder: result
DSXXXXXXXX-DTXXXXXX_ID_chewbbaca_results_statistics.tsv	metrics on *loci* encoded as EXC, INF, LNF, PLOT, NIPH, ALM, ASM	folder: result
DSXXXXXXXX-DTXXXXXX_ID_import_chewbbaca_check.csv	quality check with info on calledPerc, calledNum, annotated, new, notFound, discarded	folder: qc > result
DSXXXXXXXX-DTXXXXXX_ID_chewbbaca_results_crc32.csv	.csv table of alleles encoded with CRC32 schema. Rows: samples, columns: *loci*	folder: result

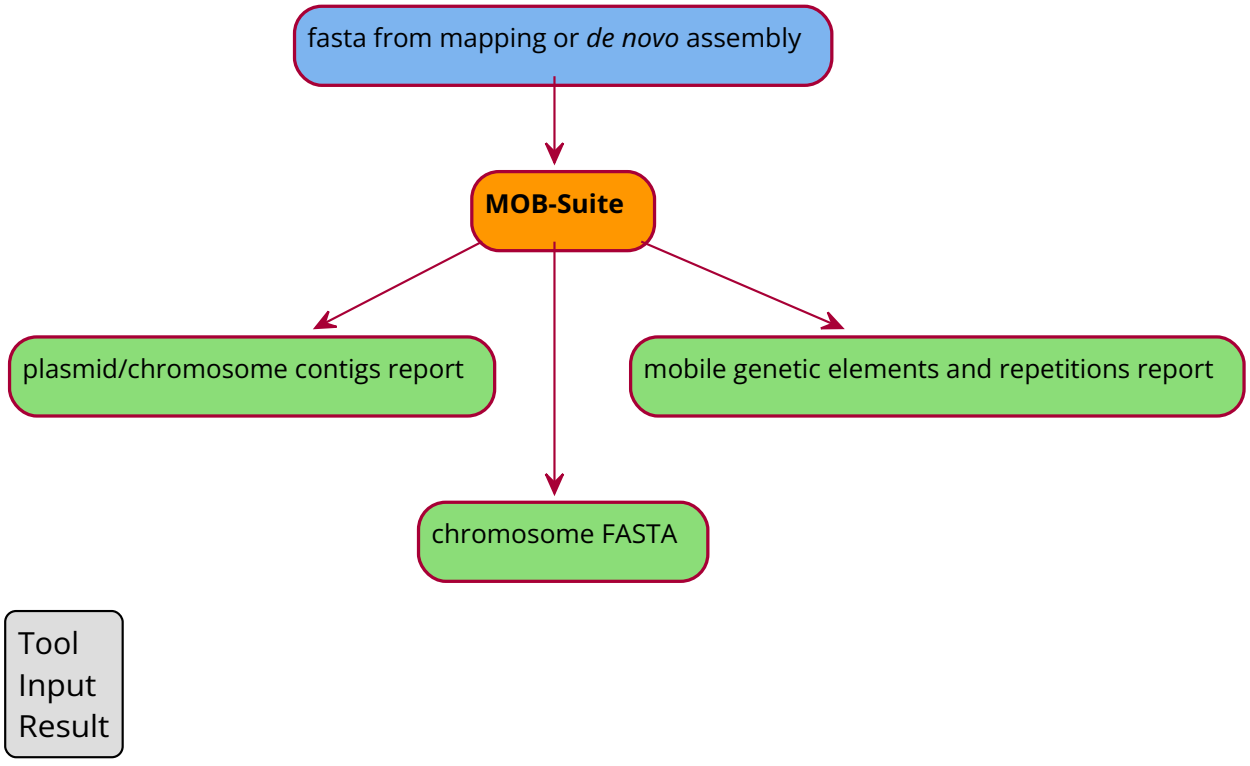
For more information on *locus* encoding and on chewBBACA output files, please refer to chewBBACA official guide.

4TY_plasmid

Introduction

4TY_plasmid runs MOB-Suite to type and reconstruct plasmid sequences, starting from the assemblies of sequences generated by Whole Genome Sequencing (WGS) assemblies.

MOB-Suite GitHub page: https://github.com/phac-nml/mob-suite



Run 4TY_plasmid analysis

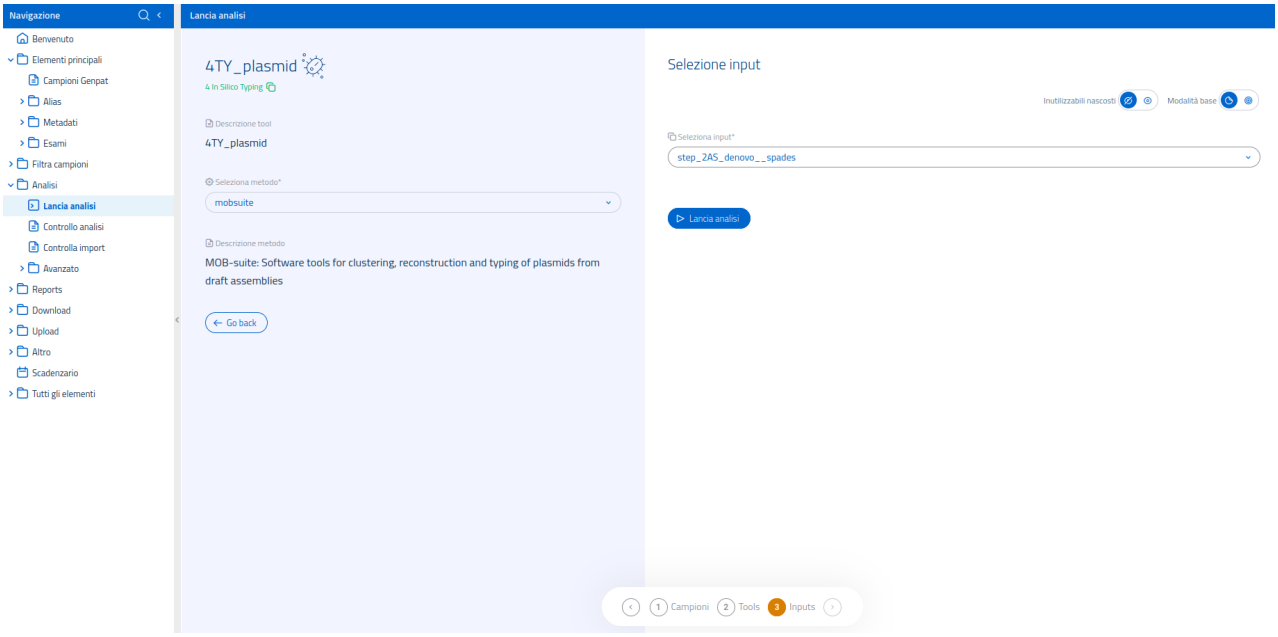
Once the analysis **4TY_plasmid** has been selected in the dedicated run analyses page, it will be possible to choose the bioinformatics software ("method") to use, among those available for the analysis. The available tool in this case is **MOB-Suite** - Software tools for clustering, reconstruction and typing of plasmids from draft assemblies.

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.

Accepted inputs can originate from:

- step_2MG_denovo
- step_2AS_denovo
- step_2AS_mapping
- step_2AS_hybrid
- step_2AS_import

It will be necessary to specify the input sequences - whether derived from *de novo* assembly or from mapping - and the reference genome used for mapping, if the latter is selected as input.



Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

Result folder

> To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are 2 subfolders:

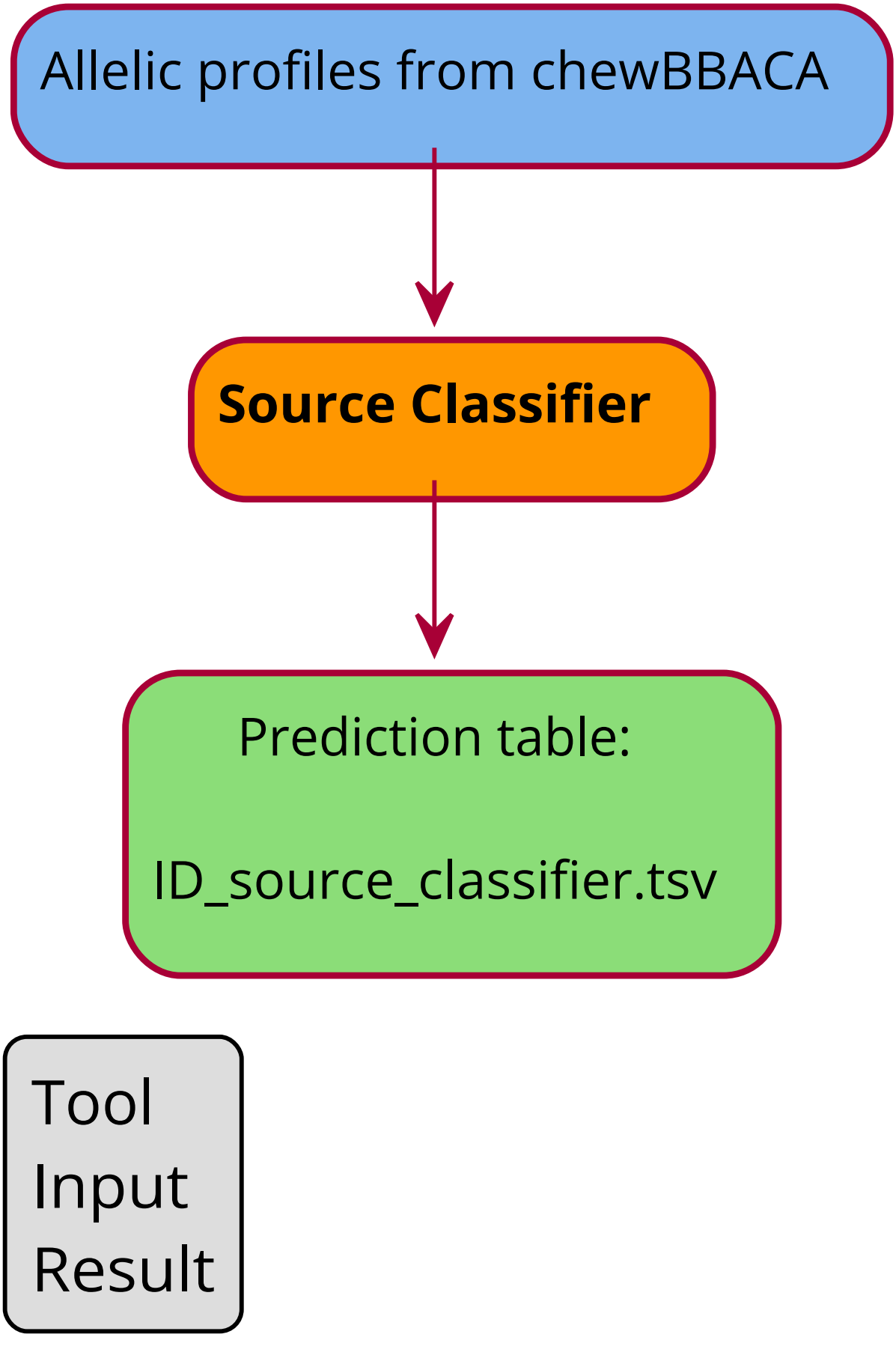
- **meta**: (*metadata*), where the log and configuration files of the executed process are saved;
- **result**: where the files containing the results produced by the analysis are saved.

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_mobsuite_chromosome.fasta	fasta with contigs belonging to the chromosome	foder: result
DSXXXXXXXX-DTXXXXXX_ID_mobsuite_contig_report.txt	tabular file with contig assignment to the chromosome or to the plasmid group	folder: result
DSXXXXXXXX-DTXXXXXX_ID_mobsuite_mge.report.txt	BLAST HSP for mobile genetic elements (MGE) and repetitive elements	folder: result

4TY_ML

Introduction

The ML (Machine Learning) analysis allows to predict the food source (Source Attribution) of *L. monocytogenes*. The analysis uses a binary classifier (predicted sources: "meat" or "non-meat"), produced as a deliverable of the current research project.


Allelic profiles from chewBBACA
Source Classifier
Prediction table:
ID_source_classifier.tsv
Tool
Input
Result

Run 4TY_ML analysis

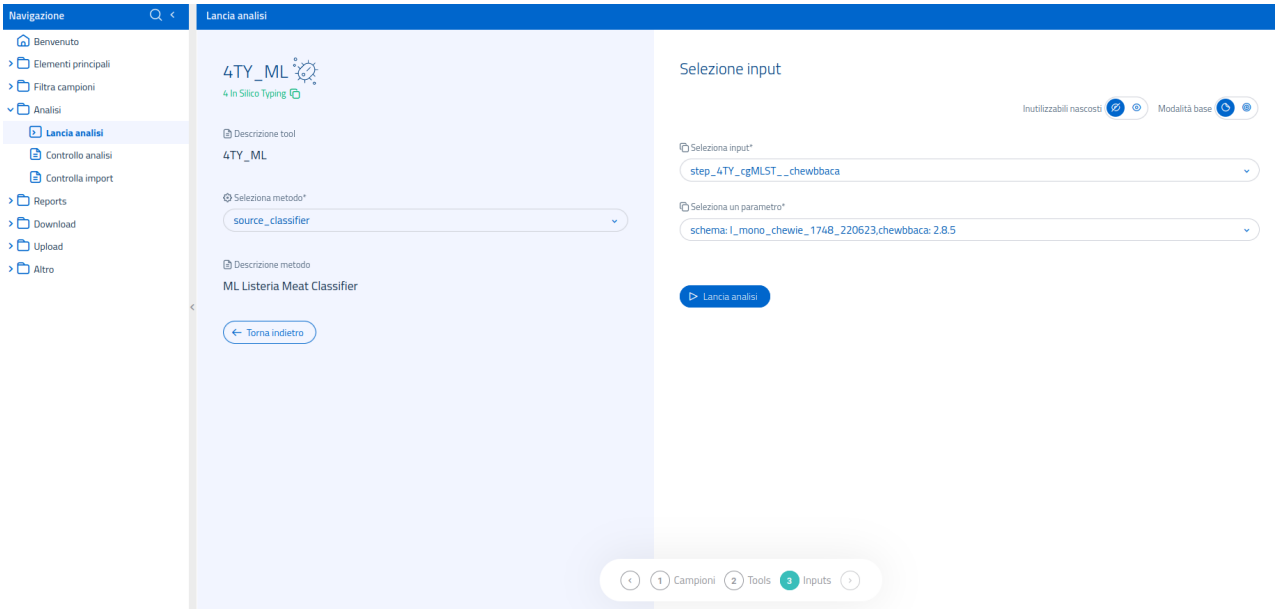
The analysis **4TY_ML** uses a pre-trained model as source-classifier, currently available only for *L. monocytogenes*. Phenotype classification will be binary: the model is trained to predict which samples, among the *L. monocytogenes* selected as input, come from a "meat" matrix and which ones from a "non-meat" matrix. Prediction is based on the cgMLST allele matrix.

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.

Accepted inputs can originate from:

- step_4TY_wgMLST
- step_4TY_cgMLST

Both input types use *chewBBACA* as execution software; moreover, for **4TY_cgMLST**, it is possible to use either the version 2.8.4 with Pasteur schema or the 2.8.5 with crc32 schema.



Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

Result folder

> To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are 2 subfolders:

- **meta**: (*metadata*), where the log and configuration files of the executed process are saved;
- **result**: where the files containing the results produced by the analysis are saved.

The table below shows the location of main output files and a brief description of them.

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_source_classifier.tsv	.tsv table with prediction results for the sample	folder: result
metadata.csv	.csv table of metadata for input samples	folder: results

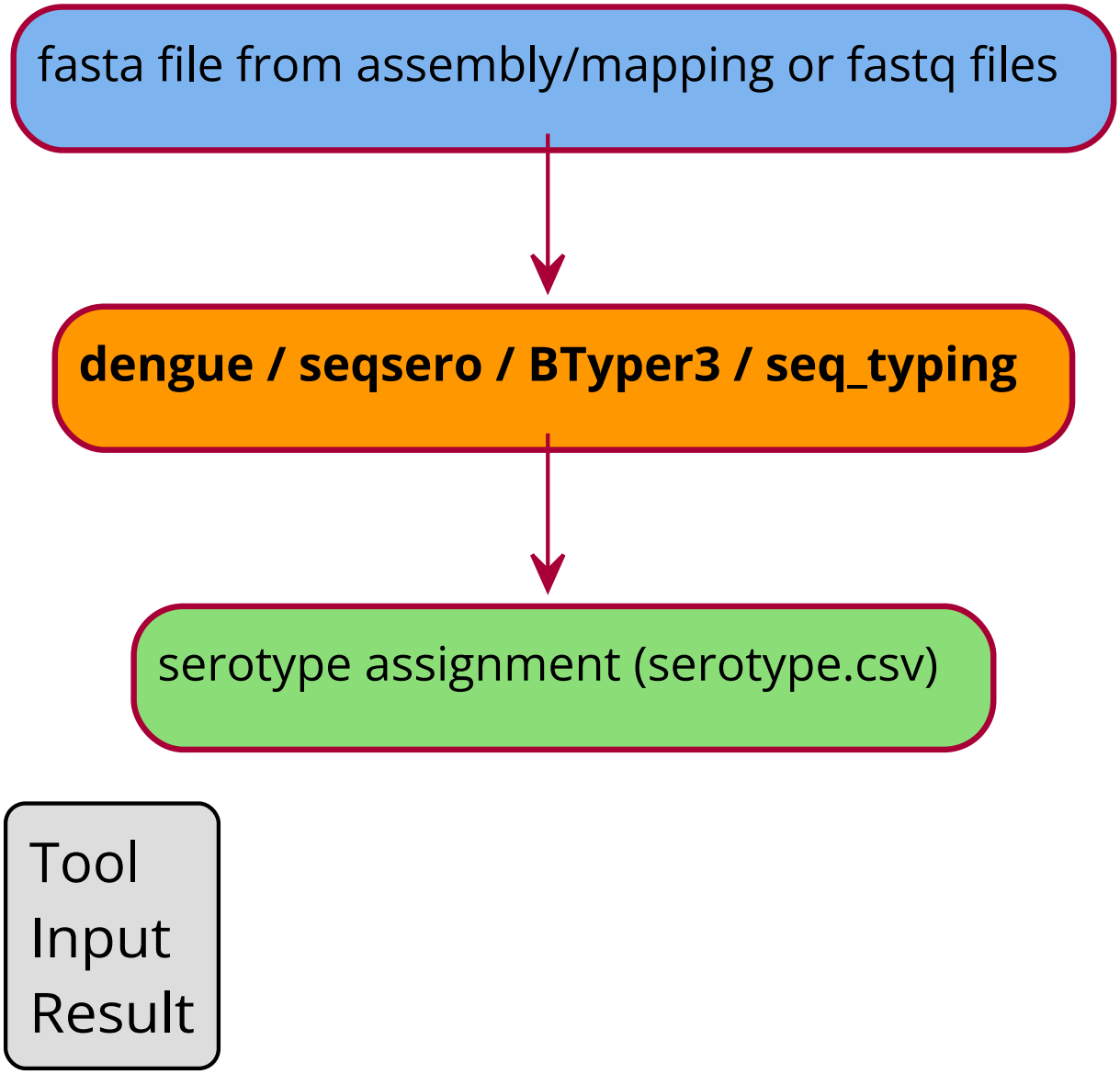
Result visualization

Results of the **4TY_ML** analysis can also be viewed through the dedicated table of checks page, **Quality checks for ML-based classification**, available under `Report > Detail > Checks/Metrics > Quality checks for ML-based classification` in the left side navigation menu.

4TY_serotype

Introduction

4TY_serotype assigns the serotype for Dengue or *Salmonella enterica*. The assignment is performed by calculating the coverage, following alignment against the available reference sequences.



Run 4TY_serotype analysis

Once the analysis **4TY_serotype** has been selected in the dedicated run analyses page, it will be possible to choose the bioinformatics software to use.

For *Salmonella enterica*:

- **Seqsero** - *Salmonella* serotype prediction from genome sequencing data.

  Seqsero GitHub page: https://github.com/denglab/SeqSero2

For *Bacillus cereus group*:

- **BTyper3** - Taxonomic classifier for Bacillus cereus based on assembled genomes.

  BTyper3 GitHub page: https://github.com/lmc297/BTyper3

For *Escherichia coli*:

- **seq_typing** - Classifier for serotype assignment in Escherichia coli.

  seq_typing GitHub page: https://github.com/B-UMMI/seq_typing

For DENV:

- **Dengue** - DENV in silico serotype calculation

The usable inputs are reads in fastq format or assembled sequences (fasta):

- step_1PP_trimming
- step_1PP_filtering
- step_1PP_hostdepl
- step_1PP_downsampling
- step_1PP_generated
- step_2AS_import
- step_2AS_denovo
- step_2AS_mapping
- step_2AS_indexing
- step_2MG_denovo

If the output of a mapping analysis is selected as input, the corresponding reference sequence must also be specified.

Result folder

To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are the subfolders:

- **meta**: (*metadata*), where the log and configuration files of the executed process are saved;
- **result**: where the files containing the results produced by the analysis are saved.

The tables below display the lists of files contained in the folders, according to the tool used, together with some useful information. Please note that the file name will include the name of the program selected for the analysis.

Seqsero

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_SeqSero_result.tsv	.tsv table containing the serotype assignment from Seqsero	folder: result
DSXXXXXXXX-DTXXXXXX_ID_extracted_antigen_alleles.fasta	multifasta of allelic sequences from Seqsero	folder: result

Dengue

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_dengue_serotype.csv	.csv table containing the serotype assignment	folder: result
DSXXXXXXXX-DTXXXXXX_ID_dengue_serotype_summary.csv	coverage summary for each reference / serotype, in .csv format	folder: result

For lineage assignment, the analysis also makes use of alignment and lineage calculation tools (Dengue). The results of these analyses are generated and stored according to the hierarchy of individual analyses (2AS_mapping and 4TY_lineage).

5.4.4 Multi sample

SNPs-based clustering

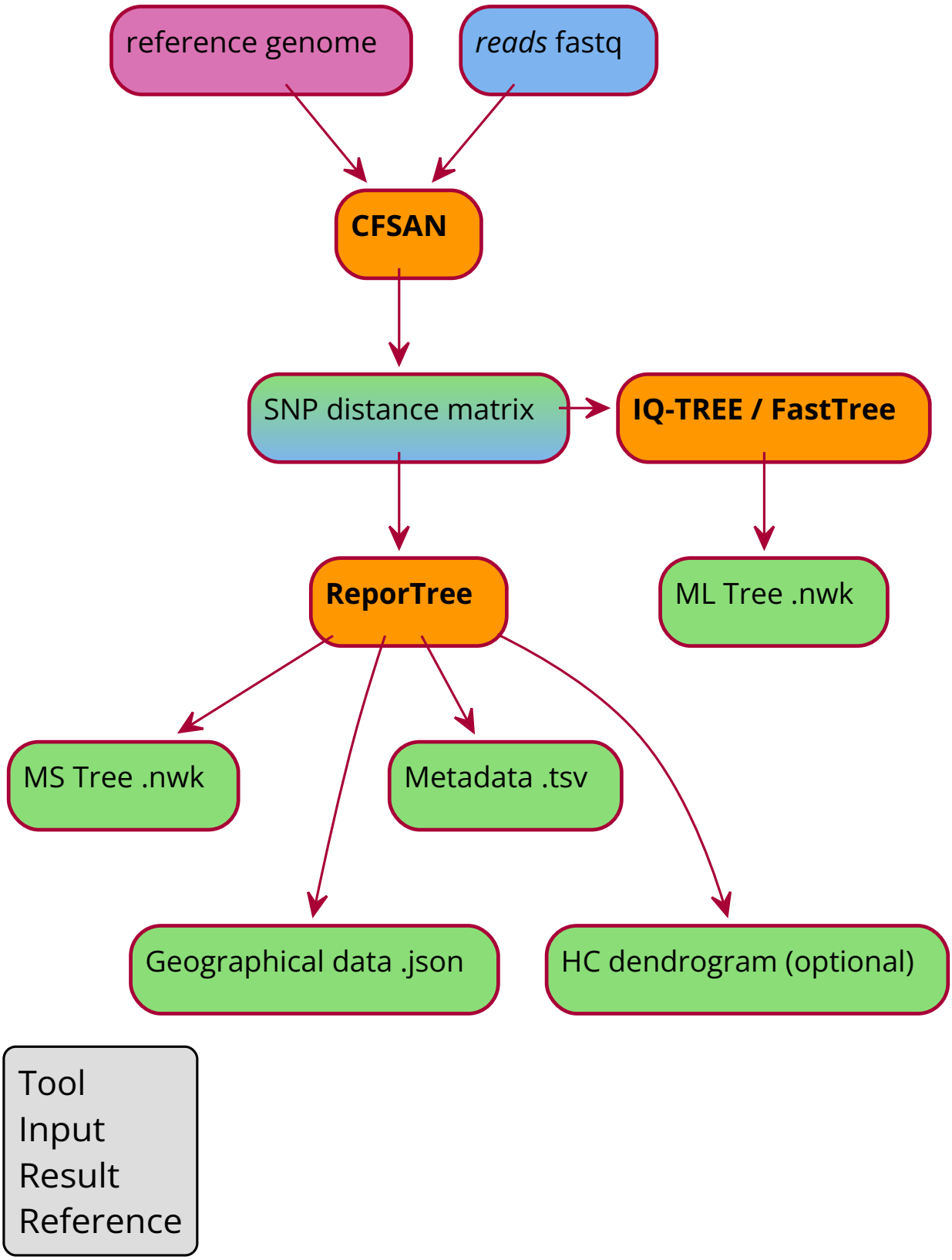
CFSAN

Introduction

CFSAN performs a reference-based phylogenetic analysis and builds a Single Nucleotide Polymorphism (SNP) matrix. The matrix is then used to generate "Maximum Likelihood" (ML) distance graphs with IQ-TREE or FastTree, and "Minimum Spanning Tree" (MST) and "Hierarchical Clustering" (HC) with **ReporTree**.

Further information on the software is available in:

- **CFSAN official guide:** <(https://snp-pipeline.readthedocs.io/en/latest/readme.html)> **CFSAN GitHub page:** https://github.com/CFSAN-Biostatistics/snp-pipeline
- **ReporTree GitHub page:** https://github.com/insapathogenomics/ReporTree
- **IQ-TREE GitHub page:** https://github.com/iqtree/iqtree2
- **FastTree GitHub page:** https://github.com/morgannprice/fasttree



Run CFSAN analysis

Once the **CFSAN** analysis has been selected in the dedicated run analysis page, the system will move to a confirmation interface.

In the parameter definition screen, it will be necessary to specify a reference genome; morevover the input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.

In the case of CFSAN, the inputs are always reads in FASTQ:

- step_1PP_trimming
- step_1PP_downsampling
- step_1PP_hostdepl
- step_1PP_generated
- step_1PP_filtering

The parameter definition screen offers a high level of customization, including:

- **Parametri di CFSAN:**

  `Pipeline version`: it allows you to choose the software version (2.0.2 "legacy" or 2.2.1);
  `Keep intermediate files`: at the end of the run, the analysis normally deletes the largest files that are not needed for downstream analyses. This option allows you to choose not to delete these files at the end of the process;
  `Remove abnormal SNPs across all vcf files`: the Snippy-core output files typically also list the reference used for mapping. This option allows you to remove the reference from the `core.vcf` file, an action necessary to prevent the reference from being included as a sample in the final tree.

- **ReporTree parameters:**

- `Partition threshold for clustering definition`: it allows you to customize the threshold values for the creation of tree partitions;

  `Minimum proportion of samples per site without missing data`: ReporTree normally considers all genomic sites where variants have been called for the calculation of the distance matrix and the tree. This option allows you to define a threshold value: only sites for which variants are called in a percentage of the dataset equal to or greater than the specified value will be considered. Sites that do not pass the filter are discarded;
  `Minimum proportion of loci/positions called for SNP/allele matrices`: ReporTree normally uses all provided samples to compute the distance matrix and the tree. This option allows you to define a threshold for the number of sites where a variant has been called. All samples with a number of sites without variants lower than the specified value will be discarded;
  `Sample of interest` options: they allow you to specify one or more samples of interest on which ReporTree recalculates a 'Zoom', i.e., a tree containing only the sample of interest and the samples belonging to the same cluster;
  `Method used in HC clustering to create a dendrogram` options: selection of the mode for optional hierarchical clustering;
  `Custom parameters` options: text fields where command-line style options can be entered, to be passed to ReporTree exclusively for the computation of the specified trees (MST produced with Grapetree or HC).

- **Clustering ML parameters:**

  `Maximum likelood phylogenetic tree calculation`: it allows you to choose the software to use for optional clustering with an ML algorithm. The available software are **IQ-TREE v2.3.6**, **IQ-TREE "legacy" v1.6.12** and **FastTree**.

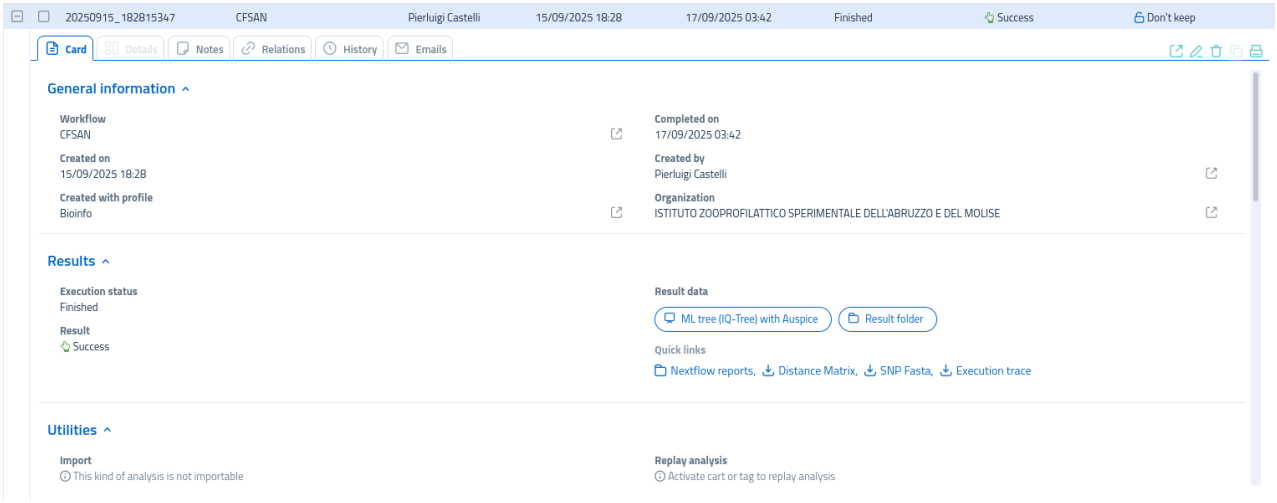
> Given the number and complexity of parameters and results of the CFSAN and ReporTree software, users are encouraged to consult their official guides: - CFSAN official guide - ReporTree official guide.

The procedure for launching the analysis is similar to other analyses that include the execution of ReporTree:

Run ReporTree analysis
Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

In the completed analysis card, several options will be available, including the possibility to browse the output folder, direct access to some log and metadata files, opening the .nwk file, and the direct visualization of the MSTree through the integration of SPREAD into the platform.

To visualize the tree, it will be necessary to navigate to the analysis card and, in the section *Result data*, select the link: - **ML tree (IQ-Tree) with Auspice** for the ML tree; - **MST with SPREAD** for the MSTree.



Note: Results from multi-sample analyses cannot be imported. Users are advised to use the dedicated functionality for non-importable analysis preservation or, alternatively, the option to directly download the newick file (`.nwk`) and the metadata file, so that the files can be stored and the tree can be visualized at any time, either within the platform or using external software.

Result folder

To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are two subfolders:

- **meta** (*metadata*), where the log and configuration files of the executed process are saved;
- **result**, where the files containing the results produced by the analysis are saved.

The table below shows the list of the main files of interest present in the folders, along with some useful information.

File	Description	Location
`metrics.tsv`	summary .tsv table with sample values and statistics and their corresponding SNPs	folder: result
`referenceSNP.fasta`	FASTA file containing the reference sequence bases for each SNP position	folder: result
`snp_distance_matrix.tsv`	.tsv table with a distance matrix, calculated on differences in SNPs, for each pair of samples	folder: result
`snp_distance_pairwise.tsv`	.tsv table containing pairwise distances, based on differences in SNPs, for each pair of samples	folder: result
`iqtree.nwk`	tree file in .nwk (Newick) format computed using a Maximum Likelihood (ML) algorithm	folder: result
`snpma.vcf`	list of SNPs and their corresponding positions, for each sample, in VCFv4.2 format	folder: result
`iqtree.mldist`	distance matrix for the ML tree	folder: result
`auspice_iqtree.json`	JSON file containing metadata and the ML tree	folder: result

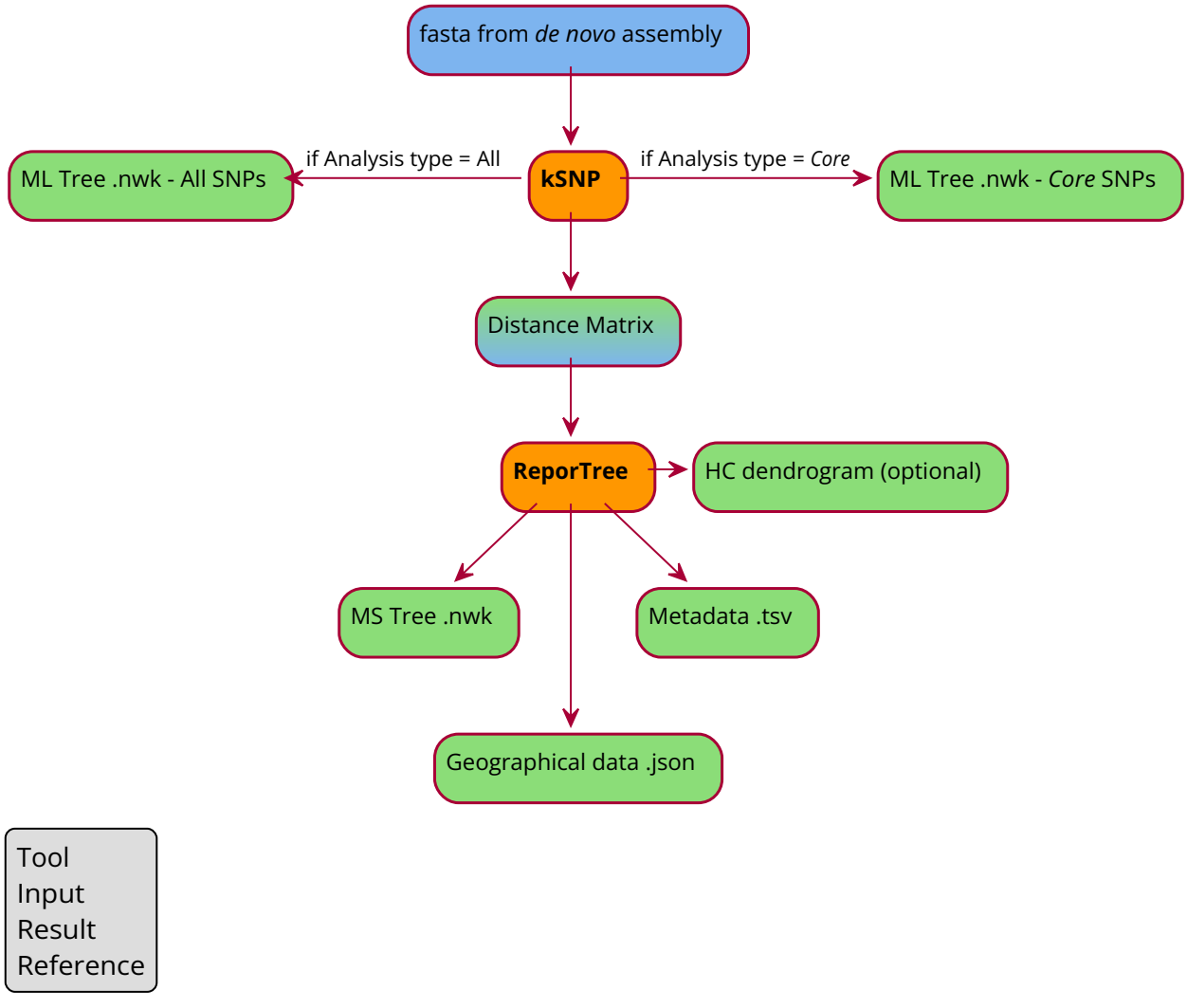
For further information regarding all CFSAN outputs, please refer to the dedicated section of the CFSAN official manual.

kSNP

Introduction

kSNP3 identifies Single Nucleotide Polymorphisms (SNPs) in the samples; it performs an alignment-free (reference-free) phylogenetic analysis and builds a phylogenetic tree using a Maximum Likelihood (ML) algorithm, based on the samples' SNP matrix. Moreover, the analyses launches **ReporTree** to calculate the "Minimum Spanning Tree" (MST).

- kSNP guide
- kSNP official site
- **kSNP GitHub page:** https://github.com/theiagen/public_health_bioinformatics
- **ReporTree GitHub page:** https://github.com/insapathogenomics/ReporTree
- **IQ-TREE GitHub page:** https://github.com/iqtree/iqtree2
- **FastTree GitHub page:** https://github.com/morgannprice/fasttree



Run kSNP analysis

Once the **kSNP** analysis has been selected in the dedicated run analysis page, the system will move to a confirmation interface.

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.

The accepted inputs for **kSNP** are:

- step_2AS_denovo
- step_2MG_denovo
- step_2AS_mapping
- step_2AS_hybrid
- **step_2AS_import**
- step_2AS_indexing

The parameter definition screen also offers a high level of customization, including:

- **kSNP parameters:**

  `Analysis type`: it allows you to choose whether to perform the phylogenetic analysis using all SNPs ("All" mode) or only core SNPs ("Core" mode);
  `Kmer length`: it allows you to manually change the lengh of k-mers;
  `kSNP version`: it allows you to choose the software version between kSNP3 ("legacy") and kSNP4 (v4.1);
- **ReporTree parameters:**
- `Partition threshold for clustering definition`: it allows you to customize the threshold values for creating tree partitions;

  `Minimum proportion of samples per site without missing data`: ReporTree normally considers, for calculating the distance matrix and the tree, all genomic sites where variants have been called. This option allows you to define a threshold value: only sites where variants are called in a percentage of the dataset equal to or higher than the specified value will be considered. Sites that do not pass the filter are discarded;
  `Minimum proportion of loci/positions called for SNP/allele matrices`: ReporTree normally uses all provided samples to compute the distance matrix and the tree. This option allows you to define a threshold value for the number of sites where a variant has been called. All samples with a number of sites without variants lower than the specified value will be discarded;
  `Sample of interest` options: they allow you to specify one or more samples of interest on which ReporTree recalculates a "Zoom", i.e., a tree containing only the sample of interest and the samples belonging to the same cluster;
  `Method used in HC clustering to create a dendrogram` options: selection of the mode for optional *Hierarchical Clustering*;
  `Custom parameters` options: text fields where command-line-style options can be entered, to be passed to ReporTree exclusively for the computation of the specified trees (MST produced with Grapetree or HC);
- **Clustering ML parameters:**

  `Maximum likelood phylogenetic tree calculation`: it allows you to choose the software to use for optional clustering with a maximum-likelihood (ML) algorithm. The available software are **IQ-TREE 2** and **FastTree**.

Given the number and complexity of parameters and results of the Snippy-core and ReporTree software, users are encouraged to consult their official guides: - kSNP official guide - ReporTree official guide

The procedure for launching the analysis is similar to other analyses that include the execution of ReporTree:

Run ReporTree analysis

Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

In the completed analysis card, several options will be available, including the possibility to browse the output folder, direct access to some log and metadata files, opening the .nwk file, and the direct visualization of the MSTree through the integration of SPREAD into the platform.

To visualize the tree, it will be necessary to navigate to the results tab and select the link: - **ML tree (IQ-Tree) with Auspice** for the ML tree; - **MST with SPREAD** for the MSTree; - **HC with Auspice** for the HC graph.

Note: Results from multi-sample analyses cannot be imported. Users are advised to use the dedicated functionality for non-importable analysis preservation or, alternatively, the option to directly download the newick file (`.nwk`) and the metadata file, so that the files can be stored and the tree can be visualized at any time, either within the platform or using external software.

Result folder

To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are two subfolders:

- **meta** (*metadata*), where the log and configuration files of the executed process are saved;
- **result**, where the files containing the results produced by the analysis are saved.

Please note that, depending on the parameter used for the **Analysis type**, the name of the output files will differ:

- SNPs_all_matrix (if "All" has been selected for *Analysis type*);
- core_SNPs_matrix (if "Core" has been selected for *Analysis type*).

The table below shows the names when running on "Core" SNPs.

File	Description	Location
`core_SNPs_matrix.fasta`	multifasta file containing headers with the sample name and a line for the concatenated SNP sequences. “N” indicates the absence of an SNP in that sample	folder: results
`gt.nwk`	tree file in .nwk (Newick) format calculated using an MST algorithm	folder: results
`gt_clusterComposition.tsv`	table listing the clusters and their composition	folder: result
`gt_metadata_w_partitions.tsv`	table of sample metadata with cluster membership based on partitions	folder: result
`gt_dist_hamming.tsv`	table with Hamming distances	folder: result
`gt_partitions_summary.tsv`	list of tree partitions and corresponding samples in the clusters	folder: result
`gt_partitions.tsv`	table of samples and their cluster membership based on the partitioning	folder: result
`gt_loci_used.txt`	List of *loci*	folder: result

For further information regarding the formatting of kSNP outputs, please refer to the dedicated section of the kSNP user guide.

VCF2MST

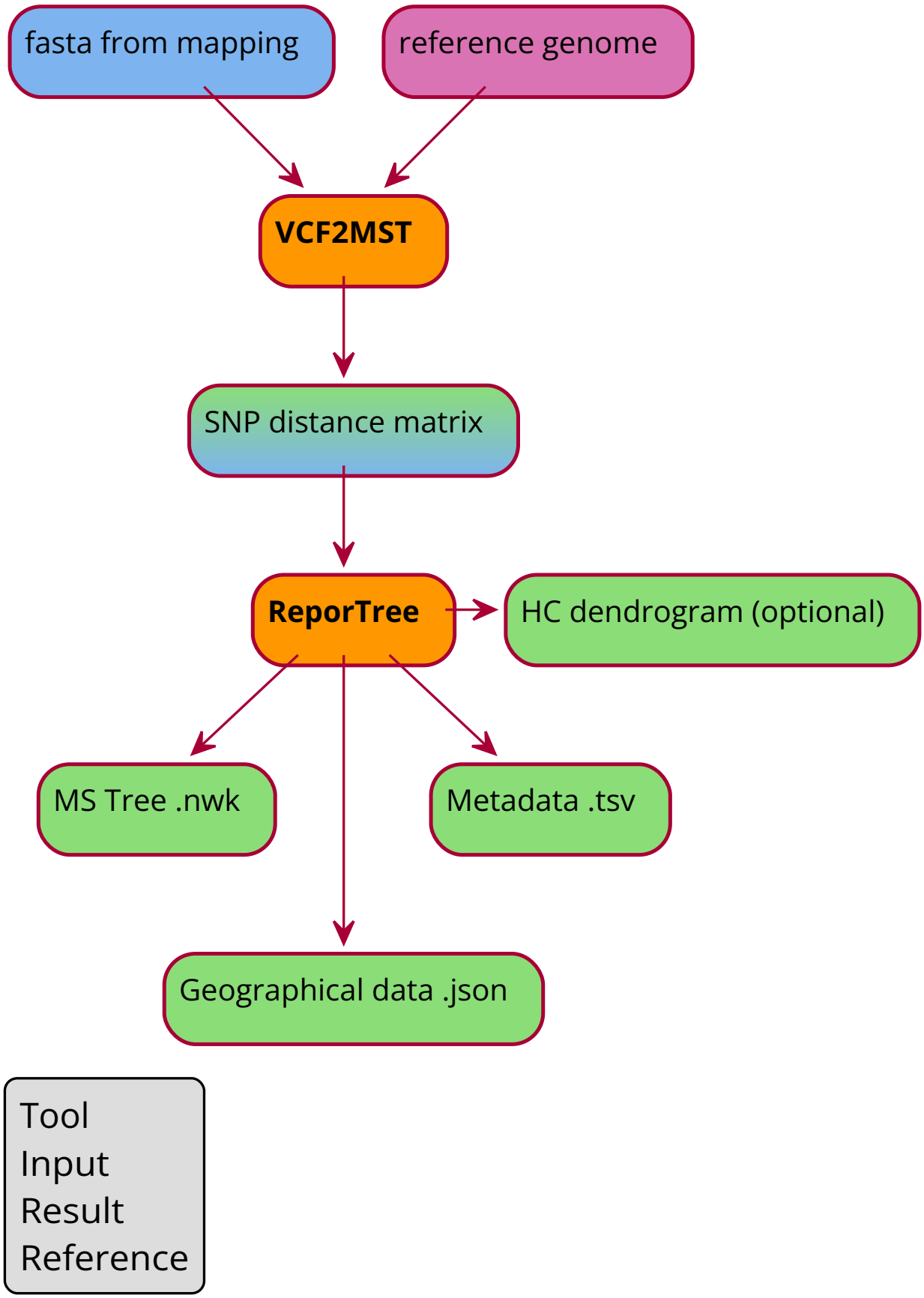
Introduction

VCF2MST analysis performs **VCF2MST** and **ReporTree** to build a phylogenetic Minimum Spanning Tree (MST) from a Single Nucleotide Polymorphism (SNP) matrix computed from mapping-derived FASTA files.

The construction of the distance matrix is performed without relying on phylogenomic inferences based on sequence alignment.

The program constructs the Minimum Spanning Tree based on Hamming distances, which measure the number of substitutions required for one sequence (string) to be transformed into another.

- **VCF2MST GitHub page:** https://github.com/genpat-it/vcf2mst
- **Paper on VCF2MST:** https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-021-08112-0
- **ReporTree GitHub page:** https://github.com/insapathogenomics/ReporTree
- **IQ-TREE GitHub page:** https://github.com/iqtree/iqtree2
- **FastTree GitHub page:** https://github.com/morgannprice/fasttree


fasta from mapping
reference genome
VCF2MST
SNP distance matrix
ReporTree
HC dendrogram (optional)
MS Tree .nwk
Metadata .tsv
Geographical data .json
Tool
Input
Result
Reference

Run VCF2MST analysis

Once the **VCF2MST** analysis has been selected in the dedicated run analysis page, the system will move to a confirmation interface.

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.

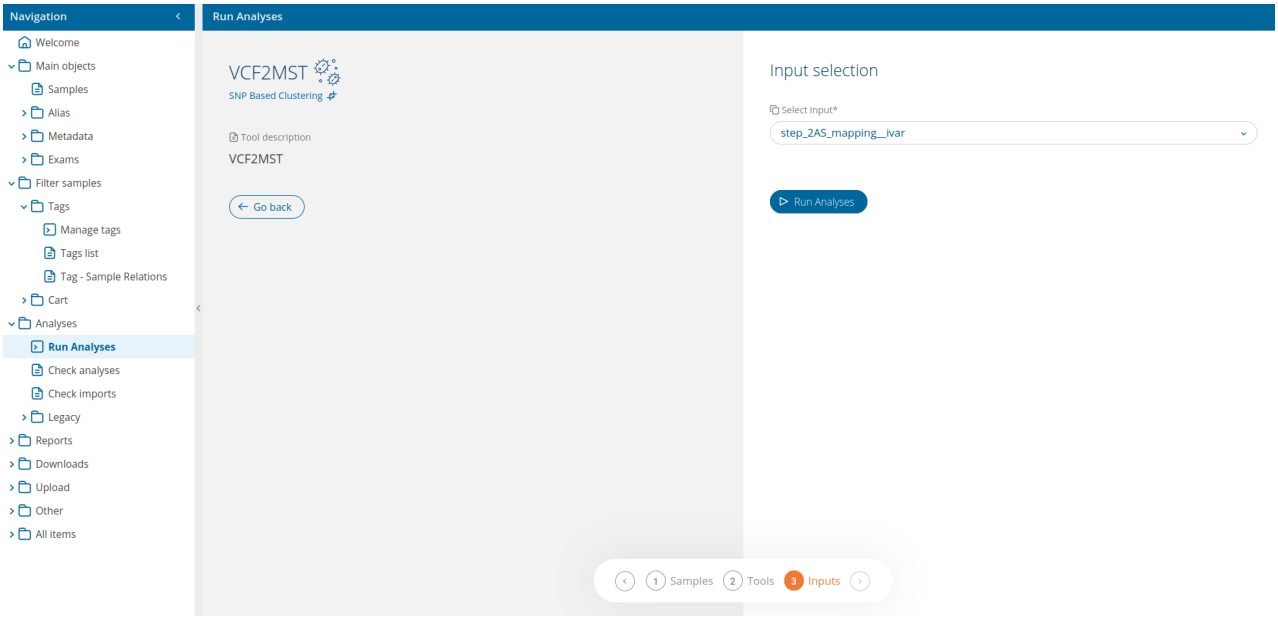
The input for VCF2MST consists of the FASTA files from mapping: step_2AS_mapping.

In the input selection interface there will also be a field to provide the reference genome (the reference is automatically populated).

The parameter definition screen offers a high level of customization, including:

- **VCF2MST parameters:**
  `VCF filtering and normalization` : it allows enabling the normalization of complex variants by breaking them down into the individual SNPs that compose them;
- **ReporTree parameters:**
  `Partition threshold for clustering definition` : it allows customizing the threshold values for the creation of tree partitions;
  `Sample of interest` options: they allow to specify one or more samples of interest on which ReporTree recalculates a "Zoom", i.e., a tree containing only the sample(s) of interest and the samples belonging to the same cluster.;
  `Method used in HC clustering to create a dendrogram` options: selection of the mode for the optional *Hierarchical Clustering*.

> Given the number and complexity of ReporTree parameters, users are encouraged to consult the official ReporTree guide, for more information on how to use the software.



The procedure for launching the analysis is similar to other analyses that involve running ReporTree:

Run ReporTree analysis

Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

In the completed analysis card, in addition to the ability to browse the output folder, several options will be available, such as direct access to some log and metadata files, opening the .nwk file, and the direct visualization of the MSTree through the integration of SPREAD within the platform.

To view the tree, you will need to navigate to the GrapeTree result card and select the link **Minimum Spanning Tree with SPREAD**.

> **Note:** Results from multi-sample analyses cannot be imported. Users are advised to use the dedicated functionality for non-importable analysis preservation or, alternatively, the option to directly download the newick file (`.nwk`) and the metadata file, so that the files can be stored and the tree can be visualized at any time, either within the platform or using external software.

Result folder

> To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are two subfolders:

- **meta** (*metadata*), where the log and configuration files of the executed process are saved;
- **result**, where the files containing the results produced by the analysis are saved.

The following table provides a list of the main files of interest in the folders, along with some useful information.

File	Description	Location
`gt.nwk`	.nwk file of MST tree	folder: result
`gt_loci_used.txt`	list of loci used for the calculation of the distance matrix (*start position*)	folder: result
`gt_partitions.tsv`	table with information on the cluster membership of the samples at each tree partitioning value	folder: result
`gt_clusterComposition.tsv`	Table listing clusters and their composition	folder: result
`gt_partitions_summary.tsv`	summary table of the tree partitions and the corresponding samples in the clusters	folder: result
`zooms.txt`	list of requested zoom levels	folder: result
`gt_dist_grapetree.tsv`	MST tree distance matrix	folder: result
`gt_dist.tsv`	pairwise distances of the MST tree	folder: result
`gt_dist_hamming.tsv`	distance matrix	folder: result
`gt_metadata_w_partitions.tsv`	metadata table with information on cluster membership	folder: result
`reportree_metadata.tsv`	sample metadata table produced by ReporTree	folder: result

The Newick format file (.nwk) contains all the tree data and can be viewed using **SPREAD** or another external software for visualizing dendrograms or distance graphs.

Snippy-core

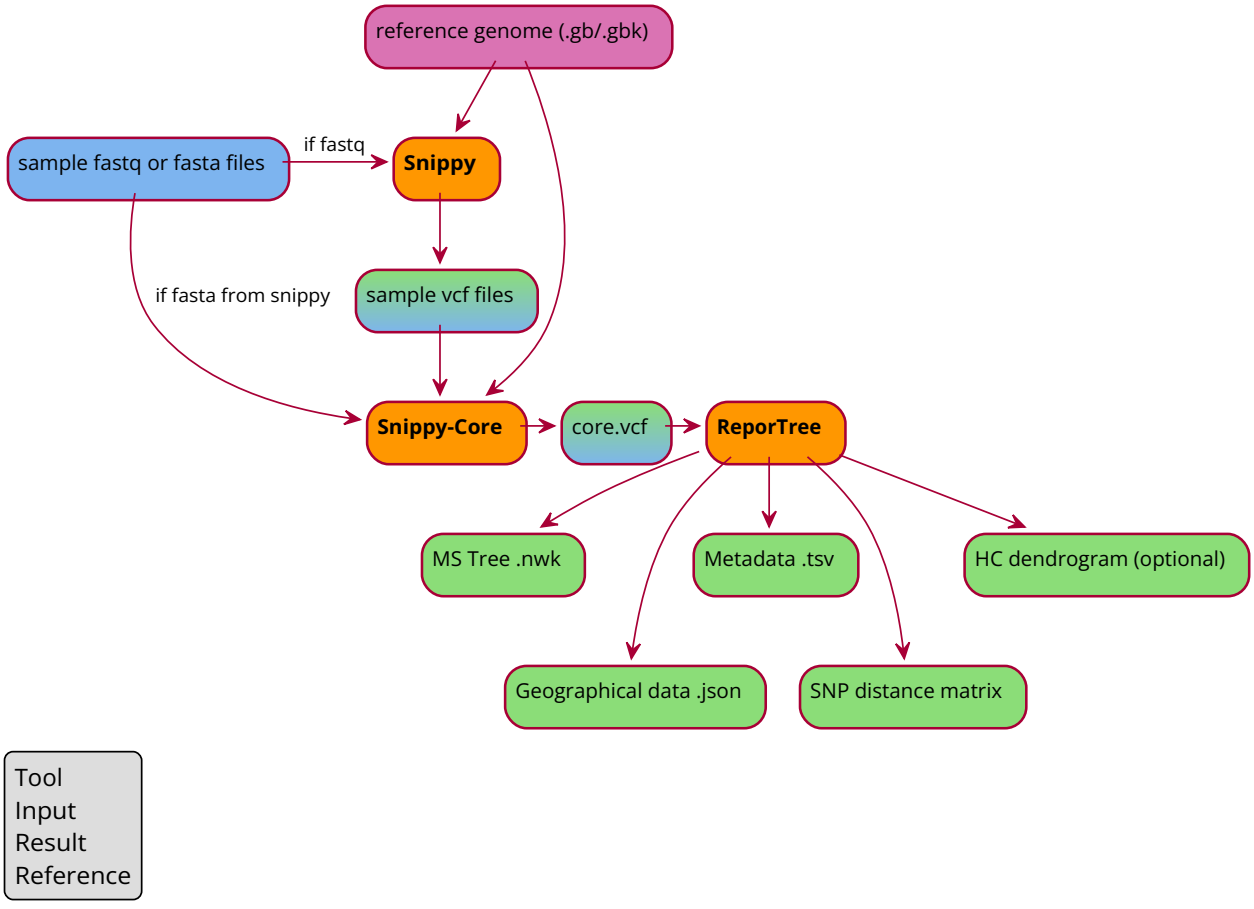
Introduction

Snippy-core performs the extraction of the *core SNPs* from the VCF files produced by **Snippy** (*Rapid haploid variant calling and core genome alignment*) and subsequent construction of the MST tree with **ReporTree**.

If Snippy results are not provided as input, the Snippy-core analysis first performs mapping using the Snippy tool. The resulting mapping FASTA files are then used as input for Snippy-core, which produces the `core.vcf` file containing the *core* dataset variants (SNPs and indels).

The VCF file is then processed with ReporTree to obtain the distance matrix and the MST tree.

- **Snippy & Snippy-core GitHub page:** https://github.com/tseemann/snippy
- **ReporTree GitHub page:** https://github.com/insapathogenomics/ReporTree
- **IQ-TREE GitHub page:** https://github.com/iqtree/iqtree2
- **FastTree GitHub page:** https://github.com/morgannprice/fasttree



Run Snippy-core analysis

Once the **Snippy-core** analysis has been selected in the dedicated run analysis page, the system will move to a confirmation interface.

The parameter definition interface requires selecting a reference, which is necessary both for Snippy-core and for optional mapping with Snippy. A reference in *genBank* (`gb`) format will always be used.

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.

The possible inputs that can be used for Snippy-core are:

- step_1PP_trimming
- step_1PP_generated
- step_1PP_hostdepl
- step_1PP_downsampling
- step_1PP_filtering
- step_2AS_mapping
- step_2AS_denovo
- step_2AS_hybrid
- step_2MG_denovo

The parameter definition screen offers a high level of customization, including:

- **Snippy/Snippy-core parameters:**

  `Alignment type` : it allows the selection between the two SNP alignment and extraction modes of Snippy e Snippy-core ("Core" e "Full");
  `Keep intermediate files` : once the execution is complete, the analysis normally removes the largest files which are unnecessary for downstream analyses. This option allows you to choose not to delete those files at the end of the process;
  `Remove reference in downstream analyses` : the output files of Snippy-core normally also include the reference used for mapping. This option allows the removal of the reference from the `core.vcf` file, a necessary step to prevent the reference from appearing as a sample in the final tree;
  `Filter out recombinations with Gubbins` : it allows the execution of the additional software “Gubbins” to remove variants identified as recombinations from the `core.vcf` file.

- **ReporTree parameters:**

- `Partition threshold for clustering definition` : it allows to customize the threshold values for the creation of the tree partitions;

  `Minimum proportion of samples per site without missing data` : ReporTree normally considers all sites of the aligned genomes for distance matrix and tree computation. This option allows the user to define a threshold value: sites with a percentage of "missing data", i.e. a percentage of Ns (%N) below the specified threshold, are discarded. Only sites that pass the filter are retained and used for tree calculation;
  `Minimum proportion (0 to 1) of ATCG in informative sites of the alignment per sample` : ReporTree normally uses all provided samples for distance matrix and tree computation. This option allows the user to define a threshold value for the percentage of Ns in the sample sequence. Samples with %N below the specified value are discarded;
  `Sample of interest` options: they allow one or more samples of interest to be specified; on these samples, ReporTree recalculates a “Zoom”, i.e. a tree containing only the sample(s) of interest and the samples belonging to the same cluster;
  `Method used in HC clustering to create a dendrogram` options: selection of the optional *Hierarchical Clustering* mode;
  `Custom parameters for reportree` options: text fields for entering Command-line style options to be passed to ReporTree, exclusively for the computation of the trees specified in the option (MST produced with Grapetree or HC).

- **Clustering ML parametres:**

  `Maximum likelood phylogenetic tree calculation` : it allows the user to choose the software to be used for optional clustering with the ML algorithm. The available software are **IQ-TREE** and **FastTree**;
  `Custom parameters for ML phylogenetic tree calculation` : text field for entering Command-line style options for ML clustering;
  `[IQ-TREE only] Try to apply CMAPLE tree search algorithm` : if IQ-TREE is selected, it allows the user to choose whether to apply the **CMAPLE** tree placement algorithm, integrated into IQ-TREE.

Given the number and complexity of the parameters and outputs of Snippy-core and ReporTree, users are encouraged to consult the official documentation: - Snippy & Snippy-core official guide - ReporTree official guide

The procedure for launching the analysis is similar to other analyses that include the execution of ReporTree:

Run ReporTree analysis

Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

In the completed analysis card, in addition to the ability to browse the output folder, several options will be available, such as direct access to some log and metadata files, opening the .nwk file, and the direct visualization of the MSTree through the integration of SPREAD within the platform.

To view the tree, you will need to navigate to the ReporTree result card and select the link **Minimum Spanning Tree with SPREAD**.

> **Note:** Results from multi-sample analyses cannot be imported. Users are advised to use the dedicated functionality for non-importable analysis preservation or, alternatively, the option to directly download the newick file (`.nwk`) and the metadata file, so that the files can be stored and the tree can be visualized at any time, either within the platform or using external software.

Result folder

> To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are two subfolders:

- **meta** (*metadata*), where the log and configuration files of the executed process are saved;
- **result**, where the files containing the results produced by the analysis are saved.

The table below shows the list of the main files of interest present in the folders, along with some useful information.

File	Description	Location
`gt.nwk`	.nwk file of MST tree	folder: result
`gt_dist_grapetree.tsv`	MST tree distance matrix	folder: result
`snippycore.full.aln`	multifasta file of the alignment of samples to the reference (Snippy-core). It contains the complete alignment, including non-variable sites	folder: result
`snippycore.aln`	multifasta file of the alignment of samples to the reference (Snippy-core). It contains the alignment of only the variable regions	folder: result
`clean.full.aln`	multifasta file of the alignment of samples to the reference (Snippy-core). This is the filtered `snippycore.full.aln` file, i.e. after removing uncertain alignment positions with a low-coverage or gaps	folder: result
`reportree_metadata.tsv`	table of sample metadata produced by ReporTree	folder: result
`gt_dist.tsv`	pairwise distances of the MST tree	folder: result
`gt_clusterComposition.tsv`	table listing the clusters and their composition	folder: result
`gt_loci_used.txt`	list of loci used for the calculation of the distance matrix (*start position*)	folder: result
`gt_dist_hamming.tsv`	distance matrix	folder: result
`gt_partitions.tsv`	table with information on the cluster membership of the samples at each tree partitioning value	folder: result
`zooms.txt`	list of requested zoom levels	folder: result
`gt_metadata_w_partitions.tsv`	metadata table with information on cluster membership	folder: result
`gt_partitions_summary.tsv`	summary table of the tree partitions and the corresponding samples in the clusters	folder: result
`snippycore.tab`	.tsv table of variants. Columns: chromosome, position, nucleotide in the reference, nucleotide in sample 1, nucleotide in sample 2...	folder: result
`snippycore.vcf`	VCFv4.2 file of variants. It includes an informative header and a table with the variant type and a binary matrix indicating the presence/absence of the variant across samples	folder: result

The .nwk file contains all the tree data and can be visualised using **SPREAD** or other external software for viewing dendrograms or distance graphs.

Gene-by-gene based clustering

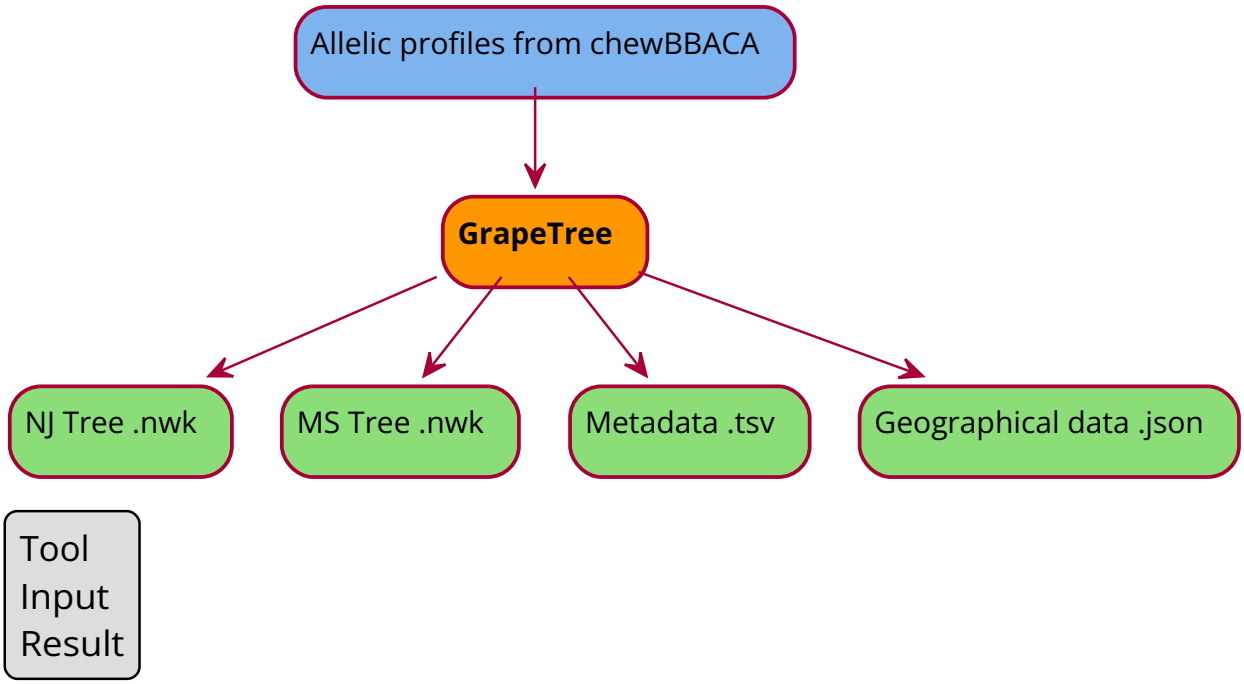
GrapeTree

Introduction

The GrapeTree analysis builds tree graphs using the "Minimum Spanning Tree" (MSTree) or "Neighbor Joining" (NJ) algorithms, based on the allelic profiles of provided samples, as identified in the cgMLST analysis.

GrapeTree GitHub page: https://github.com/achtman-lab/GrapeTree

On the platform, it is possible both to compute the trees and to visualize them thanks to the integration of the stand-alone version of the GrapeTree interactive visualization system.



A complete guide to the use and interface of the visualization system on the platform is available on the page dedicated to SPREAD.

For further information on the software, please refer to the research article and the official documentation of the EnteroBase system, of which GrapeTree is a part.

Run GrapeTree analysis

Once the **GrapeTree** analysis has been selected in the dedicated run analysis page, the system will move to a confirmation interface: there is no need to choose the tool since multi-sample analyses always use a specific software.

The interface for input selection provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.

The possible inputs that can be used for GrapeTree are:

- step_4TY_wgMLST
- step_4TY_cgMLST

For both possible inputs, it is imperative that the program used to compute the allelic profiles is **chewBBACA**. Therefore, we suggest you to be careful and select this software when running the 4TY_cgMLST or 4TY_wgMLST, if you intend to launch GrapeTree as a downstream analysis.

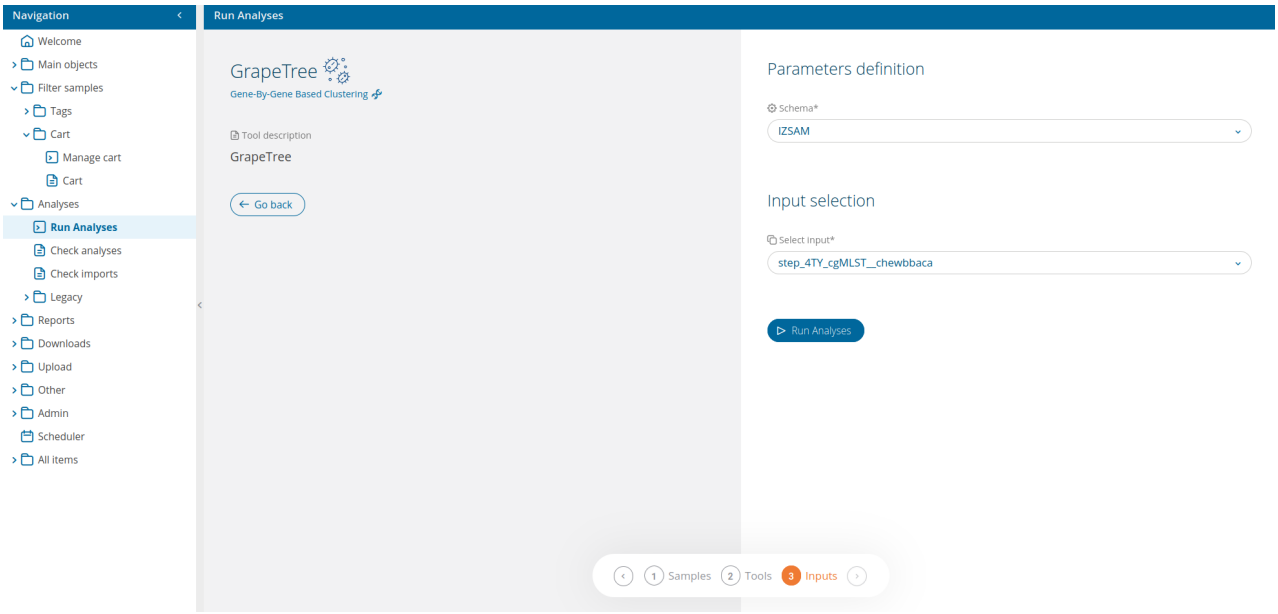
In the parameter definition screen it is possible to:

- select the encoding: the available options include the **crc32** encoding and the **izsam** encoding. Selecting **default** corresponds to *crc32*;
- specify the scheme for the microbial species on which 4TY_cgMLST__chewbbaca was performed, if allelic profiles from cgMLST are selected as input. The field is pre-filled with the appropriate scheme for the bacterial species whose codes are present in the cart, or the active tag, used to launch the analysis. No further action is therefore required from the user.

Note: by default, GrapeTree uses, as input, the output generated by **version 2.8.5** of the ChewBBACA database, encoding with **crc32**. To maintain backward compatibility with previous analyses, it is still possible to run GrapeTree on older profiles generated with ChewBBACA 2.8.4; in this case, it is necessary to select **izsam** as the encoding.

Please note that these associations must be respected for the analysis to be performed correctly. **Any other combination will result in the failure of the analysis**.

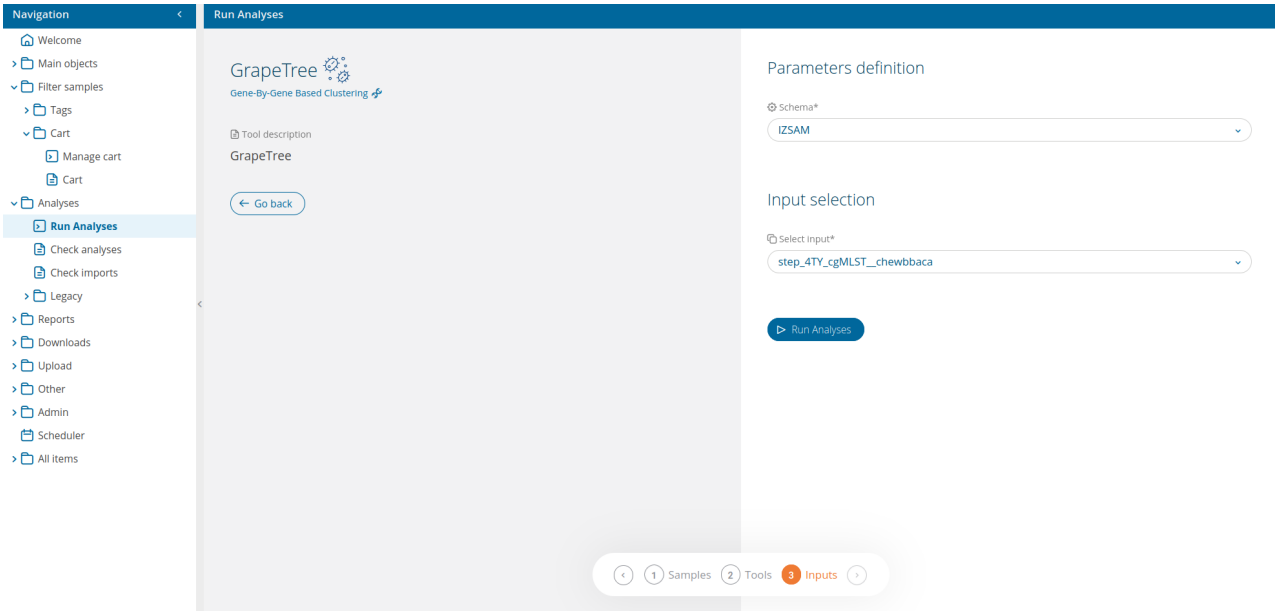
Moreover, please note that, for the analyses to be considered valid, it is necessary to use ChewBBACA version 2.8.5.



Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

In the completed analysis card, in addition to the option to explore the output folder, several features will be available, such as the access to some of the output files and the direct graph visualization, both MSTree and NJ graph, thanks to GrapeTree's integration in the platform.

Visualizing the tree will require to navigate to the Grapetree result section and select the **Minimum Spanning Tree with Grapetree** link:



Note: Multi-sample analysis results cannot be imported. Please, use the dedicated functionality for storing non-importable analyses or, alternatively, the option for directly downloading the newick (`.nwk`) files and the metadata file, so that you can preserve the files and view the tree at any time, either within the platform or using an external software.

Result folder

To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card. Inside the subsequent `Results` folder, there are two subfolders:

- **meta** (*metadata*), where the log and configuration files of the executed process are saved;
- **result**, where the files containing the results produced by the analysis are saved.

The table below shows the list of the main files of interest present in the folders, along with some useful information.

File	Description	Location
cgMLST.nwk	file of the tree in .nwk (newick) format calculated using the MSTree algorithm	folder: result
cgMLST.tsv	.tsv table of alleles identified by chewBBACA during cgMLST (allelic call). Rows: samples; Columns: *loci*.	folder: result
cgMLST_NJ.nwk	file of the tree in .nwk (newick) format calculated using the NJ algorithm	folder: result
missing_loci.tsv	list of *loci* not shared among the samples and therefore discarded for the tree construction	folder: result
cgMLST_dists_matrix.csv	allelic distance matrix from cgMLST in .csv format	folder: result

The .nwk files contain all the tree data and can be viewed using GrapeTree or other external software for dendrogram visualization.

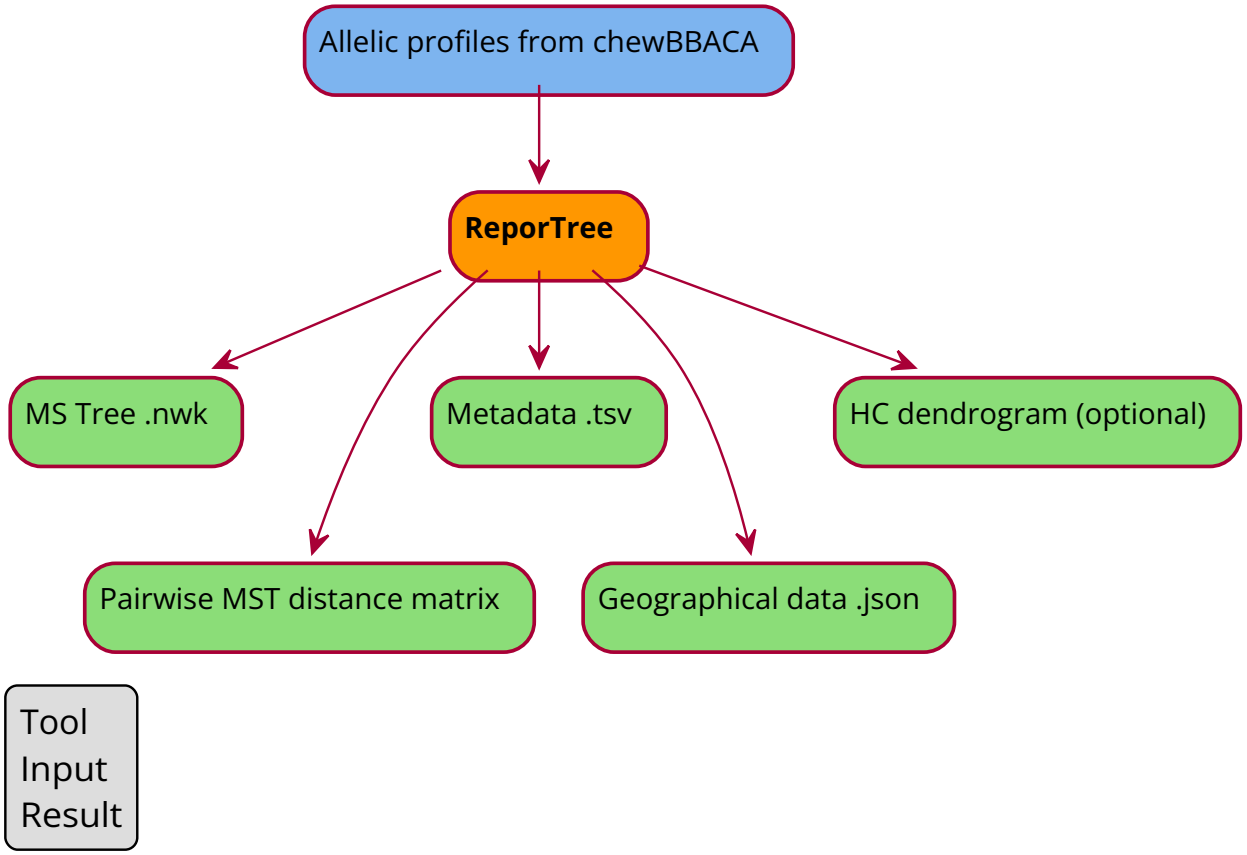
ReporTree (cgMLST)

Introduction

ReporTree builds distance graphs among microorganisms, starting from the allelic profiles of the provided samples. The tree is built using the "Minimum Spanning Tree" (MSTree or MST) algorithm.

Reportree Github page https://github.com/insapathogenomics/ReporTree - **IQ-TREE GitHub page:** https://github.com/iqtree/iqtree2 - **FastTree GitHub page:** https://github.com/morgannprice/fasttree

Once the tree has been computed, it can be viewed directly in the platform (using SPREAD), through the integrated dendrogram visualization system accessible from the completed analysis card.



Run Reportree analysis

Once the **ReporTree** analysis has been selected in the dedicated run analysis page, the system will move to a confirmation interface: there is no need to choose the tool since the analysis is specific for ReporTree use.

The interface for input selection provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.

The required inputs for ReporTree are the allelic profiles of the samples:

- step_4TY_wgMLST
- step_4TY_cgMLST

The parameter definition screen offers a high level of customization, including:

- **ReporTree parameters:**

  `Partition threshold for clustering definition` : it allows you to customize the threshold values for creating the tree partitions;
  `Minimum proportion of samples per site without missing data` : ReporTree normally considers all *loci* in the scheme when calculating the distance matrix and building the tree. This option allows you to set a threshold value: only *loci* called in a percentage of the dataset equal to or greater than the specified value will be considered. *Loci* that do not pass the filter are discarded;
  `Minimum proportion of loci/positions called for SNP/allele matrices` : ReporTree normally uses all provided samples to compute the distance matrix and build the tree. This option allows you to define a threshold for the number of *loci* called in the allelic profile of a sample. Samples with a number of called *loci* lower than the specified value are discarded;
  `Sample of interest` options: they allow one or more samples of interest to be specified on which ReporTree recalculates a "Zoom", i.e., a tree containing only the sample(s) of interest and the samples belonging to the same cluster;
  `Method used in HC clustering to create a dendrogram` options: selection of the mode for the optional *Hierarchical Clustering*;
  `Custom parameters` : text field where command-line-style options can be entered, which will be passed directly to ReporTree.
- **VCF2MST parameters:**

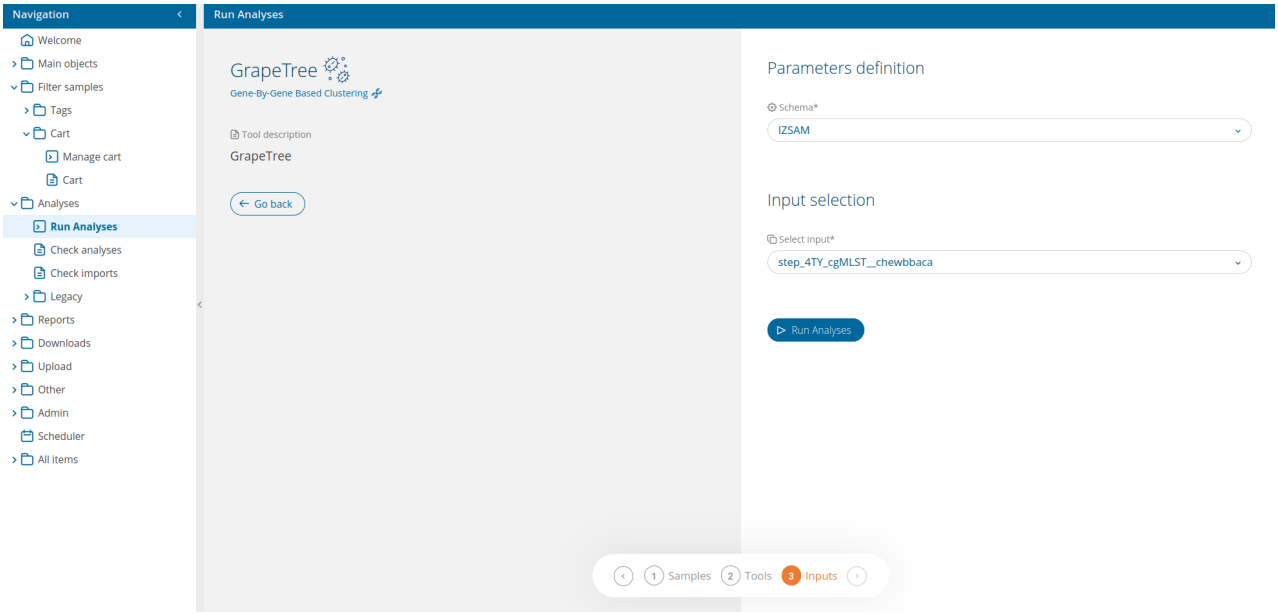
  `VCF filtering and normalization` : it enables the normalization of complex variants by decomposing them into their individual SNPs.

> Given the number and complexity of ReporTree parameters, please refer to the Reportree official guide for more information on how to use the software.

Run ReporTree analysis
Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

In the completed analysis card, several options will be available, including the possibility to browse the output folder, direct access to some log and metadata files, opening the .nwk file, and the direct visualization of the MSTree through the integration of GrapeTree into the platform.

Visualizing the tree will require to navigate to the Grapetree result section and select the **Minimum Spanning Tree with Grapetree** link:



Note: Multi-sample analysis results cannot be imported. Please, use the dedicated functionality for storing non-importable analyses or, alternatively, the option for directly downloading the newick (`.nwk`) file and the metadata file, so that you can preserve the files and view the tree at any time, either within the platform or using an external software.

Result folder

To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are two subfolders:

- **meta** (*metadata*), where the log and configuration files of the executed process are saved;
- **result**, where the files containing the results produced by the analysis are saved.

The table below shows the list of the main files of interest present in the folders, along with some useful information.

File	Description	Location
gt_clusterComposition.tsv	Table with a list of clusters and their composition	"result" folder
gt_dist_grapetree.tsv	Distance matrix	"result" folder
reportree_metadata.tsv	Table of sample metadata	"result" folder
gt_metadata_w_partitions.tsv	Table of sample metadata + cluster membership based on partitions	"result" folder
gt.nwk	MST tree in newick format	"result" folder
gt_dist_hamming.tsv	Table with Hamming distances	"result" folder
result_alleles_all.tsv	Table of allelic calls	"result" folder
gt_dist.tsv	List of pairwise distances between samples	"result" folder
gt_partitions_summary.tsv	List of the tree partitions and the corresponding samples in the clusters	"result" folder
gt_loci_report.tsv	Report on the statistics of the used *loci*	"result" folder
gt_partitions.tsv	Table of samples and their corresponding cluster membership based on partitioning	"result" folder
gt_loci_used.txt	List of *loci*	"result" folder

The .nwk file contains all the tree data and can be visualised using **SPREAD** or other external software for viewing dendrograms or distance graphs.

Pipeline Surveillance

Introduction

The ***Surveillance*** pipeline runs **ReporTree** on the allele profiles of the samples, specifying a set of **samples of interest**. The samples of interest are those provided as input (Tag or Cart); the distance matrix and the tree are computed using all samples of the specified species available on the platform.

The *Surveillance* pipeline is available, both as an automated analysis and as a manual pipeline, for the following bacteria:

- *Listeria monocytogenes*;
- *Campylobacter*;
- *Salmonella enterica*.

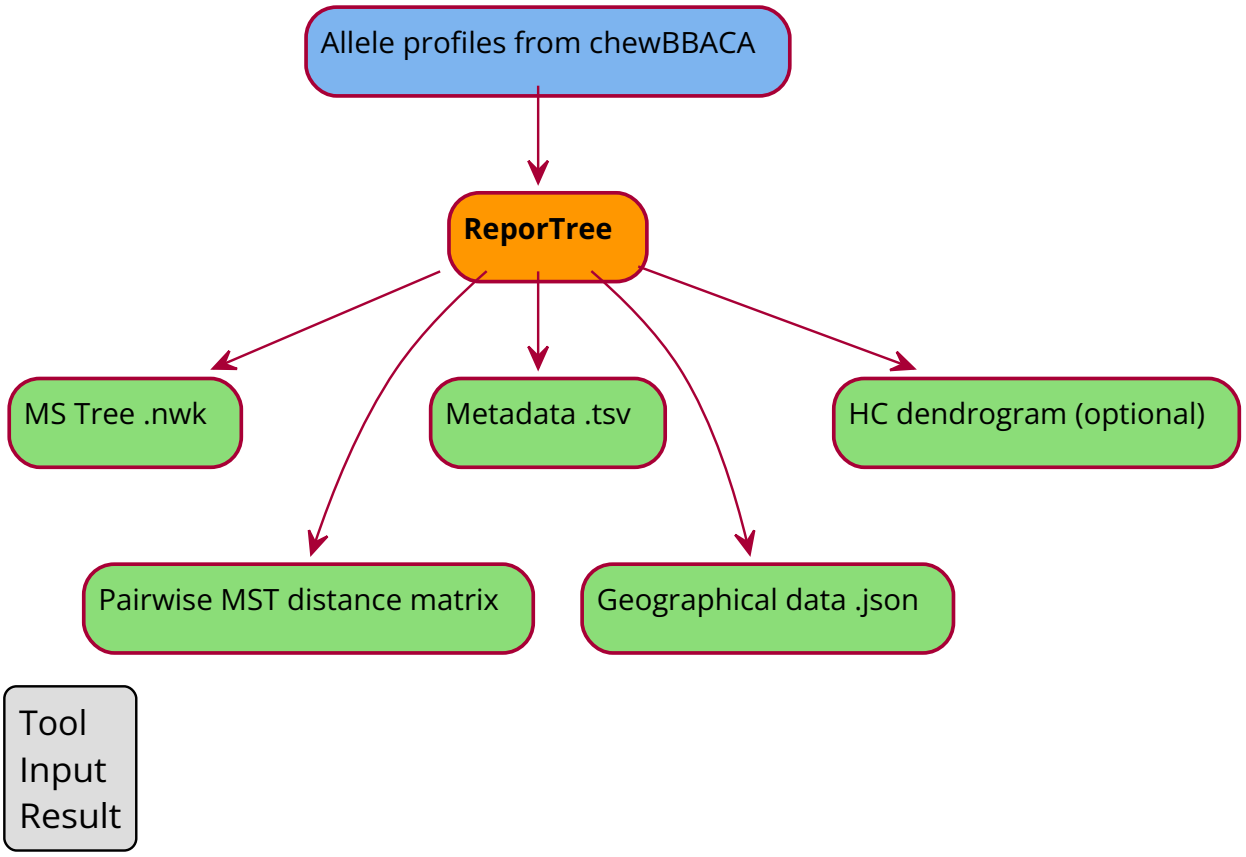
The analysis is automatically triggered by the platform whenever a new allele profile is computed for the microorganisms listed above. This includes: - allele profiles computed at the end of the NGSmanager pipeline - triggered *automatically* when new samples are uploaded to the platform; - triggered *manually* by a user; - manual executions of the 4TY_cgMLST analysis.

Surveillance can be manually launched from the Run analysis menu: - by LNR Listeria users for analyses on *L. monocytogenes* species; - by LNR Campylobacter users for analyses on *C. coli* and *C. jejuni* species; - by LNR Salmonella users for analyses on *S. enterica*.

cgMLST and wgMLST allele profiles can be used.

Reportree GitHub page https://github.com/insapathogenomics/ReporTree

Citation: *Mixão V., Pinto M., Sobral D., Di Pasquale A., Gomes J.P., Borges V., (2023), 'ReporTree: a surveillance-oriented tool to strengthen the linkage between pathogen genetic clusters and epidemiological data'. Genome Medicine. doi: 10.1186/s13073-023-01196-1*



Run Surveillance pipeline

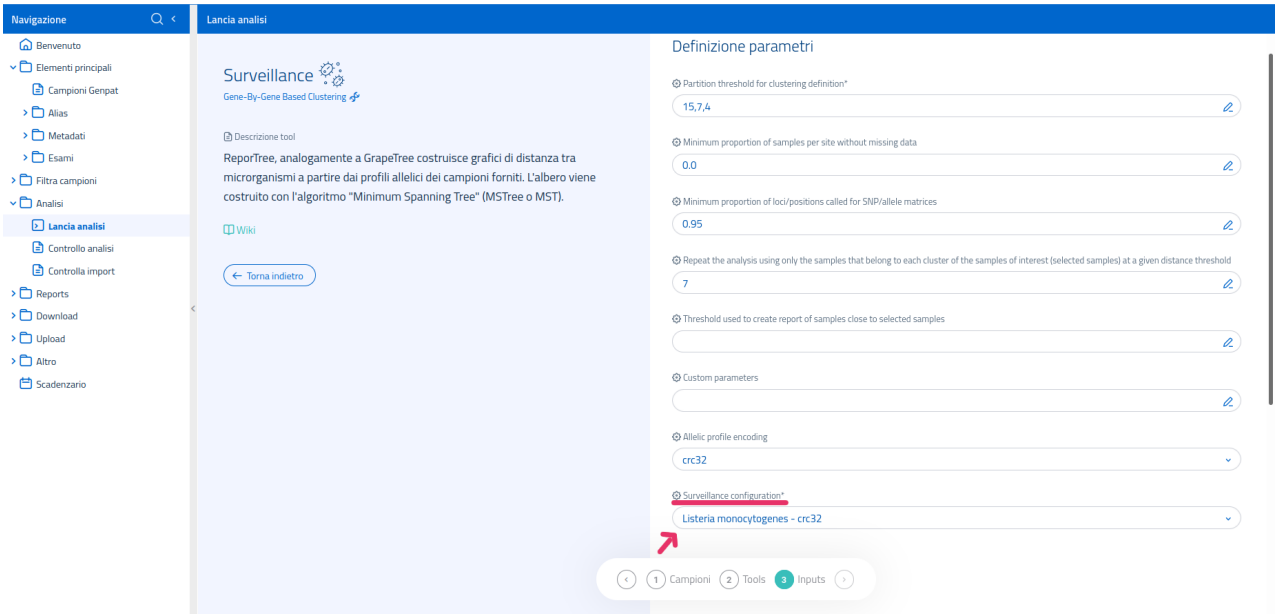
The procedure for launching the analysis is similar to that of the other analyses that include the execution of ReporTree:

Run ReporTree analyses
In the analysis launch system, it is possible to use the filter at the top to display only the pipelines. Once the **Surveillance** pipeline is selected on the run analyses page, the system will switch to a confirmation interface.

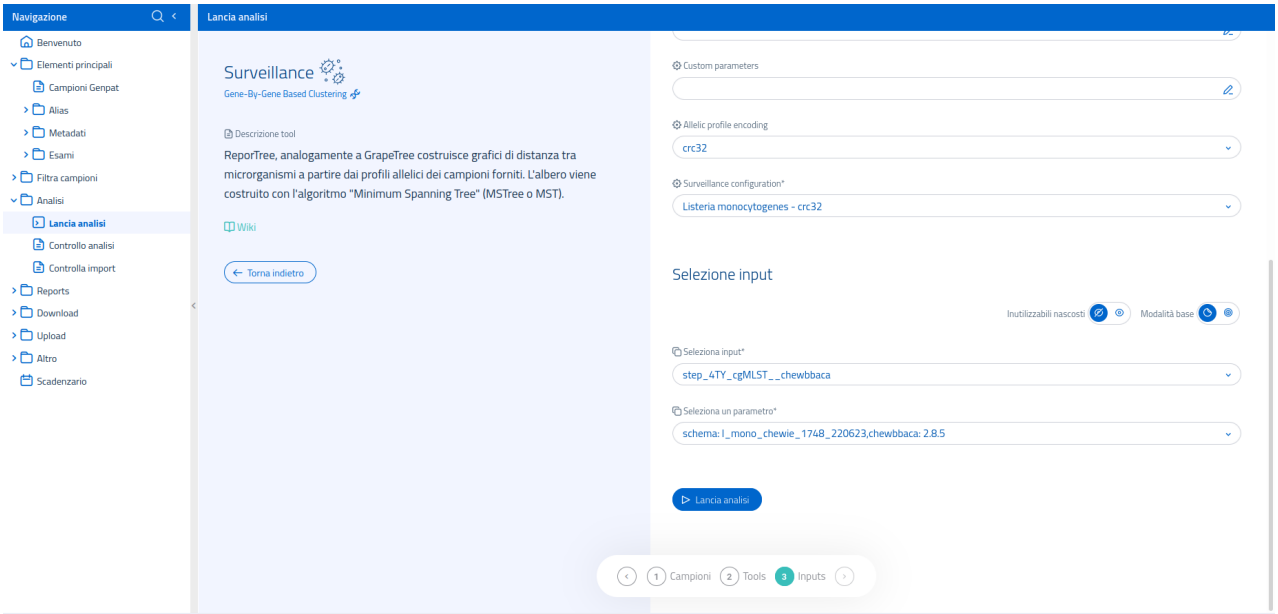
The **Surveillance** pipeline will run ReporTree with the following variations:

- While a standard ReporTree run is performed exclusively on the samples present in the cart, the ReporTree run, executed via the *Surveillance* pipeline, is performed on all samples available on the platform that belong to the species indicated in the dedicated field (image **A**).
- A standard ReporTree run requires specifying the codes of the samples of interest, among those used as input, either in the cart or in the tag; when ReporTree is run through the Surveillance pipeline, the input consists **only of the samples of interest** (that means that the activated cart or the active tag must contain *only* those samples) that, therefore, do not need to be esplicitly specified when defining the parameters.

A



B



Since the pipeline uses ReporTree, the inputs are the same:

- step_4TY_wgMLST;
- step_4TY_cgMLST.

On the parameter definition screen, the pipeline allows a high level of customization through the same additional ReporTree parameters.

For any usage and reference regarding the tool options, please refer to the ReporTree Wiki page.

For any further details, please refer to the relevant sections of the ReporTree analysis and to the official tool documentation:

- Run parameters;
- Result folder;
- Official ReporTree guide.

Pipeline Surveillance

Introduction

The ***Surveillance*** pipeline runs **ReporTree** on the allele profiles of the samples, specifying a set of **samples of interest**. The samples of interest are those provided as input (Tag or Cart); the distance matrix and the tree are computed using all samples of the specified species available on the platform.

The *Surveillance* pipeline is available, both as an automated analysis and as a manual pipeline, for the following bacteria:

- *Listeria monocytogenes*;
- *Campylobacter*;
- *Salmonella enterica*.

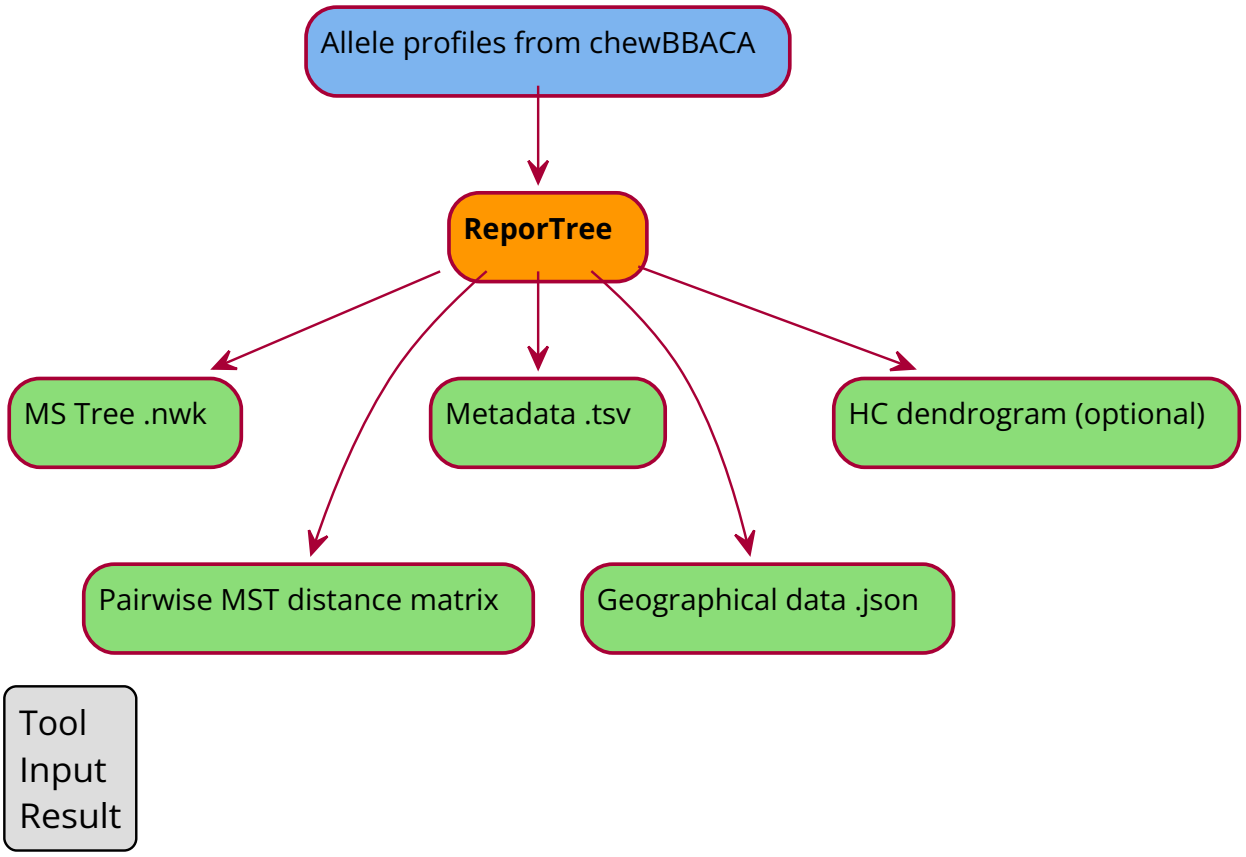
The analysis is automatically triggered by the platform whenever a new allele profile is computed for the microorganisms listed above. This includes: - allele profiles computed at the end of the NGSmanager pipeline - triggered *automatically* when new samples are uploaded to the platform; - triggered *manually* by a user; - manual executions of the 4TY_cgMLST analysis.

Surveillance can be manually launched from the Run analysis menu: - by LNR Listeria users for analyses on *L. monocytogenes* species; - by LNR Campylobacter users for analyses on *C. coli* and *C. jejuni* species; - by LNR Salmonella users for analyses on *S. enterica*.

cgMLST and wgMLST allele profiles can be used.

Reportree GitHub page https://github.com/insapathogenomics/ReporTree

Citation: *Mixão V., Pinto M., Sobral D., Di Pasquale A., Gomes J.P., Borges V., (2023), 'ReporTree: a surveillance-oriented tool to strengthen the linkage between pathogen genetic clusters and epidemiological data'. Genome Medicine. doi: 10.1186/s13073-023-01196-1*



Run Surveillance pipeline

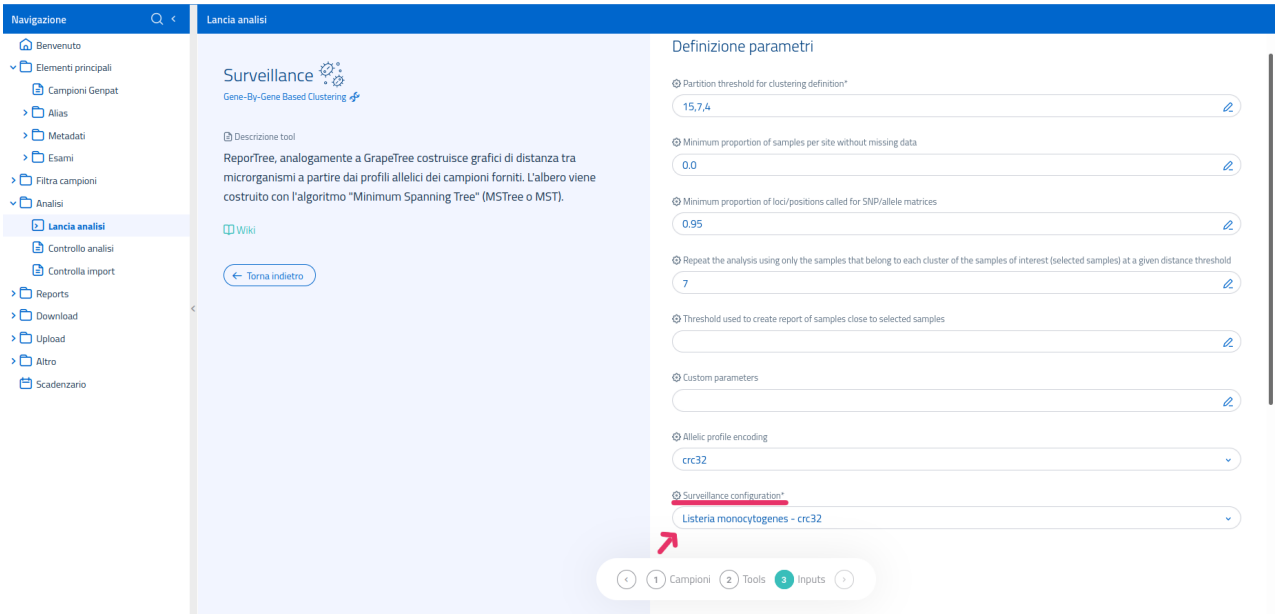
The procedure for launching the analysis is similar to that of the other analyses that include the execution of ReporTree:

Run ReporTree analyses
In the analysis launch system, it is possible to use the filter at the top to display only the pipelines. Once the **Surveillance** pipeline is selected on the run analyses page, the system will switch to a confirmation interface.

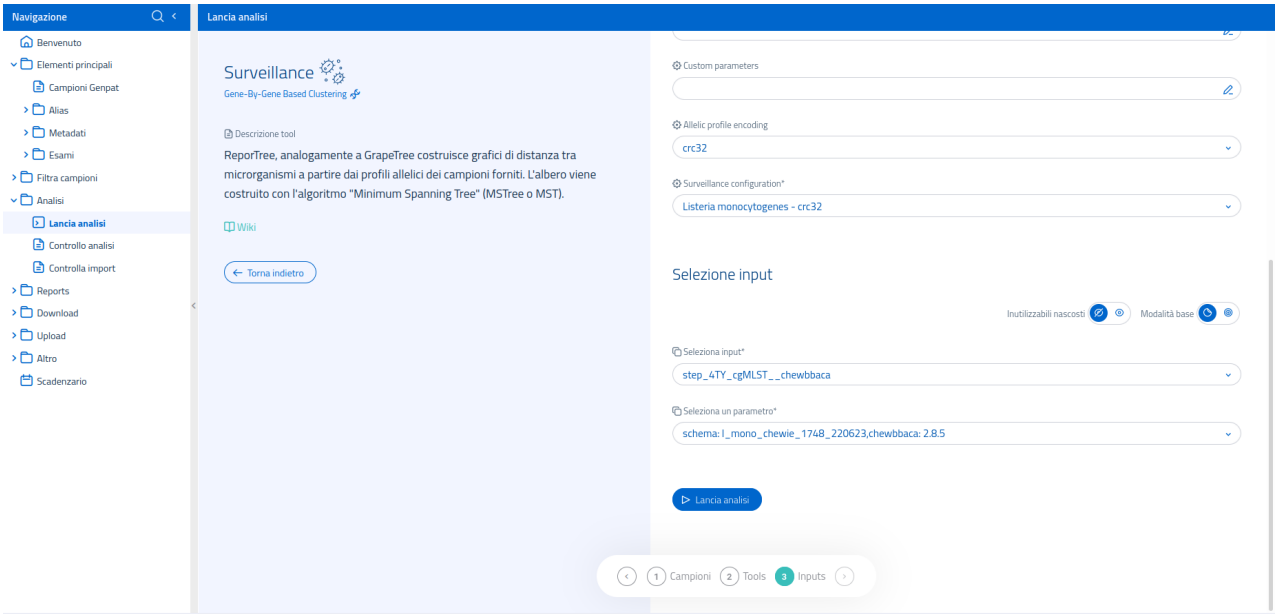
The **Surveillance** pipeline will run ReporTree with the following variations:

- While a standard ReporTree run is performed exclusively on the samples present in the cart, the ReporTree run, executed via the *Surveillance* pipeline, is performed on all samples available on the platform that belong to the species indicated in the dedicated field (image **A**).
- A standard ReporTree run requires specifying the codes of the samples of interest, among those used as input, either in the cart or in the tag; when ReporTree is run through the Surveillance pipeline, the input consists **only of the samples of interest** (that means that the activated cart or the active tag must contain *only* those samples) that, therefore, do not need to be esplicitly specified when defining the parameters.

A



B



Since the pipeline uses ReporTree, the inputs are the same:

- step_4TY_wgMLST;
- step_4TY_cgMLST.

On the parameter definition screen, the pipeline allows a high level of customization through the same additional ReporTree parameters.

For any usage and reference regarding the tool options, please refer to the ReporTree Wiki page.

For any further details, please refer to the relevant sections of the ReporTree analysis and to the official tool documentation:

- Run parameters;
- Result folder;
- Official ReporTree guide.

Pipeline Surveillance

Introduction

The ***Surveillance*** pipeline runs **ReporTree** on the allele profiles of the samples, specifying a set of **samples of interest**. The samples of interest are those provided as input (Tag or Cart); the distance matrix and the tree are computed using all samples of the specified species available on the platform.

The *Surveillance* pipeline is available, both as an automated analysis and as a manual pipeline, for the following bacteria:

- *Listeria monocytogenes*;
- *Campylobacter*;
- *Salmonella enterica*.

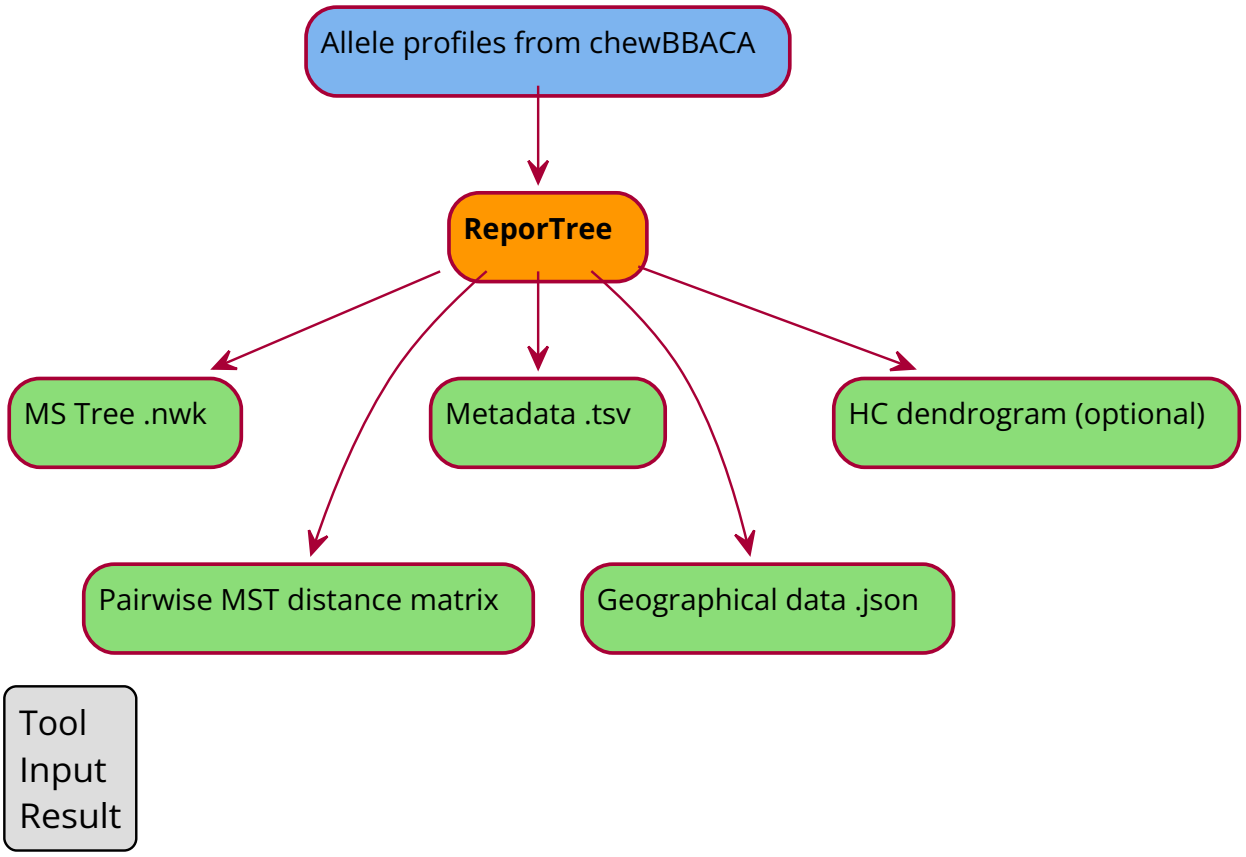
The analysis is automatically triggered by the platform whenever a new allele profile is computed for the microorganisms listed above. This includes: - allele profiles computed at the end of the NGSmanager pipeline - triggered *automatically* when new samples are uploaded to the platform; - triggered *manually* by a user; - manual executions of the 4TY_cgMLST analysis.

Surveillance can be manually launched from the Run analysis menu: - by LNR Listeria users for analyses on *L. monocytogenes* species; - by LNR Campylobacter users for analyses on *C. coli* and *C. jejuni* species; - by LNR Salmonella users for analyses on *S. enterica*.

cgMLST and wgMLST allele profiles can be used.

Reportree GitHub page https://github.com/insapathogenomics/ReporTree

Citation: *Mixão V., Pinto M., Sobral D., Di Pasquale A., Gomes J.P., Borges V., (2023), 'ReporTree: a surveillance-oriented tool to strengthen the linkage between pathogen genetic clusters and epidemiological data'. Genome Medicine. doi: 10.1186/s13073-023-01196-1*



Run Surveillance pipeline

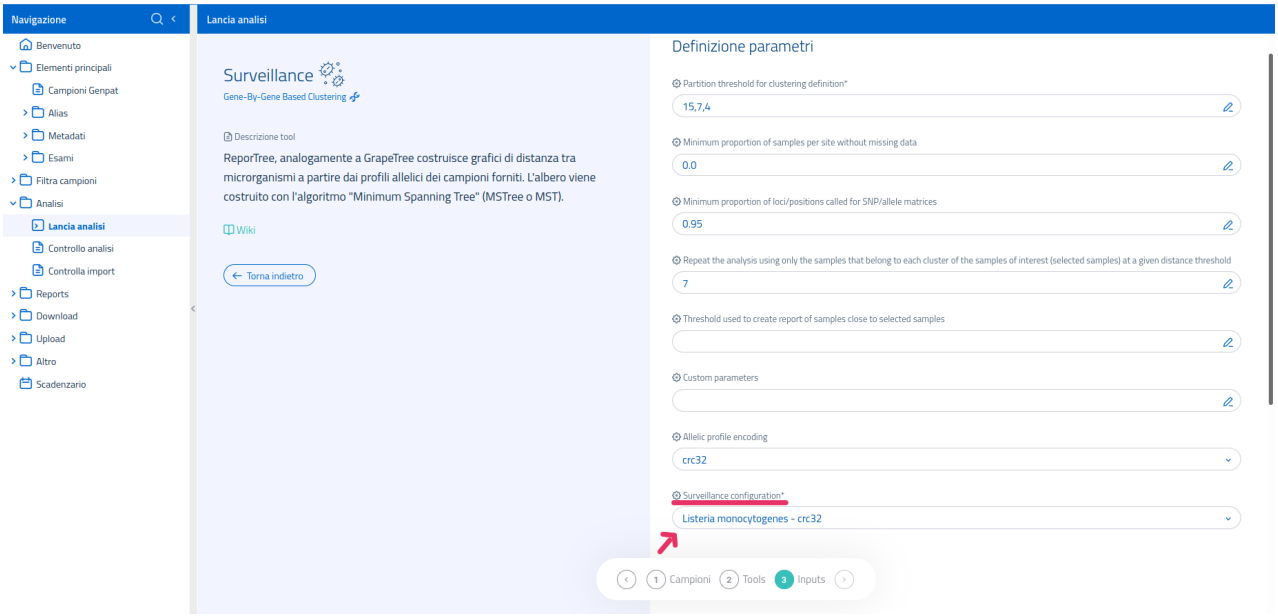
The procedure for launching the analysis is similar to that of the other analyses that include the execution of ReporTree:

Run ReporTree analyses
In the analysis launch system, it is possible to use the filter at the top to display only the pipelines. Once the **Surveillance** pipeline is selected on the run analyses page, the system will switch to a confirmation interface.

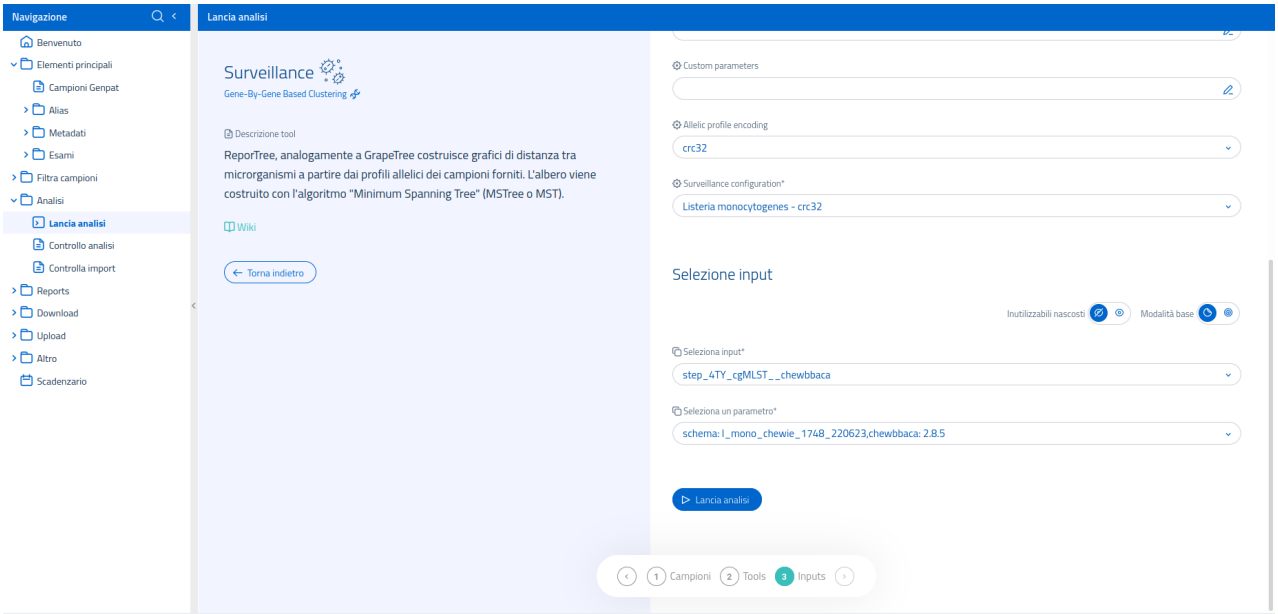
The **Surveillance** pipeline will run ReporTree with the following variations:

- While a standard ReporTree run is performed exclusively on the samples present in the cart, the ReporTree run, executed via the *Surveillance* pipeline, is performed on all samples available on the platform that belong to the species indicated in the dedicated field (image **A**).
- A standard ReporTree run requires specifying the codes of the samples of interest, among those used as input, either in the cart or in the tag; when ReporTree is run through the Surveillance pipeline, the input consists **only of the samples of interest** (that means that the activated cart or the active tag must contain *only* those samples) that, therefore, do not need to be esplicitly specified when defining the parameters.

A



B



Since the pipeline uses ReporTree, the inputs are the same:

- step_4TY_wgMLST;
- step_4TY_cgMLST.

On the parameter definition screen, the pipeline allows a high level of customization through the same additional ReporTree parameters.

> For any usage and reference regarding the tool options, please refer to the ReporTree Wiki page.

For any further details, please refer to the relevant sections of the ReporTree analysis and to the official tool documentation:

- Run parameters;
- Result folder;
- Official ReporTree guide.

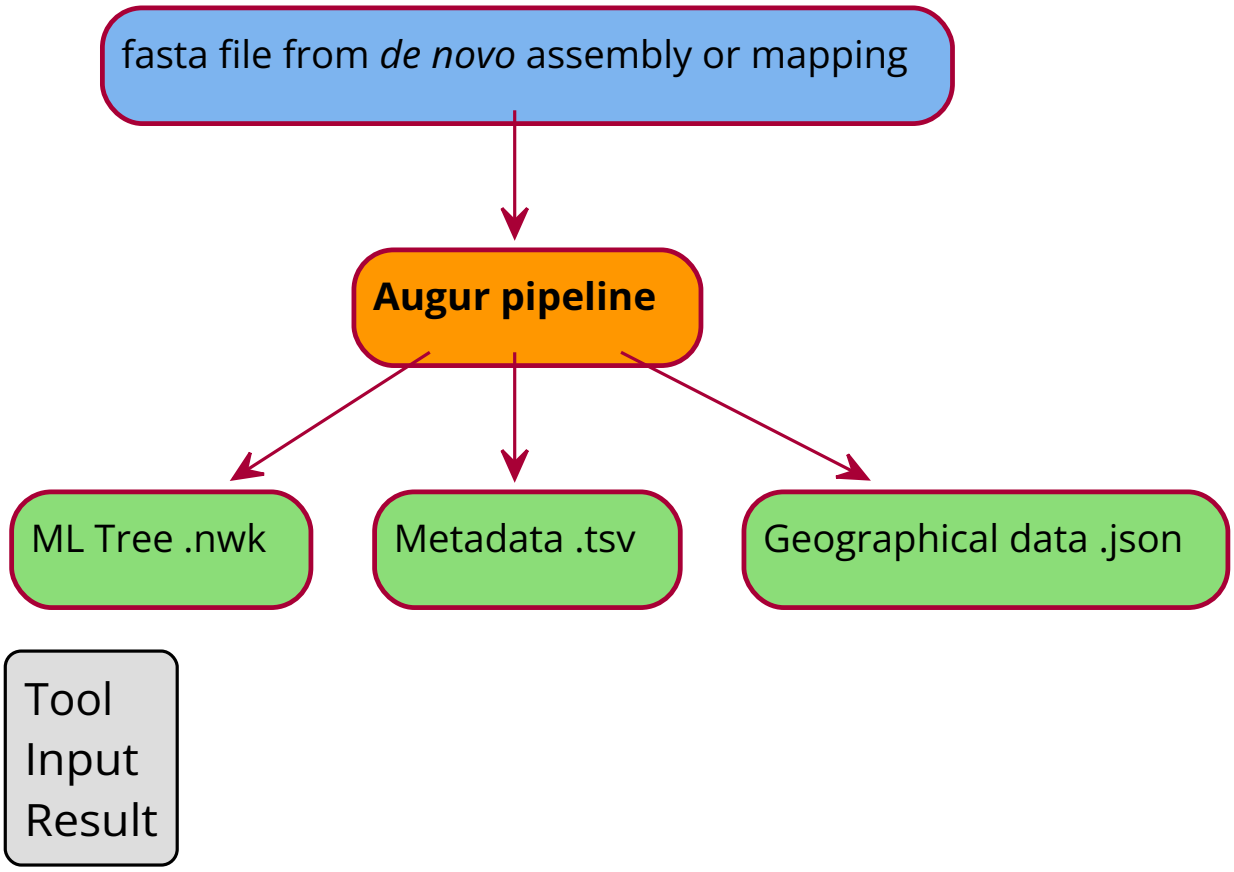
Augur

Introduction

The integration of "*Augur: A bioinformatics toolkit for phylogenetic analysis*", into the platform, allows the **Augur pipeline** to be run to produce a phylogenetic tree (using the Maximum Likelihood - ML - algorithm) starting from the FASTA files of the mapping or the *de novo* assembly.

Augur github page: https://github.com/nextstrain/augur/tree/7cb384877cb61e26781dec947f81fee29462c008
Official Augur documentation page: https://docs.nextstrain.org/projects/augur/en/stable/index.html

Once the tree has been computed, it can be viewed directly in the platform (using Auspice), through the integrated dendrogram visualization system accessible from the completed analysis card.



Run Augur analysis

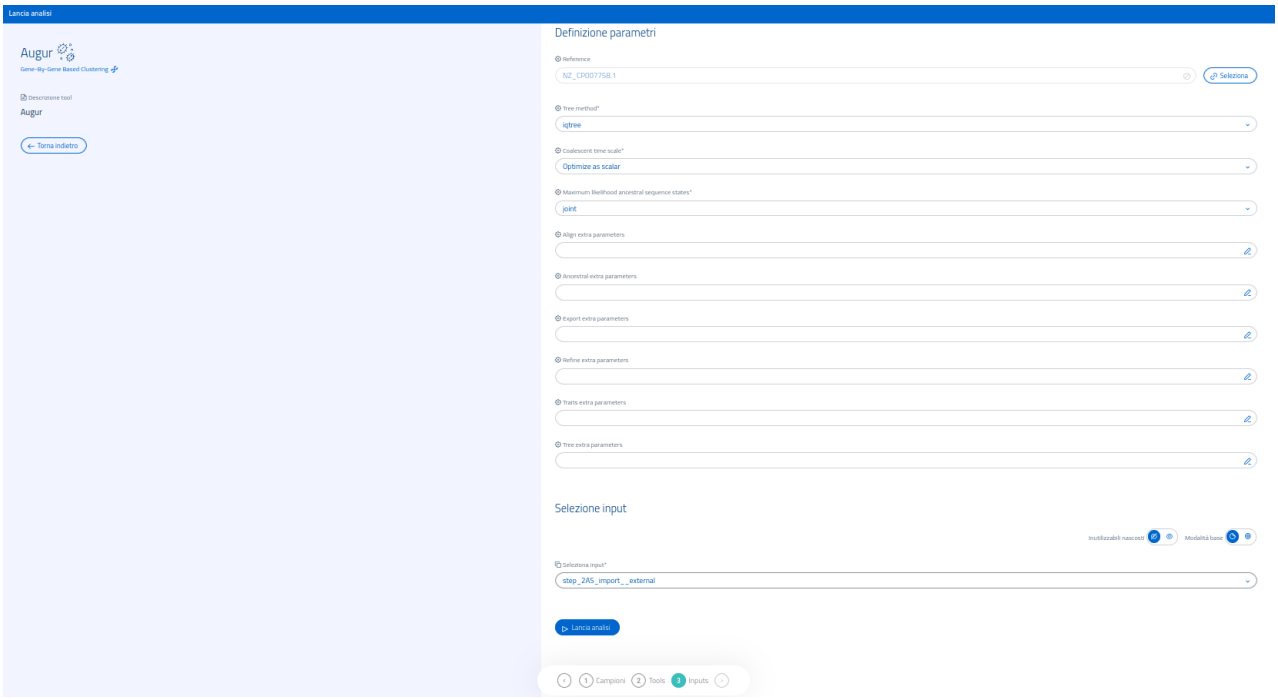
Once the **Augur** analysis has been selected in the dedicated run analysis page, the system will move to a confirmation interface: there is no need to choose the tool, since this analysis is specifically designed for the use of Augur.

The interface for input selection provides the advanced input selection mode, to allow the use of inputs processed by different methods, that were applied upstream in the analysis workflow.

The possible inputs that can be used for Augur include the results of all mapping analyses or *de novo* assembly.

In the parameter definition screen it is possible to:

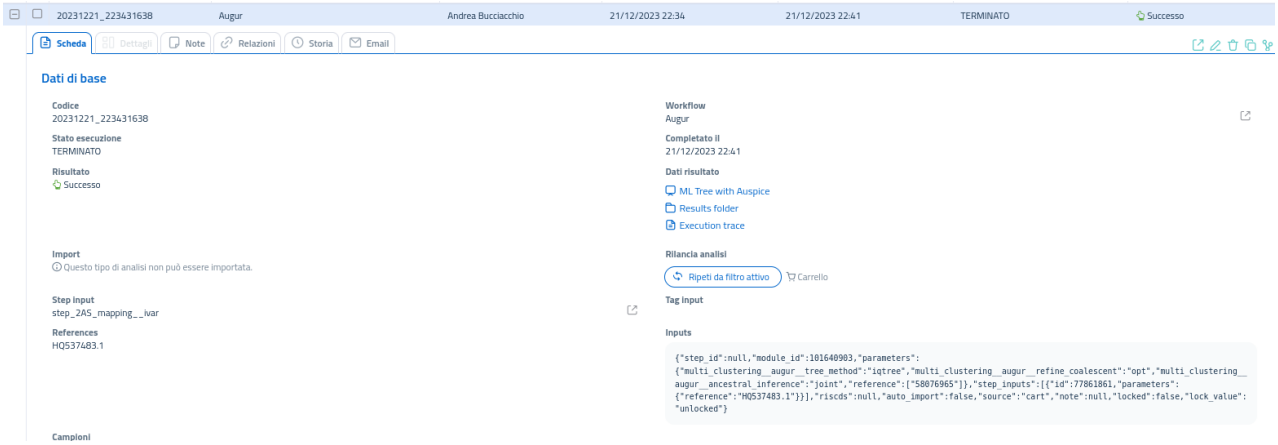
- select the reference to be used as the tree root;
- choose the useful software to build the .nwk file (iqtree, fasttree or raxm);
- select a mode for the time scale;
- select a mode for the Maximum Likelihood algorithm in handling the ancestral sequence (*reference*);
- add additional parameters as text (the fields will recognize the parameters for launching the tool from the terminal. Please, refer to the Augur official guide Augur for the list of available parameters).



Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

In the completed analysis card, in addition to the option to explore the output folder, several features will be available, such as direct access to selected log and metadata files, opening the .nwk file and the *direct visualization* of the Maximumm Likelihood tree through the integration, within the platform, of Auspice.

Visualizing the tree will require to navigate to the Grapetree result section and select the **ML Tree with Auspice** link:



Note: Multi-sample analysis results cannot be imported. Please, use the dedicated functionality for storing non-importable analyses or, alternatively, the option for directly downloading the newick (`.nwk`) files and the metadata file, so that you can preserve the files and view the tree at any time, either within the platform or using an external software.

Result folder

To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card. Inside the subsequent `Results` folder, there are two subfolders:

- **meta** (*metadata*), where the log and configuration files of the executed process are saved;
- **result**, where the files containing the results produced by the analysis are saved.

The table below shows the list of the main files of interest present in the folders, along with some useful information.

File	Description	Position
aa_muts.json	json file used by Auspice for amino acid mutations	folder: result
alignment.fasta	multi-alignment FASTA file	folder: result
augur_metadata.tsv	metadata table in .tsv format produced by Augur	folder: result
auspice.json	json file used by Auspice to map correspondences between dendrogram leaves and nodes on the map	folder: result
branch_lengths.json		folder: result
nt_muts.json	json file used by Auspice for nucleotide sequence mutations	folder: result
traits.json	json file used by Auspice to populate the metadata fields	folder: result
tree.nwk	tree file in .nwk (newick) format, calculated using the Maximum Likelihood algorithm	folder: result
metadata.csv	metadata table in .csv format	folder: result

The .nwk files contain all the tree data and can be viewed using Auspice or other external software for dendrogram visualization.

MAFFT

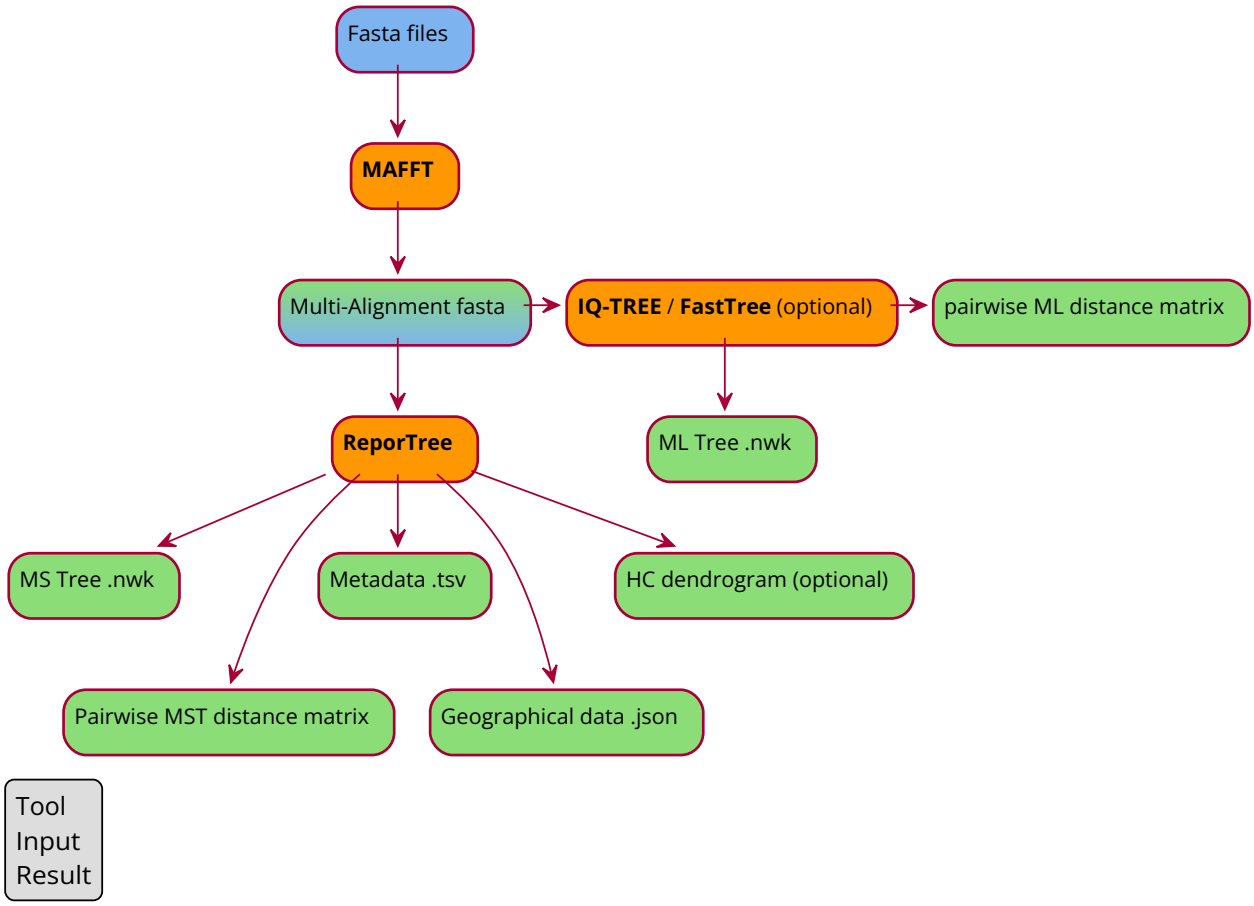
Introduction

The **MAFFT** analysis performs multiple sequence alignment; the analysis is followed by the launch of **ReporTree** as a clustering analysis.

MAFFT produces the FASTA file of the multiple sequence alignment, which is used as input for ReporTree. The latter, in turn, produces the distance matrix and the "Minimum Spanning Tree" (MSTree or MST); the analysis can also compute the *hierarchical clustering* graph.

- **Official MAFFT website:** https://mafft.cbrc.jp/alignment/software/
- **MAFFT GitHub page:** https://github.com/GSLBiotech/mafft
- **Reportree GitHub page:** https://github.com/insapathogenomics/ReporTree
- **IQ-TREE GitHub page:** https://github.com/iqtree/iqtree2
- **FastTree GitHub page:** https://github.com/morgannprice/fasttree

Once the MST tree has been computed, it can be visualized directly within the platform using the integrated tree visualization system (SPREAD), accessible from the completed analysis card.



Run MAFFT analysis

Once the **MAFFT** analysis has been selected in the dedicated run analysis page, the system will move to a confirmation interface, also providing the advanced input selection mode. This feature allows the use of inputs generated by different software upstream of the analysis workflow.

The analysis produces the multiple sequence alignment file, which is subsequently used by ReporTree. The input files for constructing the multi-FASTA are the individual FASTA files of the samples, which can be provided as:

- step_2AS_mapping
- step_2AS_denovo
- step_2AS_hybrid
- step_2AS_import

The parameter configuration screen offers a high level of customization, including:

- **MAFFT parameters:**

  `Custom parameters for MAFFT` : text field for Command-line style options, where MAFFT modifiers can be specified.

- **ReporTree parameters:**

  `Partition threshold for clustering definition` : it allows to customize the threshold values for the creation of the tree partitions;
  `Minimum proportion of samples per site without missing data` : ReporTree normally considers all sites of the aligned genomes for distance matrix and tree computation. This option allows the user to define a threshold value: sites with a percentage of "missing data", i.e. a percentage of Ns (%N) below the specified threshold, are discarded. Only sites that pass the filter are retained and used for tree calculation;
  `Minimum proportion (0 to 1) of ATCG in informative sites of the alignment per sample` : ReporTree normally uses all provided samples for distance matrix and tree computation. This option allows the user to define a threshold value for the percentage of Ns in the sample sequence. Samples with %N below the specified value are discarded;
  `Sample of interest` options: they allow one or more samples of interest to be specified; on these samples, ReporTree recalculates a “Zoom”, i.e. a tree containing only the sample(s) of interest and the samples belonging to the same cluster;
  `Method used in HC clustering to create a dendrogram` options: selection of the optional *Hierarchical Clustering* mode;
  `Custom parameters for reportree` options: text fields for entering Command-line style options to be passed to ReporTree, exclusively for the computation of the trees specified in the option (MST produced with Grapetree or HC).

- **ML clustering parameters:**

  `Maximum likelood phylogenetic tree calculation` : it allows the user to choose the software to be used for optional clustering with the ML algorithm. The available software are **IQ-TREE** and **FastTree**;
  `Custom parameters for ML phylogenetic tree calculation` : text field for entering Command-line style options for ML clustering;
  `[IQ-TREE only] Try to apply CMAPLE tree search algorithm` : if IQ-TREE is selected, it allows the user to choose whether to apply the **CMAPLE** tree placement algorithm, integrated into IQ-TREE.

Given the number and complexity of ReporTree parameters, please refer to the Reportree official guide for more information on how to use the software.

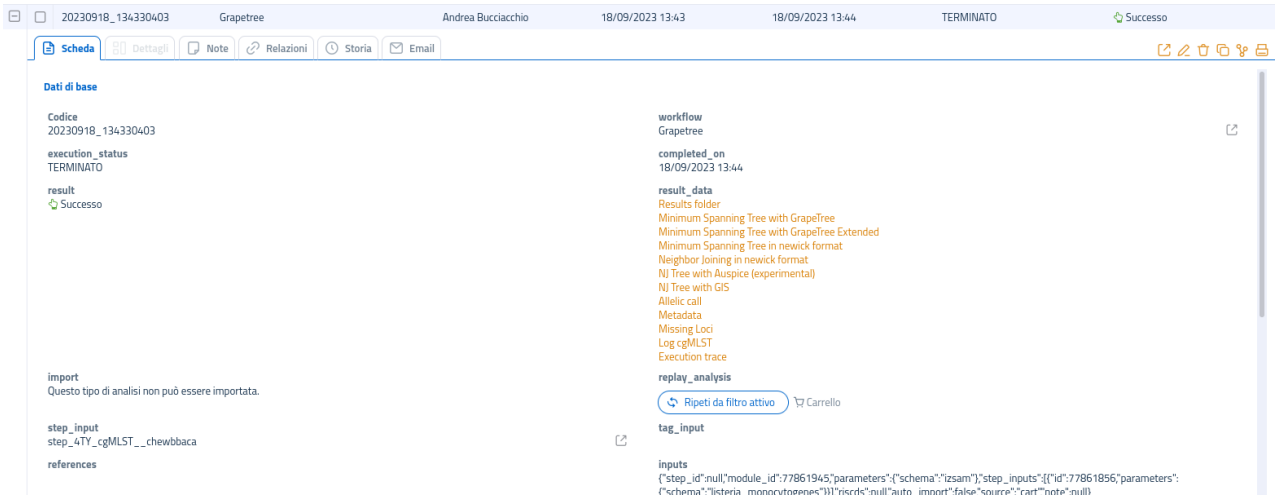
The procedure for launching the analysis is similar to other analyses that include the execution of ReporTree:

Run ReporTree analysis

Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

In the completed analysis card, in addition to the option to explore the result folder, several features will be available, such as direct access to some log files and metadata, opening the NWK file and the direct graph visualization of the MSTree through the integration of SPREAD into the platform.

To visualize the tree, it will be necessary to navigate to the GrapeTree result card and select the **Minimum Spanning Tree with SPREAD** link:



Note: Results from multi-sample analyses cannot be imported. Users are advised to use the dedicated functionality for non-importable analysis preservation or, alternatively, the option to directly download the newick file (`.nwk`) and the metadata file, so that the files can be stored and the tree can be visualized at any time, either within the platform or using external software.

Result folder

To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are two subfolders:

- **meta** (*metadata*), where the log and configuration files of the executed process are saved;
- **result**, where the files containing the results produced by the analysis are saved.

The table below shows the list of the main files of interest present in the folders, along with some useful information.

File	Description	Location
`gt.nwk`	.nwk file of MSTree	folder: result
`gt_dist_grapetree.tsv`	MST tree distance matrix	folder: result
`alignment.fasta`	multifasta of the alignment of the samples	folder: result
`reportree_metadata.tsv`	sample metadata table produced by ReporTree	folder: result
`gt_dist.tsv`	pairwise distances of the MST tree	folder: result
`gt_clusterComposition.tsv`	Table listing clusters and their composition	folder: result
`gt_loci_used.txt`	list of loci used for the calculation of the distance matrix (*start position*)	folder: result
`multifasta.fasta`	multifasta of the unaligned sample sequences	folder: result
`gt_dist_hamming.tsv`	distance matrix	folder: result
`gt_partitions.tsv`	table with information on the cluster membership of the samples at each tree partitioning value	folder: result
`zooms.txt`	list of requested zoom levels	folder: result
`gt_metadata_w_partitions.tsv`	metadata table with information on cluster membership	folder: result
`gt_partitions_summary.tsv`	summary table of the tree partitions and the corresponding samples in the clusters	folder: result

The file in Newick format (.nwk) contains all the tree data and can be visualized using **SPREAD** or another external software for displaying dendrograms or distance plots.

Pangenome extraction

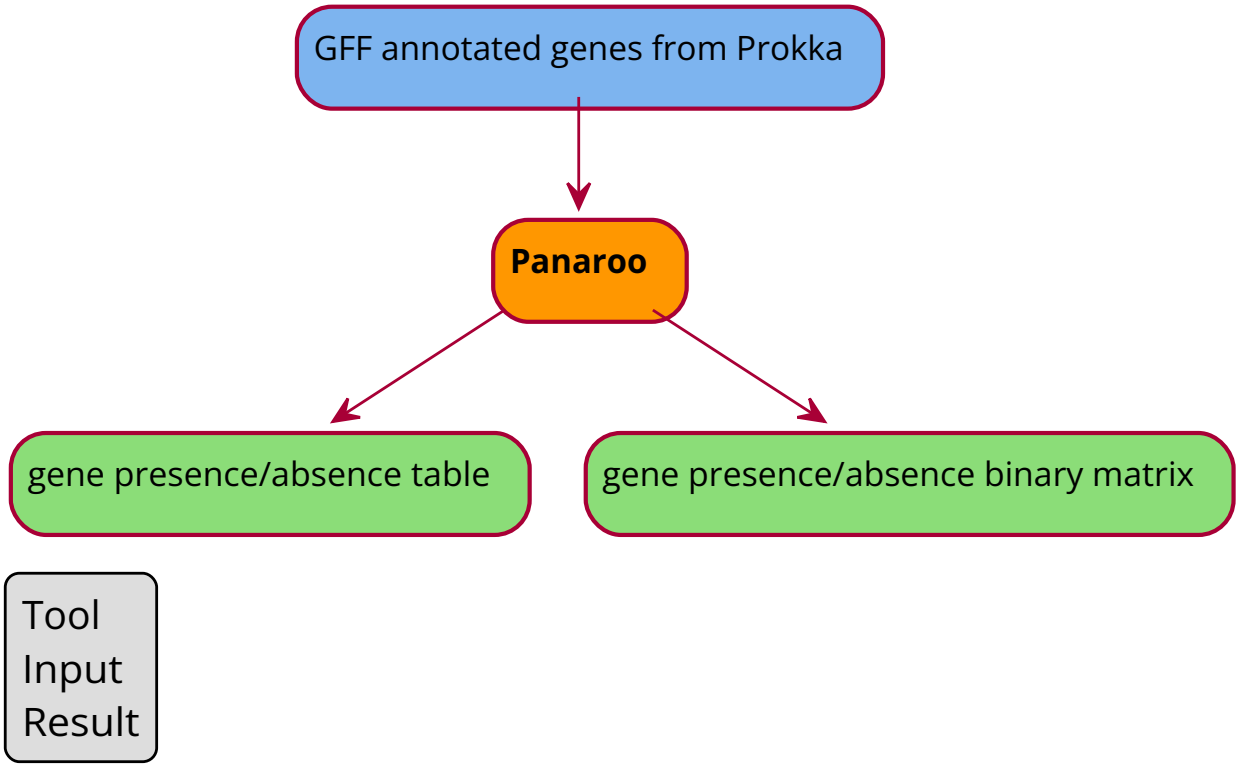
Panaroo

Introduction

Panaroo processes the pan-genome into presence/absence matrices of annotated genes across the samples, starting from the functional genome annotation performed by Prokka.

For more details on how Panaroo works, please refer to Panaroo's official guide.

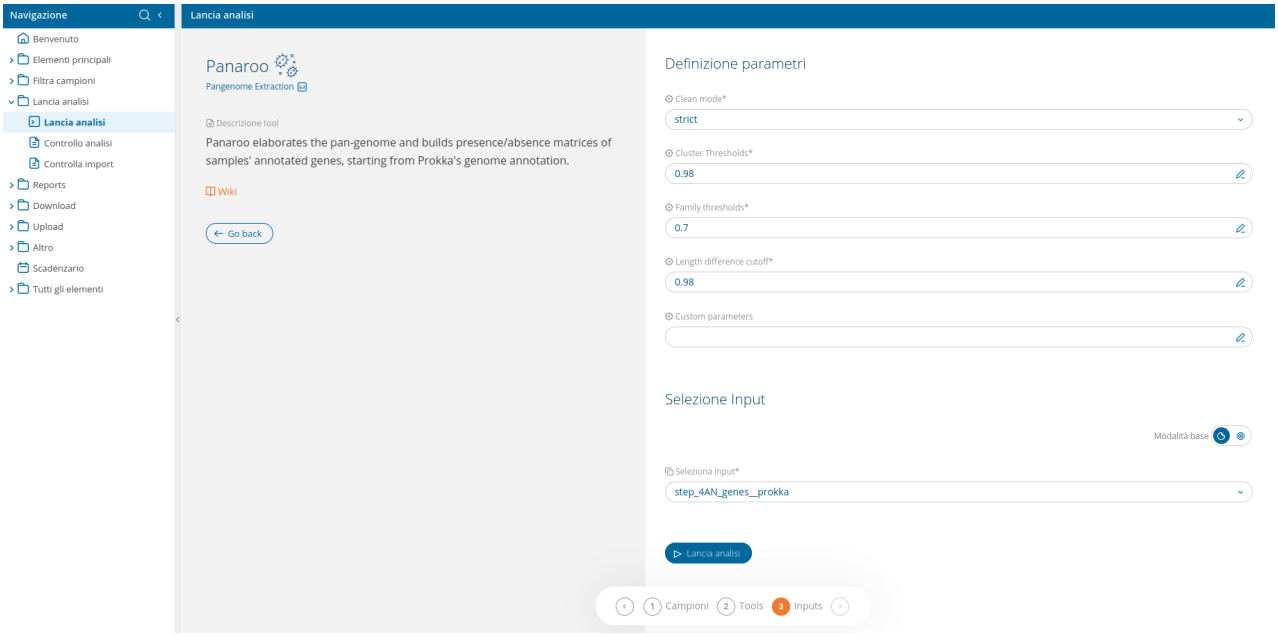
Panaroo's GitHub page: https://github.com/gtonkinhill/panaroo



Run Panaroo analysis

Once the **Panaroo** analysis has been selected in the dedicated run analysis page, the system will move to a confirmation interface: there is no need to choose the tool since the analysis is specific for the use of *Panaroo - An updated pipeline for pangenome investigation.*

In the interface for input selection, it is possible to confirm the input for the program: Panaroo is designed to process the outputs of Prokka; it is therefore necessary to use the annotation files from 4AN_genes as input. Since there is no possibility of choice, the field is pre-filled and no further user intervention is required.



Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The system will send the user a notification both when the analysis is successfully started and when the operation is completed.

Result folder

> To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are two subfolders:

- **meta** (*metadata*), where the log and configuration files of the executed process are saved;
- **result**, where the files containing the results produced by the analysis are saved.

The table below shows the list of the main files produced by Panaroo, along with some useful information.

Further information on Panaroo output files can be found in the official guide to Panaroo outputs.

File	Description	Location
combined_DNA_CDS.fasta	FASTA file containing the nucleotide sequences of the annotated genes and genes identified by Panaroo	folder: result
combined_protein_CDS.fasta	FASTA file containing the amino acid sequences of the annotated genes and genes identified by Panaroo	folder: result
combined_protein_GFF3cdhit_out.txt	log of the Panaroo CD-HIT step	folder: result
combined_protein_cdhit_out.txt.clstr	information about CD-HIT clusters	folder: result
core_alignment_header.embl		folder: result
core_gene_alignment.aln	alignment file	folder: result
final_graph.gml	pan-genome graph	folder: result
gene_data.csv	.csv table of the correspondences between annotated gene sequences and the internal codes used by Panaroo	folder: result
gene_presence_absence.Rtab	binary matrix of gene presence/absence in the samples	folder: result
gene_presence_absence.csv	.csv file describing gene presence in samples	folder: result
gene_presence_absence_roary.csv	.csv file describing gene presence in samples with additional information (following Roary's model)	folder: result
pan_genome_reference.fa	*reference pan-genome* in FASTA format for all genes found in the dataset	folder: result
pre_filt_graph.gml	raw pan-genome graph	folder: result
struct_presence_absence.Rtab	binary matrix describing the presence/absence of genome rearrangement events	folder: result
summary_statistics.txt	summary statistics file	folder: result

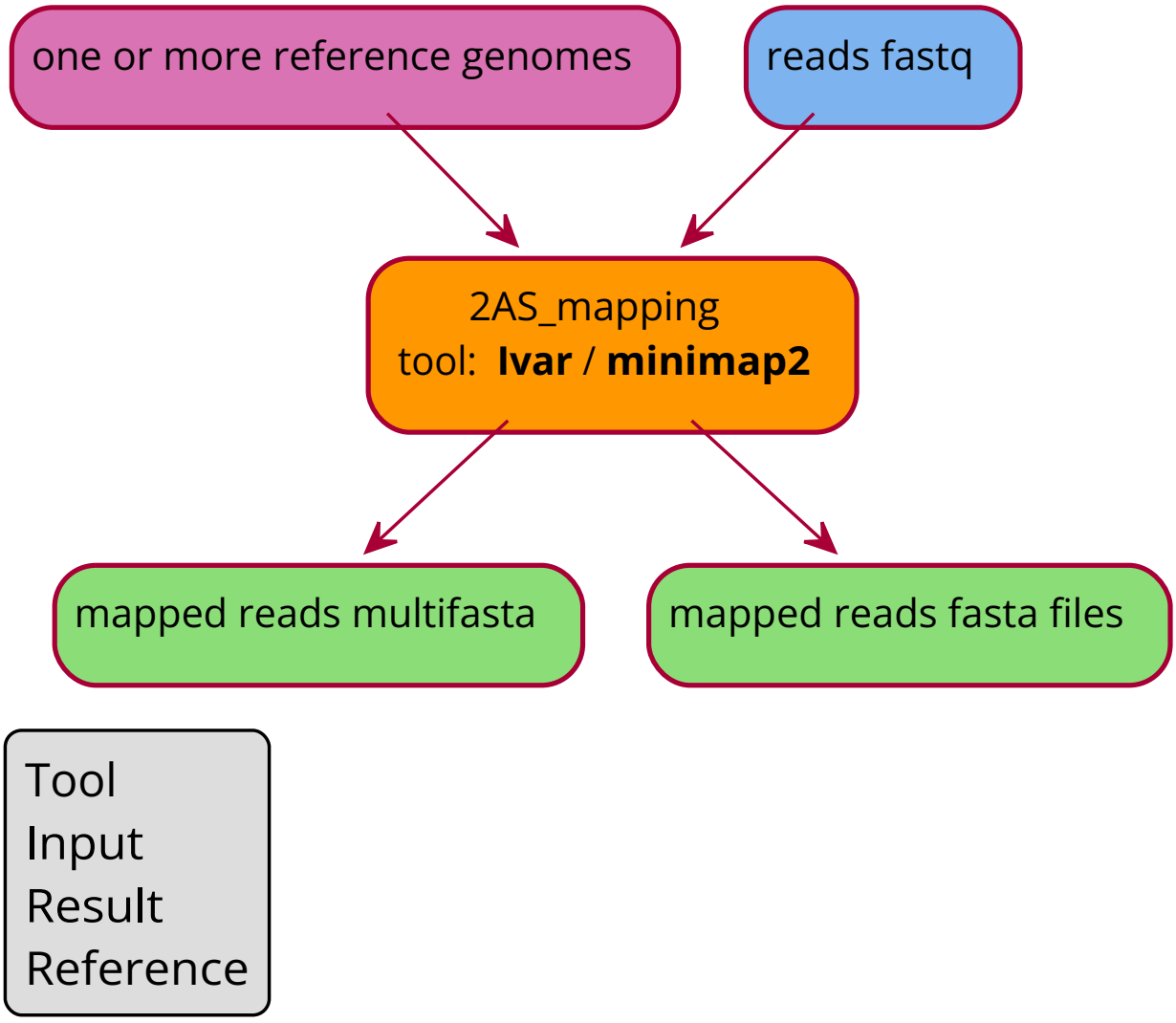
For graph visualization, Panaroo's authors recommend using Cytoscape. Further information on visualizing the pan-genome graph is available in the Panaroo official documentation section.

5.4.5 Pipelines

Mapping for Segmented Viruses Pipeline

Introduction

The **Mapping for Segmented Viruses** pipeline allows the mapping of genome fragments of segmented viruses by comparing them against multiple references. This allows each segment to be mapped to the most appropriate reference. In addition to the individual alignments, the pipeline also generates the complete multi-FASTA file.



Run Mapping for Segmented Viruses pipeline

The filter at the top of the run analyses page allows to display only the pipelines. Once the **Mapping for Segmented Viruses** pipeline has been selected, a confirmation interface will be displayed.

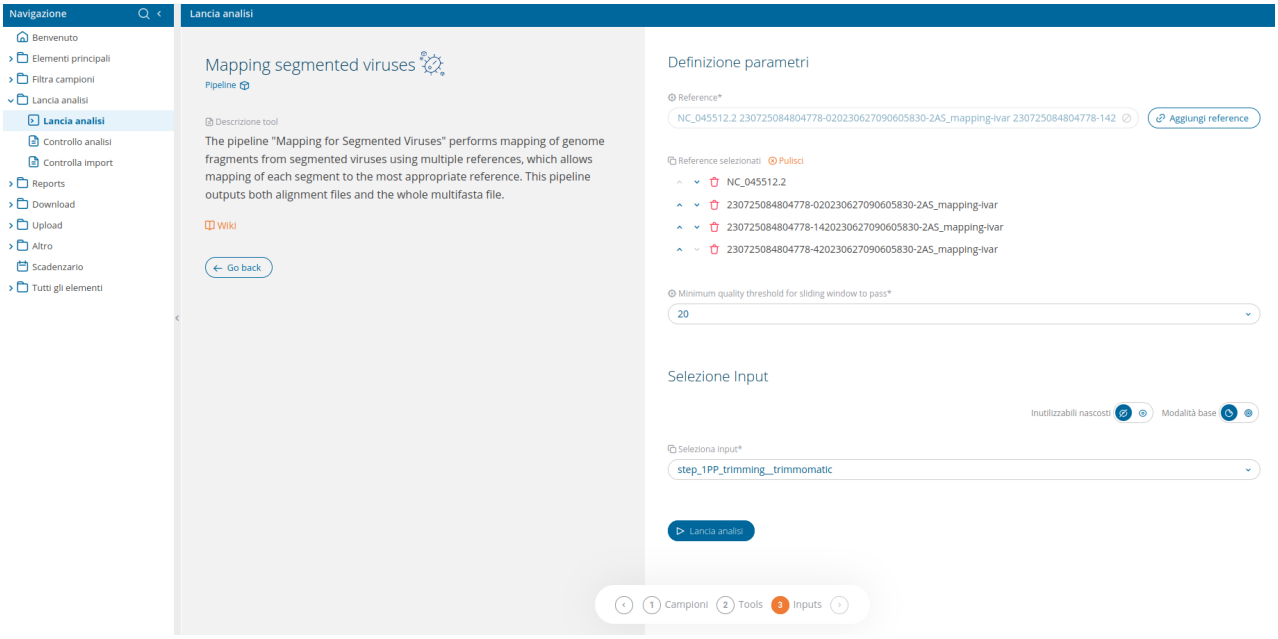
The **Mapping for Segmented Viruses** pipeline performs the mapping for all genome fragments, coming from the segmented viruses provided by the user, using multiple references. The tools used will be:

- **Ivar** (2AS_mapping__ivar) in case of Illumina short reads or IonTorrent;
- **minimap2** (2AS_mapping__minimap2) in case of Nanopore long reads.

Accepted inputs can originate from pre-processing analyses:

- depleted reads (step_1PP_hostdepl)
- downsampled reads (step_1PP_downsampling)
- trimmed reads (step_1PP_trimming)
- filtered reads (step_1PP_filtering)

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.



Once the pipeline is launched, the page generates a link to the Check analyses section to allow a check of the processus status. The user will be notified by the system both when the pipeline is successfully launched and when the operation is completed.

Multiple References

The multiple references required by the pipeline must be provided by the user through the dedicated selection system, whose user guide is available in the Wiki dedicated section on running analyses.

Results

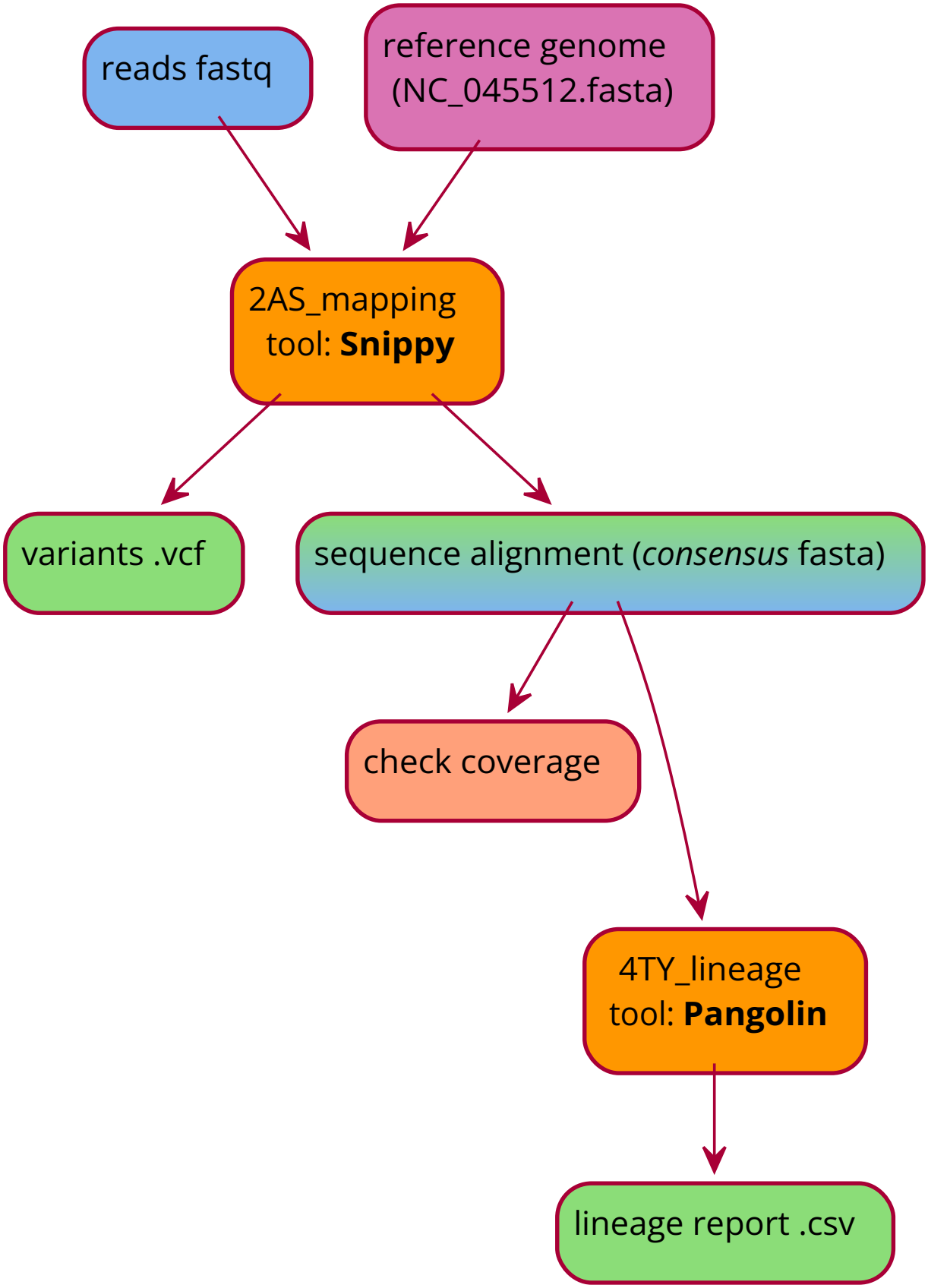
Files generated by the "Mapping for Segmented Viruses" pipeline will be the same as those produced by the individual analyses that compose it, organized according the same hierarchy. As a reference, please refer to the relevant Wiki section of the analysis 2AS_mapping. To these files, the multi-FASTA file of the mapped genome segments is added, whose information is reported in the following table:

File	Description	Location
`DSXXX-DTXXX_20XX.XX.XXXX.X_tool_aggregate.fasta`	multiFASTA with aggregated consensus in a single file	`DSXXX-DTXXX_tool > result`
`DSXXX-DTXXX_20XX.XX.XXXX.X_tool_reference.fasta`	consensus FASTA generated from the mapping to a single reference. Each individual mapping results are saved in a folder identified by an additional sequential number, added to the folder's base code	`DSXXX-DTXXX[sequential-number]_tool > result`
`DSXXX-DTXXX_20XX.XX.XXXX.X_tool_reference.bam` and `.bam.bai`	`bam` file and its corresponding index generated from the mapping to a single reference. The results are saved in the same way as the FASTA files	`DSXXX-DTXXX[sequential-number]_tool > result`
warnings.log	text file with warnings in case of an incorrect number of references	folder: results

Covid Emergency Pipeline

Introduction

The **Covid Emergency** pipeline allows a fast execution of assembly (step_2AS_mapping) with reference and the lineage assignment (4TY_lineage) of SARS-CoV2 samples. The reference genome used during the mapping step is always NC_045512 (Wuhan-Hu-1). Since the tool used for mapping is Snippy, the pipeline also produces the VCF file of variants.



Tool
Input
Result
Reference

Run Covid Emergency Pipeline

The filter at the top of the run analysis page allows to display only the pipelines. Once the **Covid Emergency** pipeline has been selected, a confirmation interface will be displayed.

The Covid Emergency pipeline consists of 2 steps:

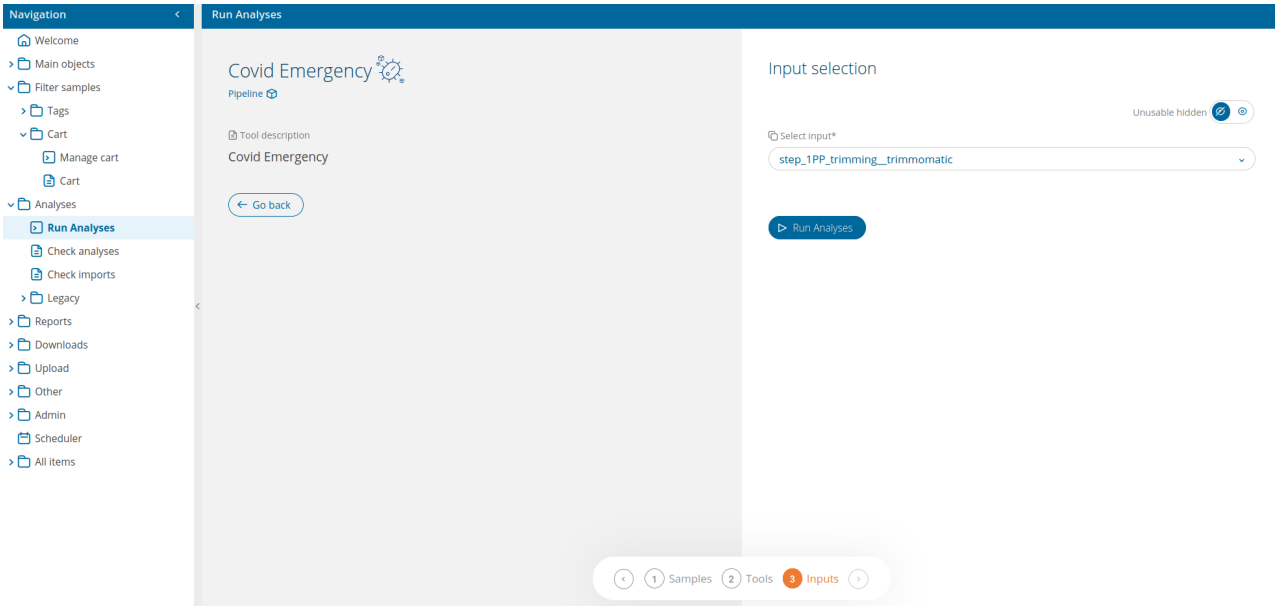
1. Mapping executed with **Snippy**;
2. Lineage assignment executed with **Pangolin**.

Therefore, usable inputs are the same as those available for a mapping with Snippy, *i.e.* reads in fastq format:

- step_1PP_hostdepl;
- step_1PP_downsampling;
- step_1PP_trimming.

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.

In the section dedicated to the parameter definition, in the quality threshold selection field, the user can specify a minimum value (20 o 14) for the pre-called bases.



Once the pipeline is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The user will be notified by the system both when the pipeline is successfully launched and when the operation is completed.

Results

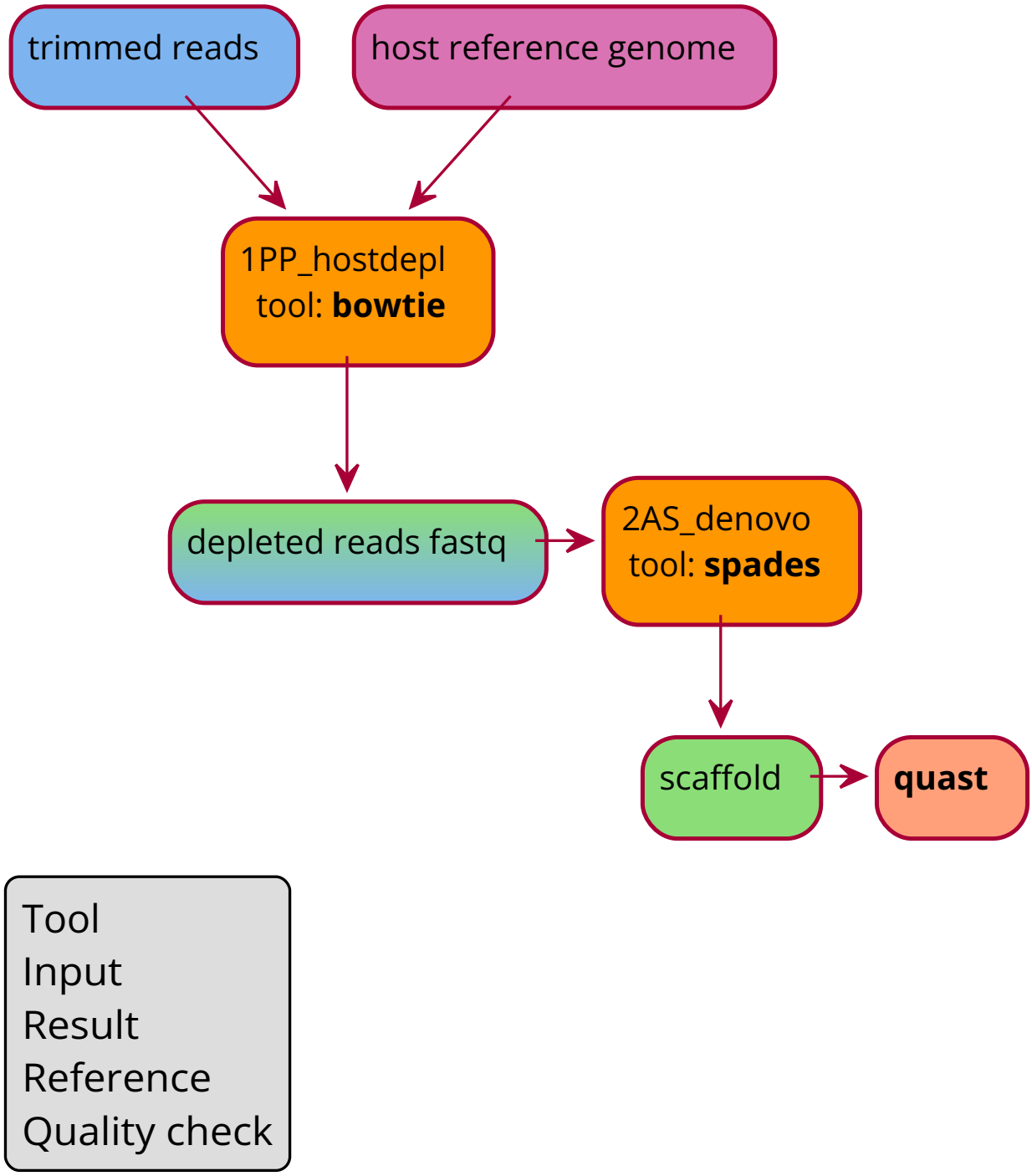
Files generated by the "Covid Emergency" pipeline will be the same as those produced by the individual analyses that compose it, organized according to the same hierarchy. As a reference, please refer to the relevant sections of the individual analyses:

- Output of mapping with Snippy;
- Output of lineage assignment with Pangolin.

Depletion & *de novo* Pipeline

Introduction

The **Depletion & *de novo*** pipeline performs consecutively the depletion (1PP_hostdepl) of host reads and the *de novo* assembly (2AS_denovo).



Run Depletion & *de novo* pipeline

The filter at the top of the run analysis page allows to display only the pipelines. Once the **Depletion & *de novo*** pipeline has been selected, a confirmation interface will be displayed.

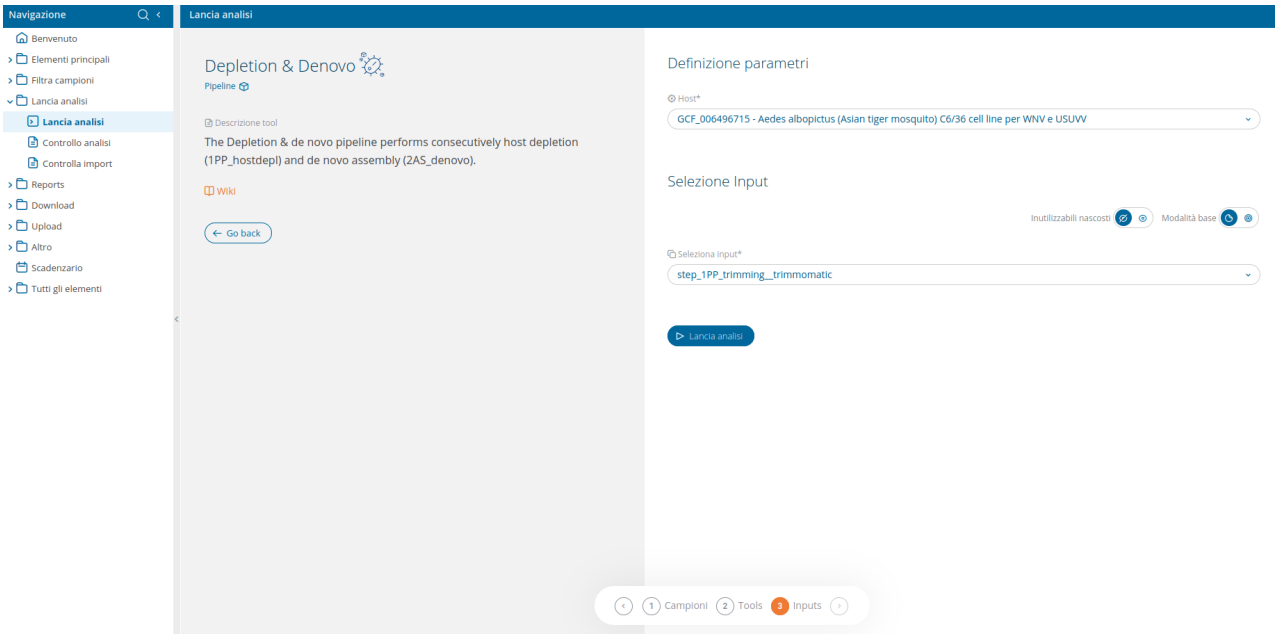
The pipeline performs the following 2 analyses:

1. Depletion of host sequences with **bowtie**;
2. *de novo* assembly with **Spades** and quality check with **Quast**.

Usable inputs are, therefore, the same as those available for host depletion with "bowtie", *i.e.* the host reference genome (*host* field at the top, in the parameter definition section) and the fastq reads:

- step_1PP_hostdepl;
- step_1PP_downsampling;
- step_1PP_trimming.

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.



Once the pipeline is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The user will be notified by the system both when the pipeline is successfully launched and when the operation is completed.

Results

Files generated by the "Depletion & *de novo*" pipeline will be the same as those produced by the individual analyses that compose it, organized according to the same hierarchy. As a reference, please refer to the relevant sections of the individual analyses:

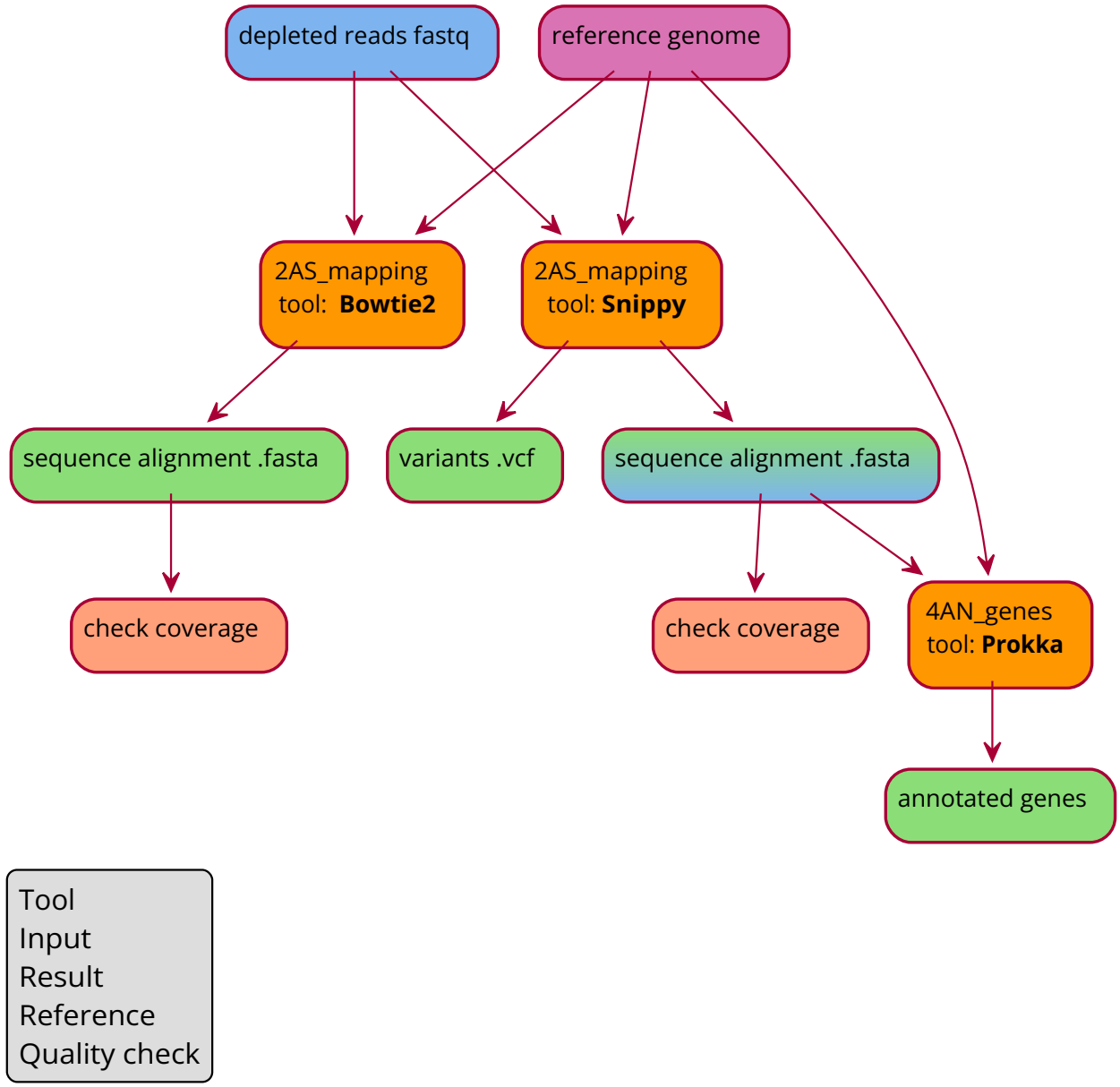
- Output of host depletion with bowtie;
- Output of *de novo* assembly with SPAdes and of quality check with Quast.

Genome Draft pipeline

Introduction

The **Genome Draft** pipeline is part of the **Vdraft** pipeline, also available separately. **Genome Draft** performs mapping of sequences (2AS_mapping) with both Bowtie2 and Snippy, followed by genome annotation (4AN_genes) using Prokka.

Genome annotation will only be executed on Snippy outputs and only if a file in genebank format (.gb or .gbk) has been used as reference; erence used, as a genebank, (.gb or .gbk) file format; in turn, Snippy can be executed only if the selected sample has the 4AN_genes analysis associated (Prokka).



Run Genome Draft pipeline

The filter at the top of the run analysis page allows to display only the pipelines. Once the **Genome Draft** pipeline has been selected, a confirmation interface will be displayed.

The **Genome draft** pipeline includes the following analyses:

1. Mapping with **Bowtie2** (2AS_mapping__bowtie);
2. Mapping with **Snippy** (2AS_mapping__snippy);
3. Genome annotation with **Prokka** (4AN_genes__prokka).

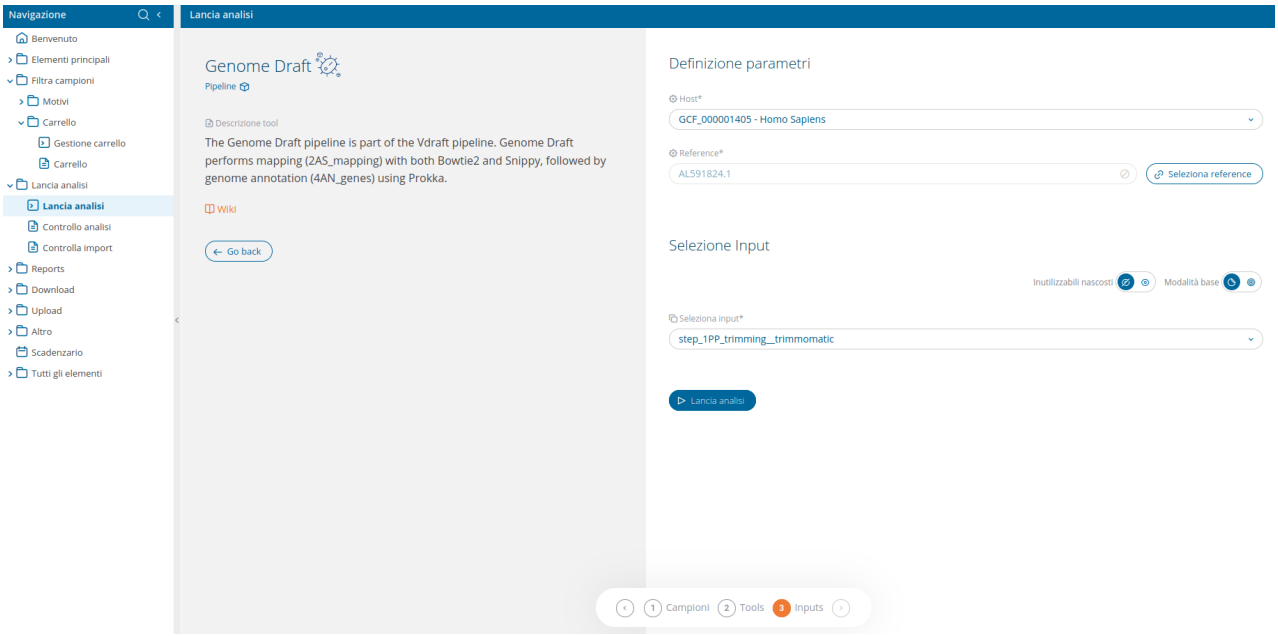
Both mapping runs also include the calculation of coverage.

The usable inputs are the outputs of pre-processing analyses:

- depleted reads (step_1PP_hostdepl);
- downsampled reads (step_1PP_downsampling);
- trimmed reads (step_1PP_trimming);
- filtered reads (step_1PP_filtering).

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.

It will also be necessary to specify the genome of the host organism and a reference genome from those available, which will be used both for read mapping and for the functional annotation.



Once the pipeline is launched, the page generates a link to the Check analyses section to allow a check of the processus status. The user will be notified by the system both when the pipeline is successfully launched and when the operation is completed.

Results

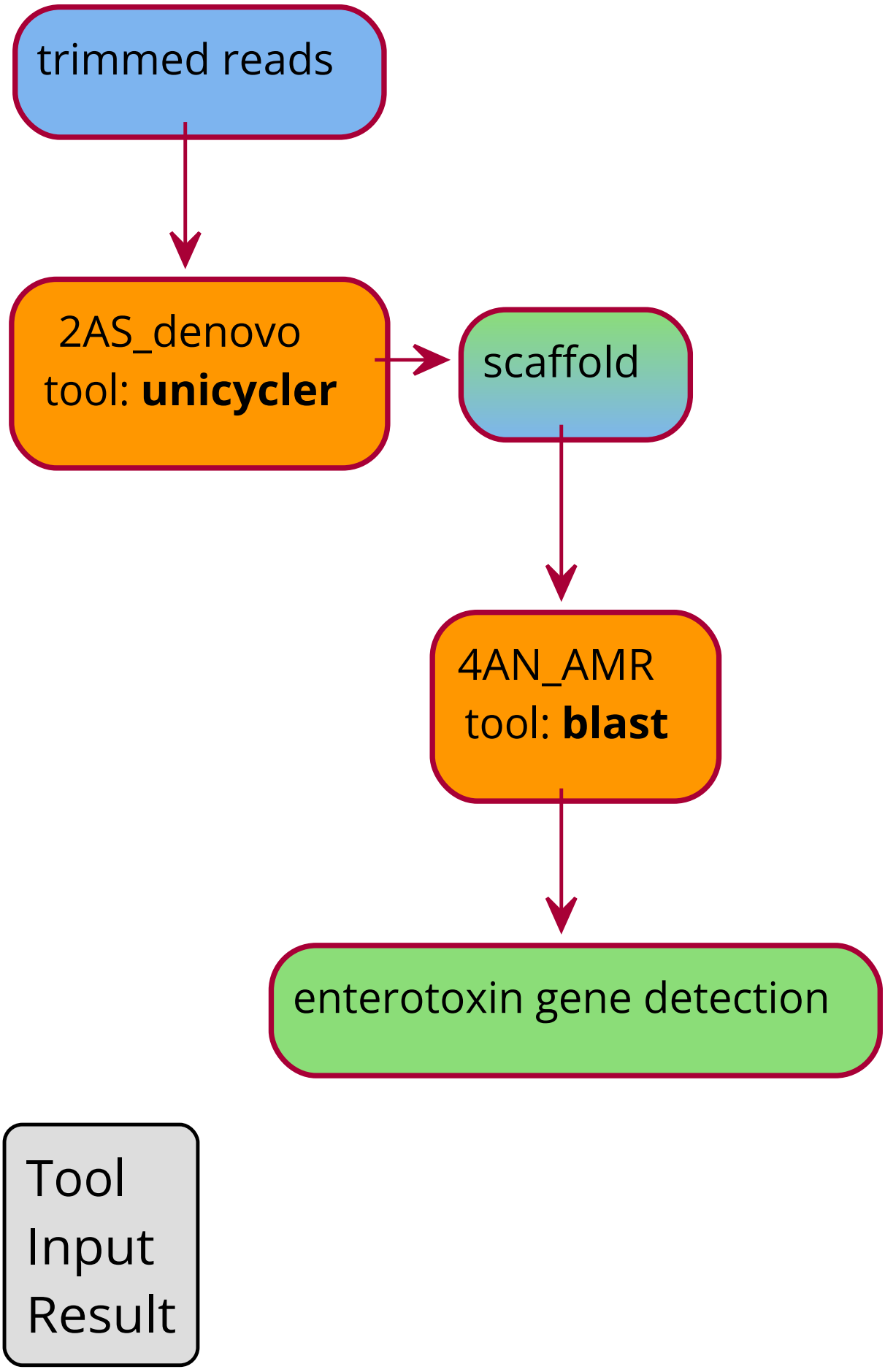
Files generated by the "Genome draft" pipeline will be the same as those produced by the individual analyses that compose it, organized according the same hierarchy. As a reference, please refer to the relevant sections of the individual analyses:

- Output of mapping;
- Output of genome annotation.

Enterotoxin *S. aureus* Finder Pipeline

Introduction

The **Enterotoxin *Staphylococcus aureus* Finder** pipeline allows to detect the presence of the gene encoding enterotoxin in the genome of *S. aureus*, running the *de novo* assembly (2AS_denovo) first, followed by typing 4AN_AMR.


trimmed reads
2AS_denovo
tool: unicycler
scaffold
4AN_AMR
tool: blast
enterotoxin gene detection
Tool
Input
Result

Run Enterotoxin *S. aureus* Finder pipeline

The filter at the top of the run analysis page allows to display only the pipelines. Once the **Enterotoxin *S. aureus* Finder** pipeline has been selected, a confirmation interface will be displayed.

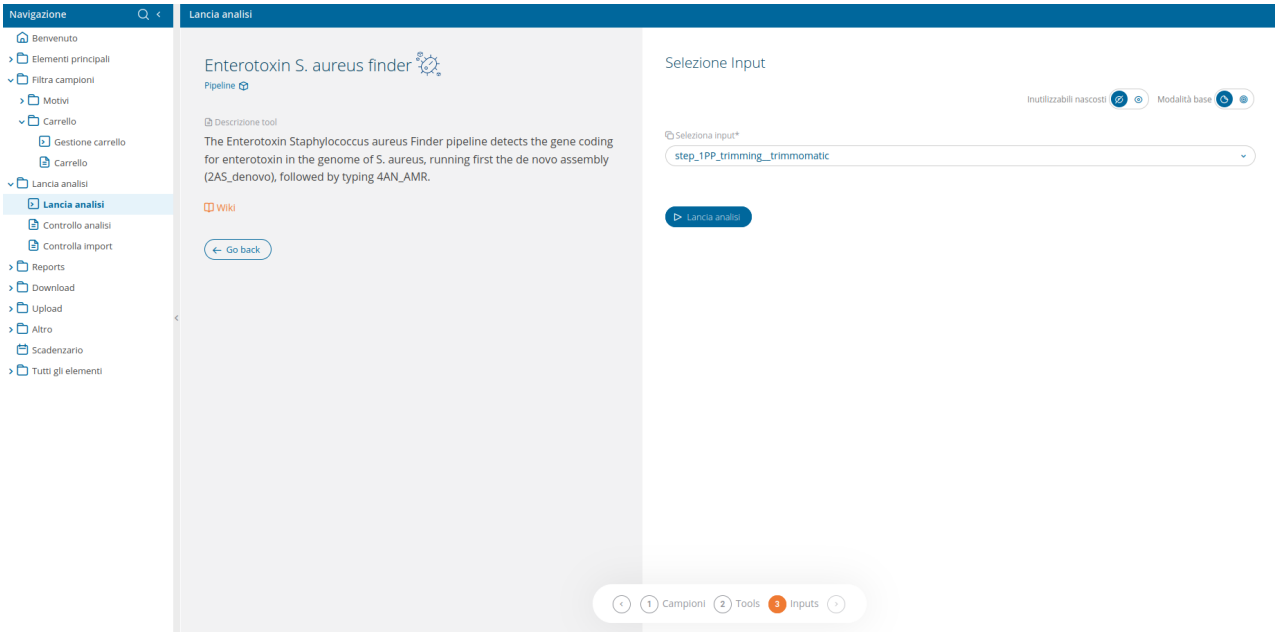
The **Enterotoxin *S. aureus* Finder** pipeline performs:

1. *de novo* assembly with **unicycler**;
2. detection of the presence of the gene encoding enterotoxin with **blast**.

The usable inputs are the trimmed reads obtained from a pre-processing analysis:

- depleted reads (step_1PP_hostdepl);
- downsampled reads (step_1PP_downsampling);
- trimmed reads (step_1PP_trimming);
- filtered reads (step_1PP_filtering).

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.



Once the pipeline is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The user will be notified by the system both when the pipeline is successfully launched and when the operation is completed.

Results

Files generated by the "Enterotoxin *S. aureus* Finder" pipeline will be the same as those produced by the individual analyses that compose it, organized according the same hierarchy. As a reference, please refer to the relevant sections of the individual analyses:

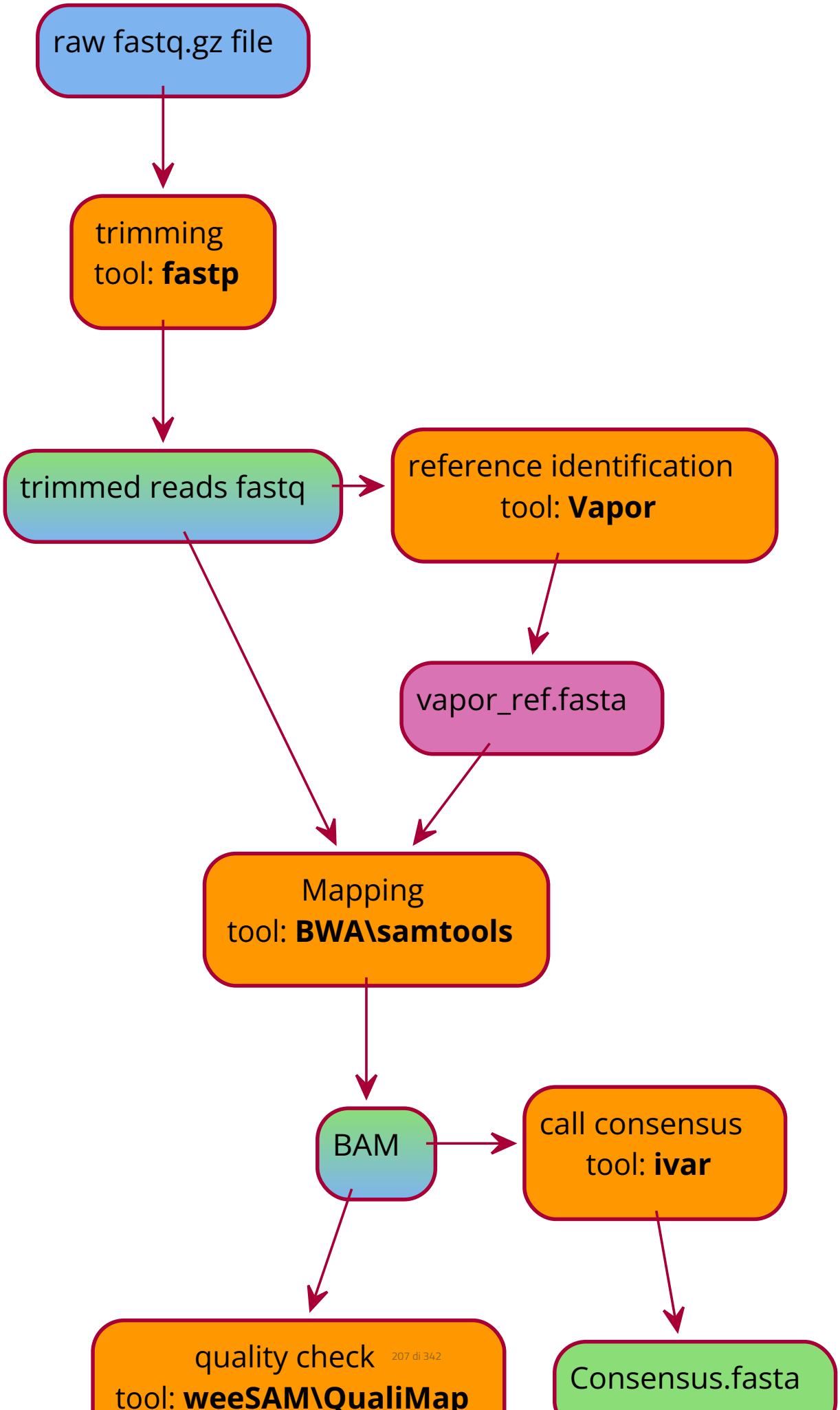
- Output of *de novo* assembly with unicycler;
- Output of the detection of the presence of the gene encoding enterotoxin.

LSDV Pipeline - IAEA

Introduction

This pipeline analyzes LSDV samples, automatically identifying the reference and reconstructing the consensus sequence.

- **Official tool page:** To be added
- **Authors:** Graham Freimanis and Richard Orton


raw fastq.gz file
trimming
tool: **fastp**
trimmed reads fastq
reference identification
tool: **Vapor**
vapor_ref.fasta
Mapping
tool: **BWA\samtools**
BAM
call consensus
tool: **ivar**
quality check
tool: **weeSAM\QualiMap**
Consensus.fasta

Run LSDV pipeline

Once the **IAEA LSDV** pipeline has been selected in the Run Analysis page, a confirmation interface will allow proceeding to the selection of parameters.

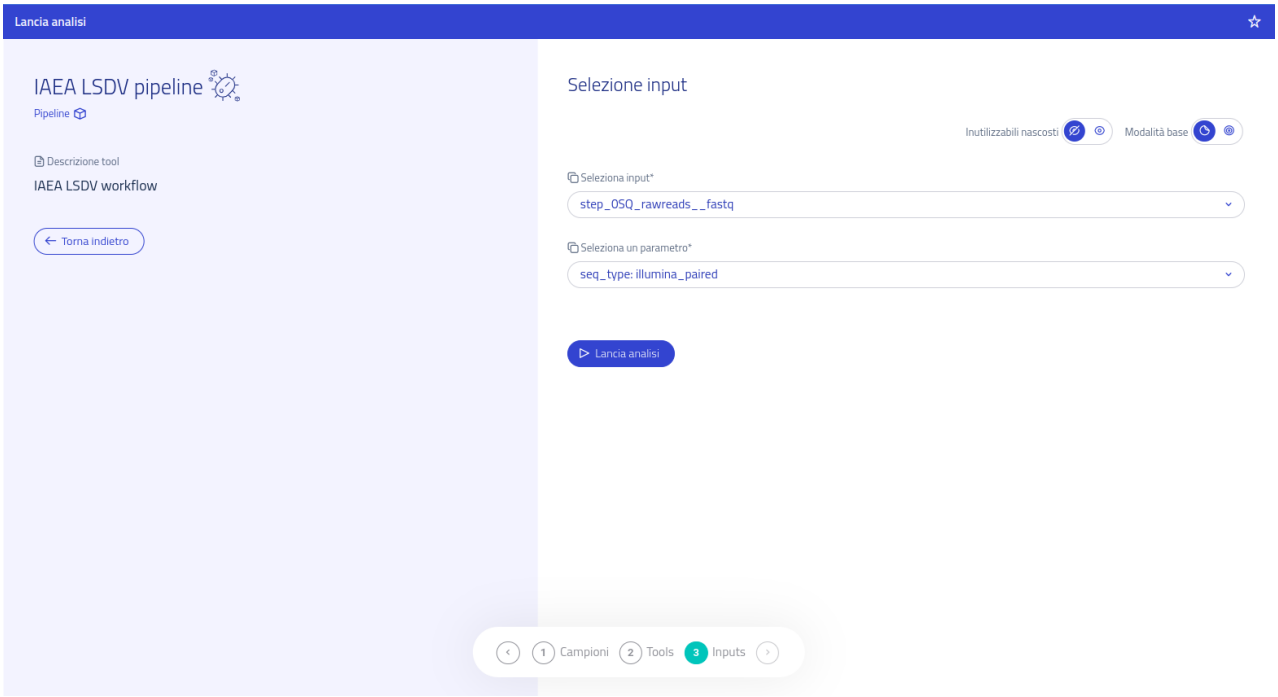
The **IAEA LSDV pipeline** pipeline consists of 3 main steps:

1. Read trimming;
2. Identification of the best reference;
3. Mapping.

The required input are raw fastq reads (`.fastq.gz`):

- **step_0SQ_rawreads**

Screenshot of input selection



Once the pipeline is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The user will be notified by the system both when the pipeline is successfully launched and when the operation is completed.

Result folder

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are 2 subfolders:

- **meta**: (*metadata*), where the log and configuration files of the executed process are saved;
- **result**: where the files containing the results produced by the analysis are saved.

All the results produced by the pipeline tools can be found in the `result > LSDV-PIPELINE-RESULTS` folder.

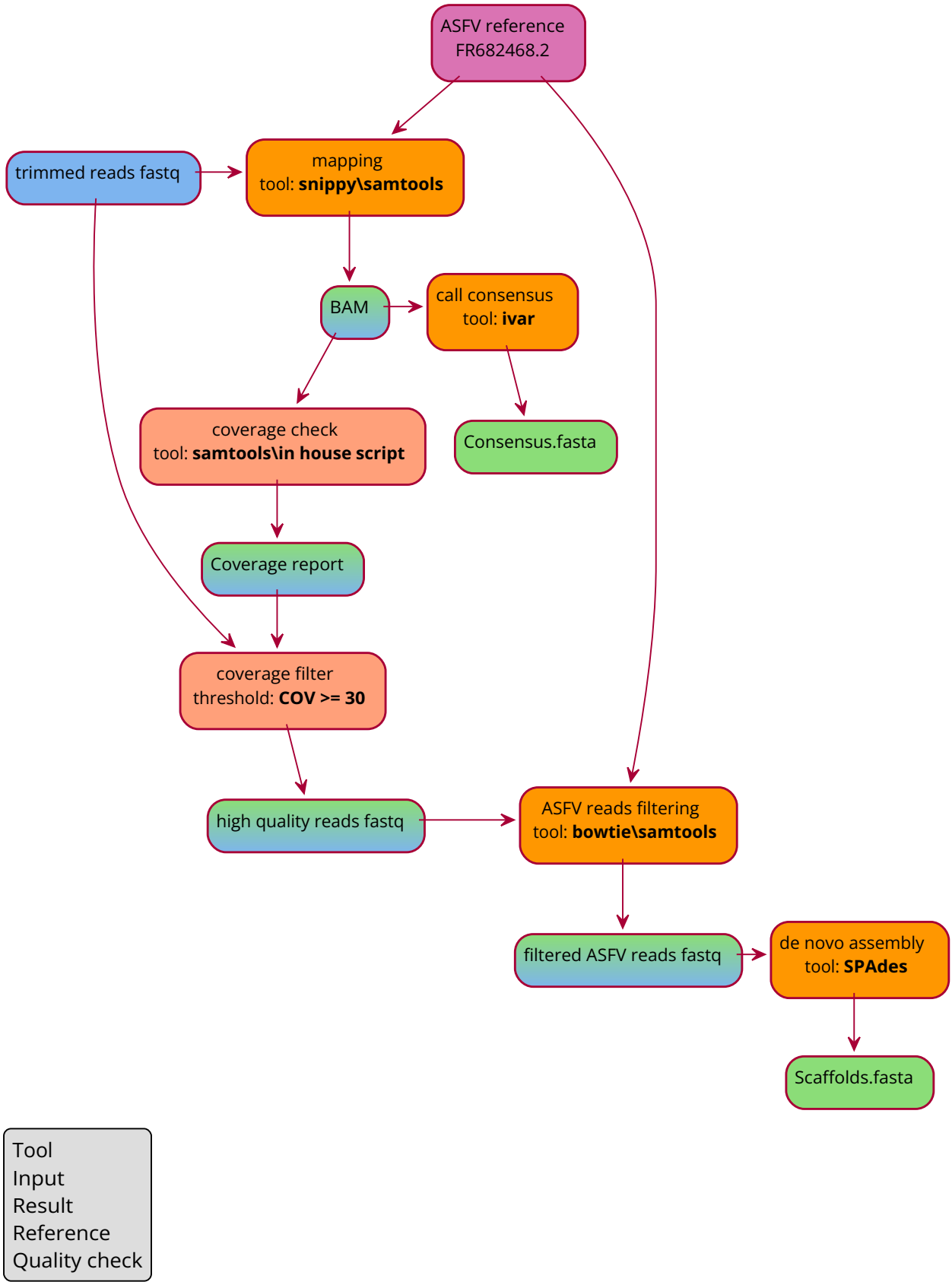
Data visualization

The pipeline generates an .html webpage that can be opened in the browser.

ASFV Pipeline

Introduction

This pipeline processes *African Swine Fever Virus* (ASFV) samples by mapping reads to the FR682468.2 reference genome and calculating sample coverage. Only samples with sufficient coverage are retained, filtered against the ASFV reference, and de novo assembled to obtain the final sequences.



How to run the pipeline

Once the **ASFV Pipeline** has been selected in the Run Analysis page, a confirmation interface will allow proceeding to the selection of parameters.

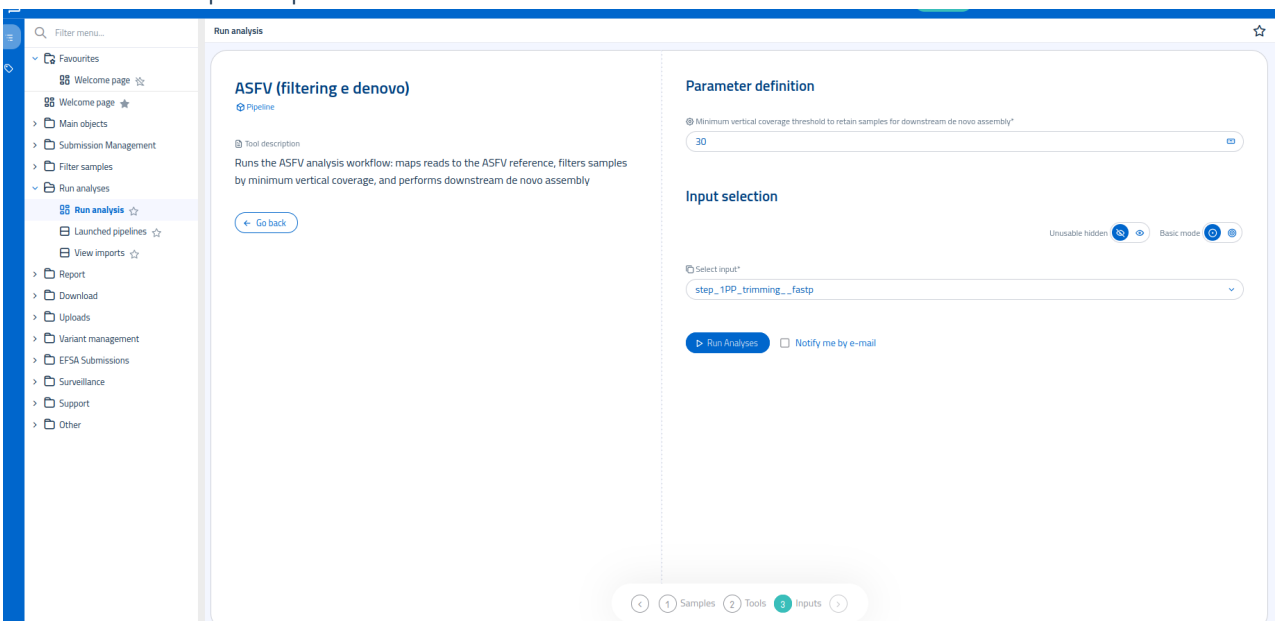
The **ASFV Pipeline** consists of 3 main steps:

1. Host-read depletion;
2. Identification of the best reference sequence;
3. Mapping and consensus generation.

From the drop-down menu, you can select the host to be used for depletion.

Next, you can select the input, which for this analysis is *trimmed reads*.

Screenshot of the input and parameter selection



Once the pipeline is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The user will be notified by the system both when the pipeline is successfully launched and when the operation is completed.

Result folder

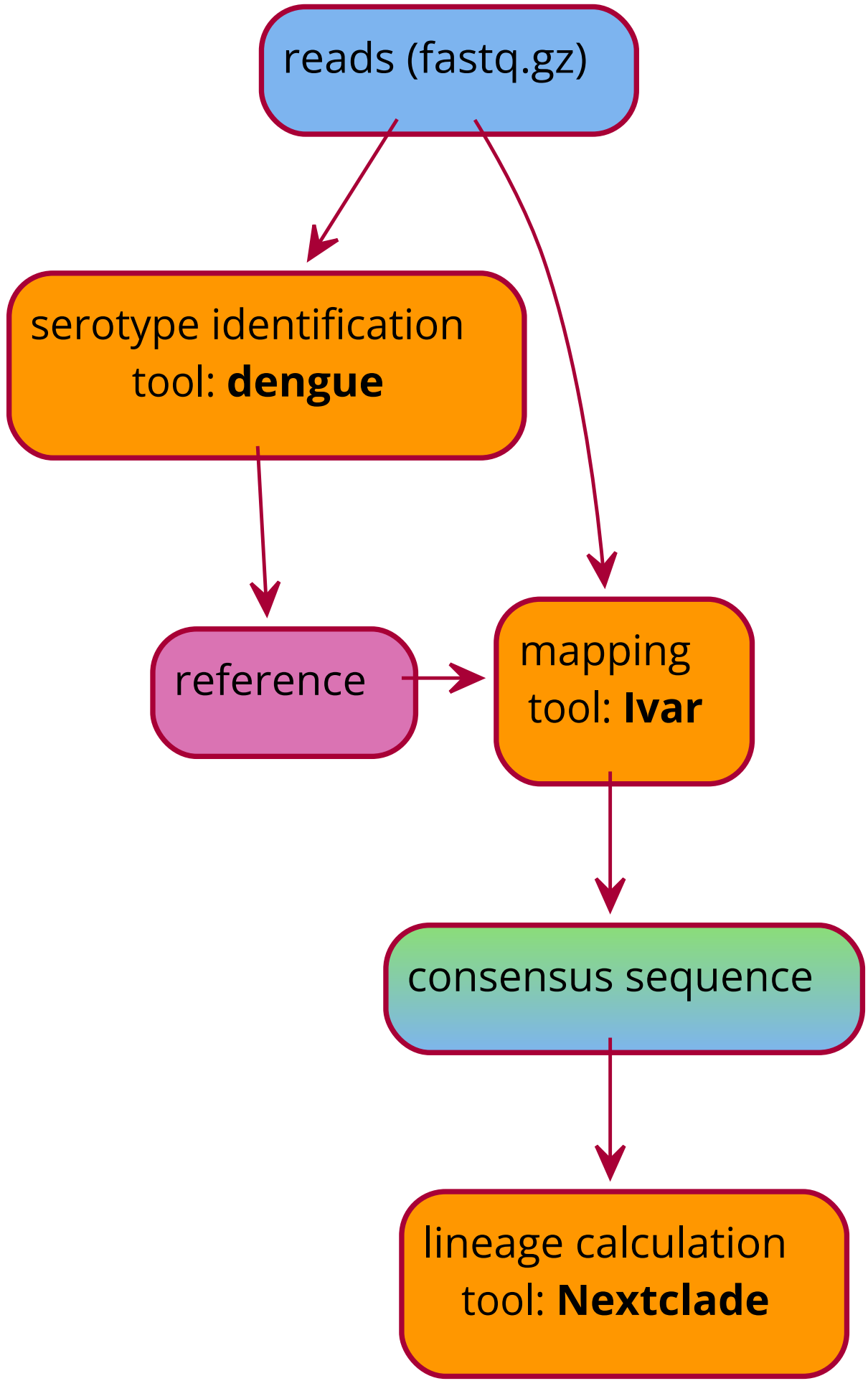
The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are 2 subfolders:

- **meta**: (*metadata*), where the log and configuration files of the executed process are saved;
- **result**: where the files containing the results produced by the analysis are saved.

DENV Pipeline - serotype calculation and mapping

Introduction

This pipeline calculates the serotype for **Dengue** virus samples, builds the consensus sequence based on mapping to the appropriate reference for the identified serotype and finally determines the lineage.


reads (fastq.gz)
serotype identification
tool: dengue
reference
mapping
tool: Ivar
consensus sequence
lineage calculation
tool: Nextclade

Run DENV pipeline

Once the **DENV pipeline** analysis is selected on the run analysis page, the system will switch to a confirmation interface that allows you to proceed with the parameter selection.

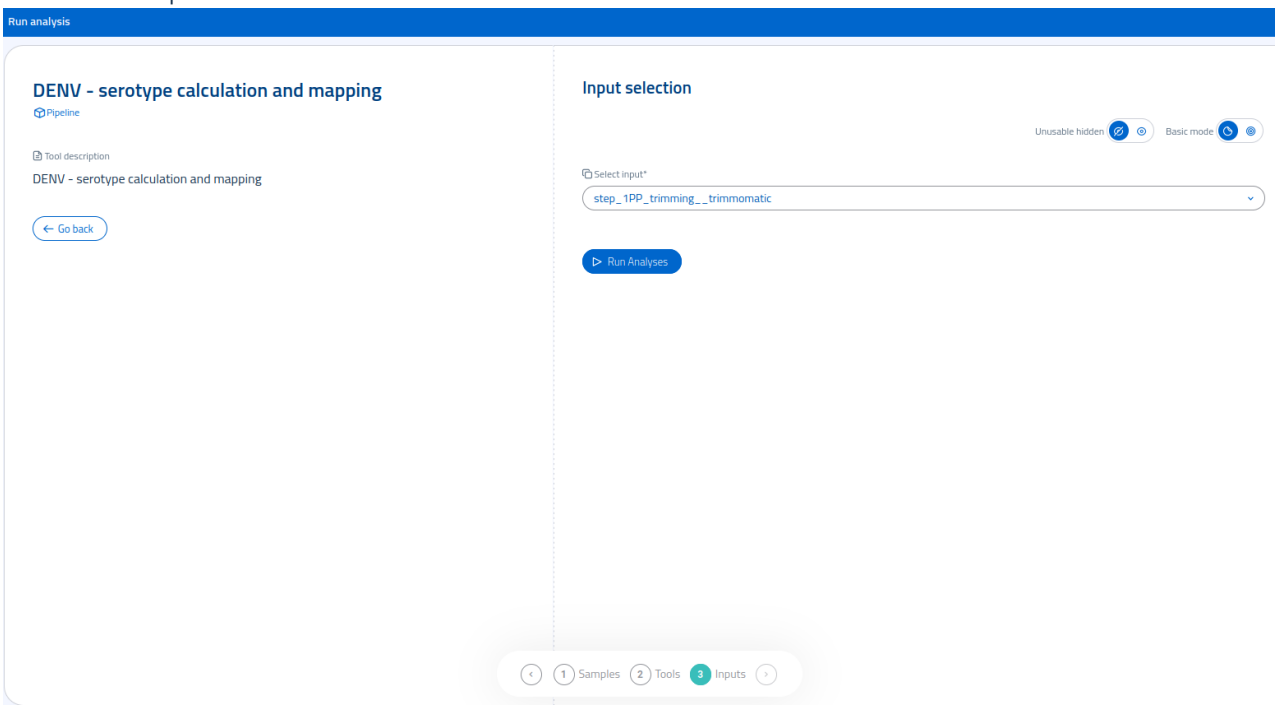
The **DENV pipeline** involves 3 main steps:

1. serotype identification;
2. mapping to the appropriate reference;
3. lineage calculation.

The required inputs are FASTQ raw reads (`.fastq.gz`):

- **step_0SQ_rawreads**
- step_1PP_trimming

Screenshot of input selection



Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The user will be notified by the system both when the pipeline is successfully launched and when the operation is completed.

Result folder

The **Result folder** can be accessed by clicking the link in the analysis card, in the *Result data* section. Within the following `results` folder, it is possible to find the files produced by the pipeline, which will be the same as those produced by the analyses that compose it:

- Serotype identification;
- Mapping with Ivar;
- Lineage calculation with Nextclade.

Pipeline Mapping *Brucella*

The pipeline **Mapping *Brucella*** performs mapping for *Brucella* samples against a fixed reference: *Brucella melitensis* bv.1 str. 16M" `GCF_000740415.1` .

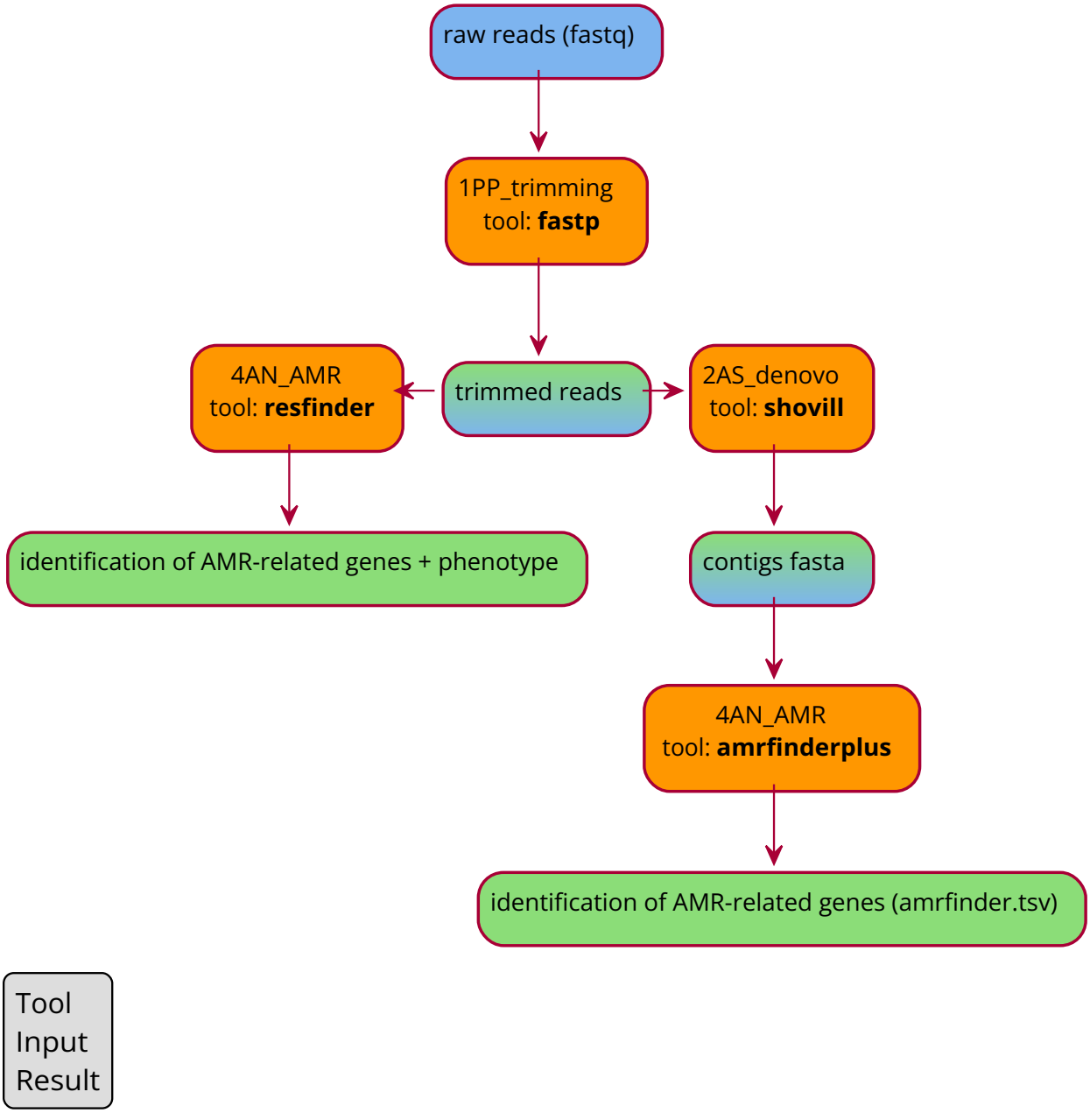
For information regarding the mapping, please refer to the page "**2AS_mapping**".

Modulo AMR

Introduction

The **AMR Module** identifies genes associated with antibiotic resistance (AMR) using both **ResFinder** (on reads) and **AMRfinderPlus** (on FASTA contigs from *de novo* assembly).

The raw reads are trimmed with **fastp** before running ResFinder and assembled with **shovill** before running AMRfinderPlus.



Run AMR Module

Once the *AMR Module* has been selected, it will be possible to choose whether to run AMRFinderPlus also for additional genes, beyond those associated with AMR, such as virulence genes or stress response genes. The module performs AMR gene identification by running the following software:

- **AMRFinderPlus** - AMRFinderPlus identifies antimicrobial resistance genes, acquired in bacterial protein sequences and/or in assembled nucleotide sequences, as well as known point mutations, associated with resistance, across different taxa.

  AMRFinderPlus GitHub page: https://github.com/ncbi/amr/wiki/Running-AMRFinderPlus#usage

- **ResFinder** - Identification of acquired antimicrobial resistance genes.

  ResFinder GitHub page: https://github.com/cadms/resfinder.

The required inputs are sequence files in FASTQ format, both raw and processed:

- step_0SQ_rawreads;
- step_1PP_trimming;
- step_1PP_hostdepl;
- step_1PP_downsampling;
- step_1PP_generated;
- step_1PP_filtering.

As prerequisites for running ResFinder and AMRfinderPlus, the module also performs the trimming (if untrimmed reads are provided) and the *de novo* assembly, using the software listed below:

- **fastp** - All-in-one pre-processing for FASTQ files.

  Fastp GitHub page: https://github.com/OpenGene/fastp

- **shovill** - pipeline for assembling genomes of bacterial isolates from Illumina paired-end reads.

  Shovill GitHub page: https://github.com/tseemann/shovill

Once the pipeline is launched, the page generates a link to the Check analyses section to allow a check of the processus status. The user will be notified by the system both when the pipeline is successfully launched and when the operation is completed.

Results

As a reference for the results, please refer to the sections dedicated to the individual analyses that compose the pipeline "AMR Module":

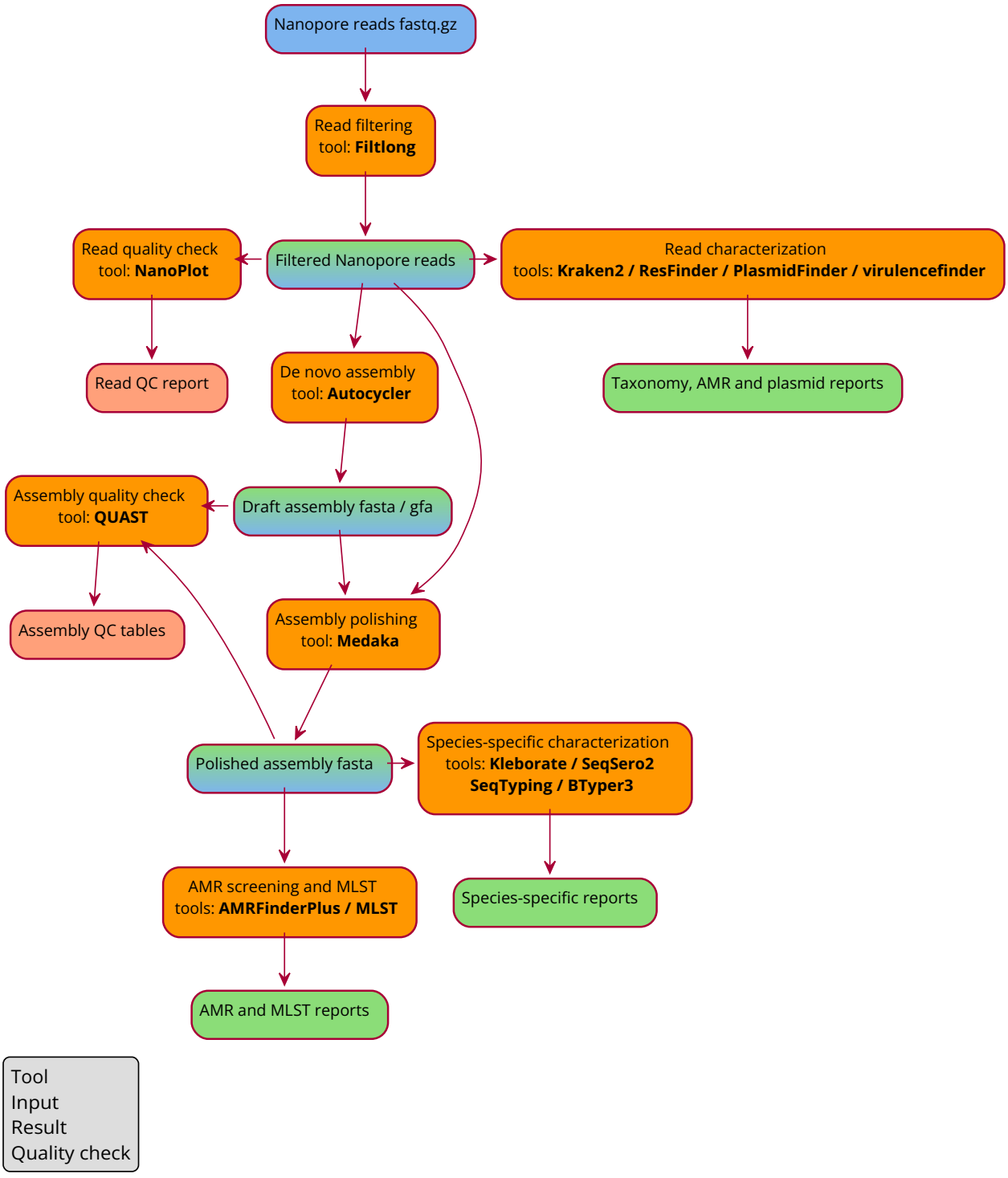
- Output of trimming performed with fastp.
- Output of AMR identification with ResFinder
- Output of assembly with Shovill
- Output of AMR identification with AMRfinderPlus.

Nanopore Pipeline

Introduction

This pipeline analyzes bacterial whole-genome sequencing (WGS) data generated with Oxford Nanopore technology. It filters long reads, reconstructs a *de novo* assembly, and polishes the assembled sequence. It also performs taxonomic classification, antimicrobial resistance (AMR) screening, plasmid detection, and typing according to the sample's genus and species.

Assembly and polishing use the filtered Nanopore reads.



AMRFinderPlus and the typing tools run according to the supported taxa. For *Klebsiella*, MLST is performed with Kleborate.

How to run the pipeline

Once the **Nanopore Pipeline** has been selected in the Run Analysis page, a confirmation interface will allow proceeding to input and parameter selection.

The **Nanopore Pipeline** consists of 5 main steps:

1. Long-read filtering with **Filtlong** and read quality assessment with **NanoPlot**;
2. Taxonomic classification with **Kraken2**, AMR screening with **ResFinder**, and plasmid detection with **PlasmidFinder** on the filtered reads;
3. *De novo* assembly with **Autocycler** and assembly quality assessment with **QUAST**;
4. Assembly polishing with **Medaka**, using the complete filtered read set, followed by **QUAST**;
5. AMR screening, MLST, and species-specific characterization of the polished assembly.

Inputs

The required input is **one basecalled Nanopore FASTQ file per sample** (`.fastq.gz`), for example from **step_0SQ_rawreads**.

Main parameters

The following table lists the main settings and their default values:

Tool	Parameter	Default	Description
Filtlong	Minimum read length	`1000` bp	Minimum length of retained reads.
Filtlong	Retained percentage	`90`	Keeps the best reads accounting for 90% of the input bases, subject to the other filters.
Filtlong	Target bases	Unset	Optional limit on the number of bases retained.
Filtlong	Trimming / splitting	Disabled / `0`	Trims or splits reads; `0` disables splitting.
Autocycler	Genome size	Automatic	Uses the Qualibact genome-size thresholds for the supplied taxon. An explicit size, such as `5000000` or `5m`, can be supplied.
Autocycler	Assemblers	`flye raven miniasm myloasm metamdbg`	Assemblers used to reconstruct the genome from read subsets.
Autocycler	Read type	`ont_r10`	Nanopore read type; supported values are `ont_r9` and `ont_r10`.
Autocycler	Subsample count	`4`	Number of read subsets used for assembly; must be at least `2`.
Medaka	Model	Automatic	Inferred from the basecaller metadata in the FASTQ headers; a compatible model can also be specified explicitly.
Medaka	Bacterial model preference	Enabled	Selects the bacterial methylation model when compatible with the basecaller.
Medaka	Batch size	`100`	Batch size used during consensus generation.

Species-specific analyses

Typing is performed on the polished assembly. The following analyses are selected automatically:

Sample taxon	Tool	Analysis
Klebsiella	**Kleborate**	Species assignment, followed by AMR, virulence, and MLST analyses using the supported species-complex modules.
Other bacteria	**mlst**	MLST scheme and sequence type identification.
Salmonella enterica	**SeqSero2**	Serotype prediction.
Escherichia coli	**SeqTyping**	Serotype prediction.
Bacillus cereus	**BTyper3**	Taxonomic and typing characterization.

Once the pipeline is launched, the page generates a link to the Check analyses section, where the execution status can be monitored. The system notifies the user when the pipeline is successfully launched and when execution is completed.

Result folder

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section.

Results are organized by sample and analysis step under `results > YEAR > CMP > STEP > DS-DATE_METHOD`, where `CMP` identifies the sample, `DS` identifies the sequencing dataset, `DATE` is the execution timestamp, and `METHOD` identifies the tool. Each analysis folder contains:

- **meta**: metadata, parameters, configuration files, and execution logs;
- **result**: files produced by the analysis;
- **qc**, where available: quality-control results in **qc/result** and their logs in **qc/meta**.

The main outputs are listed below. Folder paths are relative to `results > YEAR > CMP`; `*` represents the dataset and execution identifiers.

Analysis	Folder	Main outputs
Read filtering and QC	`1PP_trimming/*_filtlong`	Filtered reads in `result/*_filtlong.fastq.gz`; NanoPlot HTML report and NanoStats text file in `qc/result`.
De novo assembly and QC	`2AS_denovo/*_autocycler`	Draft assembly (`.fasta`) and assembly graph (`.gfa`) in `result`; Autocycler summary (`.tsv`) in `meta`; QUAST table in `qc/result`.
Assembly polishing and QC	`2AS_mapping/*_medaka_polishing`	Polished assembly in `result/*_medaka_polishing_denovo.fasta`; selected model in `meta/*_model.txt`; QUAST table in `qc/result`.
Taxonomic classification	`3TX_class/*_kraken2`	Kraken2 report (`*_kraken.tsv`) and Krona table and HTML visualization in `result`.
AMR screening on reads	`4AN_AMR/*_resfinder`	Resistance reports (`.txt`), sequence files (`.fsa`), and structured results (`.json`) in `result`.
Plasmid detection	`4AN_AMR/*_plasmidfinder`	Summary (`*_results.tsv`), alignments, and sequence files in `result`.
AMR screening on assembly	`4AN_AMR/*_amrfinderplus`	AMRFinderPlus report (`.tsv`) in `result`, for supported taxa.
MLST for other bacteria	`4TY_MLST/*_mlst`	MLST reports (`.csv` and `.tsv`) in `result`.
Klebsiella characterization	`3TX_species/*_kleborate`, `4AN_AMR/*_kleborate`, and `4TY_MLST/*_kleborate`	Species assignment, characterization reports (`.tsv`), and MLST results (`.csv` and `.tsv`) in the corresponding `result` folders.
Salmonella enterica serotyping	`4TY_serotype/*_seqsero`	Serotype report (`*_result.tsv`) and extracted antigen alleles (`.fasta`) in `result`.
Escherichia coli serotyping	`4TY_serotype/*_seq_typing`	Serotype reports (`*.report_types.tab` and `*.report.txt`) in `result`.
Bacillus cereus characterization	`4TY_serotype/*_btyper3`	BTyper3 final results in `result`.

The **polished Medaka FASTA** is the final assembly used for downstream characterization. The Autocycler draft assembly is also retained.

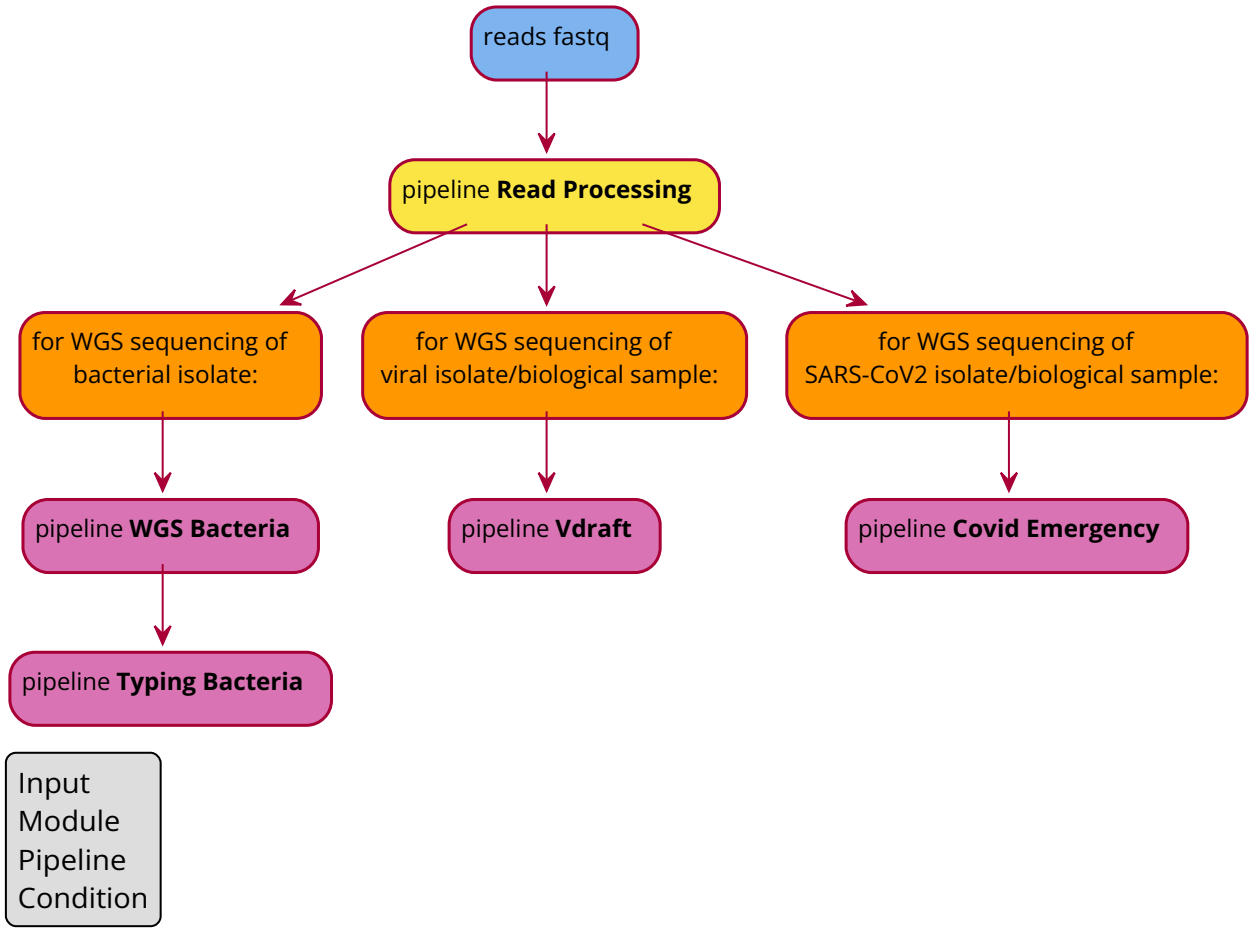
Data visualization

The **NanoPlot HTML report** can be opened in a browser to inspect filtered-read quality and length distributions. The **Krona HTML report** provides an interactive view of the taxonomic classification. **QUAST** tables summarize assembly quality before and after polishing.

NgsManager Pipeline

Introduction

The **NgsManager pipeline** allows the managemnet of reads and the execution of specific pipelines based on the sample type.



The pipelines executed by NgsManager are the following:

- Raw Read Processing pipeline - the results of 3TX_class__kraken in the Read Processing pipeline determine species identification,on the basis of which the following analyses are performed:
- For bacteria:
  WGS on Bacteria pipeline (execution of *de novo* assembly with Shovill: step_2AS_denovo__shovill);
  Typing on Bacteria pipeline.
- For viruses:
  Vdraft pipeline (execution of *de novo* assembly with Abricate: step_2AS_denovo__abricate + Genome Draft module).
- For SARS-CoV2:
  Covid Emergency pipeline.

(For information on the individual analyses, please refer to the corresponding pages, accessible by clicking on the items listed above).

Run NgsManager Pipeline

The filter at the top of the run analyses page allows to display only the pipelines. Once the **NgsManager** pipeline has been selected, a confirmation interface will be displayed.

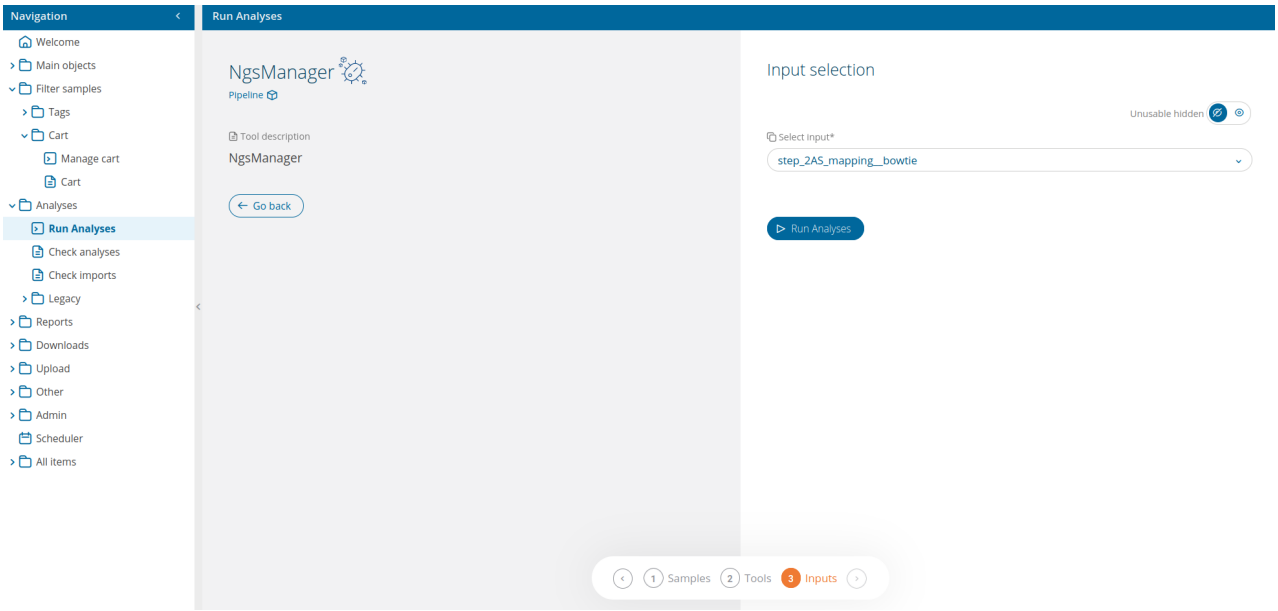
The selection of NgsManager pipelines, in the case of a manual launch, is done by choosing the sample type, among those available, within the "Parameter definition" section.

Regarding the input, NgsManager accepts:

- raw reads (step_0SQ_rawreads);
- step_1PP_trimming;
- step_1PP_hostdepl;
- step_1PP_generated;
- step_1PP_filtering;
- step_1PP_downsampling;
- step_2AS_mapping;
- step_2AS_indexing;
- imported fasta file (step_2AS_import).

If 2AS_mapping is used as input, a reference genome will also be required, automatically selected by the system based on the one used for mapping. The Raw Read Processing pipeline runs the same way for all sample types but only if the selected input consists of raw reads.

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.



Once the pipeline is launched, the page generates a link to the Check analyses section to allow a check of the processus status. The user will be notified by the system both when the pipeline is successfully launched and when the operation is completed.

Results

The files produced by "NgsManager" depend on the analyses that make up the executed pipelines; they are organized according to the same hierarchy used for individually run analyses. For reference, please consult the corresponding sections of the pipelines included in NgsManager, where all information regarding individual analyses and links to their respective result tables can be found.

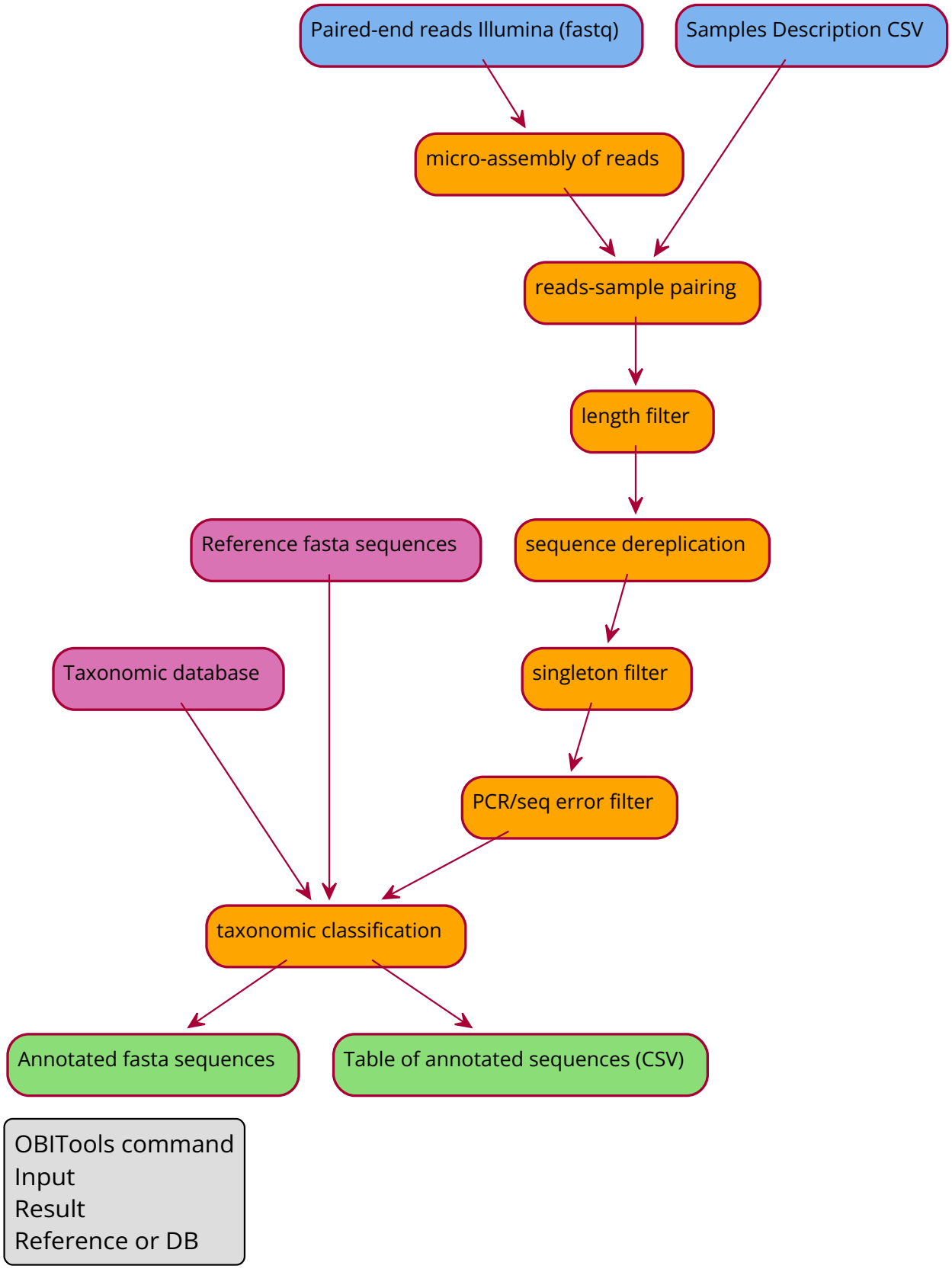
- Results of the "Raw Read Processing" pipeline;
- Results of the pipeline for bacteria:
  - Results of the "Typing on Bacteria" pipeline;
- Results of the "Covid Emergency" pipeline for SARS-CoV2;
- Results of the "Vdraft" pipeline for viruses (Genome Draft);
- Results of the "WGS on Bacteria" pipeline (2AS_denovo_shovill).

Pipeline OBITools

Introduction

The **OBITools** pipeline executes the OBITools open source software, which consists of a set of programs designed for analyzing NGS data with DNA metabarcoding approach.

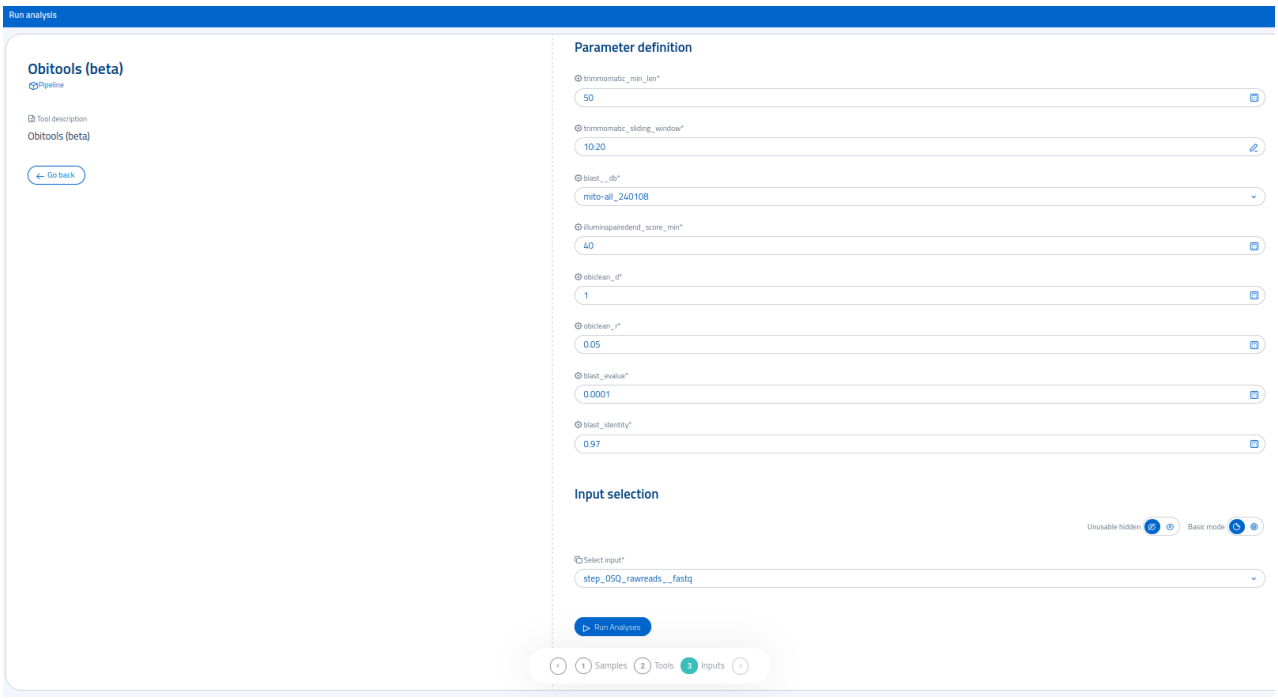
OBITools available at: http://metabarcoding.org/obitools ObiTools official guide: https://pythonhosted.org/OBITools/welcome.html ObiTools GitHub page: https://github.com/metabarcoding/



Run OBITools pipeline

The filter at the top of the run analyses page allows to display only the pipelines. Once the **OBITools** pipeline has been selected, a confirmation interface will be displayed.

OBITools exclusively accepts **raw reads from Illumina paired-end technology**. The parameter definition page allows users to modify all arguments of the OBITools software. For a detailed guide on the subject, please refer to the official software resources: https://pythonhosted.org/OBITools/welcome.html.



Once the pipeline is launched, the page generates a link to the Check analyses section to allow a check of the processus status. The user will be notified by the system both when the pipeline is successfully launched and when the operation is completed.

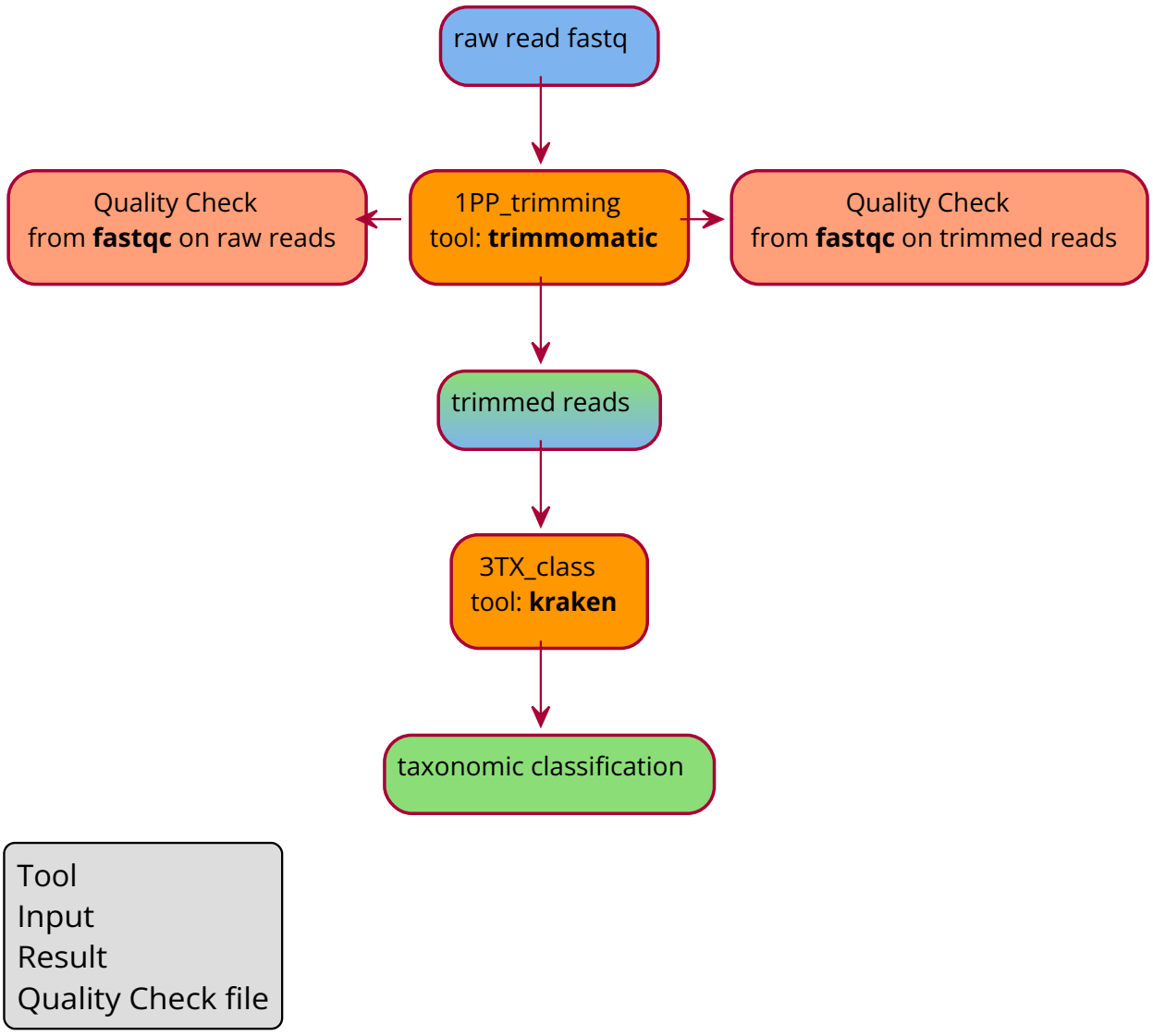
Raw Read Processing pipeline

Introduction

The **Raw Read Processing** pipeline performs the preprocessing of reads (**1PP**), followed by the taxonomic classification.

The reads produced by the sequencer are first subjected to quality control with *fastQC* and to trimming with *Trimmomatic* (1PP_trimming); the subsequent taxonomic analysis and the **calculated species assignment** are then performed using Kraken (3TX_class).

Please, note that in the 1PP_trimming analysis, quality control is performed with **fastQC** on both the raw reads and the trimmed reads obtained.



The pipeline performs the following 2 analyses:

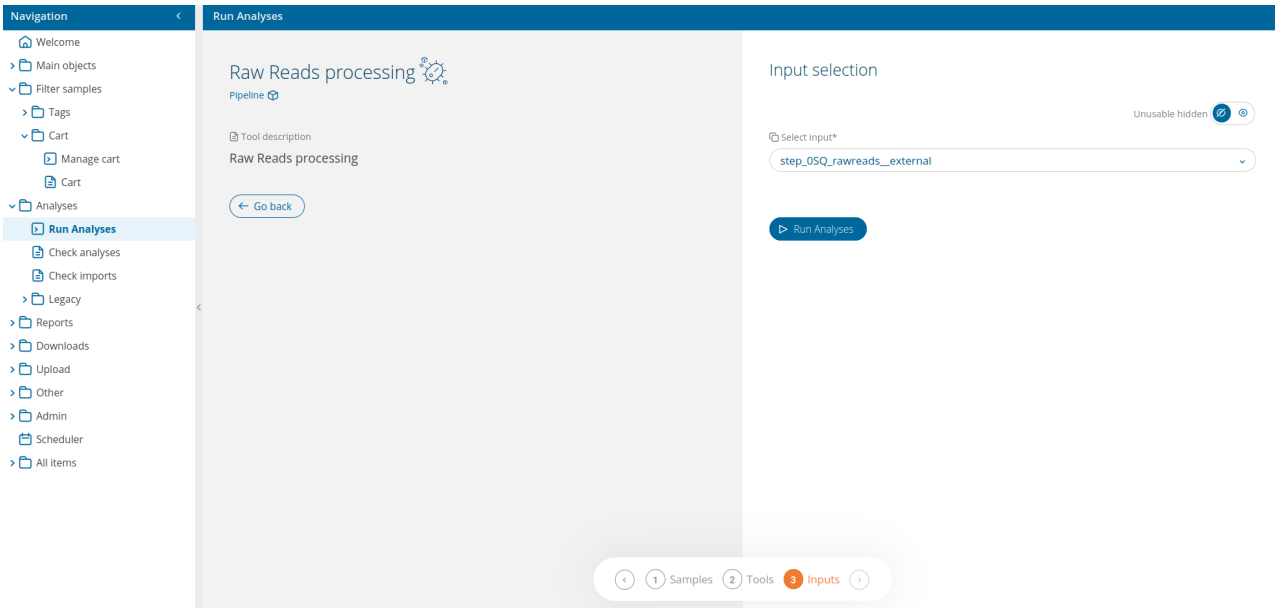
1. Trimming of raw reads to remove low-quality bases with **trimmomatic** (1PP_trimming__trimmomatic) and **fastQC**;
2. Taxonomic classification of identified bacteria or viruses and contamination check with **Kraken** (3TX_class__kraken) and **Quast**. Classification with **Kraken** is responsible for assigning the **calculated species**, on the basis of which the downstream pipelines are automatically selected when *Raw Read Processing* is executed by NGSmanager.

Run Raw Read Processing pipeline

The filter at the top of the run analyses page allows to display only the pipelines. Once the **Raw Read Processing** pipeline has been selected, a confirmation interface will be displayed.

Accepted inputs are raw reads in fastq format (internal or imported).

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.



Once the pipeline is launched, the page generates a link to the Check analyses section to allow a check of the processus status. The user will be notified by the system both when the pipeline is successfully launched and when the operation is completed.

Results

Files generated by the "Raw Read Processing" pipeline will be the same as those produced by the individual analyses that compose it, organized according the same hierarchy. As a reference, please refer to the relevant sections of the individual analyses:

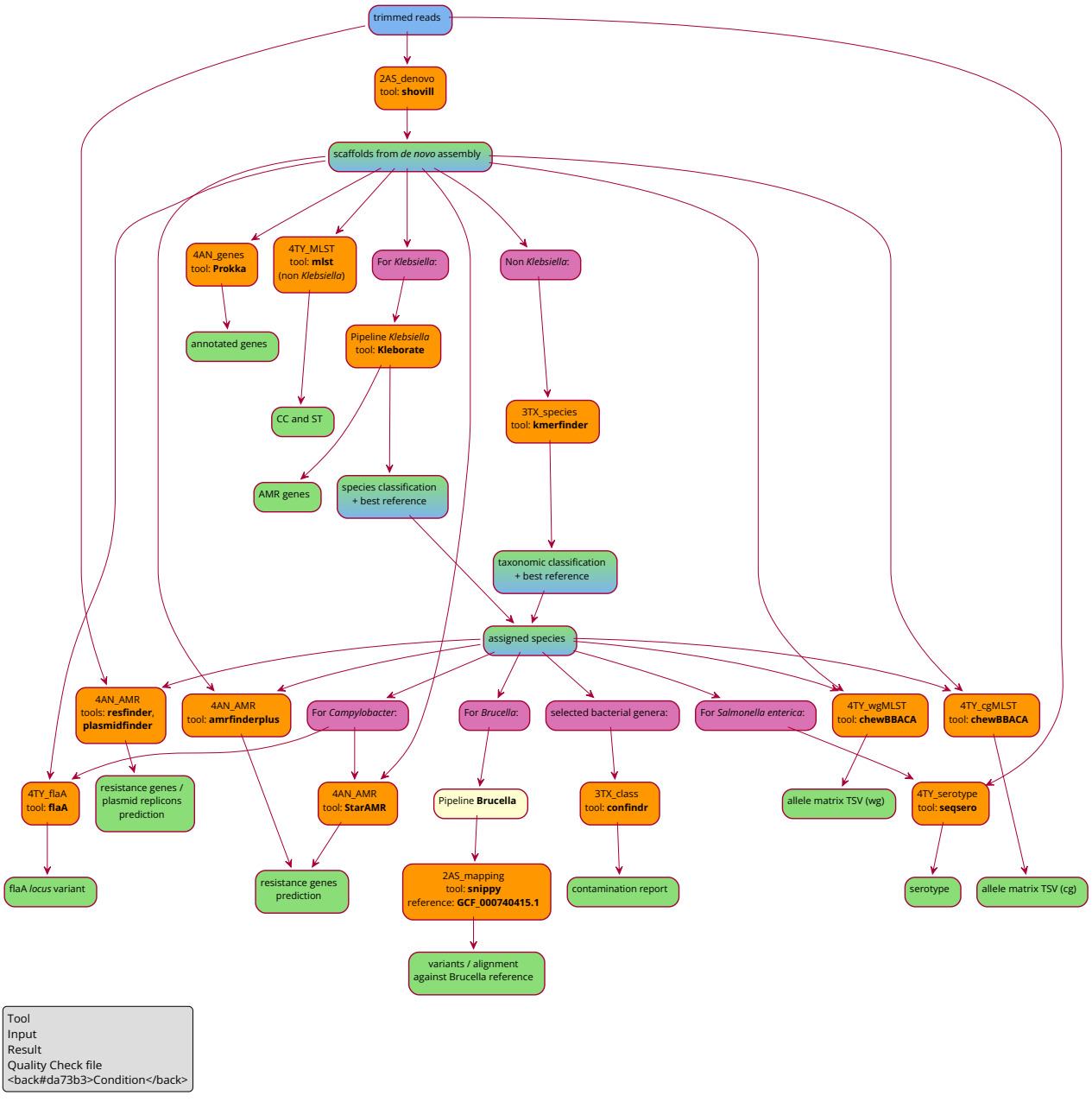
- Output of trimming with trimmomatic and fastQC quality check;
- Output of taxonomic classification with Kraken and Quast quality check.

Pipeline Typing on Bacteria

Introduction

The **Typing on Bacteria** pipeline performs, in a single run, all the typing analyses to which bacterial samples are usually subjected, including:

- species determination;
- identification of genes associated with antibiotic resistance and virulence;
- identification of the flaA *locus* variant (*Campylobacter*);
- MLST;
- core-genome MLST;
- whole-genome MLST;
- genome annotation;
- calculation of horizontal and vertical coverage.



The **Typing on Bacteria** pipeline runs the following analyses:

1. 3TX_species: species determination using **kmerfinder**, to assign the best reference to use for downstream mapping. **When executed from NGSmanager, the inputs are the scaffolds from the *de novo * assembly generated by Shovill (WGS on Bacteria pipeline)**;
2. 4AN_AMR: identification of genes associated with antibiotic resistance and virulence;
3. 4TY_flaA: identification of the flaA *locus* variant (only for *Campylobacter*);
4. 4TY_MLST: *in silico* Multi-Locus Sequence Typing using MLST software;
5. 4TY_cgMLST: core-genome MLST using ChewBBACA software;
6. 4TY_wgMLST: whole-genome MLST using ChewBBACA software;
7. 4TY_serotype: serotype determination;
8. 4AN_genes: functional annotation of the genome using **Prokka**;
9. Kleborate on *Klebsiella* samples.

Depending on **kmerfinder** results (3TX_species), the following analyses are also performed:

- 4TY_cgMLST;
- 2AS_mapping using *Bowtie*: the reference used is the one identified as the "best reference" by kmerfinder, among the genomes available in the static database (kmerfinder).

The software used, for **4TY_cgMLST** and **4TY_wgMLST**, is selected based on the available schemes for each tool, as indicated in the dedicated Wiki pages:

- Available schemas for cgMLST;
- Available schemas for wgMLST.

Similarly: - **4TY_MLST_mlst** is performed only if the appropriate scheme is available for the microorganism (the list of schemes can be consulted in the official mlst guide on GitHub); - **4TY_flaA** is performed only on Campylobacter samples.

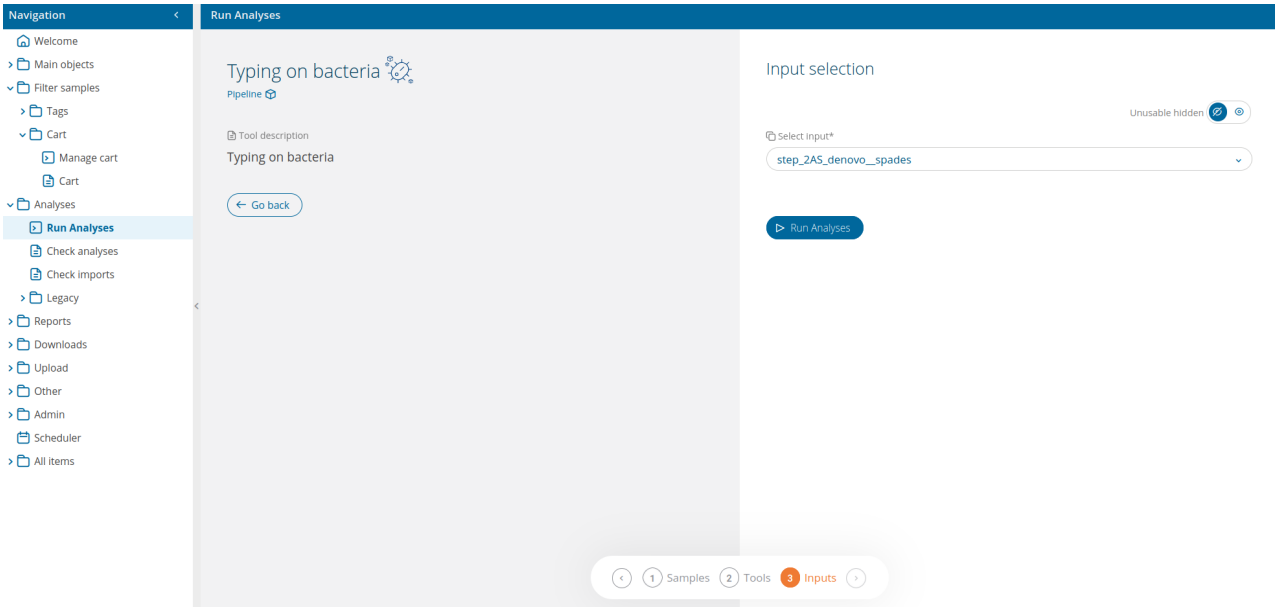
Run Typing on Bacteria pipeline

The filter at the top of the run analyses page allows to display only the pipelines. Once the **Typing on Bacteria** pipeline has been selected, a confirmation interface will be displayed.

The usable inputs are *consensus* sequences (.fasta format) obtained from mapping (2AS_mapping) or scaffolds from the *de novo* assembly (2AS_denovo).

In the case of *consensus* sequences, originating from mapping, are used as input, it will also be necessary to specify a reference genome among the available ones.

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.



Once the pipeline is launched, the page generates a link to the Check analyses section to allow a check of the processus status. The user will be notified by the system both when the pipeline is successfully launched and when the operation is completed.

Results

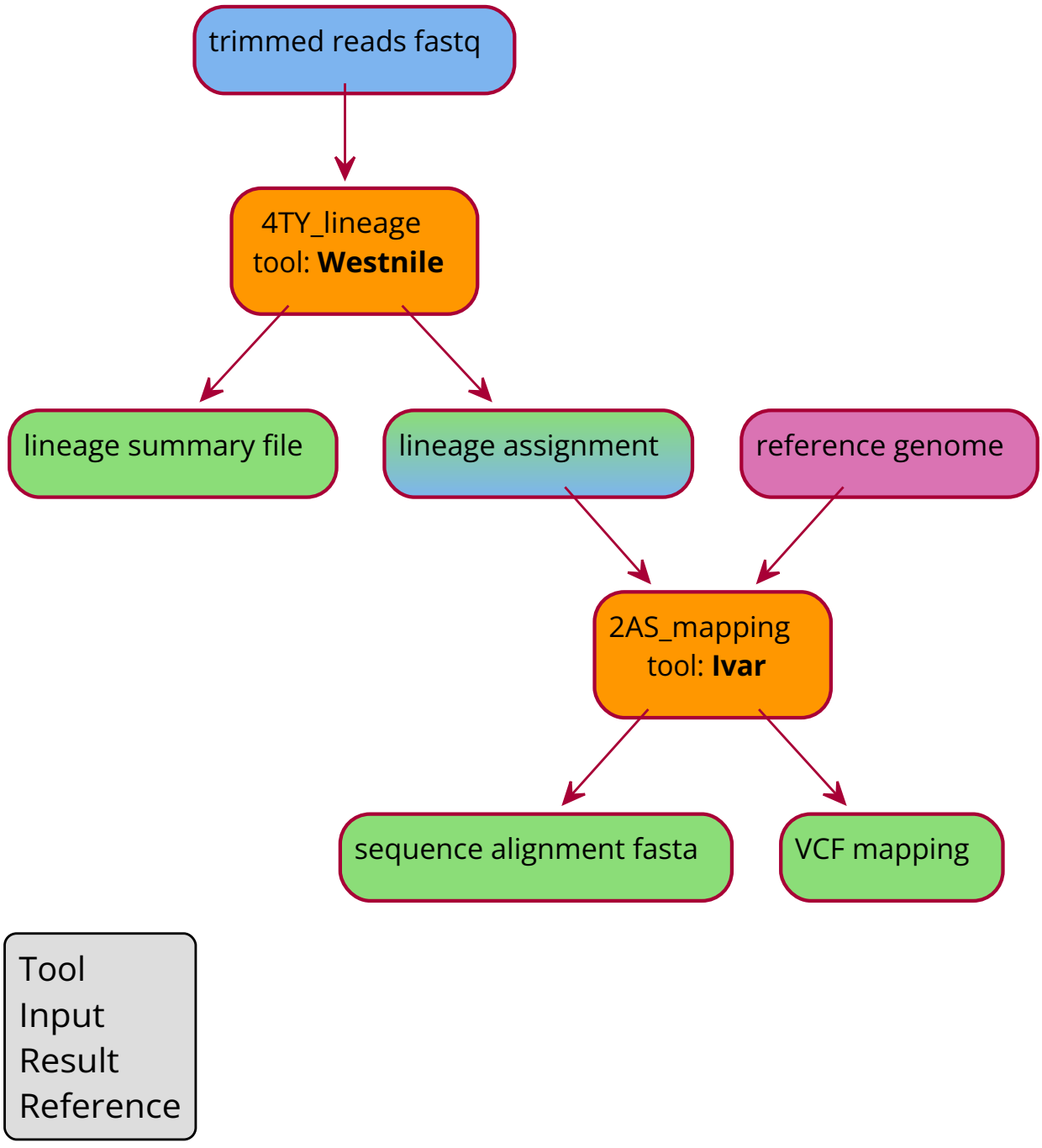
Files generated by the *Typing on Bacteria* pipeline will be the same as those produced by the individual analyses that compose it, organized according the same hierarchy. As a reference, please refer to the relevant sections of the individual analyses:

- Output of species assignment;
- Output of antibiotic-resistance and virulence prediction;
- Output of *flaA* locus variant determination;
- Output of Multi-*Locus* Sequence Typing (MLST);
- Output of core-genome MLST (cgMLST);
- Output of whole-genome MLST (wgMLST);
- Output of genome annotation;
- Output of mapping on kmerfinder's results.

WNV (West Nile Virus) Pipeline

Introduction

The **WNV - Lineage Calculation and Mapping** pipeline calculates the lineage of the West Nile Virus and maps the sample against the corresponding reference.



Run WNV Pipeline

The filter at the top of the run analyses page allows to display only the pipelines. Once the **WNV** pipeline has been selected, a confirmation interface will be displayed.

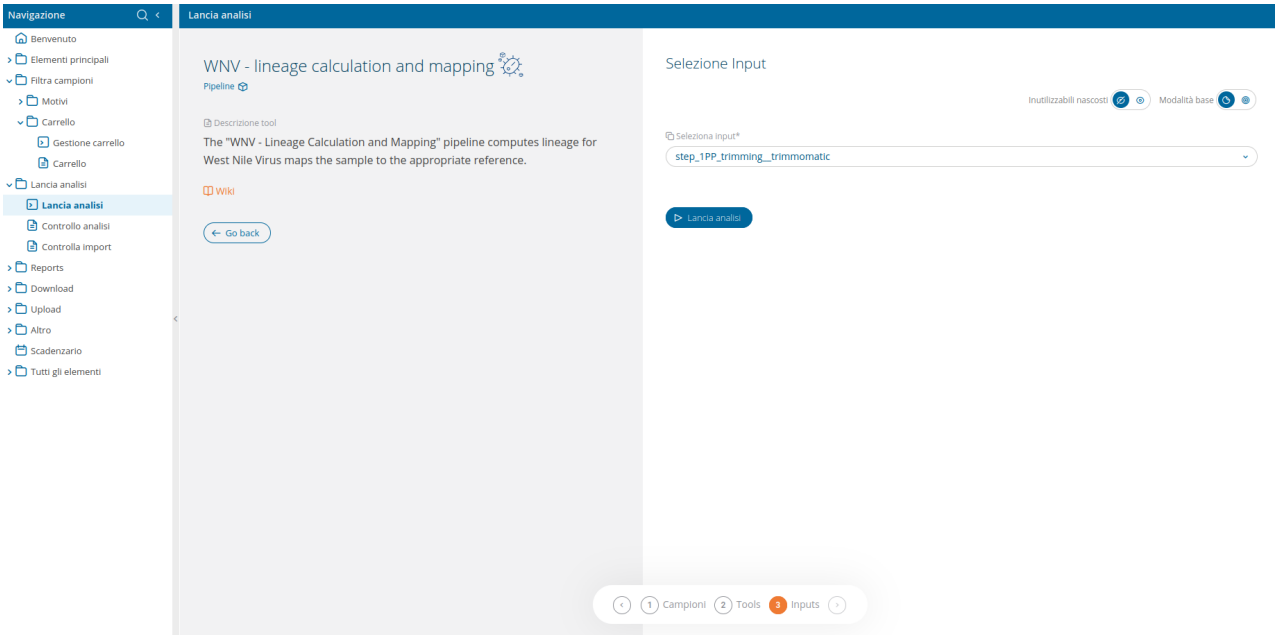
The **WNV** performs the following analyses:

1. West Nile Virus lineage computation with **Westnile** (4TY_lineage__westnile);
2. Mapping, with **Ivar** (2AS_mapping__ivar), of each genome segment against its appropriate references.

Accepted inputs can originate from pre-processing analyses:

- depleted reads (step_1PP_hostdepl);
- downsampled reads (step_1PP_downsampling);
- trimmed reads (step_1PP_trimming);
- filtered reads (step_1PP_filtering).

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.



The reference genome used for the mapping step will be the base genome associated with the lineage assigned by the **Westnile** tool, according to the correspondences shown in the table below:

assigned lineage	**reference**
lin. 1	FJ483548.1
lin. 2	HQ537483.1
lin. 3	AY765264.1
lin. 4	AY277251.1
lin. 5	GQ851605.1
lin. 6	GU047875.1
lin. 7	KY703855.1
lin. 8	KY703856.1
lin. 9	KJ831223.1

Once the pipeline is launched, the page generates a link to the Check analyses section to allow a check of the processus status. The user will be notified by the system both when the pipeline is successfully launched and when the operation is completed.

Results

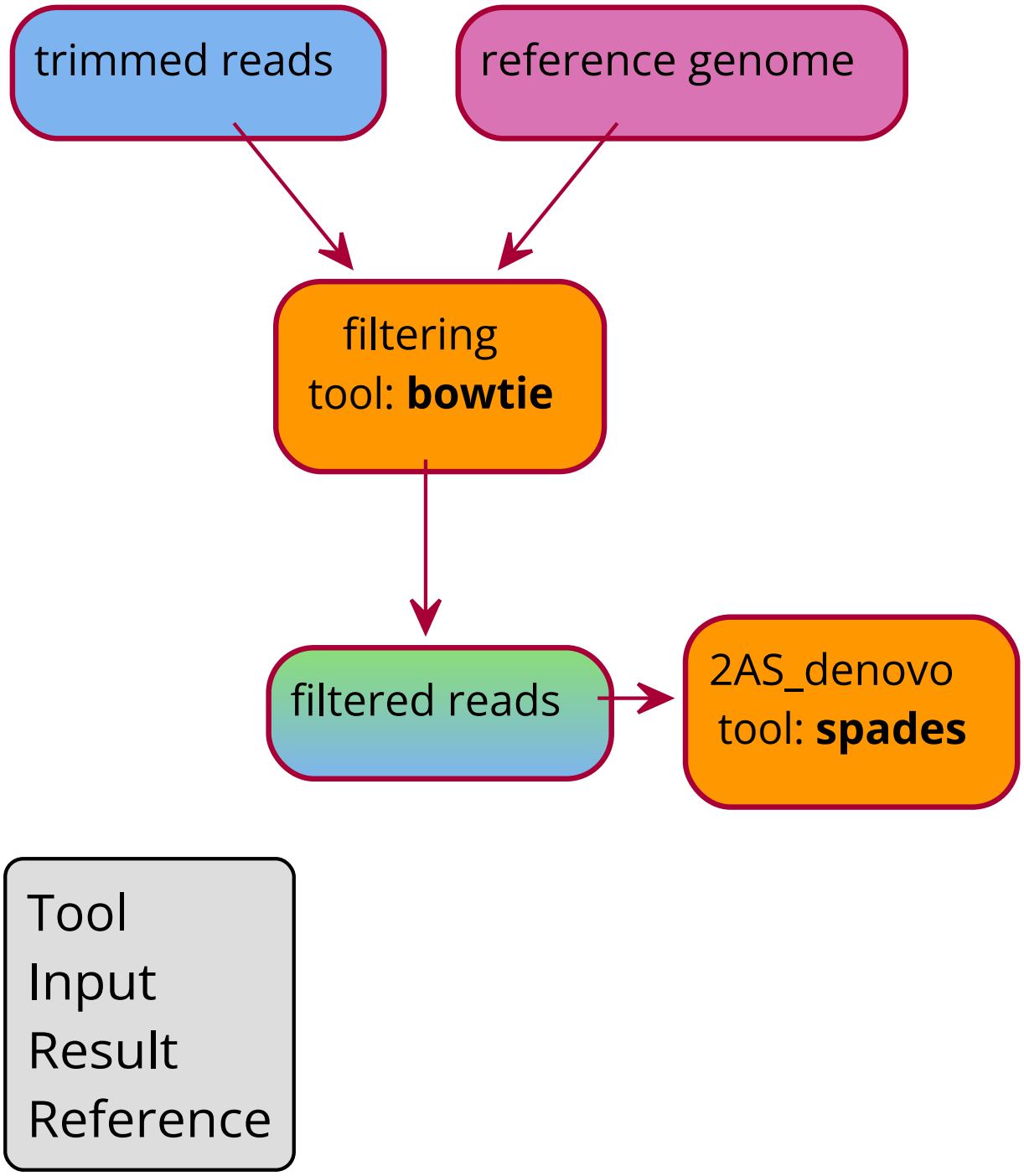
Files generated by the *WNV* pipeline will be the same as those produced by the individual analyses that compose it, organized according the same hierarchy. As a reference, please refer to the relevant sections of the individual analyses:

- Output of lineage assignment with westnile
- Output of mapping with ivar.

Filtering & *de novo*

Introduction

The **Filtering & *de novo*** pipeline removes reads that do not map to the selected reference (1PP_filtering) and performs the *de novo* assembly (2AS_denovo).



Run Filtering & *de novo* pipeline

The filter at the top of the run analysis page allows to display only the pipelines. Once the **Filtering & *de novo*** pipeline has been selected, a confirmation interface will be displayed.

Filtering & *de novo* performs:

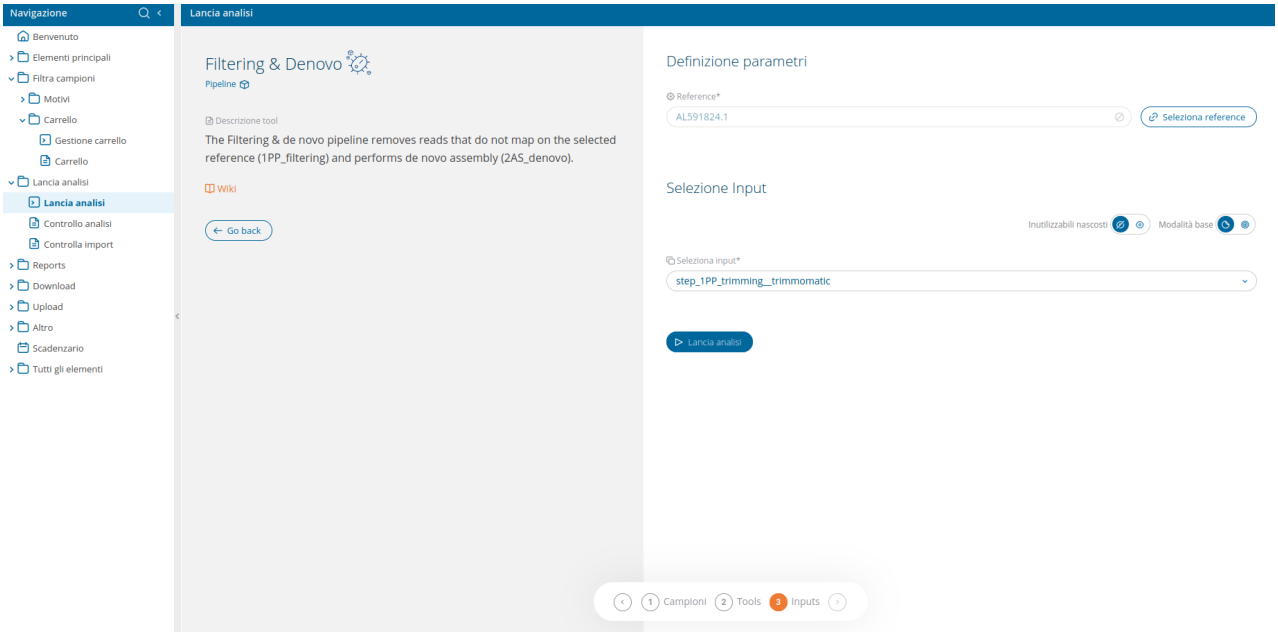
1. removal of reads that do not map to the reference with **Bowtie**;
2. *de novo* assembly with **SPAdes**.

The usable inputs are the trimmed reads obtained from a pre-processing analysis:

- depleted reads (step_1PP_hostdepl);
- downsampled reads (step_1PP_downsampling);
- trimmed reads (step_1PP_trimming);
- filtered reads (step_1PP_filtering);
- reads generated from fasta (step_1PP_generated).

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.

In the section dedicated to the parameter definition, it will be necessary to specify a reference genome, which is required to carry out the first step of the pipeline, the *filtering* (1PP_filtering).



Once the pipeline is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The user will be notified by the system both when the pipeline is successfully launched and when the operation is completed.

Results

Files generated by the "Filtering & *de novo*" pipeline will be the same as those produced by the individual analyses that compose it, organized according the same hierarchy. As a reference, please refer to the relevant sections of the individual analyses:

- Output of filtering of the reads using Bowtie;
- Output of *de novo* assembly with SPAdes.

Pipeline Klebsiella - Kleborate

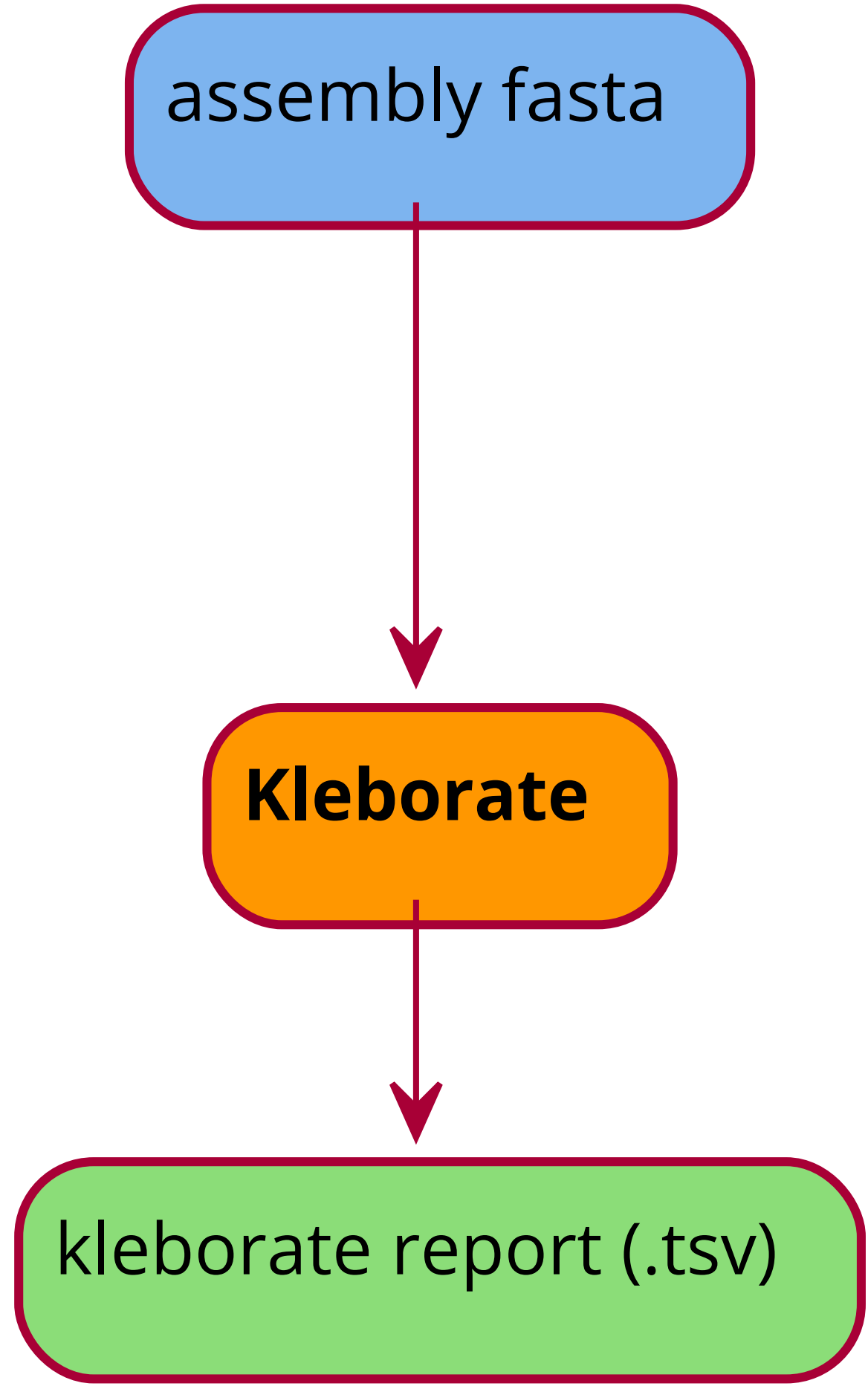
Introduction

The screening of *Klebsiella pneumoniae* genomes and of the *Klebsiella pneumoniae species complex* (KpSC) is carried out using the **Kleborate** pipeline.

Kleborate analyses the genome to obtain data regarding:

- Species;
- MLST of K. pneumoniae species complex;
- ICEKp associated virulence *loci*: yersiniabactin (ybt), colibactin (clb), salmochelin (iro), hypermucoidy (rmp);
- Plasmid associated Virulence *loci*: salmochelin (iro), aerobactin (iuc), hypermucoidy (rmp, rmpA2);
- Antimicrobial resistance (AMR) determinants: acquired genes, SNPs, gene truncation and intrinsic β-lactamases, K (capsular) and O (LPS) antigens;
- serotype, through the wzi alleles and the Kaptive software.

Kleborate GitHub Page: https://github.com/klebgenomics/Kleborate

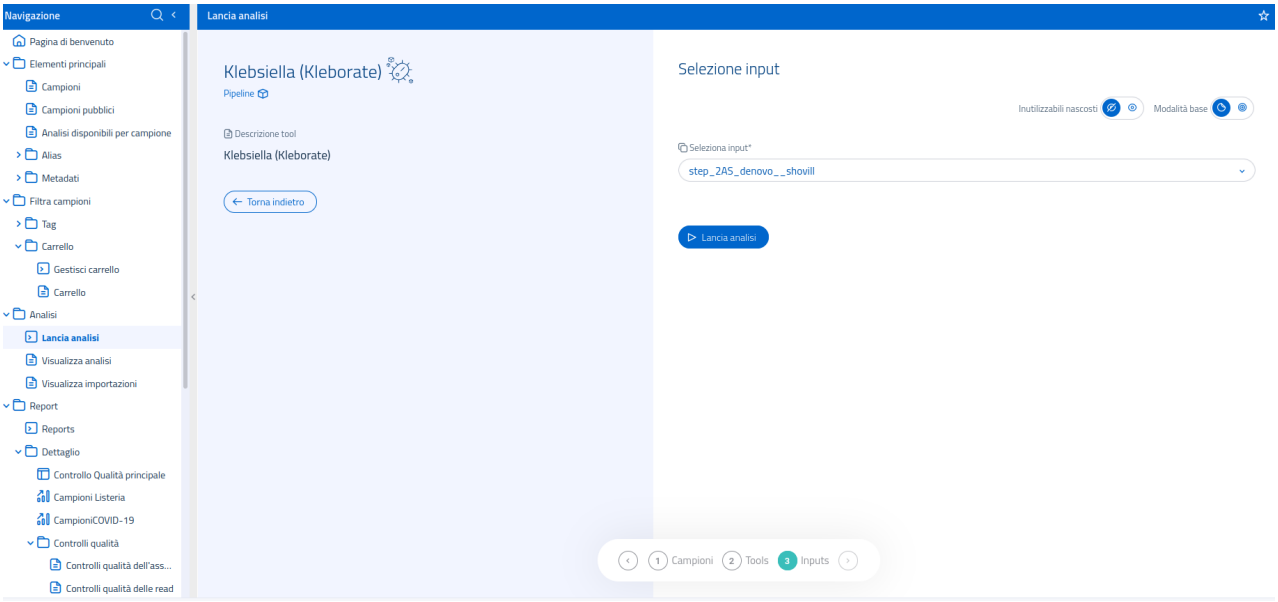

assembly fasta
Kleborate
kleborate report (.tsv)

Run Kleborate pipeline

The filter at the top of the run analysis page allows to display only the pipelines. Once the **Klebsiella (Kleborate)** pipeline has been selected, you can run the analysis using *de novo* assembly as input:

- step_2AS_denovo
- step_2MG_denovo
- step_2AS_hybrid

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.



Once the pipeline is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The user will be notified by the system both when the pipeline is successfully launched and when the operation is completed.

Results

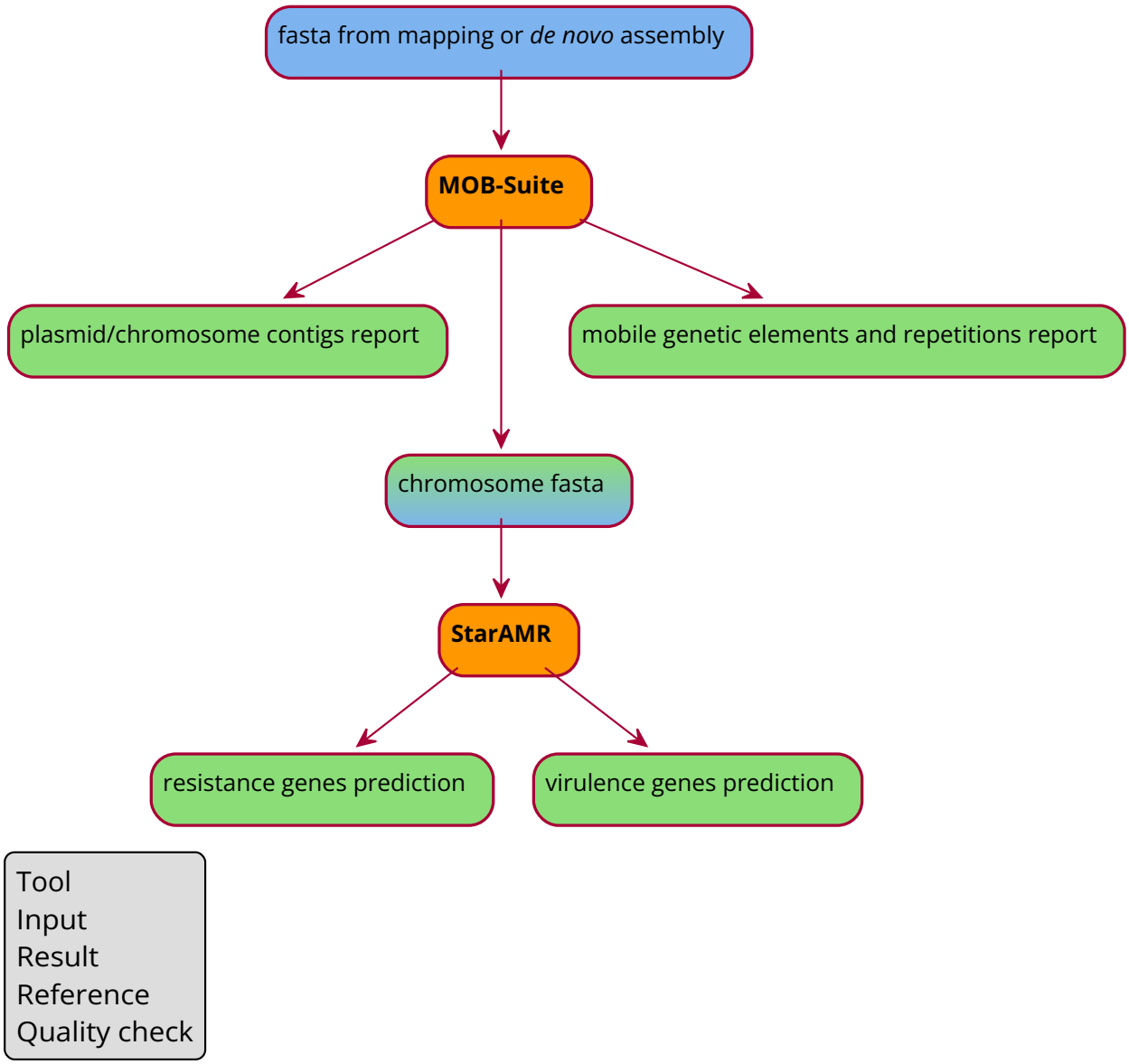
The "Kleborate" pipeline produces a single output: a .tsv table with all typing data generated by the software.

File	Description	Location
`*_kleborate.tsv`	.tsv table containing all typing results	folder: result

Pipeline Plasmids AMR

Introduction

The **Plasmids AMR** pipeline allows plasmid typing by running *4TY_plasmid* with MOB-Suite and, subsequently, *4AN_AMR* with StarAMR on the FASTA file generated by 4TY_plasmid.



Run Plasmids AMR pipeline

The filter at the top of the run analyses page allows to display only the pipelines. Once the **Plasmids AMR** pipeline has been selected, a confirmation interface will be displayed.

The **Plasmids AMR** pipeline includes the following analyses:

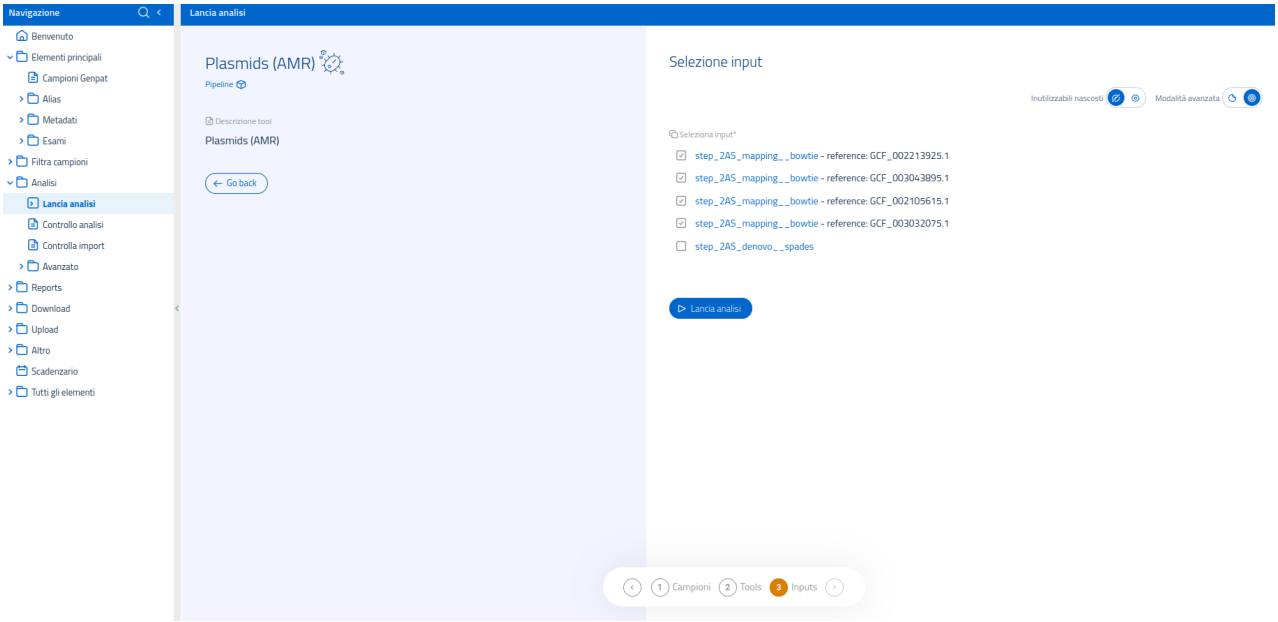
1. Plasmid identification with **MOB-Suite** (4TY_plasmid__mobsuite);
2. Prediction of AMR-related genes with **Snippy** (4AN_AMR).

The usable inputs are the same as those of the first analysis performed (*4TY_plasmid*), namely the outputs of the sequence reconstruction analyses:

- fasta file from *de novo* assembly (step_2AS_denovo);
- fasta file from mapping (step_2AS_mapping);
- fasta file from hybrid assembly (step_2AS_hybrid);
- fasta file from *de novo* assembly for metagenomic samples (step_2MG_denovo);
- imported reads (step_2AS_import).

In the case of input originating from mapping, it will also be necessary to provide the reference used for the mapping (pre-filled field).

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.



Once the pipeline is launched, the page generates a link to the Check analyses section to allow a check of the processus status. The user will be notified by the system both when the pipeline is successfully launched and when the operation is completed.

Results

Files generated by the *Plasmids AMR* pipeline will be the same as those produced by the individual analyses that compose it, organized according the same hierarchy. As a reference, please refer to the relevant sections of the individual analyses:

- Output of plasmid typing;
- Output of antibiotic resistance prediction.

nf-core/ampliseq pipeline

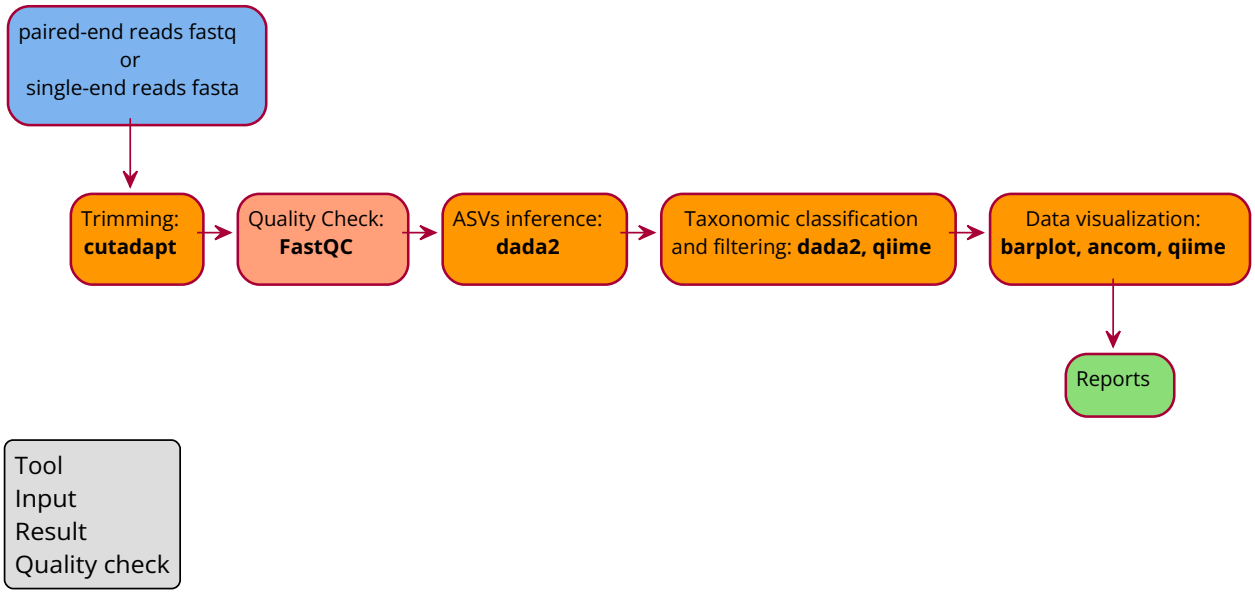
Introduction

The **nf-core/ampliseq** pipeline originates from the nf-core community, which is dedicated to building and sharing bioinformatics pipelines managed with Nextflow. It performs amplicon sequencing and denoising; in addition, it can use various databases for taxonomic assignment (including databases for 16S, ITS, CO1, and 18S sequences). The pipeline supports data originating from Illumina, PacBio and IonTorrent devices.

The version of the *nf-core/ampliseq* pipeline available on GenPat is **2.7.1**.

Official page of version 2.7.1 of *nf-core/ampliseq* on the "nf-core community" website: https://nf-co.re/ampliseq/2.7.1.

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



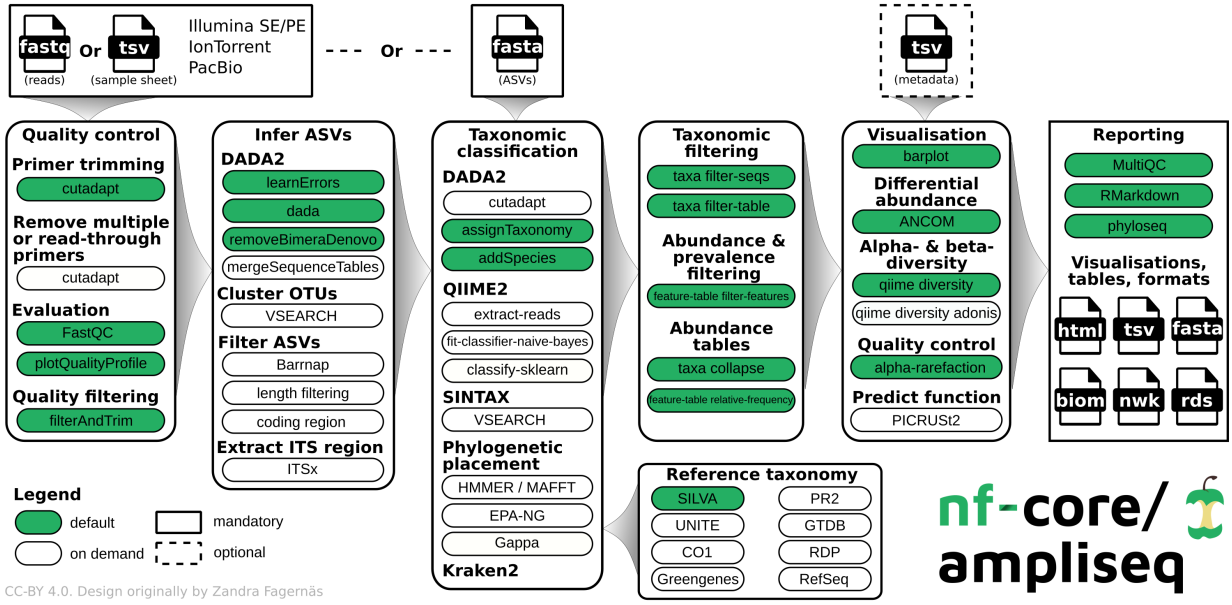


Image property of ©nf-core: https://nf-co.re/ampliseq/2.7.1.

The default analyses of nf-core/ampliseq (shown in green in the image above) are the ones executed in GenPat. Some of the on-demand analyses can be added using the additional parameters in Launch Analysis; moreover, some default analyses can be skipped, also using the additional parameters. It is recommended to consult the "parameters" and "parameters: skipping specific steps" sections of the *nf-core/ampliseq official guide* to make the best use of the additional parameters and the text field for shell-like parameters.

Run nf-core/ampliseq pipeline

TThe filter at the top of the run analysis page allows to display only the pipelines. Once the **nf-core/ampliseq** pipeline has been selected, a confirmation interface will be displayed.

The steps of the pipeline are summarized in the diagram at the top of the page.

The pipeline supports sequence files generated by Illumina, IonTorrent or PacBio sequencing devices:

- step_1PP_trimming;
- step_1PP_hostdepl;
- step_1PP_downsampling;
- step_1PP_filtering.

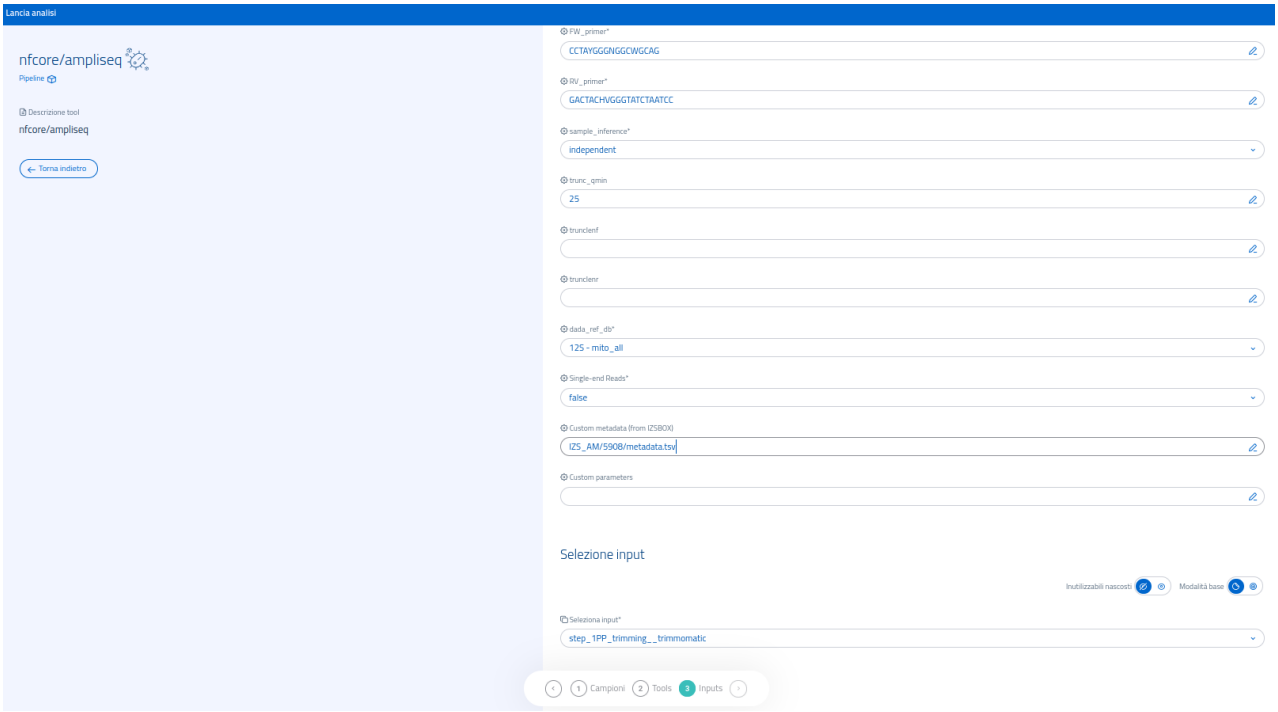
The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.

Additional parameters

The following list refers to the additional parameters, available in the dedicated section of the launch system:

- `FW_primer` and `RV_primer` : text fields in which to paste the sequence of the forward primers and reverse primers, respectively; this is a necessary step for the trimming of the primers;
- `sample_inference` : inference mode for ASV calculation:
  - `independent` : fast mode, faster and lighter, but with lower sensitivity";
  - `pooled` : "performance" mode, with a higher sensitivity but heavier to run;
  - `pseudo` : (*e.i.* pseudo-pooled), "balanced mode";
- `trunclenf` , `trunclenr` and `trunc_qmin` : the denoising step requires reads of equal length. The truncation parameters are text fields where you have to manually specify the nucleotide position to be used as the cutoff value for the forward read (`trunclenf`) and for reverse read (`trunclenr`). Longer sequences are truncated whereas shorter ones are discarded. `trunc_qmin` allows to specify a quality minimum value for the nucleotide call: if `trunclenf` and `trunclenr` are not provided, cutoff values are determined automatically, truncating when the median value of the quality score falls below the specified cutoff threshold.
- `dada_ref_db` : selection of the database used by DADA2 for taxonomic assignment (in GenPat, the 12S rRNA "mito-all" database and the 16S rRNA "silva" database are currently available).
- `Reads` : selection of the read type (paired-end or single-end).
- `Custom metadata (from IZSBOX)` : it allows you to specify a metadata file present in your IZSbox (to paste the file path).
- `Custom parameters` : the other parameters of nf-core/ampliseq are available as modifiers to be entered in the text field as terminal commands. The official guide contains all the necessary information.

> Please consult the parameter descriptions directly in the official nf-core/ampliseq guide: https://nf-co.re/ampliseq/2.7.1/parameters.



Once the pipeline is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The user will be notified by the system both when the pipeline is successfully launched and when the operation is completed.

Result folder

The summary card of a successfully completed pipeline allows access to the results folder, enabling you to retrieve tabular files and graph images; it also provides direct links for viewing images and reports.

The **Result folder** can be accessed by clicking the link in the analysis card, in the *Result data* section. Within the following `results` folder, it is possible to find the files produced by the nf-core/ampliseq pipeline.

The table below shows the list of the main folders and their contents, along with some useful information.

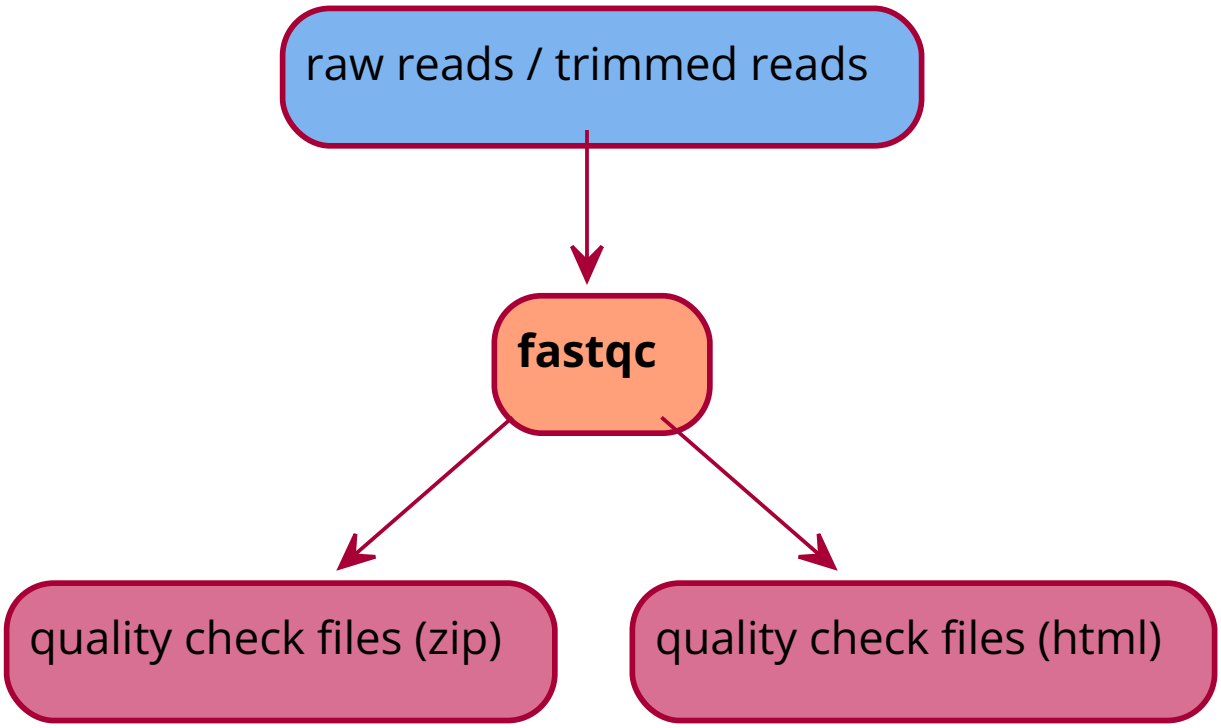
Folder	Content
`results/`	software subfolders and summary table (`overall_summary.tsv`)
`results/cutadapt`	cutadapt logs (`DSXXXXXXXX.trimmed.cutadapt.log`) and summary table (`cutadapt_summary.tsv`)
`results/dada2`	.tsv tables of DADA2 statistics, the identified ASVs, and the FASTA file of the ASV sequences (`ASV_seqs.fasta`)
`results/dada2/QC`	quality check graphs in .pdf format
`results/input`	metadata file (`metadata.tsv`)
`results/multiqc`	multiQC report in .html format, output data folder and MultiQC graph folder
`results/qiime`	folders of the data produced by QIIME: `abundance_tables` , `alpha-rarefaction` , `barplot` , `diversity` , `phylogenetic_tree` , `rel_abundance_tables` , `representative_sequences`
`results/summary_report`	graphs in .svg image format and report in .html format

QC fastqc

Introduction

QC fastqc is part of the analysis subcategory for quality control (Quality Check, QC); it runs the **fastQC** software on raw reads or trimmed reads, providing the corresponding quality metrics.

FastQC official GitHub page: https://github.com/s-andrews/FastQC.



Input
Quality Check tool
Quality Check result

Run FastQC pipeline

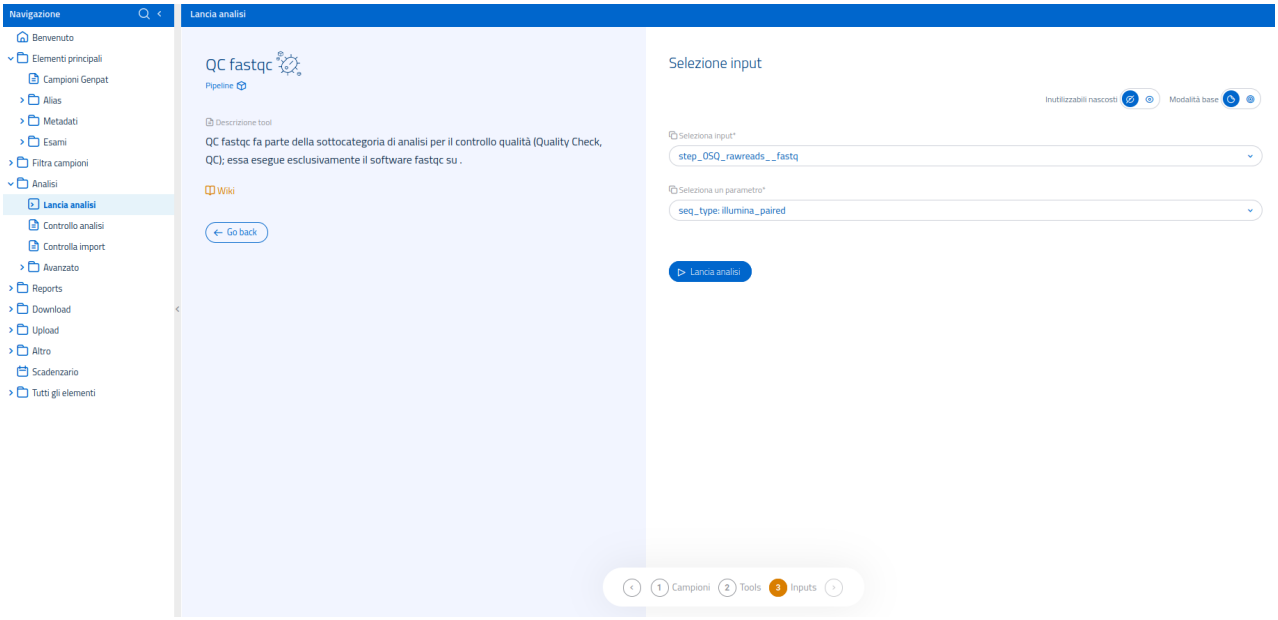
The filter at the top of the run analyses page allows to display only the pipelines. Once the **QC fastqc** pipeline has been selected, a confirmation interface will be displayed.

Accepted inputs can be raw reads in fastq format (internal or external) or trimmed reads (or those originating from preprocessing analyses):

- raw reads (step_0SQ_rawreads);
- trimmed reads (step_1PP_trimming);
- filtered reads (step_1PP_filtering);
- fastq reads generated from fasta (step_1PP_generated);
- depleted reads (step_1PP_hostdepl);
- downsampled reads (step_1PP_downsampling).

> **Note:** fastQC can be executed on paired or unpaired reads. In the first case, each fastqc result will be produced for each read and will be marked with the suffixes "R1" and "R2".

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.



Once the pipeline is launched, the page generates a link to the Check analyses section to allow a check of the processus status. The user will be notified by the system both when the pipeline is successfully launched and when the operation is completed.

Results

> To consult the guide on how to download files from the platform, please refer to the dedicated page.

The files produced by the *QC fastqc* pipeline will be the same as those produced by the same tool within the analyses that use it, as software, for quality check. To consult the fastqc results, please refer to the table available at the following link to the outputs of 1PP_trimming:

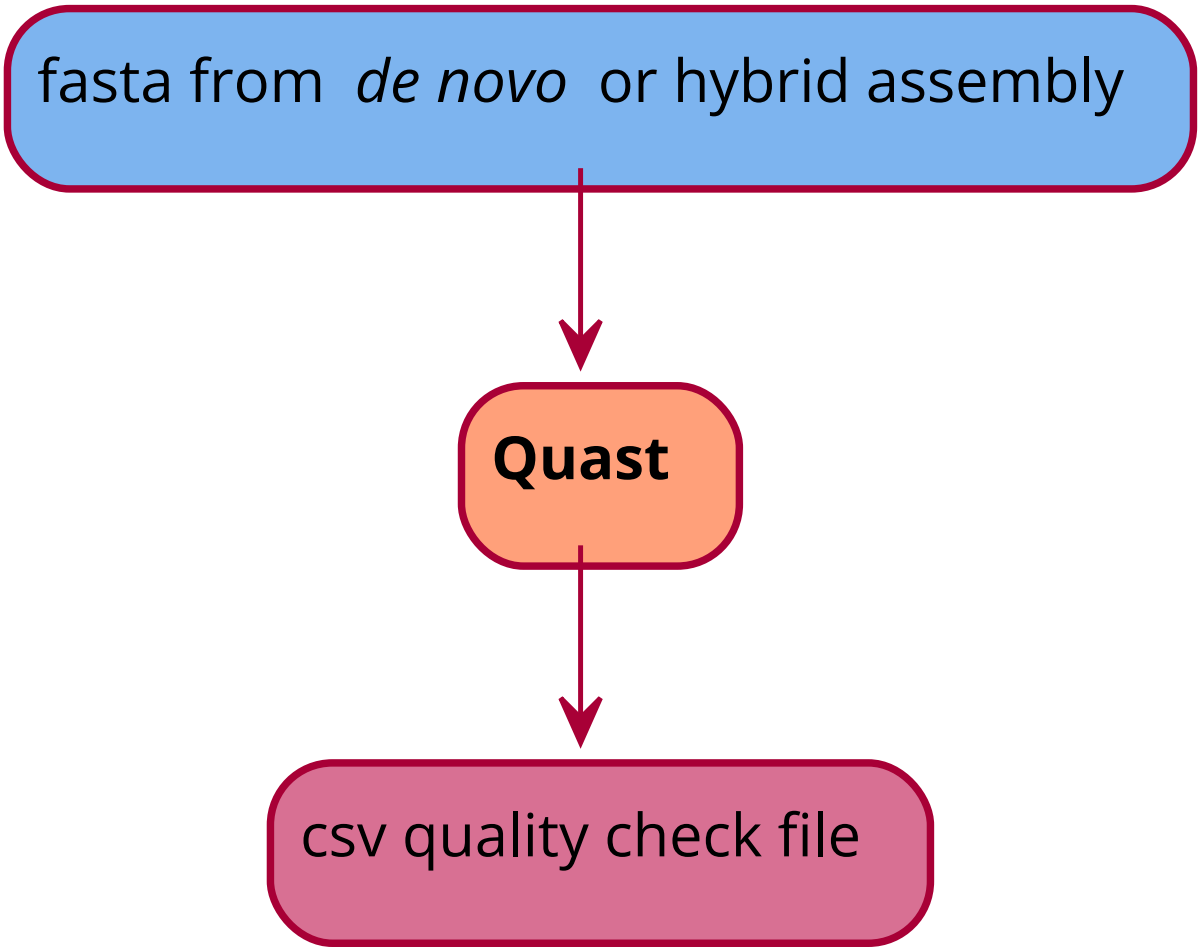
- FastQC quality check outputs.

QC Quast

Introduction

QC Quast is part of the analysis subcategory for quality control (Quality Check, QC); it runs the **Quast** software on assembly files (or hybrid assembly).

Quast official GitHub page https://github.com/ablab/quast



Input
Quality Check tool
Quality Check result

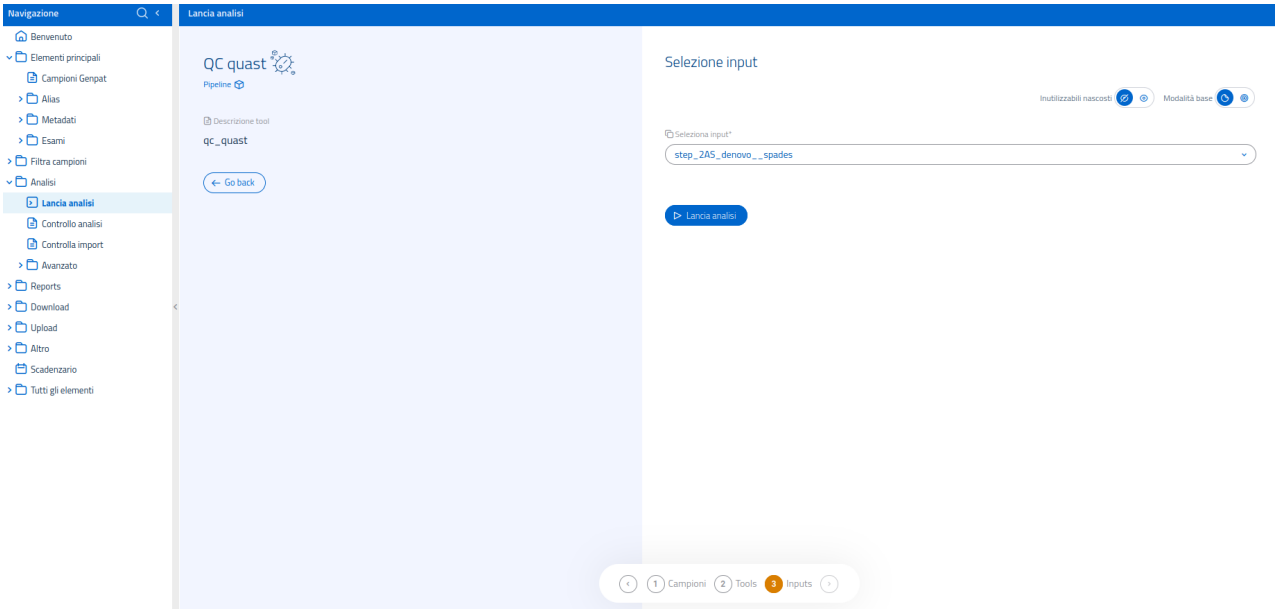
Run Quast pipeline

The filter at the top of the run analyses page allows to display only the pipelines. Once the **QC quast** pipeline has been selected, a confirmation interface will be displayed.

Accepted inputs are assembly files:

- *de novo* assembly (step_2AS_denovo);
- hybrid assembly (step_2AS_hybrid);
- *de novo* assembly from metagenomic samples (step_2MG_denovo).

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.



Once the pipeline is launched, the page generates a link to the Check analyses section to allow a check of the processus status. The user will be notified by the system both when the pipeline is successfully launched and when the operation is completed.

Results

> To consult the guide on how to download files from the platform, please refer to the dedicated page.

The files produced by the *QC Quast* pipeline will be the same as those produced by the same tool within the analyses that use it, as software, for quality check. To consult the Quast results, please refer to the table available at the following link to the outputs of 2AS_denovo analysis:

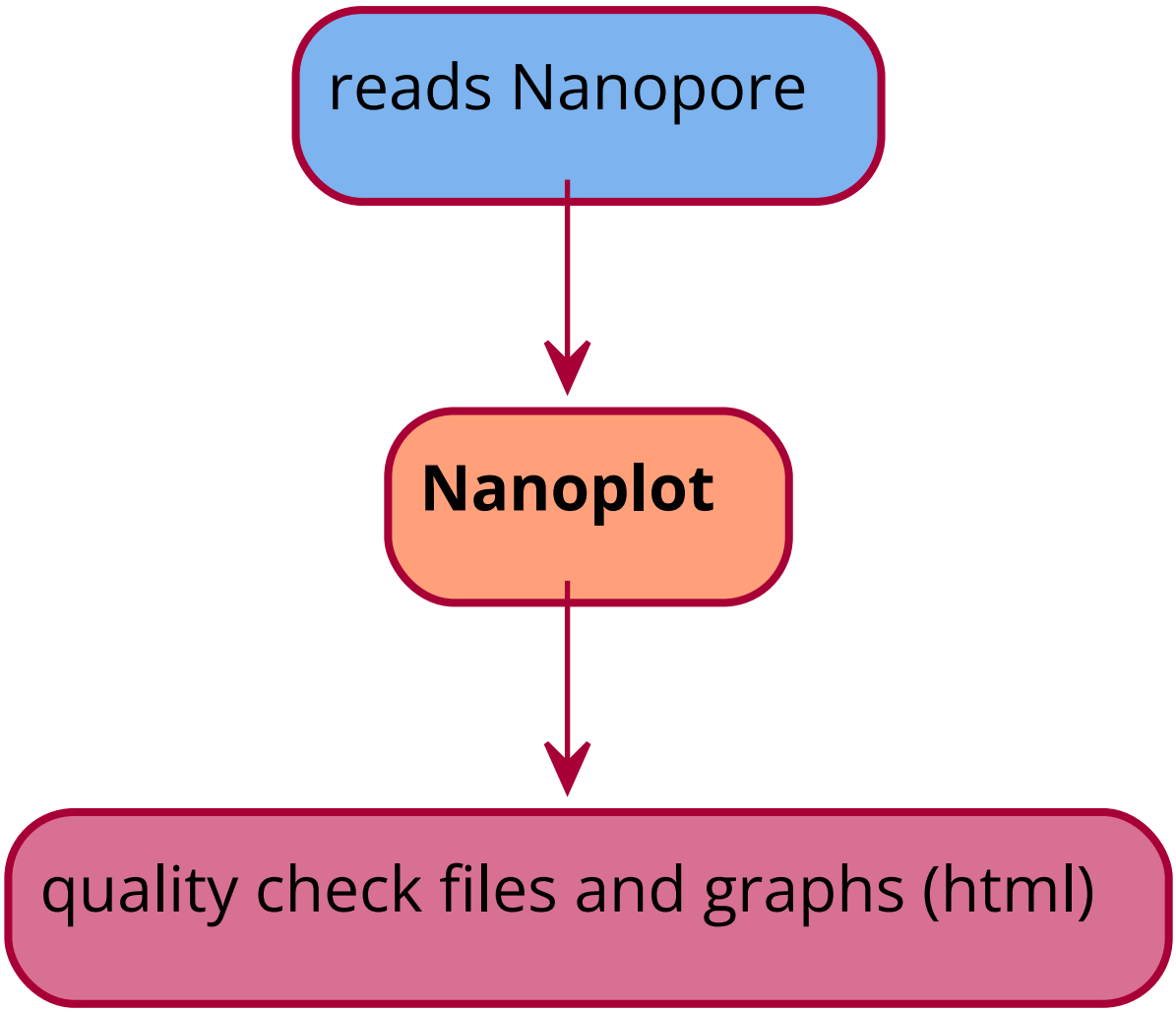
- Quast quality check.

QC Nanoplot

Introduction

QC Nanoplot is part of the analysis subcategory for quality control (Quality Check, QC); it runs the **Nanoplot** software on raw reads originating from Nanopore devices.

Nanoplot official GitHub page https://github.com/wdecoster/NanoPlot



Input
Quality Check tool
Quality Check result

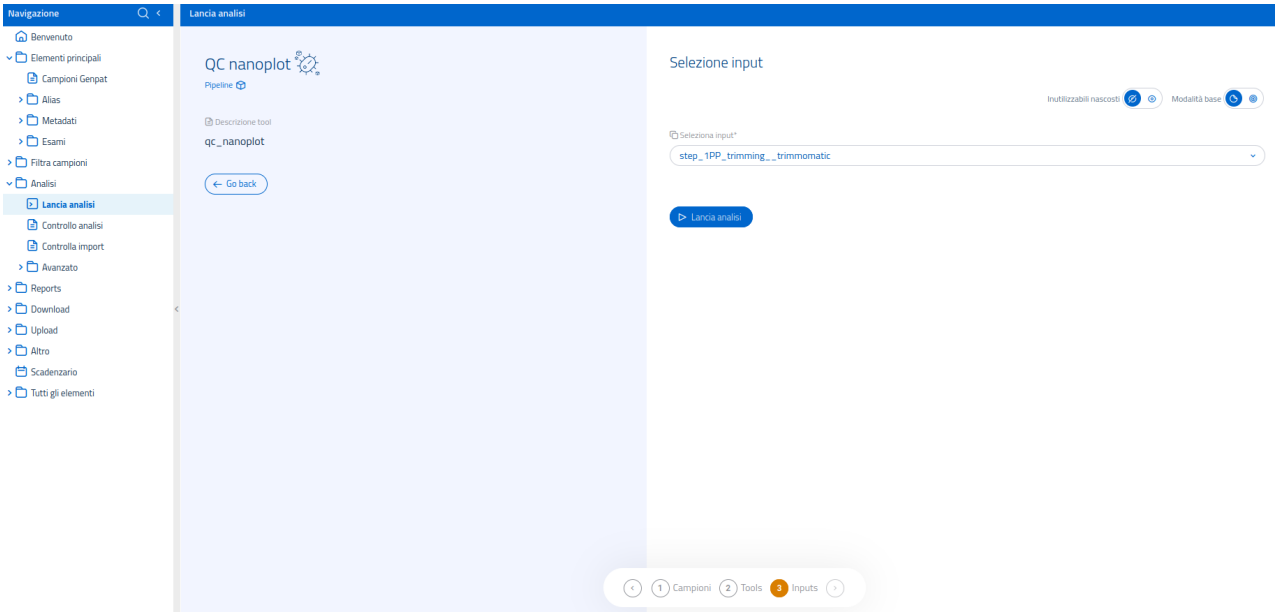
Run Nanoplot pipeline

The filter at the top of the run analyses page allows to display only the pipelines. Once the **Nanoplot** pipeline has been selected, a confirmation interface will be displayed.

The usable inputs are raw reads in FASTQ format (internal or external) or trimmed reads (or those originating from preprocessing analyses), **provided that they originate from Nanopore devices**:

- raw reads (step_0SQ_rawreads);
- trimmed reads (step_1PP_trimming);
- filtered reads (step_1PP_filtering);
- generated reads (step_1PP_generated);
- depleted reads (step_1PP_hostdepl);
- downsampled reads (step_1PP_downsampling).

The input selection interface provides the advanced input selection mode, to allow the use of inputs processed by different methods that were applied upstream in the analysis workflow.



Once the pipeline is launched, the page generates a link to the Check analyses section to allow a check of the processus status. The user will be notified by the system both when the pipeline is successfully launched and when the operation is completed.

Result folder

To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are two subfolders:

- **meta** (*metadata*), where the log and configuration files of the executed process are saved;
- **result**, where the files containing the results produced by the analysis are saved.

The table below shows the list of files produced by Nanoplot, along with some useful information.

File	Description	Location
DSXXXXXX-DTXXXXXX_ID_LengthvsQualityScatterPlot_dot.html	scatterplot of read length VS read medium quality	folder: result
DSXXXXXX-DTXXXXXX_ID_LengthvsQualityScatterPlot_kde.html	scatterplot of read length VS read medium quality	folder: result
DSXXXXXX-DTXXXXXX_ID_NanoPlot-report.html	metrics table in .html format	folder: result
DSXXXXXX-DTXXXXXX_ID_NanoStats.txt	metrics table in text format	folder: result
DSXXXXXX-DTXXXXXX_ID_Non_weightedHistogramReadlength.html	histogram of read lengths	folder: result
DSXXXXXX-DTXXXXXX_ID_Non_weightedLogTransformed_HistogramReadlength.html	histogram of read lengths after logarithmic transformation	folder: result
DSXXXXXX-DTXXXXXX_ID_WeightedHistogramReadlength.html	weighted histogram of read lengths	folder: result
DSXXXXXX-DTXXXXXX_ID_WeightedLogTransformed_HistogramReadlength.html	weighted histogram of read lengths after logarithmic transformation	folder: result
DSXXXXXX-DTXXXXXX_ID_Yield_By_Length.html	cumulative yield curve with respect to length of reads	folder: result

Note: most of tables and graphs produced by NanoPlot are in .html format. It is recommended to use a web browser for visualization.

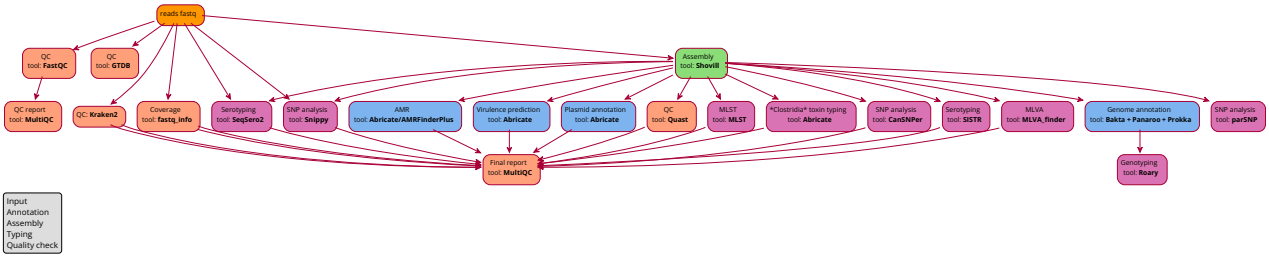
IAEA

WGSBAC Pipeline - IAEA (Salmonella)

Introduction

The WGSBAC pipeline performs genotyping and characterisation of bacteria isolates, using as input fastq.gz files from whole-genome-sequencing (WGS) for *Salmonella* samples. For a full list of available analyses in the workflow, please look up WGSBAC's GitLab page.

- **Official WGSBAC page:** https://gitlab.com/FLI_Bioinfo/WGSBAC
- **Authors:** Jörg Linde (@joerg.linde) and Mostafa Abdel-Glil (@Mostafa.Abdel-Glil)



How to run

Other microorganisms that could be analyzed with the original pipeline **are not supported in this installation**.

The WGSBAC pipeline runs a suite of tools that can be subdivided into:

1. Reads Quality Check (QC) and trimming
2. Sequence reconstruction
3. Assembly QC and contamination check
4. Genome annotation
5. Typing
6. Hierarchical clustering

A detailed list of all software in the pipeline is available at the official WGSBAC GitLab page.

After selection of the **WGSBAC Pipeline** for the desired microorganism, a confirmation interface will allow to go on to the parameter selection.

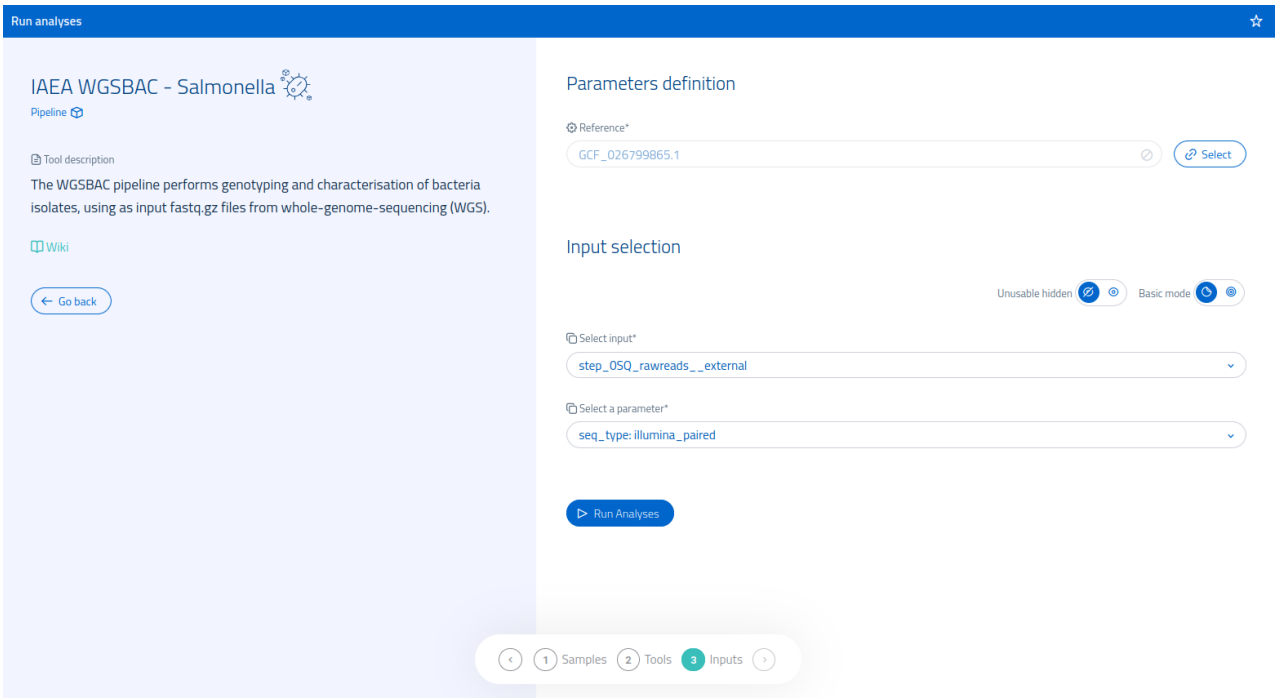
The required input are compressed fastq reads (`.fastq.gz`), which can be provided as:

- **step_0SQ_rawreads**
- **step_1PP_trimming**

Please note that the **step_1PP_trimming** input option is kept for the user's convenience, however the WGSBAC pipeline already performs trimming with **fastp**.

Furthermore, in order to perform some of the analyses of the pipeline, it is also required to specify a reference genome for the selected microorganism.

Important note: please remember that **the WGSBAC pipeline can only be used on Illumina paired end reads**. There is currently no support for other sequencing technologies.



After confirmation, the Run Analysis page will generate a link to the Check Analysis section. From there, users can visualize the execution status of the analysis. Automatic notifications will be sent after execution has finished.

Outputs

Output files from the WGSBAC Pipeline include outputs for each single analysis. Directories with results of each analysis are listed below. Please refer to WGSBAC's official GitLab page: (https://gitlab.com/FLI_Bioinfo/WGSBAC) for details on each software's results.

The result directories can be found by navigating into `work > 0 > XX > XXXXXXX` (XX and XXXXXXX represent variable names of directories: there will be 2 of each, one storing input files and the second storing the `wgsbac_results` direcory).

- For execution on *Salmonella* samples:

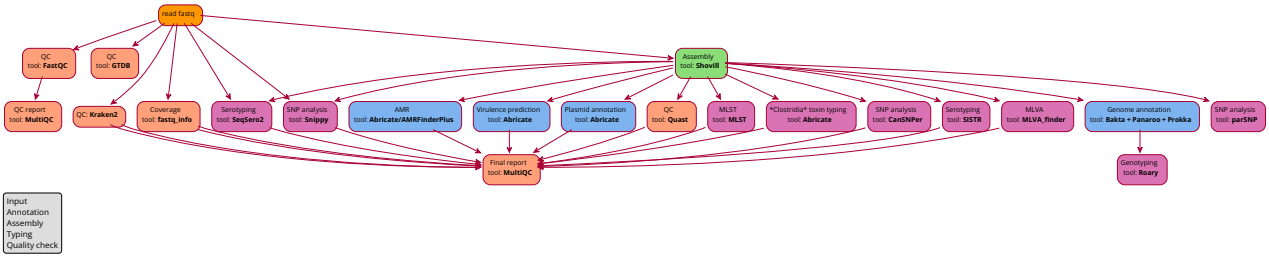
Analysis - Tool/DB	**Results Directory**
contamination check - confindr	`wgsbac_results > confindr`
coverage calculation	`wgsbac_results > coverage`
trimming and QC - fastp	`wgsbac_results > fastp`
decompressed fastq files	`wgsbac_results > fastq`
raw reads quality check - fastqc	`wgsbac_results > fastqc`
reference fasta	`wgsbac_results > finalAssembly`
input fastq.gz files	`wgsbac_results > input`
contamination check - kraken2	`wgsbac_results > kraken`
Multi Lucus Sequence Typing - mlst	`wgsbac_results > mlst`
QC - multiQC	`wgsbac_results > multiqc`
QC - fastQC	`wgsbac_results > multiqc_fastqc`
plasmid identification - plasmidfinder/platon	`wgsbac_results > plasmids`
assembly QC - Quast	`wgsbac_results > quast`
de novo assembly - Shovill	`wgsbac_results > rawassembly_res`
variant calling - Snippy	`wgsbac_results > snippy`
species identification - Sourmash	`wgsbac_results > sourmash`
virulence genes identification - SPI	`wgsbac_results > virulence`
Salmonella serotype prediction - seqSero	`seqSero`
Salmonella serovar prediction - sistr	`sistr`

WGSBAC Pipeline - IAEA (Brucella)

Introduction

The **Wgsbac** pipeline performs genotyping and characterisation of bacteria isolates, using as input fastq.gz files from whole-genome-sequencing (WGS). Here are implemented workflows for *Brucella* samples. For a full list of available analyses in the workflow, please consult the WGSBAC GitLab page.

- **WGSBAC official GitLab page:** https://gitlab.com/FLI_Bioinfo/WGSBAC
- **Authors:** Jörg Linde (@joerg.linde) and Mostafa Abdel-Glil (@Mostafa.Abdel-Glil).



Run the Wgsbac - IAEA pipeline

Other microorganisms, that could be analyzed with the original pipeline, **are not supported by the installation on this platform**.

The WGSBAC pipeline runs a suite of tools that can be divided into:

1. read quality check (QC) and trimming;
2. sequence reconstruction;
3. assembly QC and contamination check;
4. genome annotation;
5. typing;
6. hierarchical clustering.

A detailed list of all software in the pipeline is available at the WGSBAC GitLab official page.

After selecting the **WGSBAC** version for the microorganism of interest, a confirmation interface will allow you to proceed with the selection of the analysis parameters.

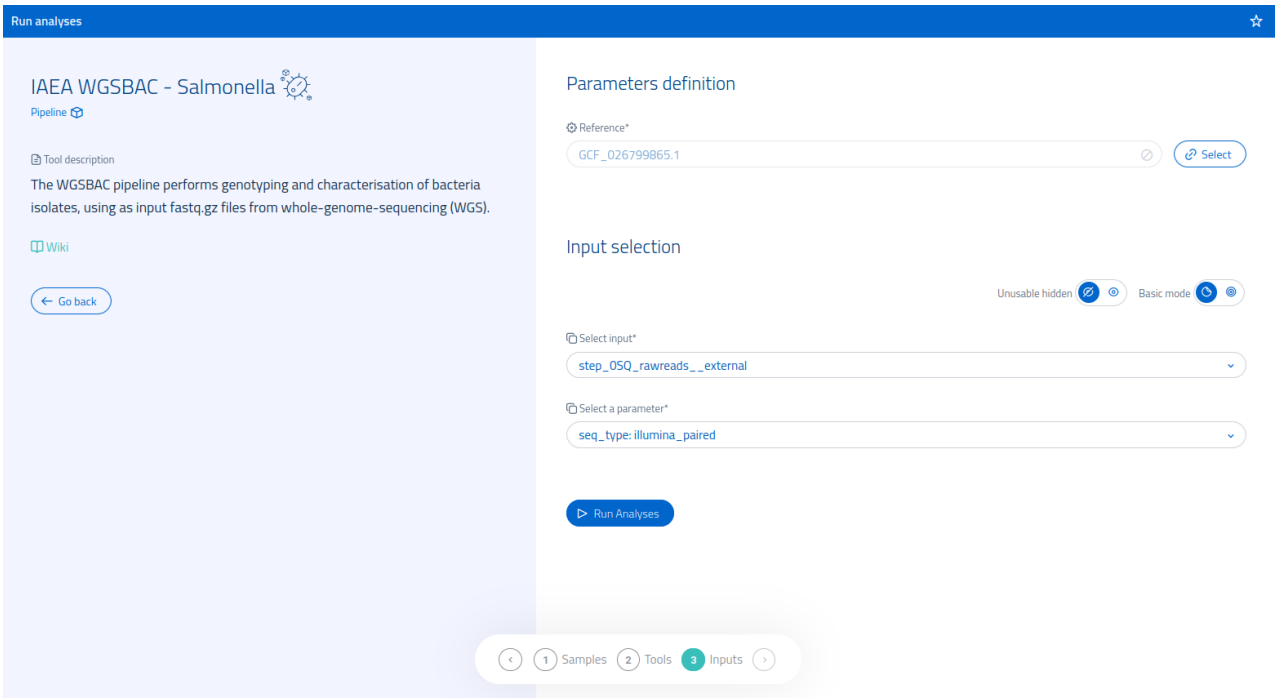
The required input are compressed fastq reads (`.fastq.gz`), which can be provided as:

- **step_0SQ_rawreads**
- **step_1PP_trimming**

Please note that the **step_1PP_trimming** is retained, as an input option, solely for user convenience as the WGSBAC pipeline performs trimming by default using **fastp**.

Furthermore, in order to perform some of the analyses of the pipeline, it is required to specify a *reference genome* for the selected microorganism.

Important note: please remember that **the WGSBAC pipeline can only be used on Illumina paired end reads**. Other sequencing technologies are currently not supported.



Once the pipeline is launched, the page generates a link to the Check analyses section to allow a check of the processus status. The user will be notified by the system both when the pipeline is successfully launched and when the operation is completed.

Result folder

Output files from the WGSBAC Pipeline include outputs for each single analysis. The result folders of each analysis are listed below. Please refer to WGSBAC's official GitLab page: (https://gitlab.com/FLI_Bioinfo/WGSBAC) for details on each software result.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, it is possible to find the outputs for each single analysis.

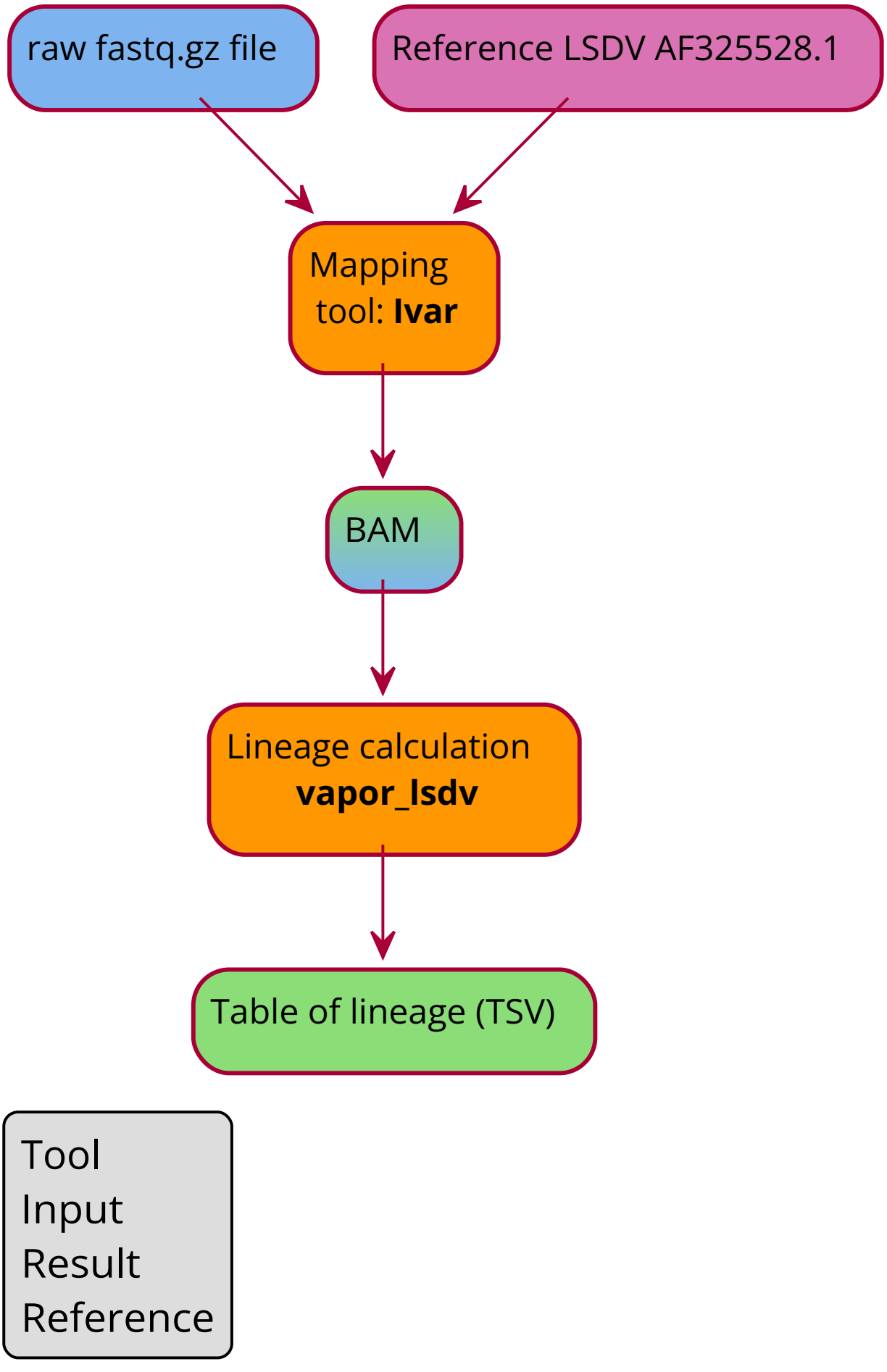
- For execution on *Brucella* samples:

Analysis - Tool/DB	Results Directory
contamination check - confindr	`wgsbac_results > confindr`
coverage calculation	`wgsbac_results > coverage`
trimming and QC - fastp	`wgsbac_results > fastp`
decompressed fastq files	`wgsbac_results > fastq`
raw reads quality check - fastqc	`wgsbac_results > fastqc`
reference fasta	`wgsbac_results > finalAssembly`
input fastq.gz files	`wgsbac_results > input`
contamination check - kraken2	`wgsbac_results > kraken`
Multi Lucus Sequence Typing - mlst	`wgsbac_results > mlst`
QC - multiQC	`wgsbac_results > multiqc`
QC - fastQC	`wgsbac_results > multiqc_fastqc`
plasmid identification - plasmidfinder/platon	`wgsbac_results > plasmids`
assembly QC - Quast	`wgsbac_results > quast`
de novo assembly - Shovill	`wgsbac_results > rawassembly_res`
variant calling - Snippy	`wgsbac_results > snippy`
species identification - Sourmash	`wgsbac_results > sourmash`

LSDV Characterization Pipeline

Introduction

This pipeline performs mapping and lineage calculation for Lumpy Skin Disease Virus (LSDV) samples; the reference used for mapping is fixed (`AF325528.1`).

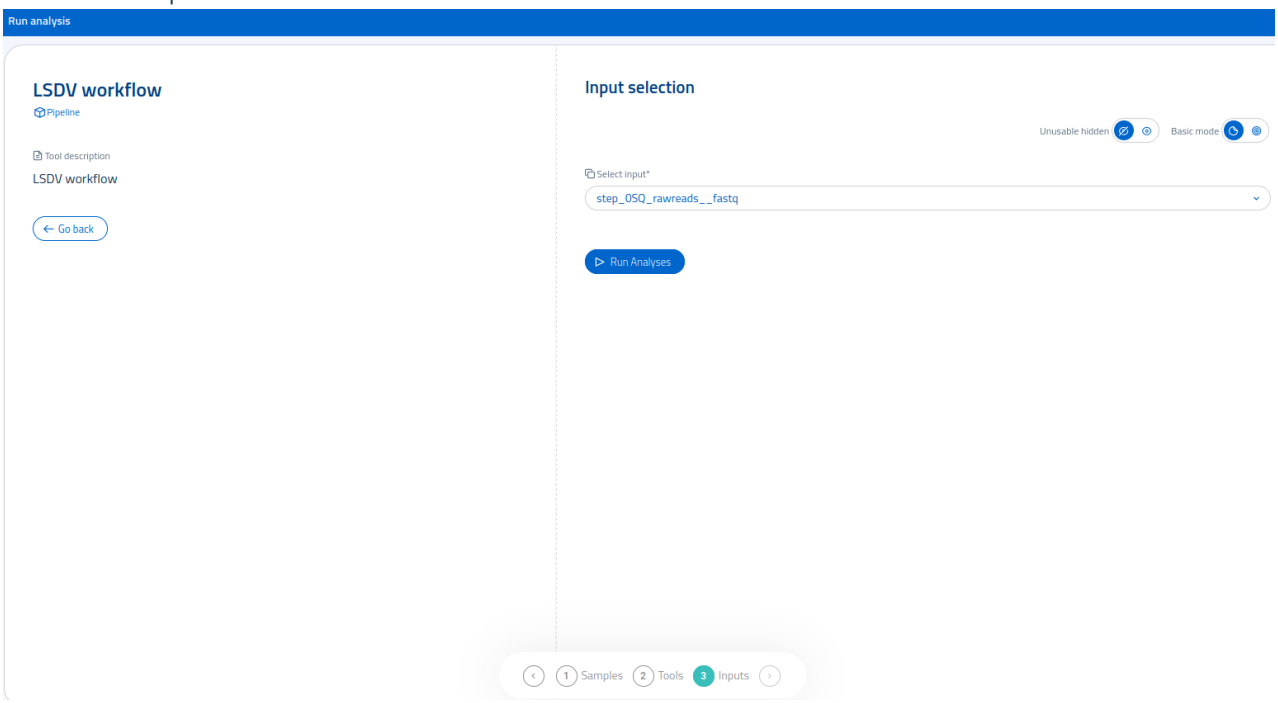

raw fastq.gz file
Reference LSDV AF325528.1
Mapping tool: **Ivar**
BAM
Lineage calculation **vapor_lsdv**
Table of lineage (TSV)
Tool
Input
Result
Reference

Run LSDV Characterization pipeline

The filter at the top of the run analyses page allows to display only the pipelines. Once the **LSDV Characterization** pipeline has been selected, a confirmation interface will allow proceeding to the selection of parameters.

The required input are **fastq reads** (`.fastq.gz`).

Screenshot of input selection



Once the pipeline is launched, the page generates a link to the Check analysis section to allow a check of the process status. The user will be notified by the system both when the pipeline is successfully launched and when the operation is completed.

Result folder

> To consult the guide on how to download files from the platform, please refer to the dedicated page.

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, there are 2 subfolders:

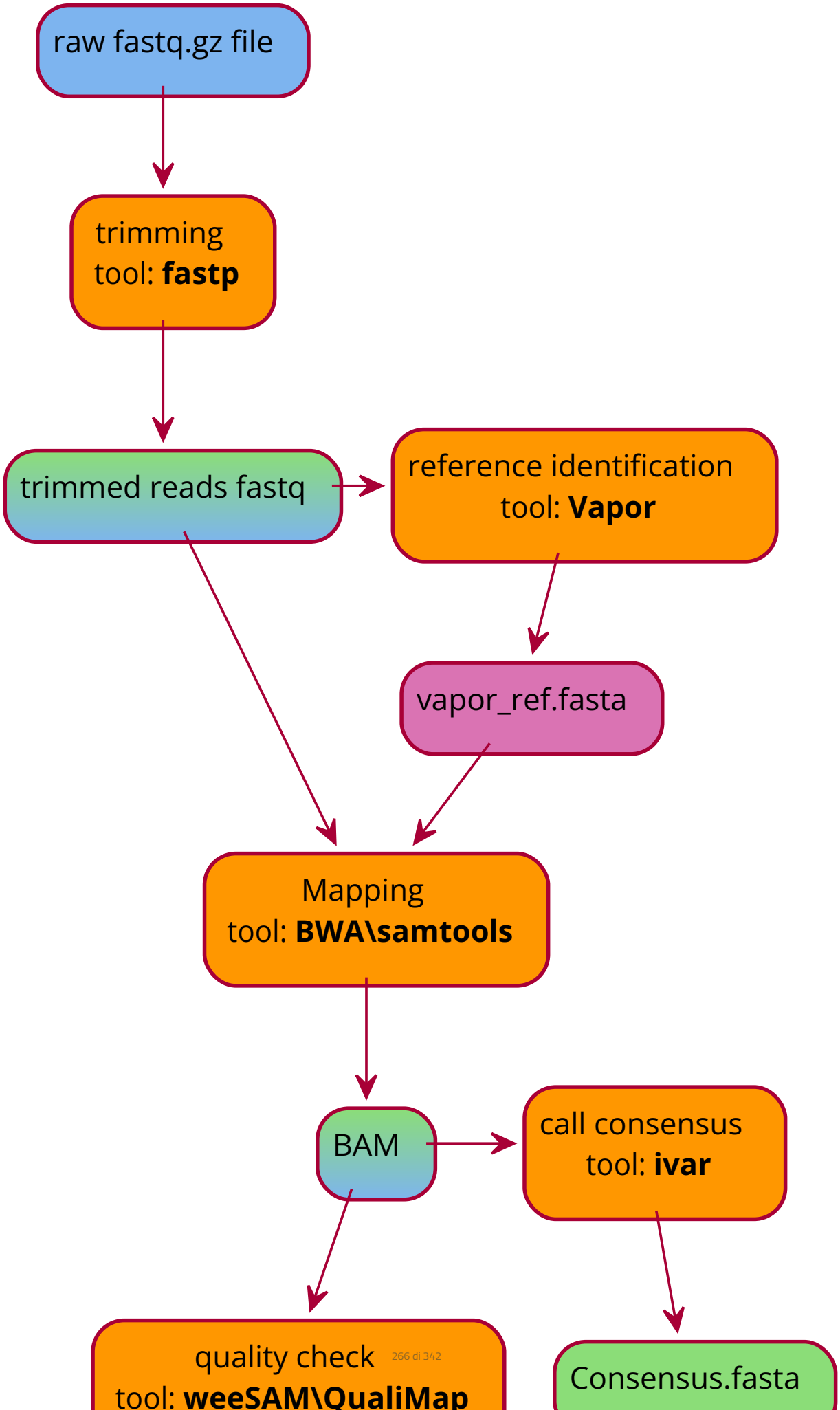
- **meta**: (*metadata*), where the log and configuration files of the executed process are saved;
- **result**: where the files containing the results produced by the analysis are saved.

AIV Pipeline - IAEA

Introduction

This pipeline processes *Avian Influenza Virus* samples, automatically determining the reference sequence for each segment and reconstructing the corresponding consensus sequence.

- **Official tool page:** To be added
- **Authors:** Graham Freimanis and Richard Orton


raw fastq.gz file
trimming
tool: **fastp**
trimmed reads fastq
reference identification
tool: **Vapor**
vapor_ref.fasta
Mapping
tool: **BWA\samtools**
BAM
call consensus
tool: **ivar**
quality check
tool: **weeSAM\QualiMap**
Consensus.fasta

Run IAEA AIV pipeline

Once the **IAEA AIV pipeline** analysis is selected on the run analysis page, the system will switch to a confirmation interface that allows you to proceed with the parameter selection.

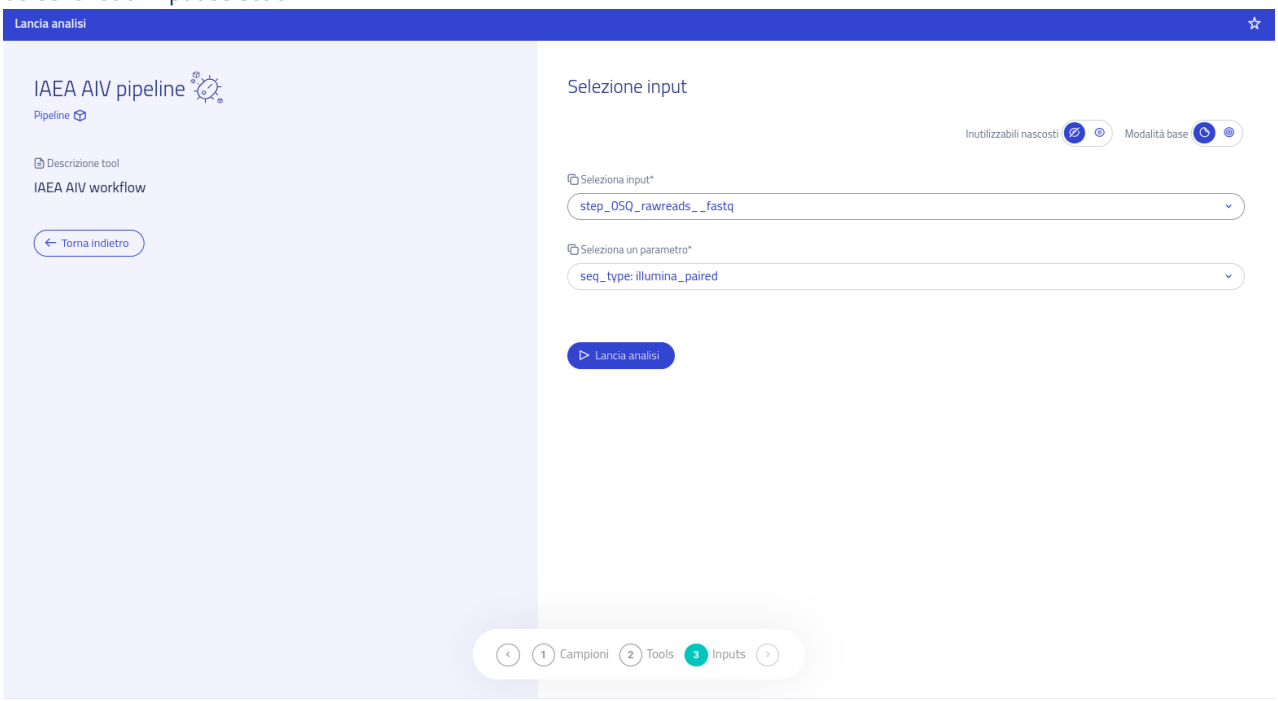
The **IAEA AIV pipeline** involves 3 main steps:

1. read trimming;
2. identification of the best reference;
3. lineage calculation and mapping.

The required inputs are FASTQ raw reads (`.fastq.gz`):

- **step_0SQ_rawreads**

Screenshot of input selection



Once the analysis is launched, the page generates a link to the Check analysis section to allow a check of the processus status. The user will be notified by the system both when the pipeline is successfully launched and when the operation is completed.

Result folder

The **Result folder** can be accessed by clicking the link in the analysis card, in the *Result data* section. Within the following `results` folder, it is possible to find the files produced by the "IAEA AIV" pipeline, which will be the same as those produced by the analyses that compose it:

- *Trimming* with fastp;
- Mapping with Ivar;
- Lineage calculation with Vapor.

Data visualization

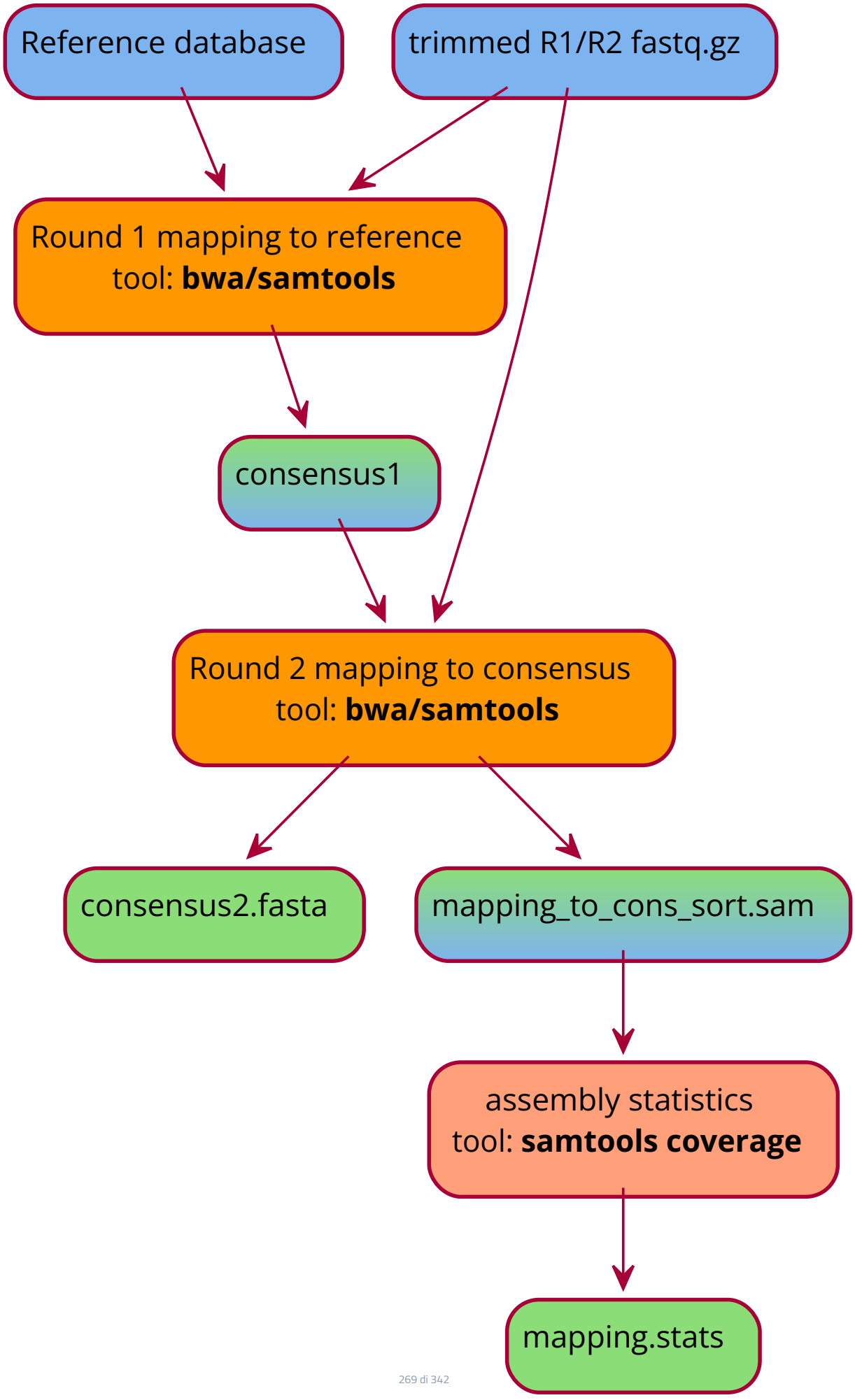
The pipeline produces some report files in .html format that can be viewed in the browser.

PPRV Pipeline - IAEA

Introduction

This pipeline processes *Peste des Petits Ruminants Virus* (PPRV) samples, automatically determining the reference sequence and reconstructing the corresponding consensus sequence.

- **Official tool page:** https://github.com/CourcelleM/PPRreads_ASSEMBLY/tree/main
- **Authors:** CIRAD laboratories https://www.ppr-labs-oie-network.org/en/materials-and-protocols/pprv-sequence-datasets


Reference database
trimmed R1/R2 fastq.gz
Round 1 mapping to reference
tool: **bwa/samtools**
consensus1
Round 2 mapping to consensus
tool: **bwa/samtools**
consensus2.fasta
mapping_to_cons_sort.sam
assembly statistics
tool: **samtools coverage**
mapping.stats

How to run the pipeline

Once the **PPRV** analysis is selected on the run analysis page, the system will switch to a confirmation interface that allows you to proceed with the parameter selection.

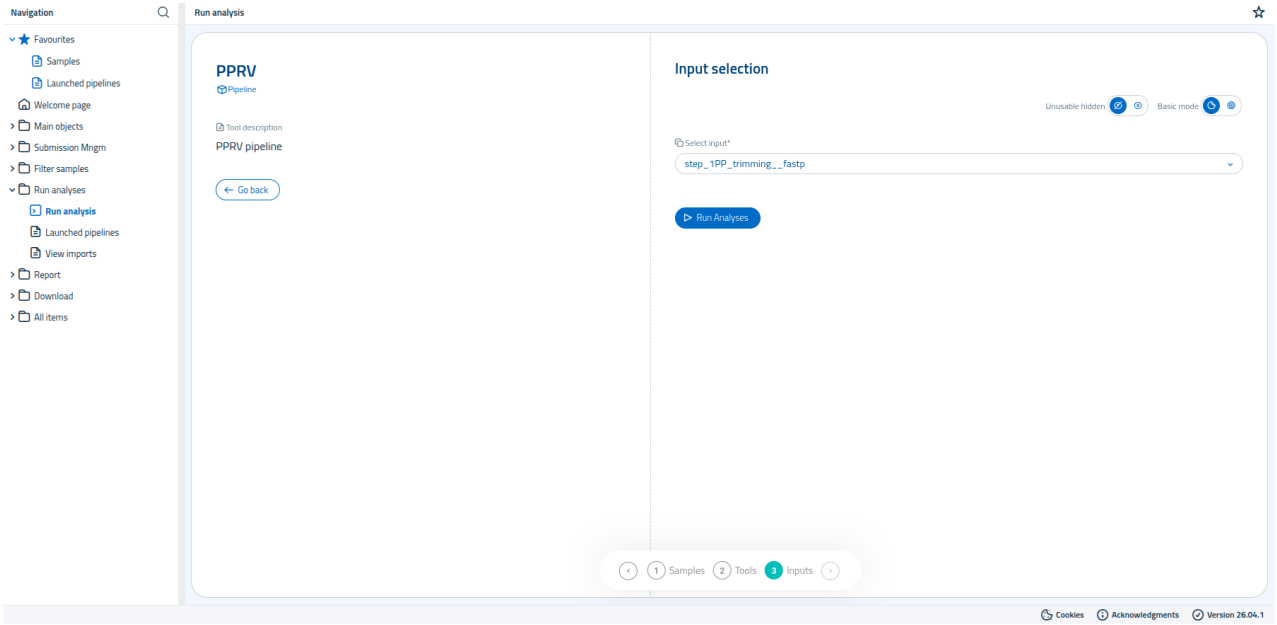
The **PPRV pipeline** consists of two main stages:

1. identification of the best reference sequence;
2. read mapping and consensus sequence generation.

The required inputs are *trimmed reads* in FASTQ format (`.fastq.gz`):

- **step_1PP_trimming**

Screenshot of input selection



Once the pipeline is launched, the page generates a link to the Check analyses section to allow a check of the processus status. The user will be notified by the system both when the pipeline is successfully launched and when the operation is completed.

Result folder

The **Result folder** can be browsed by clicking the link available within the analysis card, in the *Result data* section. Inside the subsequent `results` folder, it is possible to find the outputs for the "PPRV" pipeline.

6. Reports

6.1 Reports

From the side menu, tools are available to view reports. The standard reporting mode requires either a collection of IDs in the cart or an active tag to specify the items on which to generate reports.

The **Load utilities** button (top-right corner) provides a quick tool to fill the cart using samples from a .csv file or by specifying single codes in the designated field.

Displaying reports with an active tag

The following videotutorial shows the procedure for viewing reports by using an active tag.

Note: For privacy and data security reasons, the video was created using the Cohesive demo application; therefore, the interface may show some minor differences.

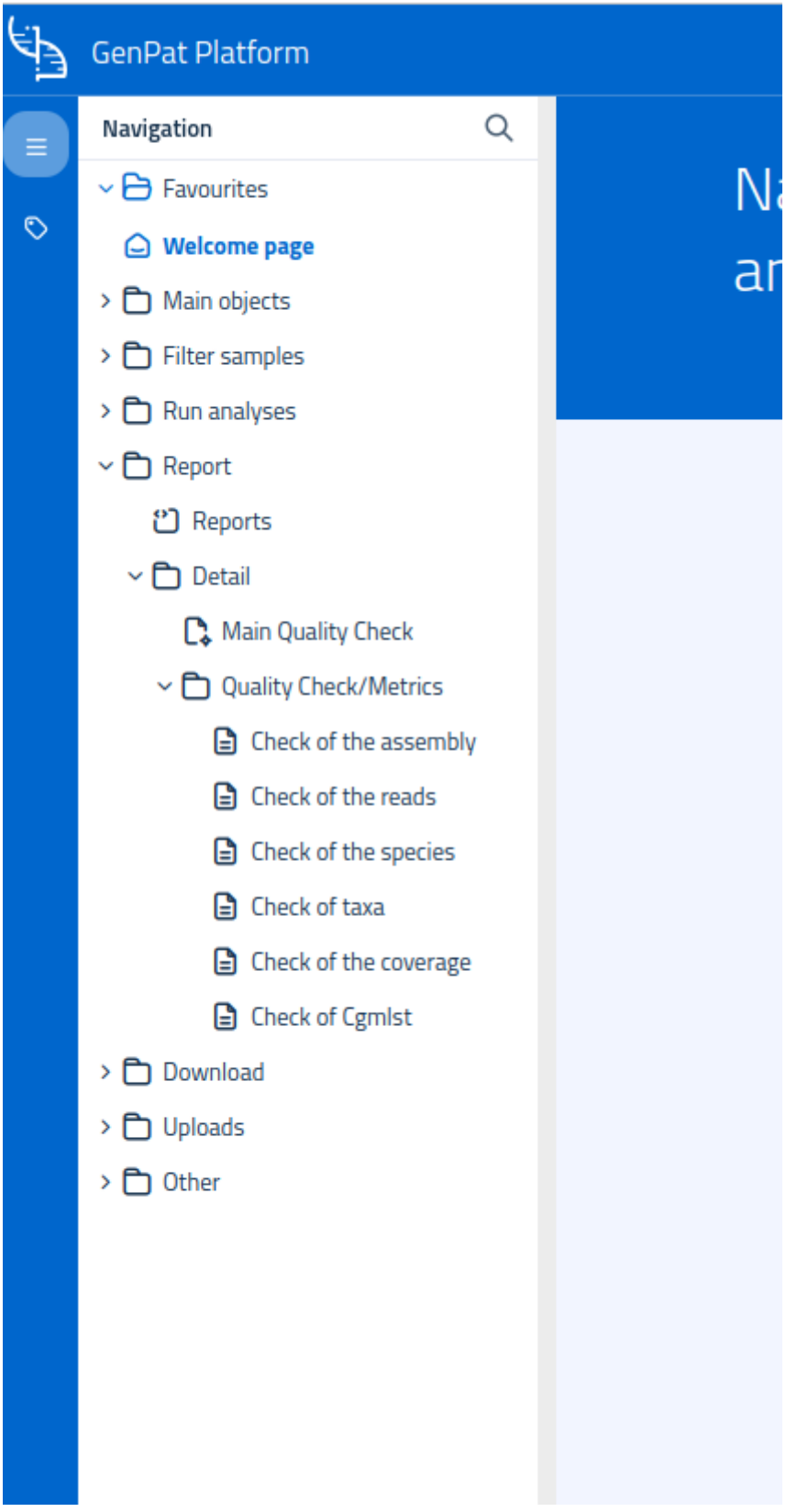
Viewing reports with an active tag

6.2 Checks

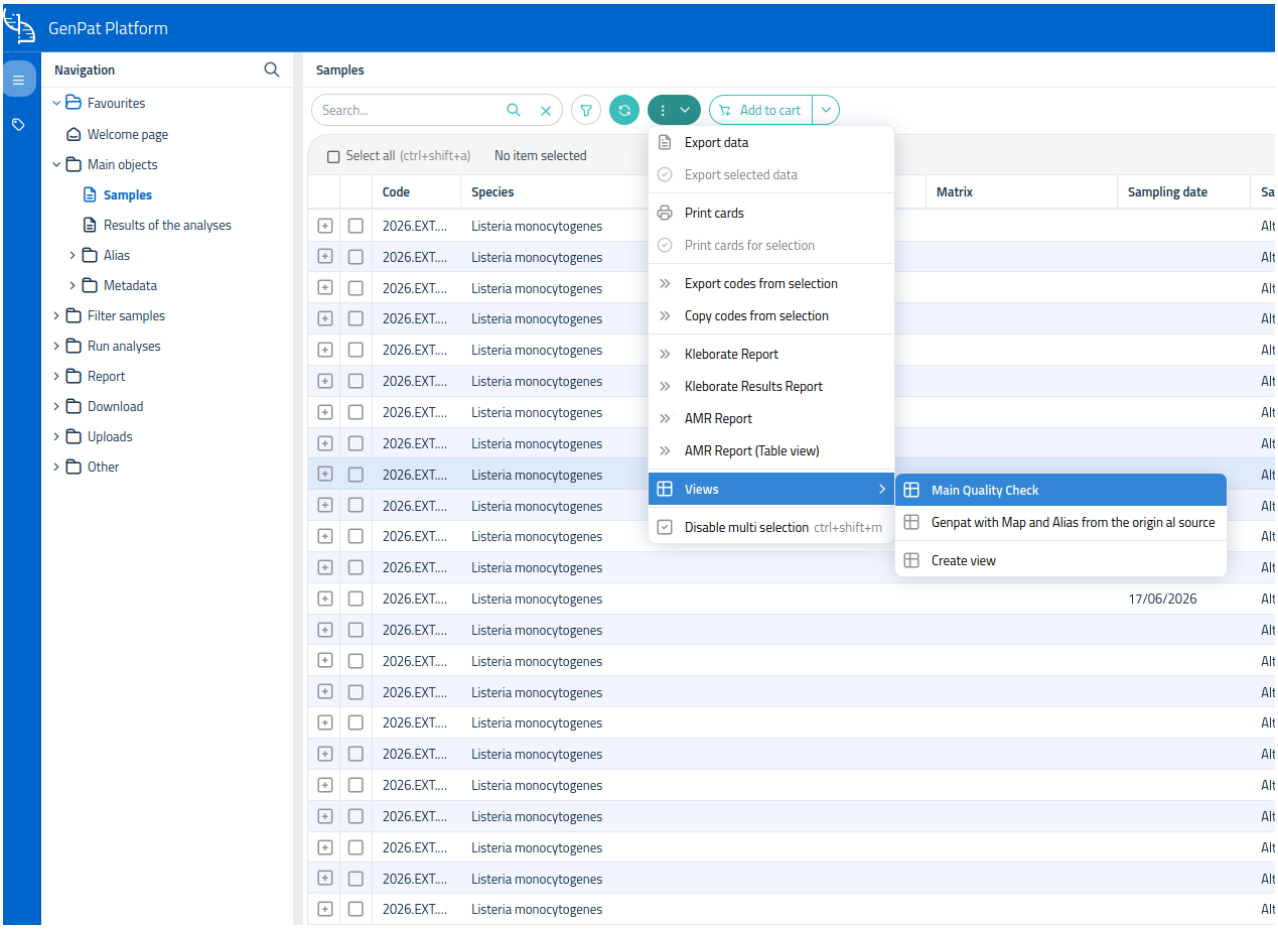
6.2.1 Tables of checks

Quality check tables are available from the *Report* section of the navigation menu.

For each available quality check, information can be displayed as a *Report* through the following path: `Report > Detail > Checks/Metrics`.



Alternatively, you can use the context menu available on the *Samples* page (`Main objects > Samples`), select `Views` , and choose *Main Quality Check*.

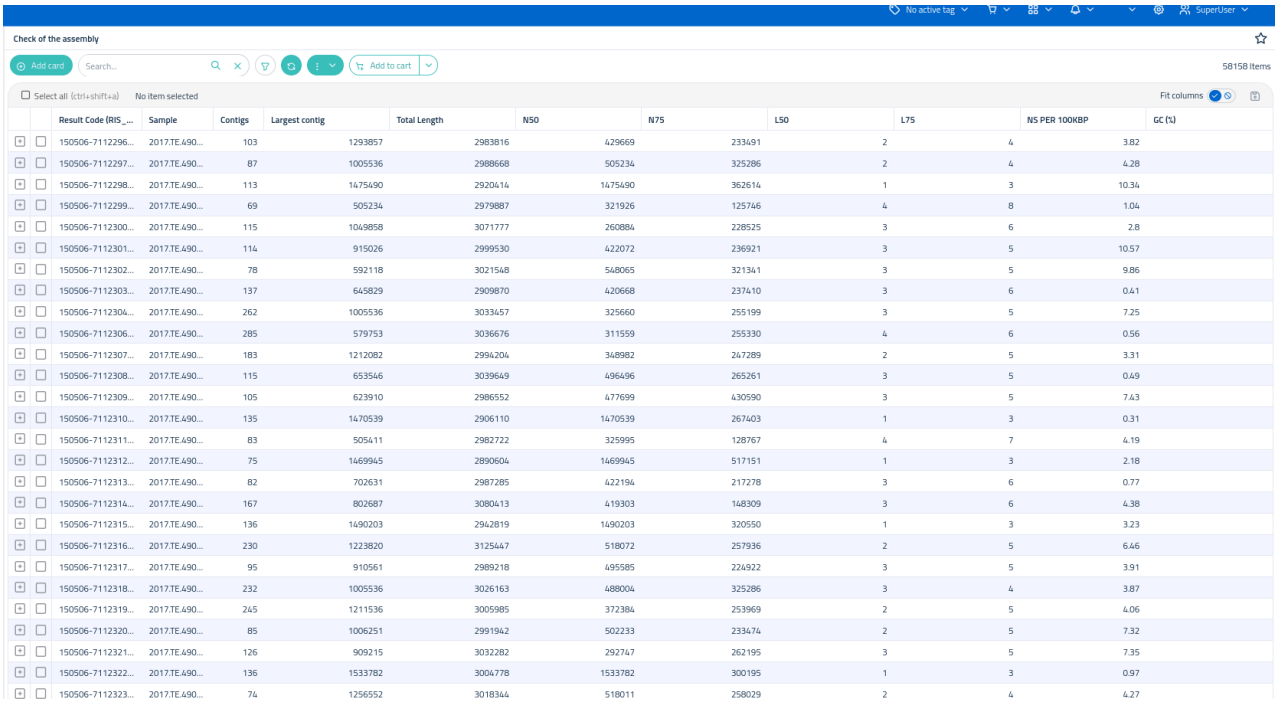


The available quality checks are:

- Check of the Assembly
- Check of the Reads
- Check of the Species
- Check of Taxa
- Check of the Coverage
- Cgmlst check
- Kleborate
- Quality checks for ML-based classification.

6.2.2 Check of the assembly

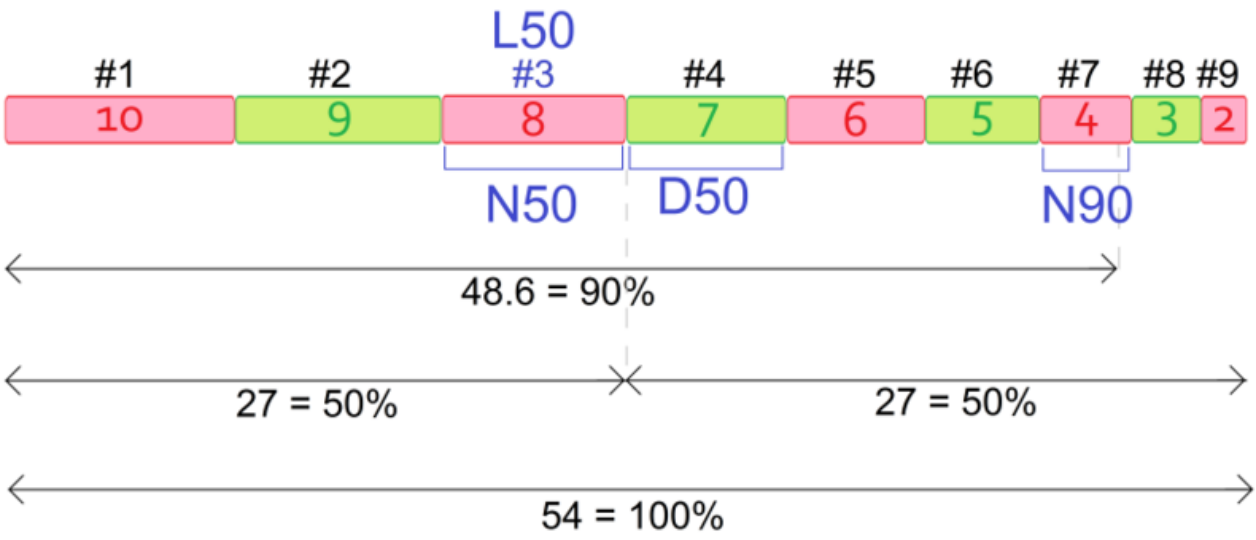
The *Check of the Assembly* table provides information about the quality of *de novo* assemblies.



Result Code (RIS_...	Sample	Contigs	Largest contig	Total Length	N50	N75	L50	L75	NS PER 100KBP	GC (%)
150506-7112296...	2017.TE.490...	103	1293857	2983816	429669	233491	2	4	3.82	
150506-7112297...	2017.TE.490...	87	1005536	2988668	505234	325286	2	4	4.28	
150506-7112298...	2017.TE.490...	113	1475490	2920414	1475490	362614	1	3	10.34	
150506-7112299...	2017.TE.490...	69	505234	2979887	321926	125746	4	8	1.04	
150506-7112300...	2017.TE.490...	115	1049858	3071777	260884	228525	3	6	2.8	
150506-7112301...	2017.TE.490...	114	915026	2999530	422072	236921	3	5	10.57	
150506-7112302...	2017.TE.490...	78	592118	3021548	548065	321341	3	5	9.86	
150506-7112303...	2017.TE.490...	137	645829	2909870	420668	237410	3	6	0.41	
150506-7112304...	2017.TE.490...	262	1005536	3033457	325660	255199	3	5	7.25	
150506-7112306...	2017.TE.490...	285	579753	3036676	311559	255330	4	6	0.56	
150506-7112307...	2017.TE.490...	183	1212082	2994204	348982	247289	2	5	3.31	
150506-7112308...	2017.TE.490...	115	653546	3039649	496496	265261	3	5	0.49	
150506-7112309...	2017.TE.490...	105	623910	2986552	477699	430590	3	5	7.43	
150506-7112310...	2017.TE.490...	135	1470539	2906110	1470539	267403	1	3	0.31	
150506-7112311...	2017.TE.490...	83	505411	2982722	325995	128767	4	7	4.19	
150506-7112312...	2017.TE.490...	75	1469945	2890604	1469945	517151	1	3	2.18	
150506-7112313...	2017.TE.490...	82	702631	2987285	422194	217278	3	6	0.77	
150506-7112314...	2017.TE.490...	167	802687	3080413	419303	148309	3	6	4.38	
150506-7112315...	2017.TE.490...	136	1490203	2942819	1490203	320550	1	3	3.23	
150506-7112316...	2017.TE.490...	230	1223820	3125447	518072	257936	2	5	6.46	
150506-7112317...	2017.TE.490...	95	910561	2989218	495585	224922	3	5	3.91	
150506-7112318...	2017.TE.490...	232	1005536	3026163	488004	325286	3	4	3.87	
150506-7112319...	2017.TE.490...	245	1211536	3005985	372384	253969	2	5	4.06	
150506-7112320...	2017.TE.490...	85	1006251	2991942	502233	233474	2	5	7.32	
150506-7112321...	2017.TE.490...	126	909215	3032282	292747	262195	3	5	7.35	
150506-7112322...	2017.TE.490...	136	1533782	3004778	1533782	300195	1	3	0.97	
150506-7112323...	2017.TE.490...	74	1256552	3018344	518011	258029	2	4	4.27	

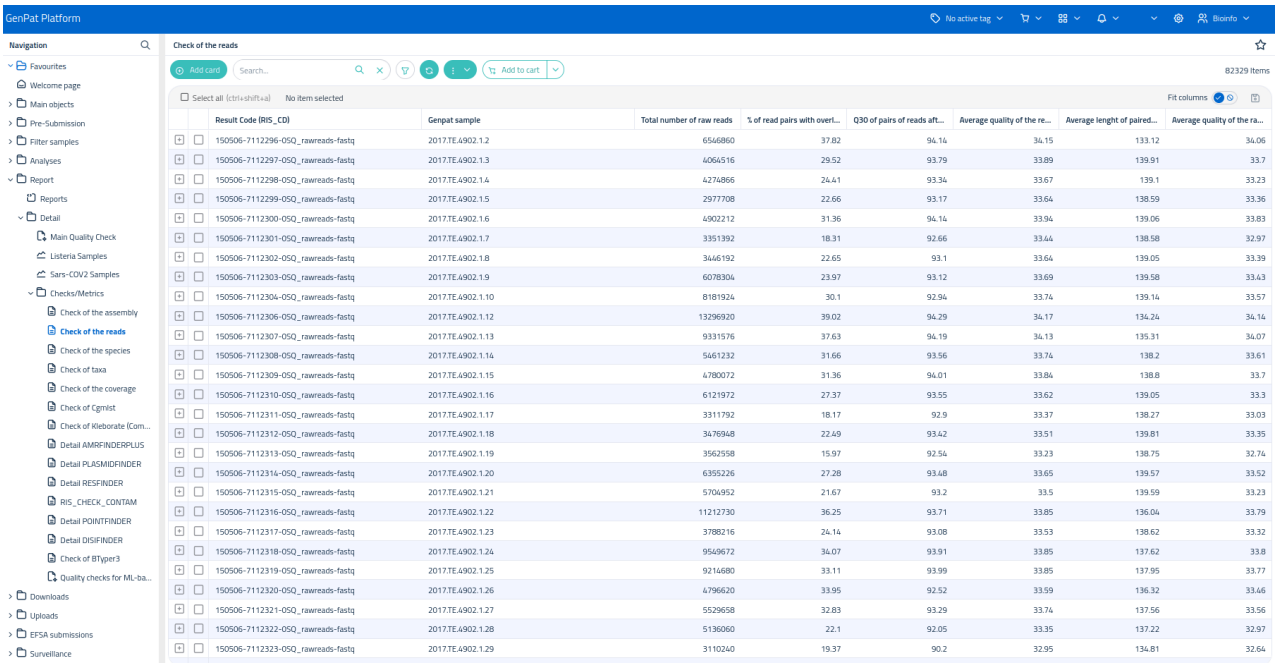
The metadata columns, in addition to **Code** and **Description** which refer to the identifier of the analysis result, are: * **Result**: the code (RIS_CD) of the analysis result to which the statistics refer; * **Sample**: sample platform code (GENPAT_CD); * **Contigs**: total number of contigs produced by the assembler; * **Longest contig**: length of the longest contig produced by the assembler; * **Total length**: total number of nucleotides that make up the sequence reconstructed by the assembly; * **N50**: lenght of the contig at which 50% of the total assembly lenght is reached, when contigs are ordered by decreasing lenght; * **N75**: lenght of the contig at which 75% of the total assembly lenght is reached, when contigs are ordered by decreasing lenght; * **L50**: number of contigs necessary to reach 50% of the sequence, when contigs are ordered by decreasing lenght; * **L75**: number of contigs necessary to reach 75% of the sequence, when contigs are ordered by decreasing lenght; * **NS per 100Kbp**: number of unknown nucleotides (Ns) in the assembled sequence per 100Kbp.

N50 and related metrics: the N50 is the sequence length of the shortest contig at 50% of the total genome length (N50 on Wikipedia). Similarly, N75 is the sequence length of the shortest contig at 75% of the total genome length.

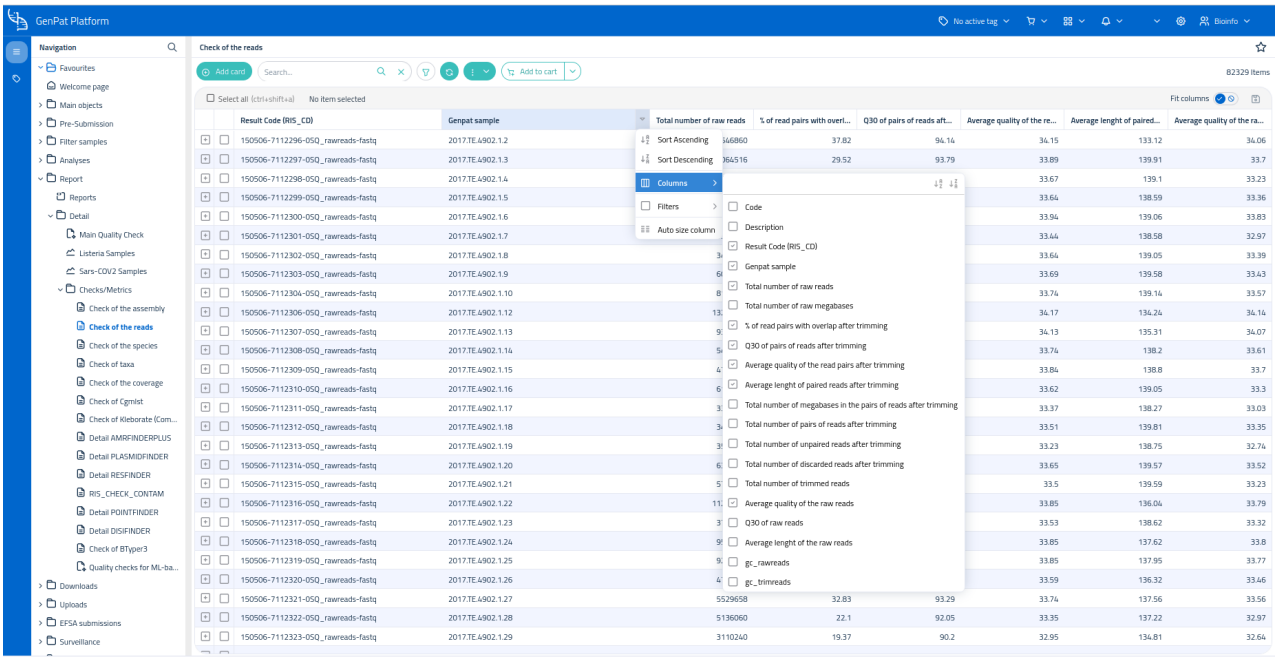

L50
#1
#2
#3
#4
#5
#6
#7
#8
#9
10
9
8
7
6
5
4
3
2
N50
D50
N90
48.6 = 90%
27 = 50%
27 = 50%
54 = 100%

6.2.3 Check of the reads

The *Check of the reads* table provides information and statistics about the quality of raw and trimmed reads.

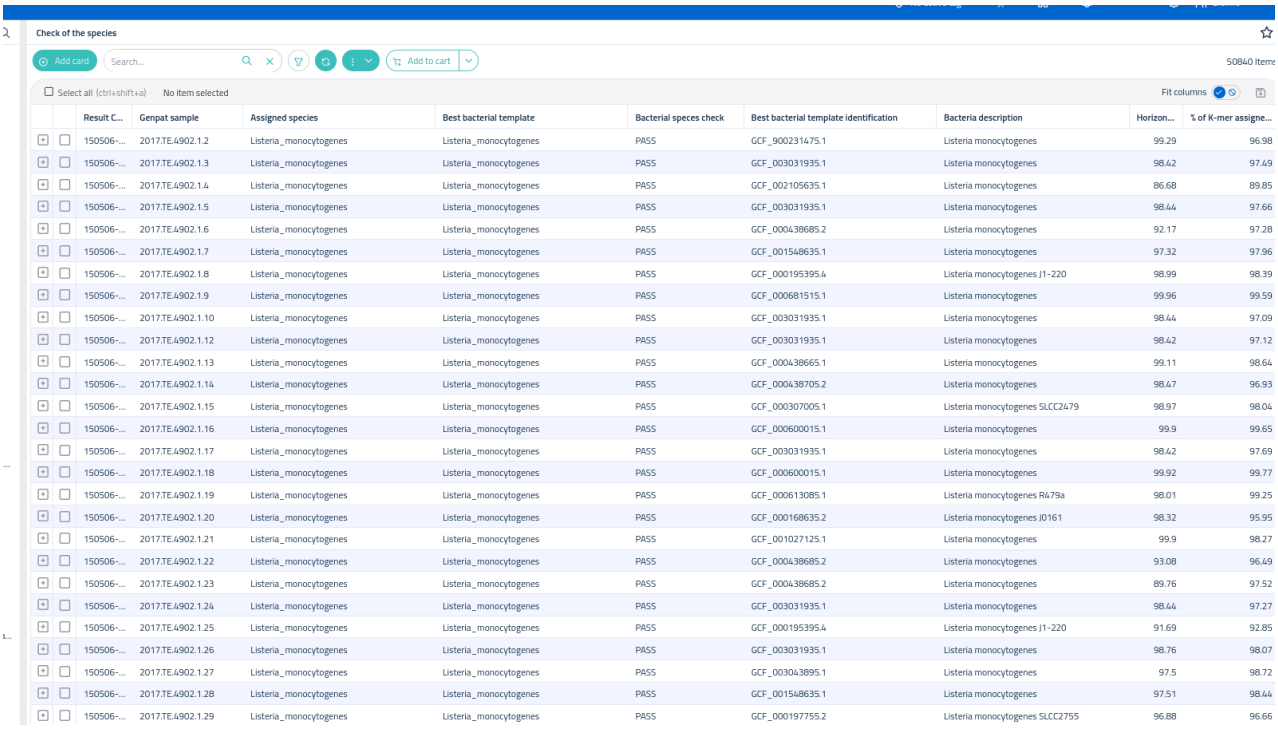


The metadata columns, in addition to **Codes** and **Description** which refer to the identifier of the analyses result, are: * **Result**: the code (RIS_CD) of the analysis result; * **Genpat sample**: sample Genpat code (GENPAT_CD); * **Total number of raw reads**: total number of reads before trimming; * **Total number of raw megabases**: total number of nucleotides in the raw sequence (in Mbp); * **% of read pairs with overlap after trimming**: percentage of partial or complete overlap between read pairs after trimming; * **Pairs of reads with Q30 after trimming**: read pairs with a quality score (Phred score, Q) >= 30 after trimming; * **Average quality of the read pairs after trimming**: average quality of the paired trimmed reads;
*Average length of paired reads after trimming: average lenght of the pairs of trimmed reads; * Total number of megabases in the pairs of reads after trimming: total number of sequenced bases in paired reads remaining after trimming (in Mbp); * Total number of pairs of reads after trimming: total number of trimmed paired-end reads; * Total number of unpaired reads after trimming: total number of unpaired reads remaining after trimming; * Total number of discarded reads after trimming: amount of the discarded reads after trimming; * Total number of trimmed reads: total amount of reads after trimming; * Average quality of the raw reads: average quality of the reads before trimming (raw reads); * Raw reads with Q30: percentage of the raw reads with a quality score (Phred score, Q) >= 30; * Average length of the raw reads; * Warnings: code of the checks whose results are below the threshold values; * Failed**: failed control code.


GenPat Platform
Navigation
Check of the reads
Add card
Add to cart
82329 items
Select all (ctrl+shift+a)
No item selected
Fit columns
Result Code (RIS_CD)
Genpat sample
Total number of raw reads
% of read pairs with overl...
Q30 of pairs of reads aft...
Average quality of the re...
Average lenght of paired...
Average quality of the ra...
Sort Ascending
Sort Descending
Columns
Filters
Auto size column
Code
Description
Result Code (RIS_CD)
Genpat sample
Total number of raw reads
Total number of raw megabases
% of read pairs with overlap after trimming
Q30 of pairs of reads after trimming
Average quality of the read pairs after trimming
Average lenght of paired reads after trimming
Total number of megabases in the pairs of reads after trimming
Total number of pairs of reads after trimming
Total number of unpaired reads after trimming
Total number of discarded reads after trimming
Total number of trimmed reads
Average quality of the raw reads
Q30 of raw reads
Average lenght of the raw reads
gc_rawreads
gc_trimreads

6.2.4 Check of the species

The *Check of the species* table provides information about the species checks performed *in silico* typing.



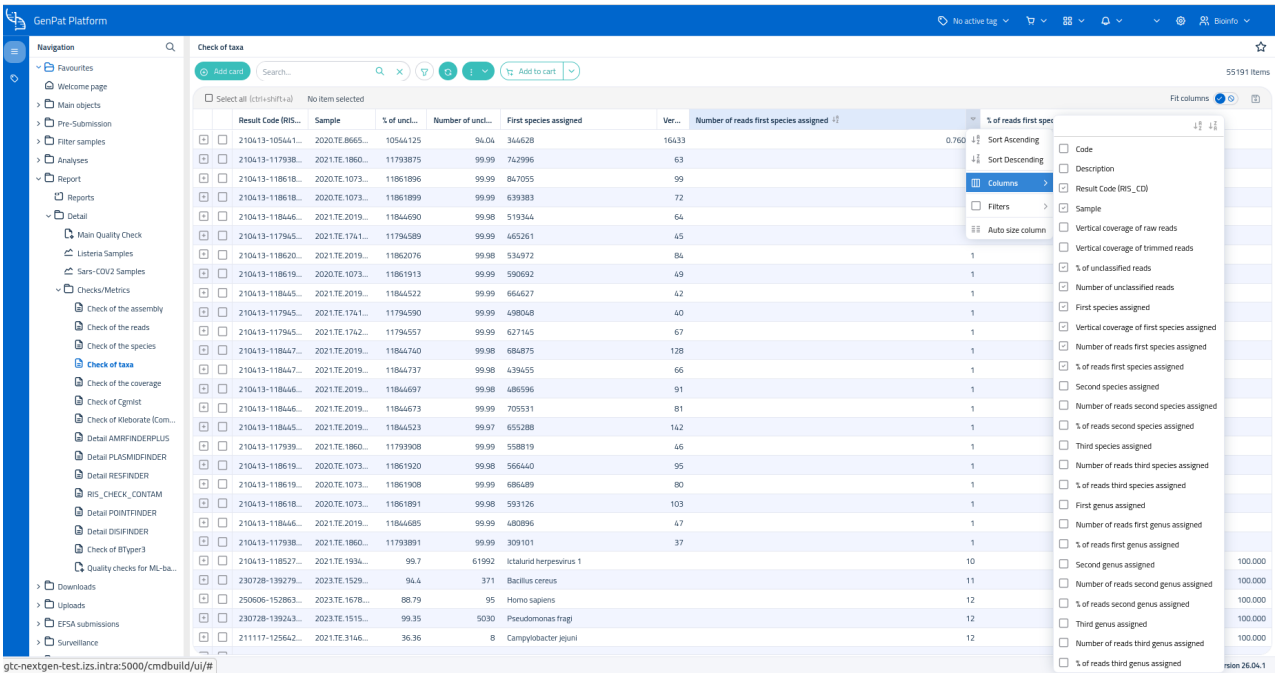
Result C...	Genpat sample	Assigned species	Best bacterial template	Bacterial speces check	Best bacterial template identification	Bacteria description	Horizon...	% of K-mer assigne...
150506-...	2017.TE.4902.1.2	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_900231475.1	Listeria monocytogenes	99.29	96.98
150506-...	2017.TE.4902.1.3	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_003031935.1	Listeria monocytogenes	98.42	97.49
150506-...	2017.TE.4902.1.4	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_002105635.1	Listeria monocytogenes	86.68	89.85
150506-...	2017.TE.4902.1.5	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_003031935.1	Listeria monocytogenes	98.44	97.66
150506-...	2017.TE.4902.1.6	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_000438685.2	Listeria monocytogenes	92.17	97.28
150506-...	2017.TE.4902.1.7	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_001548635.1	Listeria monocytogenes	97.32	97.96
150506-...	2017.TE.4902.1.8	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_000195395.4	Listeria monocytogenes J1-220	98.99	98.39
150506-...	2017.TE.4902.1.9	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_000681515.1	Listeria monocytogenes	99.96	99.59
150506-...	2017.TE.4902.1.10	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_003031935.1	Listeria monocytogenes	98.44	97.09
150506-...	2017.TE.4902.1.12	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_003031935.1	Listeria monocytogenes	98.42	97.12
150506-...	2017.TE.4902.1.13	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_000438665.1	Listeria monocytogenes	99.11	98.64
150506-...	2017.TE.4902.1.14	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_000438705.2	Listeria monocytogenes	98.47	96.93
150506-...	2017.TE.4902.1.15	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_000307005.1	Listeria monocytogenes SLCC2479	98.97	98.04
150506-...	2017.TE.4902.1.16	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_000600015.1	Listeria monocytogenes	99.9	99.65
150506-...	2017.TE.4902.1.17	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_003031935.1	Listeria monocytogenes	98.42	97.69
150506-...	2017.TE.4902.1.18	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_000600015.1	Listeria monocytogenes	99.92	99.77
150506-...	2017.TE.4902.1.19	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_000613085.1	Listeria monocytogenes R479a	98.01	99.25
150506-...	2017.TE.4902.1.20	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_000168635.2	Listeria monocytogenes J0161	98.32	95.95
150506-...	2017.TE.4902.1.21	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_001027125.1	Listeria monocytogenes	99.9	98.27
150506-...	2017.TE.4902.1.22	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_000438685.2	Listeria monocytogenes	93.08	96.49
150506-...	2017.TE.4902.1.23	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_000438685.2	Listeria monocytogenes	89.76	97.52
150506-...	2017.TE.4902.1.24	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_003031935.1	Listeria monocytogenes	98.44	97.27
150506-...	2017.TE.4902.1.25	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_000195395.4	Listeria monocytogenes J1-220	91.69	92.85
150506-...	2017.TE.4902.1.26	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_003031935.1	Listeria monocytogenes	98.76	98.07
150506-...	2017.TE.4902.1.27	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_003043895.1	Listeria monocytogenes	97.5	98.72
150506-...	2017.TE.4902.1.28	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_001548635.1	Listeria monocytogenes	97.51	98.44
150506-...	2017.TE.4902.1.29	Listeria_monocytogenes	Listeria_monocytogenes	PASS	GCF_000197755.2	Listeria monocytogenes SLCC2755	96.88	96.66

The metadata columns, in addition to **Code** and **Description** which refer to the identifier of the analysis result, are:

- **Result**: the code (RIS_CD) of the analysis result;
- **Genpat sample**: sample Genpat code (GENPAT_CD);
- **Assigned species**: the bacterial species assigned by KmerFinder;
- **Best bacterial template**: the best bacterial template assigned by kmerfinder;
- **Bacterial speces check**: check for the presence of bacterial-origin sequences;
- **Best bacterial template identification**: the ID of the best bacterial reference in Genpat;
- **Bacteria description**: name and type of the bacteria;
- **Horizontal coverage of the best bacterial template**: percentage of the sequence obtained using the best bacterial template identified as reference;
- **% of K-mer assigned to the best bacterial template**: percentage of k-mer assigned to the best bacterial template.

6.2.5 Check Taxa

The *Check of taxa* table provides information related to possible contaminations present in the sample, assessed through specific *in silico* analyses.

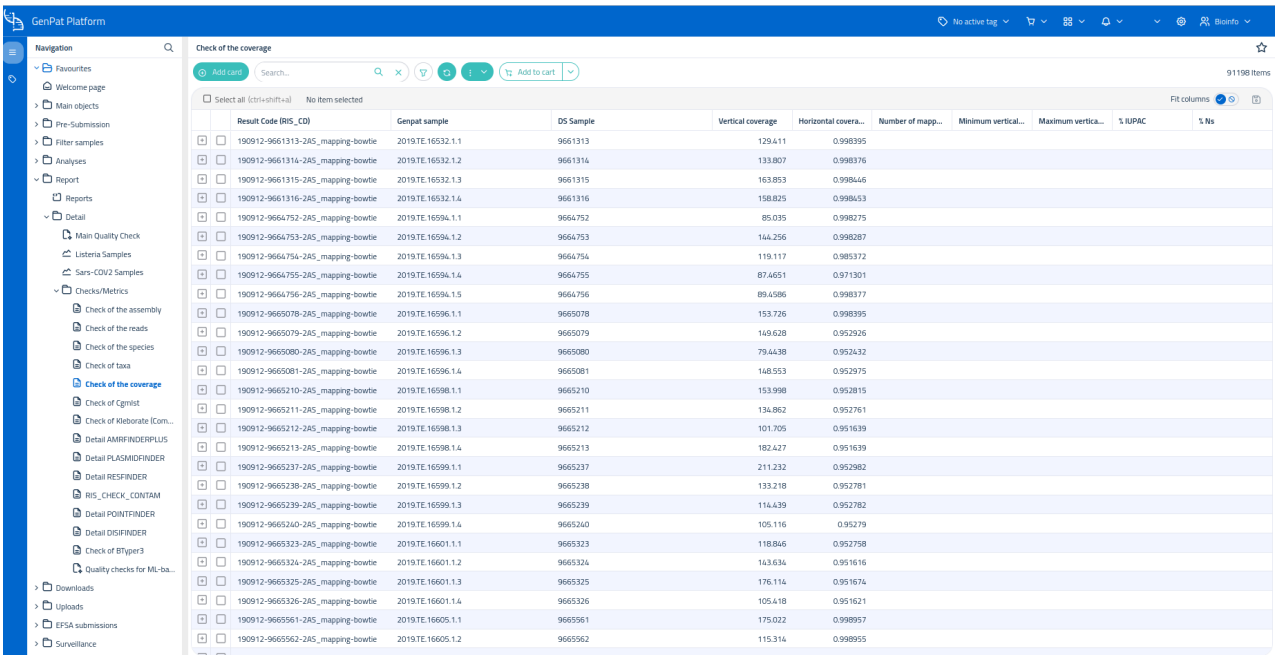


The metadata columns, in addition to **Codes** and **Description** which refer to the identifier of the analyses result, are:

- **Result**: the code (RIS_CD) of the analysis result;
- **Sample**: sample Genpat code (GENPAT_CD);
- **Vertical coverage of raw reads**: hypotetical vertical coverage of raw reads;
- **Vertical coverage of trimmed reads**: hypotetical vertical coverage of trimmed reads;
- **% of unclassified reads**: percentage of reads not assigned to any micro-organism;
- **Number of unclassified reads**: amount of reads not assigned to any micro-organism;
- **First species assigned**: first identified species (considering the highest number of reads);
- **Vertical coverage of first species assigned**: theoretical vertical coverage assigned to the first identified species;
- **Number of reads first species assigned**: amount of reads assigned to the first identified species;
- **% of reads first species assigned**: percentage of reads assigned to the first identified species;
- **Second species assigned**: second identified species;
- **Number of reads second species assigned**: amount of reads assigned to the second identified species;
- **% of reads second species assigned**: percentage of reads assigned to the second identified species;
- **Third species assigned**: third identified species;
- **Number of reads third species assigned**: amount of reads assigned to the third identified species;
- **% of reads third species assigned**: percentage of reads assigned to the third identified species;
- **First Genus Assigned**: the first identified genus;
- **Number of reads first genus assigned**: amount of reads assigned to the first identified genus;
- **% of reads first genus assigned**: percentage of reads assigned to the first identified genus;
- **Second genus assigned**: the second identified genus;
- **Number of reads second genus assigned**: amount of reads assigned to the second identified genus;
- **% of reads second genus assigned**: percentage of reads assigned to the second identified genus;
- **Third genus assigned**: the third identified genus;
- **Number of reads third genus assigned**: amount of reads assigned to the third identified genus;
- **% of reads third genus assigned**: percentage of reads assigned to the third identified genus.

6.2.6 Check Coverage

The *Check of the Coverage* table provides information about the quality of mappings.



The metadata columns, in addition to **Code** et **Description** which refer to the identifier of the analysis result, are:

- **Analysis result**: the code (RIS_CD) of the analysis result;
- **Genpat sample**: sample Genpat code (GENPAT_CD);
- **DS sample**: internal unique identifier of the sequenced sample;
- **Notes**: code of the reference sequence used;
- **Vertical Coverage**: (or *depth of coverage*) it is the number of times the genome is covered by the mapped reads, considering the average across all positions (*average vertical coverage*);
- **Horizontal Coverage**: (or *breadth of coverage*) it is the percentage of genome covered by the reads mapping, compared to the length of the reference sequence;
- **Number of mapped reads**: total amount of reads mapped on the reference;
- **Minimum vertical coverage**: minimum number of times a single position is covered by the mapped reads;
- **Maximum vertical coverage**: maximun number of times a single position is covered by the mapped reads;
- **%IUPAC**: it is the percentage of IUPAC characters within the reconstructed/assembled sequence;
- **%Ns**: it is the percentage of "N" IUPAC characters within the reconstructed/assembled sequence;
- **Consensus lenght**: lenght of the obtained consensus sequence.

Useful links:

- Coverage on Wikipedia
- Coverage on sequencing.com
- Coverage on illumina.com

6.2.7 Check of cgMLST

The *Check of cgMLST* table gives information about the quality of the core genome Multi Locus Sequence Typing.

The metadata columns include the **Code** of the analysis and of the **Genpat sample** to which the data refer. The other columns list the aggregated statistics of alleles called by **chewBBACA** (https://github.com/B-UMMI/chewBBACA): * **% of called**: percentage of alleles called by chewBBACA relative to the total of *loci* in the scheme; * **Number of called**: number of alleles called by chewBBACA; * **Annotated**, **New**, **Not found**, **Discarded**: each of these columns shows the number of alleles based on how they were called or classified by chewBBACA; - Annoted: allelic variants already present in the scheme; - New: new alleles not present in the scheme to which new codes have been assigned; - Not found: (*Locus* Not Found - LNF), the *loci* that were not found in the sample's core genome; - Discarded: *loci* for which the allelic call is not sufficiently reliable.

Note: the number of *loci* depends on the chewBBACA scheme **for the specific micro-organism**.

6.2.8 Check of Kleborate

The *Check of Kleborate* table reports the outcomes of the Kleborate pipeline for *Klebsiella pneumoniae* screening.

The data are organized by type. Since the number of columns is large, it is recommended to view additional information by opening the relevant tab or using the column management options. Please, remind that opening an individual item allows you to display all the available columns with the corresponding values.

- The **Species quality control** type includes columns such as **Species** and **Species match** which report, respectively, the species identified by Kleborate and the degree of association;
- the **Assembly quality control** type includes columns of quality data pertaining the assembly such as the **Total lenght**, the **Number of contigs**, the **Largest contig** and **N50** (please, refer to the page check of the assembly);
- the **MLST** type includes columns with information derived from the Multi *Locus* Sequence Typing, such as the Sequence Type (ST);
- the **Resistance** type includes columns with data concerning the resistance genes of the sample to various compounds;
- the **Virulence** type includes columns with data concerning the virulence genes of the sample;
- the **Base data** type includes columns containing additional information and data concerning other genes encoding the proteins of interest.

6.2.9 Quality checks for ML-based classification

The table of the *quality checks for ML-based classification* contains the main report of the binary prediction of food source for *Listeria* samples (through 4TY_ML meat classifier).

The table is composed of the following fields: * **Sample**: Genpat sample code; * **GenPat matrix**: origin matrix; * **Prediction**: result of the binary prediction of food source ('MEAT'/'NOT_MEAT'); * **Probability**: percentage value of the classification that determined the prediction.

7. Downloads

7.1 Download procedure

From the **Download** section of the side menu, the user can access the corresponding features.

The *Download* page shares interface and behavior with the analysis launch page; therefore the standard method to download files from the platform requires using a collection of codes already present in the *cart* or in an active *tag*. The choice between tag or cart occurs before proceeding and is necessary to specify the samples associated with the files to download.

The download system allows the user to:

- add samples to the cart or activate a tag via shortcuts;
- use the **Quick selection** function to quickly fill the cart using a CSV file or by listing codes in the dedicated field;
- choose between two download categories:
  - **Single sample download**, which allows you to freely choose results of all analyses run with any tool;
  - **Preset download**, which provides shortcuts for quick download of specific analysis results;
- filter the files to download.

The following video provides a brief guide to performing a download:

Download execution

> Analysis results can be downloaded only if previously **imported**.

Similarly to launching an analysis, the cart is activated to proceed with the download; **Quick selection** and **Load RIS_CD** features are also available.

Quick selection allows you to quickly populate the cart with samples of interest. **Load RIS_CD** allows you to use the unique code of one or more analyses, giving an easy access to results of older analyses.

> Additional information on using RIS_CD to launch analyses and for download functions is available in the guide section "RIS_CD export".

Launching analyses generates a link to the process card of the requested analysis and, in some cases, from that card it is possible to access some files and visualization or download options. For more information on the workflow, you can consult the following Wiki pages:

- Analysis launch system steps;
- Check analysis.

7.1.1 Preset

Procedures for "**Single sample**" and "**Preset**" downloads are identical. **Preset** downloads are shortcuts that:

- allow you to avoid specifying all required parameters;
- are available only for a selection of combinations (the most used) among the possible analyses.

The **Preset** feature is designed to facilitate the workflow for users who routinely download specific files: parameters are pre-set, which allows you to move faster and with fewer clicks through the interfaces.

7.1.2 View downloads

In the **View downloads** section of the side menu you can consult the cards of launched download processes.

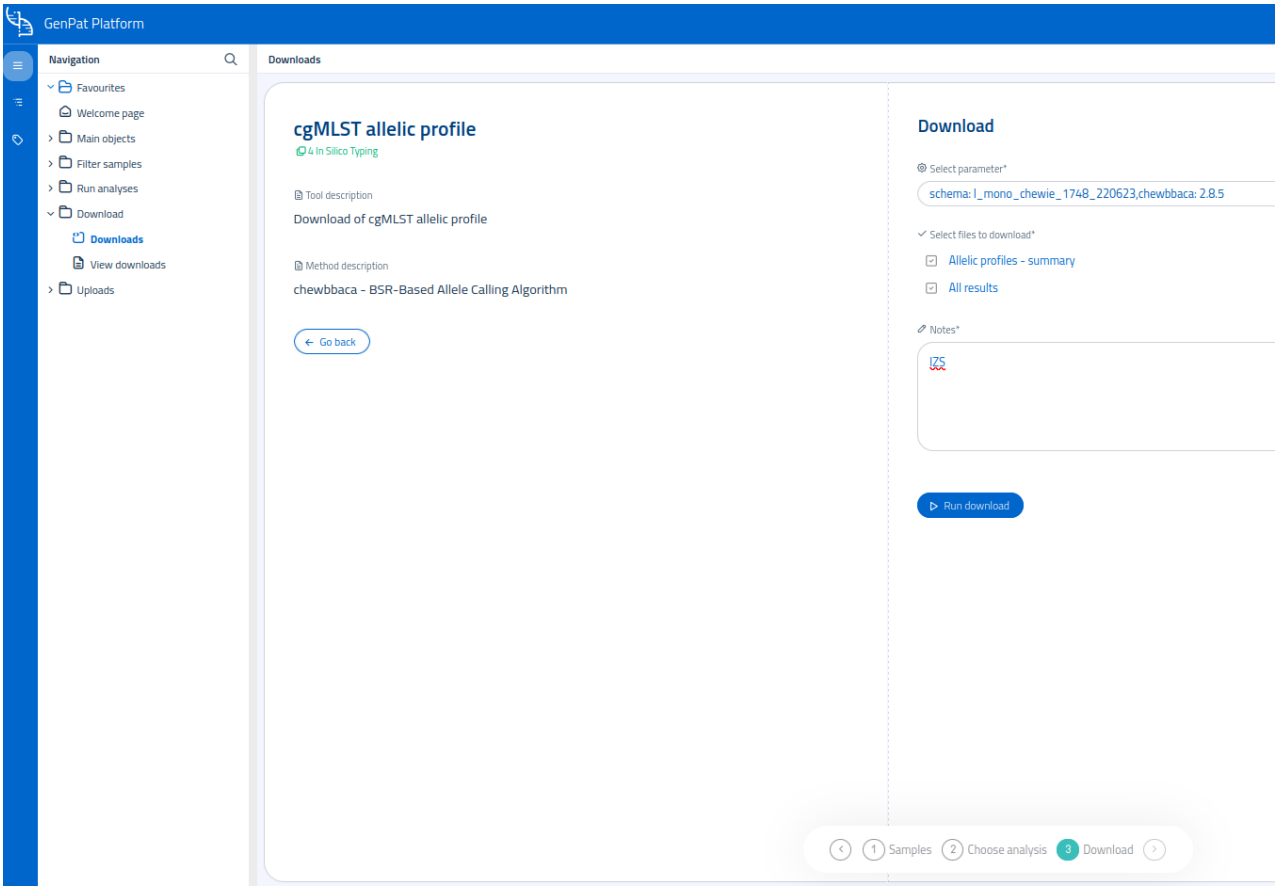
The **View Downloads** page keeps the history of the user's downloads. Each element (card) contains useful information about the performed process, including identifiers, links and a `wget` command to repeat the download.

As with the elements in other platform tables, each card is expandable and contains additional details.

Notes for the user

Both *Preset* download features and the selection of files to download are intended to evolve according to user needs. If there is a need to add a **Preset** download or a **filter** on the files to download for an analysis, users can request it from the Bioinformatics department.

Note for users with IZS profile: only for users with an "IZS" profile (e.g. "IZS Listeria"), the last phase for launching the download of allelic profiles and *de novo* assemblies will display a "Note" field, as shown in the image below. The download button will be disabled and such users will have to write a short note to communicate the reason for the download. Once the mandatory note is added, the "*Launch download*" button will become clickable.



7.1.3 wget for Windows users

This page lists some solutions to overcome the inability, for Windows users, to natively run the `wget` command generated by the platform to download all files.

Method 1: wget.exe

The method below consists of a workaround to execute the UNIX `wget` command, to download all files of an analysis in a single action.

1. Download the "wget.exe" program for Windows from the link https://eternallybored.org/misc/wget/1.21.4/64/wget.exe;
2. Save (or move) the downloaded wget.exe file to your "Download" folder;
3. Open the Windows command line (**cmd**, available in the Start menu) and move to the "Download" folder (type the command below and press "Enter"):

   ```
   cd Download
   ```

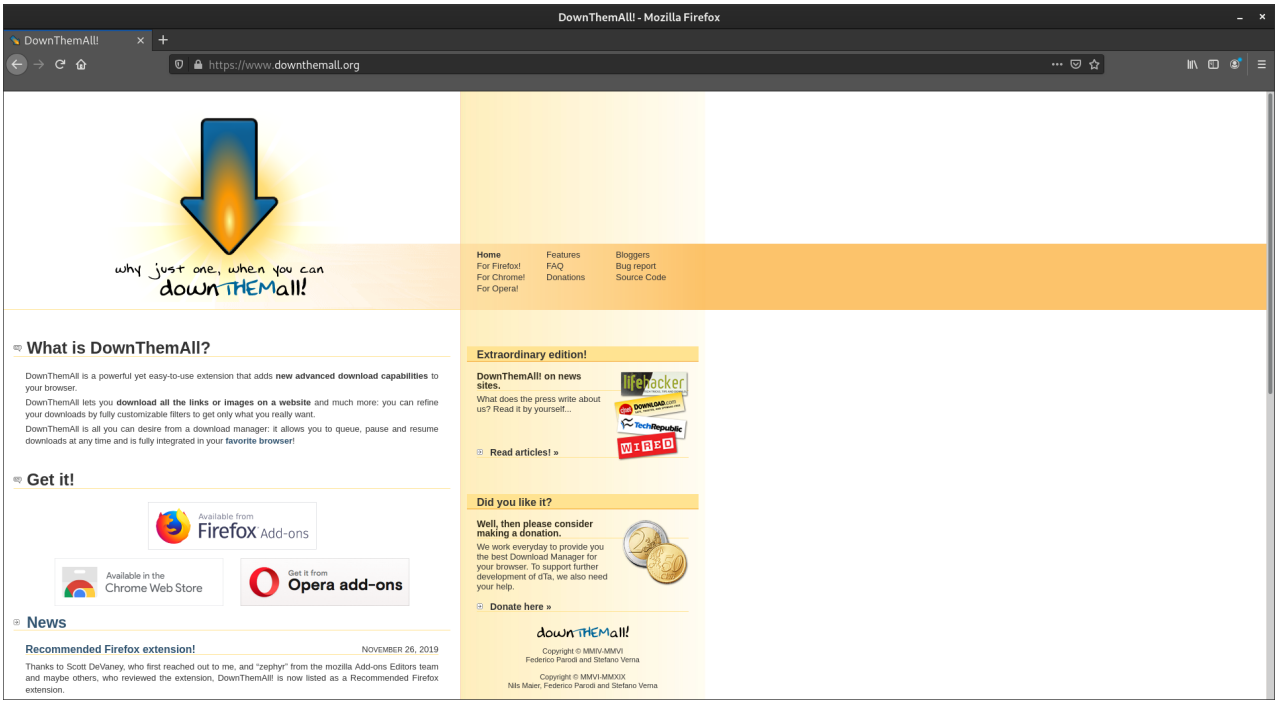
4. Copy the `wget` command generated by the platform download page into the Command Prompt and run it;
5. At the end of the download procedure, the files will be located in a folder with the name provided by the launched command (e.g. "f0086d8f-948d-4753-ad85-5171f1a5dda8"), located in the "Download" folder.

Method 2: Browser extension "DownThemAll"

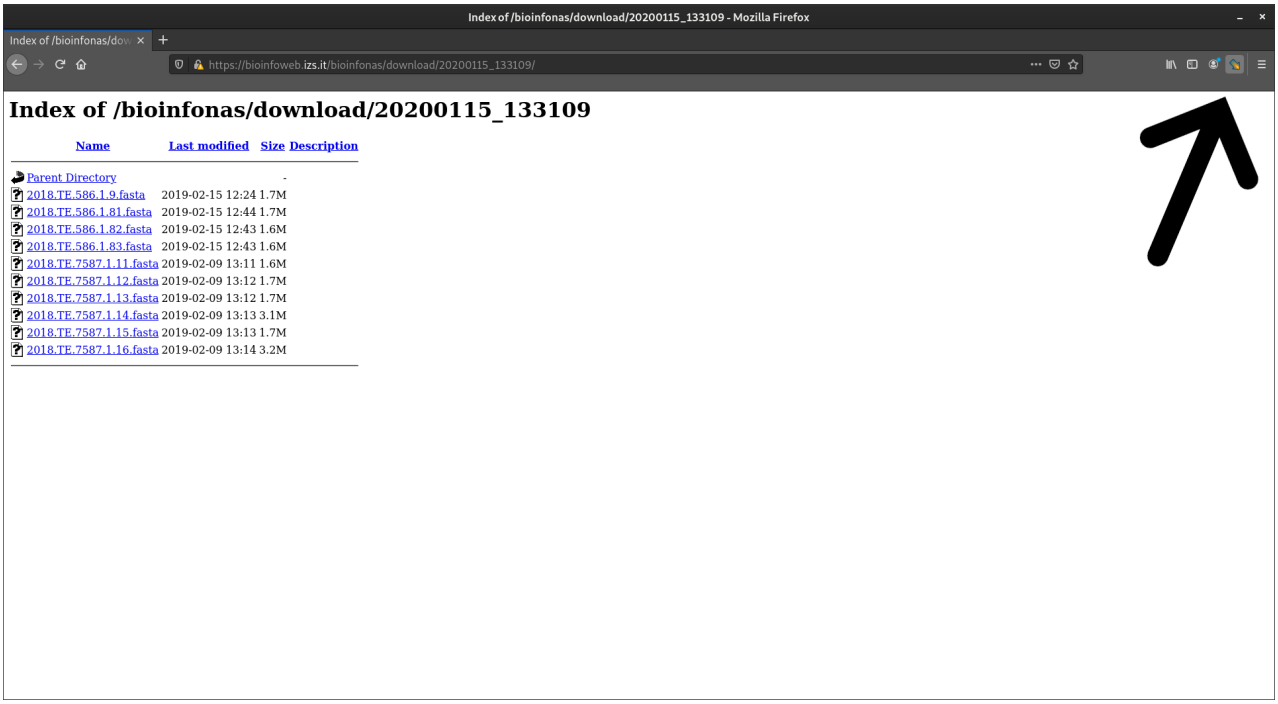
The web browser extension "DownThemAll!" can be added to Firefox, Chrome and Edge and allows downloading multiple files from a web page.

1. "DownThemAll!" is available at: https://www.downthemall.org/;

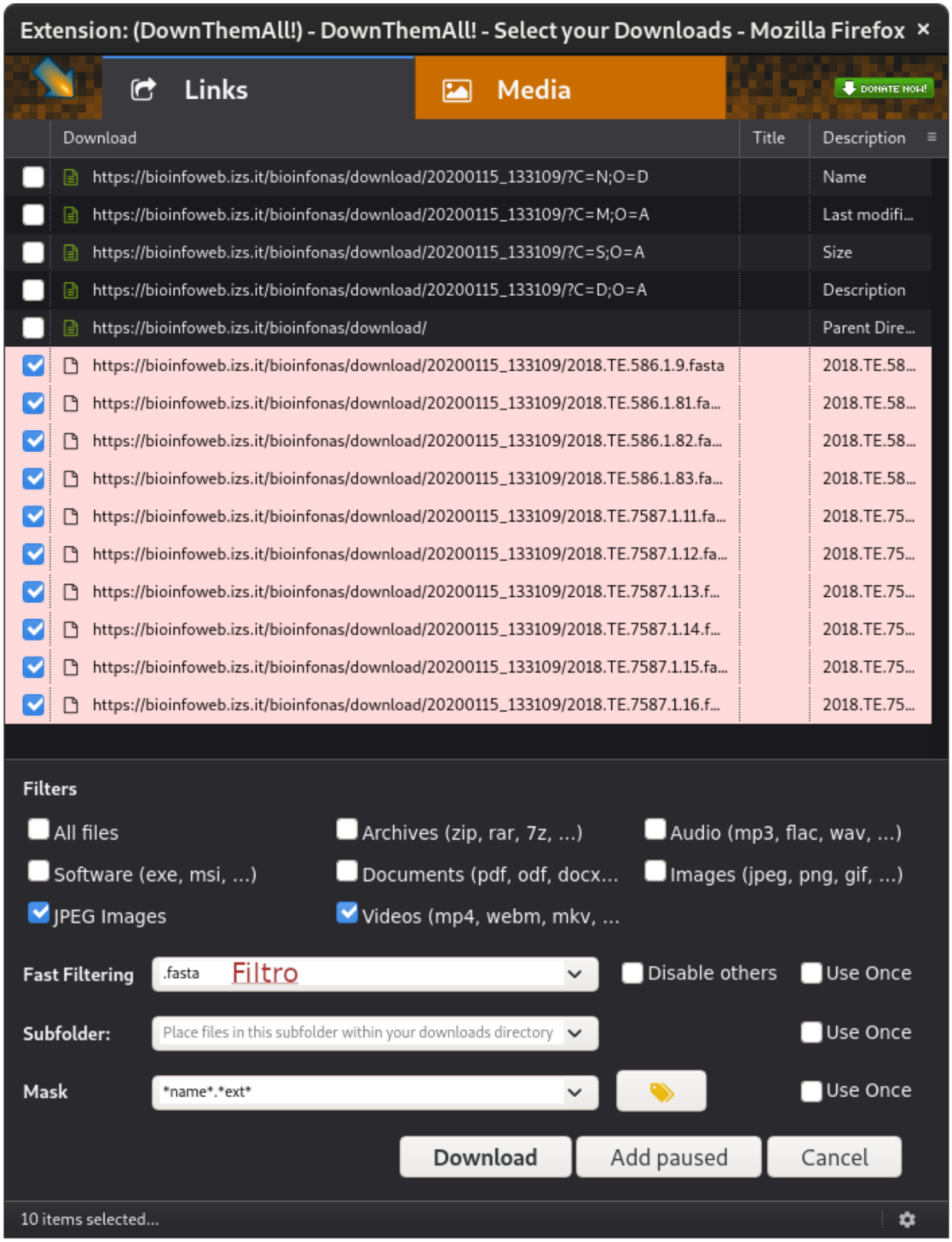
2. In the "Get it!" section you can choose the version for your browser and install it:



1. Once the extension is installed, when you browse a results folder you will be able to download the files listed on the page using the "DownloadThemAll!" function (reachable by clicking the software icon, usually at the top right of the browser toolbar):



1. The window that opens allows you to filter or select files before downloading them via the corresponding button:


Extension: (DownThemAll!) - DownThemAll! - Select your Downloads - Mozilla Firefox
Links
Media
DONATE NOW!
Download
Title
Description
https://bioinfoweb.izs.it/bioinfonas/download/20200115_133109/?C=N;O=D
Name
https://bioinfoweb.izs.it/bioinfonas/download/20200115_133109/?C=M;O=A
Last modifi...
https://bioinfoweb.izs.it/bioinfonas/download/20200115_133109/?C=S;O=A
Size
https://bioinfoweb.izs.it/bioinfonas/download/20200115_133109/?C=D;O=A
Description
https://bioinfoweb.izs.it/bioinfonas/download/
Parent Dire...
https://bioinfoweb.izs.it/bioinfonas/download/20200115_133109/2018.TE.586.1.9.fasta
2018.TE.58...
https://bioinfoweb.izs.it/bioinfonas/download/20200115_133109/2018.TE.586.1.81.fa...
2018.TE.58...
https://bioinfoweb.izs.it/bioinfonas/download/20200115_133109/2018.TE.586.1.82.fa...
2018.TE.58...
https://bioinfoweb.izs.it/bioinfonas/download/20200115_133109/2018.TE.586.1.83.fa...
2018.TE.58...
https://bioinfoweb.izs.it/bioinfonas/download/20200115_133109/2018.TE.7587.1.11.fa...
2018.TE.75...
https://bioinfoweb.izs.it/bioinfonas/download/20200115_133109/2018.TE.7587.1.12.fa...
2018.TE.75...
https://bioinfoweb.izs.it/bioinfonas/download/20200115_133109/2018.TE.7587.1.13.f...
2018.TE.75...
https://bioinfoweb.izs.it/bioinfonas/download/20200115_133109/2018.TE.7587.1.14.f...
2018.TE.75...
https://bioinfoweb.izs.it/bioinfonas/download/20200115_133109/2018.TE.7587.1.15.fa...
2018.TE.75...
https://bioinfoweb.izs.it/bioinfonas/download/20200115_133109/2018.TE.7587.1.16.f...
2018.TE.75...
Filters
All files
Archives (zip, rar, 7z, ...)
Audio (mp3, flac, wav, ...)
Software (exe, msi, ...)
Documents (pdf, odf, docx...
Images (jpeg, png, gif, ...)
JPEG Images
Videos (mp4, webm, mkv, ...
Fast Filtering
.fasta
Filtro
Disable others
Use Once
Subfolder:
Place files in this subfolder within your downloads directory
Use Once
Mask
name.*ext*
Use Once
Download
Add paused
Cancel
10 items selected...

8. Uploads

8.1 Introduction

The platform provides two main functions for sequence uploading:

- an **upload** function that allows users to upload **their own sequences**;
- a function that allows sequences to be imported directly from **NCBI**.

Regarding the uploading of user-provided sequences, it is required to complete and validate a **metadata** file along with the upload of a compressed `.zip` file containing `.fasta` or `.fastq` files, referenced by the validated metadata. For NCBI upload, users only need to know the **Accession numbers**.

8.1.1 Topics

- **Upload**;
- **Upload from NCBI**;
- **Upload PT Genpat**;
- **Upload check**;
- **Metadata glossary**;
- **Metadata and zip rules**.

8.2 Upload

The upload procedure allows the user to upload sequences to the platform from their computer. It consists of three phases:

1. **preparation and upload of metadata**;
2. **uploading sequences**.
3. **file processing**.

The first two phases require user actions, while the third occurs automatically once the file upload is complete.

Through the **Upload** page (`Uploads > Upload` in the side navigation menu), it is possible to fill in and validate **metadata** and upload `.fasta` or `.fastq` files, compressed into a single `.zip` file.

8.2.1 Preparation and upload of metadata

On the **Upload** page (`Uploads > Upload` in the side navigation menu), it is possible to add (`+` at the top left) cards containing sample metadata, by choosing between two different options:

- Add from form
- Add from file

The uploaded metadata will be validated by the system which will also check for the required minimum metadata; the validation status will be visible in the corresponding column.

Add metadata from form

The **Add from form** option allows the user to quickly and manually create a card to associate with a single sample. The created card must be filled in, by the user, with the metadata of the corresponding sample, entering at least the mandatory metadata.

> **A tip!** Check the **metadata glossary** to understand the type of required metadata and their descriptions. You will also find a guide for filling out the form.

The following video provides a quick demonstration of how to fill out a card with metadata of a single sample, along with some useful features of the metadata management through the card.

Upload from form

> **Note:** Metadata cards can be edited from the interface, regardless of the method originally used to upload the metadata to the platform.

Add a sampling point

Regarding the `sampling_point` field, if you have the necessary permissions, a feature is available that allows you to add a *sampling point from external sources* when it is not present on the platform. For more details, please refer to the dedicated section of this wiki: Sampling Point.

Add a sampling point

Add metadata from file

The **Add from file** option allows the user to upload a tabular file, containing the list of samples to upload, along with their metadata.

The following video shows how to upload a metadata file for one or more samples.

Adding metadata from a file

Metadata file templates

Below are the metadata file templates available for download:

- template.tsv
- template.xls

A description of the metadata columns is available in the metadata glossary.

> **Tip!** If you do not know the metadata identification codes, you can manually upload a record and then export the template using the functions available in the table's context menu .

Record upload and template export

8.2.2 Uploading sequences

To access the file upload form, it is necessary that **valid metadata**, corresponding to the files to be uploaded to the platform, are present. If this condition is met, the user will be able **to proceed with the upload**.

It will be sufficient to select a `.zip` file containing the sequences to be uploaded (using the drag-and-drop function or by clicking in the designated area).

> You can find more details on the rules for preparing the sequence files to be uploaded in the sequence files paragraph of the metadata and .zip rules section.

In addition, the user can specify whether to automatically run the corresponding pipeline for the sample type, using the **Process** upload toggle.

The following video shows how to use the upload features.

Upload features

> **Important!** File uploading may take time. Closing the window or leaving the page, during the upload process, would interrupt the upload.

Accepted sequences

The accepted sequences fall into one of the following 4 typologies:

- **Fasta**, for the direct upload of sequences in **fasta** format;
- **Illumina raw reads**, for uploading **fastq** raw reads, specifically **short paired-end**, generated by Illumina devices;
- **Iontorrent raw reads**, for uploading **fastq** raw reads, specifically **long single-end**, generated by Iontorrent devices;
- **Nanopore raw reads**, for uploading **fastq** raw reads, specifically **long single-end**, generated by Nanopore devices.

The file type and the sequencing technology or device, used to produce it, must be specified in the appropriate columns of the metadata file, as described in the **metadata glossary**.

8.2.3 File processing

At the end of the file upload, the upload process is not yet complete and its status can be checked in **Check uploads**.

When all processes related to file uploading have been completed, you will be notified via a notification.

8.3 Upload from NCBI

The **Upload from NCBI** page allows the user to upload one or more sequences present in the NCBI databases (National Center for Biotechnology Information) on the platform. The databases from which data can be sourced are represented by 3 different types of upload:

- **Assembly fasta**, to upload fasta files from the NCBI "Assembly" database;
- **Nucleotide fasta**, to upload fasta files from the NCBI "Nucleotide" database;
- **Reads fastq SRA**, to upload fastq files from the NCBI "Sequence Read Archive - SRA" database.

To run an upload from NCBI, only a list of **Accession Numbers** is required: you can choose which of the 3 databases to import files from, select the species of origin and paste the *Accession Numbers* into the appropriate field.

The sequences will be retrieved directly from NCBI, without needing to upload files from your computer.

> **Important!** Multiple sequences can be imported only if they belong to the same species; for sequences belonging to different species, separate uploads will be required.

The following video shows how to use the *Upload from NCBI* features. The 3 upload types share the same modes and the same requirements.

Upload from NCBI

8.4 PT GenPat Upload

8.4.1 Upload Instructions

To upload sequences for **PT GenPat 2026**, the basic procedure is the same as the standard upload on GenPat, as illustrated in the video below.

The only key differences compared to a standard upload are:

- You must log in using one of the following profiles: **LAB Campylobacter**, **LAB Listeria**, or **LAB Salmonella**;
- You must provide the PT-specific metadata listed in the required metadata table;
- You must provide a ZIP file containing:
- the **raw reads**, or
- the assemblies in FASTA format.

Data generated using **Illumina, IonTorrent, ONT,** or **PacBio** platforms are supported.

Reads must be renamed as follows: `"Code + _R1.fastq.gz"` (for paired-end Illumina data, both R1 and R2 files must be provided and renamed accordingly). Example: `XX-YY-GENPAT26-ZZ_R1.fastq.gz` and `XX-YY-GENPAT26-ZZ_R2.fastq.gz`.

Please refer to the **required metadata table** below.

Upload from form

Please note that:

1. **Reads must be compressed** in `gz` format and therefore must have the `.fastq.gz` **extension**.
2. **The platform does not perform any compression:** as specified on this and the other upload guide pages, the file to be uploaded must be a `.zip` **file already compressed by the user on their own computer**, using any compression tool of their choice.
3. **PT2026 uploads can only be performed by users with the "LAB Campylobacter", "LAB Listeria", or "LAB Salmonella" profile who have requested access.**

Samples must be uploaded using the corresponding profile (for example: Campylobacter samples **LAB Campylobacter**).

8.4.2 Metadata Files

The metadata file templates are available below:

- template.xls
- template.tsv

8.4.3 Description of Required Metadata

The following table lists the columns that must be completed in the metadata file, explaining their meaning and the expected value for each submitted sample.

Metadata	Expected Value	Description
code	Code in the format `XX-YY-GENPAT26-ZZ`	The code identifies the laboratory (`XX-YY`), the PT (GenPat 2026), and the sample number (`ZZ` , where `ZZ` is a two-digit progressive numeric code corresponding to the DNA aliquot received).
species	One of the following codes: `110341G` , `82G` , or `86G`	This field indicates the sample species and must contain the corresponding code: `110341G` for *Campylobacter jejuni*, `82G` for *Salmonella enterica*, and `86G` for *Listeria monocytogenes*.
note	Text string	This field can be used for any textual note regarding the sample and is optional.
metadata_type	Text string	This field must contain the string `pt2026` .
file_format	Text string	This field must contain one of the following values: `fasta` or `fastq` , indicating the format of the uploaded file.
seq_type	Text string	If `file_format` is set to `fastq` , this field is mandatory and must contain one of the following values according to the sequencing platform used: `ion` (IonTorrent), `nanopore` , `illumina_paired` (Illumina short-read paired-end technology).

8.4.4 Additional Notes

For further details on the manual upload procedure, please refer to the GenPat Wiki page: Manual Upload.

For further details on preparing the metadata file, please refer to the GenPat Wiki pages on metadata: Metadata Preparation and Upload and Metadata and ZIP File Rules.

8.5 Upload check

The **Check uploads** page (`Uploads > Check uploads` in the side navigation menu) contains the list of uploads made by the user, organized in a table.

The items in **Check uploads** have the same features as the other platform tables.

An upload can have 3 possible outcomes:

- **Success**,
- **Failed**,
- **Warnings**

8.5.1 Success

An upload is completed with **success** when all uploaded files correctly match their respective metadata. In this case, the metadata cards are removed from the **Uploads** page and associated with the uploaded files.

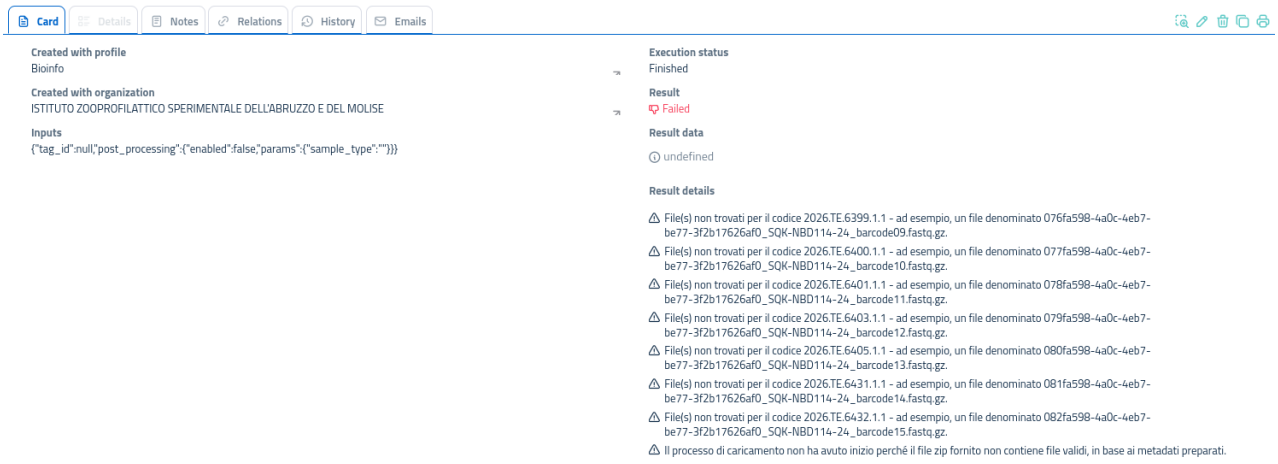
The summary cards of successfully completed uploads display the sequence identifiers generated by the platform (field “Samples”) and those that the files had at the time of upload (field “Code”). These codes can be added to the cart directly from the card using the "Add to cart" button.



8.5.2 Failed

An upload may be marked as **failed** when all the files the user is trying to upload cause an error. In the event of failure, the metadata pending upload remain available and visible on the **Upload** page, so that they can be corrected if necessary (as shown in this video).

The card of a failed upload provides information about the reason(s) for the failure, grouped in the 'result_details' field. In the image below, the process failed because the .zip file contained only files with incorrect names or names that did not match those provided in the metadata:

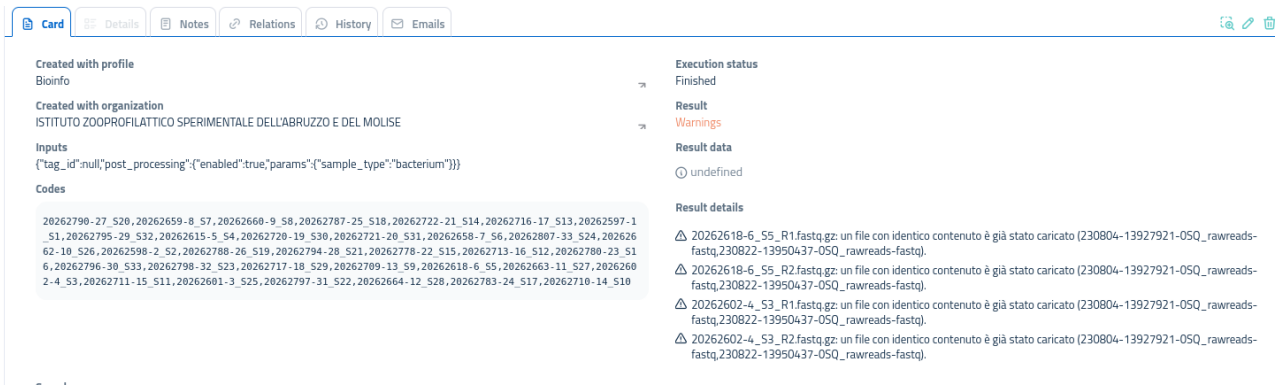


8.6 Warnings

An upload results in failure only when all the files being uploaded generate errors: if even just one of the provided files is valid, that file will be uploaded and the card will show **warnings**.

In upload cards with warnings you will find both the codes of successfully uploaded samples, which can be added to the cart, and the reports of errors (in the "result_details" field) for samples that were not uploaded. On the **Upload** page, only the metadata of successfully uploaded samples are associated and removed from the list; the others remain in the queue, awaiting correction or removal.

In the example below, some files in the provided .zip had invalid names; however, two files were correct. The two valid files were uploaded, while the others triggered the warnings.



8.7 Further information

8.7.1 Metadata glossary

The table below shows the columns of the metadata template. For each column in the downloadable template, the required type of metadata is specified, along with a description and a guidance for completing the template.

For an explanation of the limits and requirements concerning the metadata file and the files to be uploaded, please refer to the Metadata and zip rules page.

Note: each profile can upload samples classified as `official` and/or `research` (see the metadata_type row in the table below). The system will not accept uploads of types other than those permitted for that profile.

TSV field title	Metadata	Description	Mandatory
`code`	Sample code	Sample identifier code in the data provider system	yes
`species`	Pathogen species	Indicate the pathogen species with GENPAT coding	yes
`material`	Material	Indicate the sample material with GENPAT coding	yes
`sampling_point`	Sampling point	Indicate a Sampling Point code present in GENPAT	yes
`sampling_point_address`	Sampling point	Address	no
`sampling_point_note`	Sampling point	Use to indicate the type of structure	no
`sampling_date`	Sampling date	Specify date detail to the day	yes
`matrix`	Isolation matrix	Use EFSA FOODEX2 code for the isolation matrix. It is necessary to specify at least the basic FOODEX2 code for food, feed, animal and environment; in particular: **Food (A0B6Z)**, **Feed (A0BB9)**, **Animals (A04SF)**, **Environment (A0C5Y)**	yes
`host`	Host species	Use EFSA FOODEX2 code. Host animal species. Use in case of animal matrix	no
`note`	Additional notes	Optional free text. Useful to describe anomalous situations or deviations	no
`md5_file_r1`	raw read file	MD5 checksum of the R1 raw reads file or the single file in case of single-end sequencing	Mandatory if `read_format` =fasta and submission type = Official
`md5_file_r2`	raw reads file	MD5 checksum of the R2 raw reads file	Required in case of paired-end sequencing and `read_format` =fasta and submission type = Official
`filename1`	raw read file	Fasta file name in case `read_format` =fasta; or, name of the raw read R1 file or name of the single file, in case of single-end sequencing. If the `code` field represents a prefix of the file name, this field can be omitted. Examples of file names corresponding to different patterns with `code` ="code": `read_format` =illumina_paired: "code*R1*.fastq.gz" `read_format` =fastq_singleend: "code.*fastq.gz*' `read_format` =*fasta: 'code*.fasta"	no
`filename2`	raw read file	Name of R2 file in case of paired-end sequencing. If the `code` field represents a prefix of the file name, this field can be omitted. Examples of file names corresponding to different patterns with `code` ="code": `read_format` =illumina_paired: "code*R2*.fastq.gz"	no
`file_format`	Type of file submitted	It can take one of the following values: `fasta` , `fastq`	yes

TSV field title	Metadata	Description	Mandatory
`seq_type`	Sequencing technology	It can take one of the following git pullvalues: "illumina_paired", "ion", "nanopore". Mandatory for submission type = Official	conditional
`metadata_type`	Type of metadata sent. Different validations are performed based on this field	It can take different values, including "official" and/or "research", based on the user profile	yes

8.7.2 Metadata and .zip rules

In this section of the guide, the characteristics that files and metadata used for upload must have are described, together with the rules they must follow in order for validation to be successful.

Metadata

Please refer to the metadata glossary for a specific discussion of the individual fields in the metadata templates.

Metadata added from form

If metadata are added manually via the Add from form option, or if the cards of one or more samples are edited manually, the user is guided in selecting metadata values. Therefore, there are only 2 rules to follow to ensure proper validation of the metadata card:

- Use a correct **Code** (file name);
- Fill in at least the **required fields** for metadata.

Use a Code

The text to be entered in the "Code" field is the name of the fasta or fastq file you intend to upload.

- The code consists of the file name, **without extension**, that is in the zip archive being uploaded;
- The code and the file name must not contain spaces;
- For fastq files with *paired-end reads*, the code **must not** contain the *forward* or *reverse* read identifier (e.g. "_R1" or "_1") and must be **unique** (the metadata card applies to both reads).

Note The metadata template also provides the option to specify full file names (both reads for paired-end files, single file for single-end reads or fasta). If those columns are filled, the check will be performed on the full file names rather than only on the code.

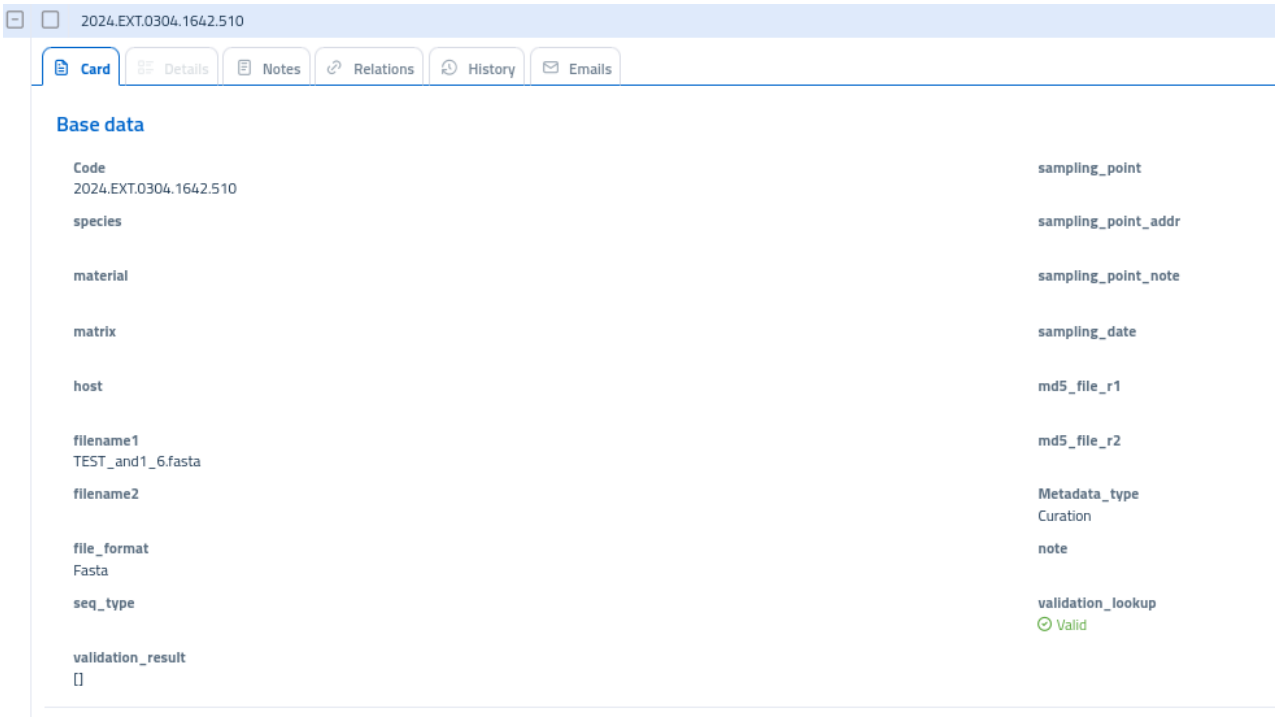
Fill in required fields

Required fields can vary based on the type of sample being uploaded; for example, a research sample will have fewer required fields than an official sample, for which more metadata are required.

In general, the minimum required metadata may require you to:

- fill in the `code` field;
- specify the sample type (official, research, *etc.*), the file type (fasta or fastq) and, in case of fastq, the sequencing platform technology (Illumina *paired-end*, Nanopore *single-end*, *etc.*);
- indicate species or material.

During validation, required fields with missing or incorrect values are highlighted with error messages.



Tip By entering metadata from form or using the platform edit functions, fields such as species or material can be set through a dropdown menu or a specific table. In these cases, the selection is guided and ensures the correct code is entered.

Metadata added via file

Templates for tabular metadata submission files can be downloaded directly from the **Metadata file templates** section of the wiki page dedicated to Upload. The templates are available in both `.tsv` and `.xls` format.

Tip Templates can also be downloaded from the platform via the metadata upload modal (button `Add from file`) and from the context menu of the upload page.

Upload metadata from file

Required metadata

Each column of the metadata file corresponds to a field in the metadata card.

In the metadata file template, the second row serves as an example for formatting and the type of text to enter in each cell.

The following indications apply to the metadata fields required by the template:

- The `metadata_type` field can take the following values:
  `official` - (official samples upload);
  `research` - (samples uploaded for research purposes).

Each user profile can upload specific types among those listed above and **the system will not accept uploads of types different from those allowed**. Some metadata are required only if the upload is for official samples (`metadata_type` = `official`).

- The `species` , `material` , `host` , `matrix` and `sampling_point` fields must **not** be filled with taxonomic names or textual descriptions, but with **the corresponding code in the platform tables**.

Tip Metadata code tables can be consulted under `Main items > Metadata` in the platform side menu. Additional information is available in the **Metadata** section of this Wiki.

- The `read_format` field can take the values:
- `fasta` ;
- `illumina_paired` (paired-end fastq files from Illumina devices);
- `ion` (single-end fastq files from Iontorrent device);
- `nanopore` (single-end fastq files from Nanopore device).

For all other indications and restrictions, refer to the information provided in the section Metadata added from form - Required fields and in the metadata glossary.

Sequence files

Files in **fasta** format are recognized if they have the `.fasta` or `.fa` extension. Similarly, **fastq** files are recognized if they have the `.fastq.gz` or `.fq.gz` extension.

Note: Compression of fastq files in `.gz` format is mandatory. Fasta files are accepted with or without `.gz` extension.

Zip archive

The archive must be in `.zip` format. During upload, the system checks metadata cards and associates them with the corresponding samples. For the upload to succeed, the name of each fasta or fastq file (excluding the file extension) must match the value of the `code` metadata entered during metadata upload.

If at least one or only some of the samples in the zip archive match the codes in the validated cards, the upload process will be completed only for those files and with **warnings**.

If the zip archive contains files whose names do not match any of the codes entered in the validated metadata cards, the upload request will result in **failure**.

Tip Files in the archive can also be inside a folder/subfolder structure, provided it does not exceed 10 levels of depth.

9. Dashboards

9.1 SPREAD, Spatiotemporal Pathogen Relationships and Epidemiological Analysis Dashboard

Formerly known as GrapeTree Extended.

9.1.1 Description

The dashboard is based on GrapeTree, a fully interactive tree visualization program from Newick (nwk) files, within EnteroBase system. Grapetree supports manipulation of both tree layout and metadata. It generates GrapeTree figures using the Neighbor-Joining (NJ) and the Minimal Spanning Tree (MSTree) algorithms or an improved Minimal Spanning Tree algorithm called MSTree V2. The original version is an integral part of EnteroBase and, for further details, you can refer to documentation.

For a formal description, please see the accepted manuscript in Genome Research.

The standalone version of GrapeTree has been integrated into the platform as the main tool for manipulating tree graphs and metadata, and it has been extended to conduct spatio-temporal analyses with an integrated geographic information system (GIS) and a time-based data visualization system. The web application allows users to upload geographic coordinates and temporal data related to each sample and allows to display them, reflecting the selection in the tree on the map and vice versa. Additionally, it enables users to reproduce a timelapse visualization both in the map and in the tree.

9.1.2 Using SPREAD

The dashboard is integrated with the platform, so you can launch analyses and view data on it. However, by directly accessing the address https://cohesive.izs.it/spread/, you can upload and visualize your own dataset.

> **Please note** The uploaded dataset is displayed and managed exclusively client-side in the browser, which means no data is transmitted, and no tracking cookies are used. The only data downloaded from the internet are the visualization code (JavaScript), fonts and any necessary tiles for map. Therefore, you can also upload sensitive data.

You can upload the files by dragging them directly over the initial drop area or using the upload button. You can drag or load a `.nwk` file followed by a `.tsv` file containing the metadata and, optionally, a `.geojson` file containing geospatial information related to the samples.

> **Important!**
> The `.nwk` file must always be uploaded before the metadata or the geoJson files.

The dashboard allows you to download a complete JSON file that includes metadata and configurations. The generated JSON file can be loaded using the same drag and drop or upload functionalities mentioned earlier, which is very useful if you want to save your work and/or share it.

Grapetree upload
Please find an example dataset on the project repository at https://github.com/genpat-it/spread/tree/main/datasets/test.

9.1.3 Overview of General Functions

Whether you choose to upload data from your computer or you are visualizing a pre-processed dataset from the platform, you will be in the same configurable workspace. The main components that make up the dashboard along with the **Tree** are:

- **Settings**
- **Map**
- **Metadata**
- **Legend**
- **Video**

These components can be shown or hidden using the respective toggles located at the top of the main workspace, depending on your work needs. Before delving into the details of these components, let's quickly mention two functions present in the header. One is related to *language change*, which can be set by clicking on the **globe** icon, and the other, accessible by clicking on the **info** icon, displays *useful information and links*.

Available Languages and Translations

Currently, the available languages are English, French and Italian. However, the application can be easily extended with new translations. If you would like to contribute, you can use the `en.js` or `it.js` file as a template from the `i18n` folder on the public repository of the project and translate the values. For any further information or support in this regard, please contact `bionformatica@izs.it` .

9.1.4 Components

Let's take a closer look at the various parts that make up the dashboard.

Tree tools

From the context menu, placed on the high right corner of the tree's working area, you can access this section containing some useful tools to modify the appearence and layout of the tree, as well as an option to revert the tree to previous state and undo changes and modifications. The available functions in this section are:

- **Centre tree:** adjusts view settings so that the tree's core is at the centre of the working area.
- **Static redraw:** refreshes the tree and redraws it with the preset layout.
- **Original tree:** reloads the page and reverts the tree to the state and layout it was in when it was first loaded. You will be asked for confirmation before leaving the page. **Caution**: this function will cause you to lose all the changes you made to the tree.

Settings

By activating the settings, you will enter an `editable section` with a series of `cards` divided into thematic areas, which allow you to configure the tree. Each card can be expanded or collapsed based on your needs. Below, you will find descriptions of the various available configurations, also visible within the application by clicking on the `?` icon of the respective card.

Node style

Under this category you can find some options to change the node's appearance or displayed information.

- **Colour by:** from the dropdown menu of this section it's possible to choose the fill-in colour of the nodes according to metadata;
- **Show/Hide labels:** the switch will allow you to choose whether to show the node labels or hide them. You can choose which label to show from the *Labels* dropdown menu just below.
- **Font Size:** the font size of the labels can be chosen either by dragging the slider or by specifying a value in the box.
- **Highlight Labels:** use the text box to provide a search term for the labels to highlight the corresponding nodes. It supports Regular Expressions.
- **Individual segments:** use the switch to turn each node into a pie chart showing the breakdown of the members contained in it. Please note that the attivation of the breakdown function is based on the category selected in *Colour by*.

Node size

- **Radius size (%)** nodes size can be increased or reduced by dragging the slider or by specifying the value in the box.
- **Kurtosis (%):** by default the area of the nodes correlate with the members in them. Using the *Kurtosis (%)* section's slider or box, the node's size can be increased or decreased based on the distribution's kurtosis of the nodes. By increasing this value, you raise the percentage by which the node size grows per member in the node, thus nodes with a larger number of members will have a larger area and will stick out more.

Branch style

In this section you'll find options to customize the tree's branches.

- **Show/Hide Labels:** this option allows you to choose whether to show the branch labels or hide them.
- **Font Size:** this option allows you to choose the font size for the labels either by using the slider or by entering a value in the designated field.
- **Scaling (%):** this option allows you to increase/decrease the overall branches' length, by using the slider or by specifying a value in the box below it.
- **Collapse Branches:** use the slider or enter a value in the box to collapse all branches shorter than the specified length. Nodes falling in that interval will be merged together. Branch length values are scaled to the branch lengths defined in the original tree data.
- **Log Scale:** by checking the box, all lengths of all branches will be scaled logarithmically.

Branch cutoffs

The settings in this section allow you to render the branches in different ways depending on their length (branch length values are scaled to the branch lengths defined in the original tree data). To increase/decrease the scale you can enter a value in *Cut off branches longer than* field. You can then choose whether to display them, hide them, or shorten them:

- **Display:** show longer branches as normal (default).
- **Hide:** makes longer branches transparent. Even though the branches are not displayed, it's still possible to interact with them.
- **Shorten:** branches longer than the specified cutoff will be cropped back to the specified cutoff. Branches will be displayed as dashed lines to indicate those affected by the reduction.

Rendering

This section contains some options that allow to control how nodes are positioned in the tree. Two modes are available: *Static* and *Dynamic*.

- **Static:** in this mode, the tree layout is calculated when the chart is generated and remains static (but interactive). The relative length of the branches and their scaling (from the original tree data) will always be maintained as long as the **Real Branch Length** option is checked.
- **Dynamic:** nodes are positioned dynamically in a way similar to a Force Directed Layout and will try to fan out in order to distance themselves from the neighbors. The tree can be moved around freely by dragging it and the nodes will spread out to maintain the distances among themselves. The *Dynamic* mode can be used to improve the aesthetics of the tree but the branch length scaling will be modified compared to the original: branches will NOT be to scale when in *Dynamic* mode. Toggle on the **Selected Only** option to apply only to selected nodes.

Grapetree settings

Map

This area allows you to view an interactive map if geographic coordinates have been defined for the samples (either through metadata files or geoJSON files). On the map, some `markers` are drawn with varying radius sizes, based on the number of samples present in a particular area. By clicking on each marker, you can open an informative popup containing the list of samples within.

The markers reflect the colors of the chosen theme for the tree visualization and selection. You can interact with the map through its **context menu**, which can be accessed by right-clicking on it or by using the three vertical dots button located at the top-right corner.

Map tools

You can control the appearance of the map through the options available in this section.

- **Join by coordinates/by metadata:** The toggle allows you to aggregate points on the map based on the geographical proximity of the provided coordinates or on the values of the uploaded metadata. In the first case, you can **set a value** for *delta*, which determines the degree of aggregation.
- **Minimum/Maximum marker radius:** With these two fields, you can adjust the size of the points on the map by defining a minimum and a maximum radius. The initial size is relative to the number of nodes located at the same geographical coordinates.
- **Pie Chart mode/Heatmap mode:** This toggle allows you to change the visualization mode of the points on the map. With the first option (selected by default), the points will be displayed with pie charts based on the categories of the tree nodes while, with the second option, a blur effect is applied to the nodes.

Grapetree map

Metadata

By activating the **Metadata** toggle, you will be able to view the metadata within an interactive grid. The table allows you to move columns, expand and/or reduce them, sort them or filter them based on search keys. By clicking on a row, you can select/deselect the samples within the tree.

Similar to the map, the metadata table also has tools that can be accessed by right-clicking on the table or by using the button located at the top-right corner with three vertical dots.

Grapetree metadata

Legend

This is a key tool for both the interaction with the tree and for the temporal visualization/selection of data. Through the **Legend tools**, you can change the category and color of the tree, customize the colors or select samples from a specific category.

The following video demonstrates how to access the context menu of the legend and its options, how to interact with the legend, and how to modify preferences and the color palette.

Grapetree legend

Video

In **Video** mode, the legend turns into a proper timeline. The video player, placed at the bottom right, allows you to iterate through the legend elements at a configurable speed. The video functions are completely metadata-driven, so if temporal information is present in the metadata, you can color the legend based on that information, sort it in ascending order and, consequently, have a timeline to measure the spread and development of a pathogen over time.

In the following video, you will find a possible use case.

Video mode of the legend

Macros

The Macros component can be accessed via the `fast-forward` icon located at the top-right corner of the application. This section provides a growing collection of action sequences (macros), which can be searched and categorized by tags.

The component allows you to search for macros and, by using specific keys (*title:*, *tag:*, *desc:*), you can narrow your search to focus on specific elements.

If you have suggestions for new macros, we warmly welcome your input!

Cluster zoom

The application can be used to display the results generated by using ReporTree. In short, ReporTree allows the generation of a series of zooms on the clusters by specifying the samples of interest and some parameters such as thresholds and distance; in this context, it is possible to indicate their presence to the dashboard, so they can be displayed accordingly. On the tree's interface, you can click on **Available zooms** you can access the **Zoom tools** to highlight any clusters that are available. You can select the *zoom type*, the *threshold* and the *cluster* to which the sample can be associated as well as the **involved nodes**. Additionally, the functions allow you to **load in a new tab** a modified cluster for which you can choose if keeping **current settings** or not.

In the following video, you will find a possible use case.

Cluster zoom

Tree

This is, of course, the main component and, therefore, always visible. It displays the tree visualization of the data and is fully interactive. You can interact with this component directly, through the context menu or through all the other components.

You can interact directly with the tree by moving individual nodes or hovering over them to see tooltips; you can also select nodes by pressing `Shift` and clicking on them or drawing an area with the mouse. The selection will also be reflected in the **Map** and **Metadata** components.

> **TIP**
> To enable the selection in the chart workspace, press and hold the `Shift` key on the keyboard: as long as the slider appears as a crosshair, you can select or deselect individual nodes by clicking on them. You can also select multiple nodes by drawing a selection area (dragging the cursor while holding down the left mouse button).

More interesting things can certainly be done through the **context menu** or, as mentioned, through interactions with the other components (which we have discuss in their respective sections). You can access the **context menu** by right-clicking on the tree or by using the button at the top-right corner with three vertical dots. The menu options are quite self-explanatory and we invite you to try them out to better understand their functionality, through interaction. Basically, they support you in modifying the tree view.

Here, we will focus on the *save* function.

- **Save as complete json (.json)**: it allows you to save the current work in a complete `.json` file containing all the information about the tree, its metadata and the current configurations of the different components. This functionality is crucial, as it is the only way to save your progress in editing SPREAD, resume it at a later time or share it with others. The `.json` file stores information such as: node placement and color, current category, node visualization modes on the map, workspace layout configuration and even which settings cards are expanded or collapsed.
- **Export newick (.nwk)**: it exports data in the `.nwk` format.
- **Export metadata (.tsv)**: it exports metadata in the `.tsv` format.
- **Export geoJson (.geoJson)**: it exports the spatial metadata in the `.geojson` format.

All files produced are compatible to be reloaded later in the dashboard, as mentioned in Using SPREAD.

Tree tools and function Save SPREAD

9.2 Auspice

The Auspice dashboard, integrated into the platform, allows you to view phylogenetic trees, associated metadata and geographic data produced by Augur.

- Official documentation

Access to the Auspice integrated dashboard is allowed from the "view analyses" section which displays the list of all the analyses and their related *workflows*. Access to Auspice dashboard is available in the *Result data* section of the card for each successfully completed **Augur** analysis.

9.2.1 Quick guide

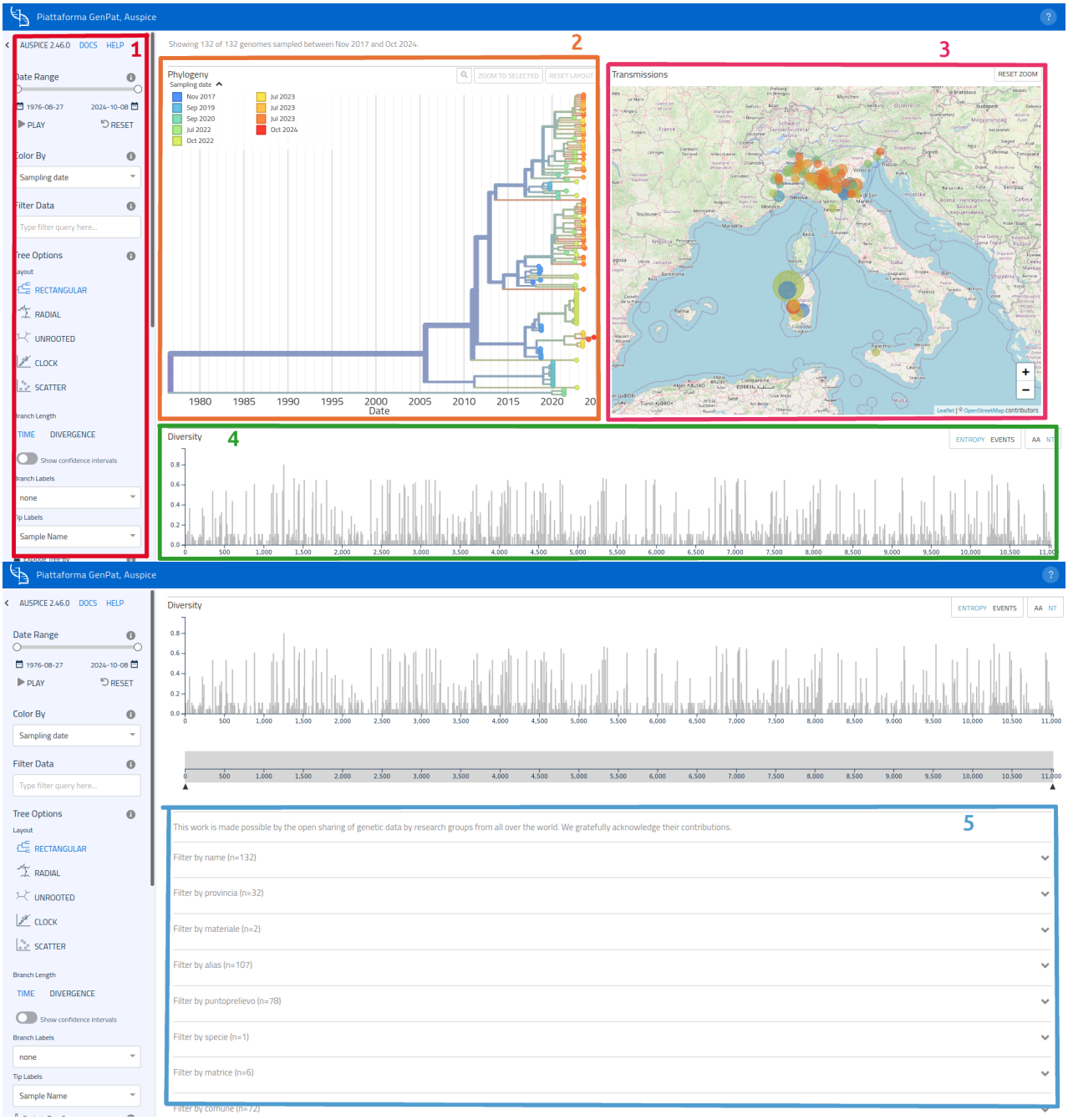
The following video provides a quick guide to Auspice's interface, with demonstration for menu, map and graph use.

Auspice Demo

9.2.2 Main user interface

Auspice's user interface is composed of 5 main areas:

- an options side bar on the left side (**1**);
- the phylogenetic tree on the left (**2**);
- a map on the right side, where *clusters* can be displayed depending on geographic data (**3**);
- the diversity graph (which can be toggled to display entropy/events), below the map and phylogenetic tree (**4**);
- a filters area, just below the graph (**5**).



Options bar

The options bar allows the user to access the Auspice's official manual pages and docs and it is divided in sub-categories, in order to manage different aspects of how data is displayed:

- View, colors and data filter;
- Tree options;
- Map options;
- Animation options;
- Layout and panels view.

The menu is thoroughly described in Auspice's official guide; furthermore, each element of the side menu has an associated **info** button ( ), which displays a description of the respective option when hovering the mouse pointer above it.

The use of the options bar is showed in the introductory video whereas, in this manual page, the dedicated sections will show some of the most important features of each dashboard element. For in-depth guides descriptions please consult Auspice's official guide.

Phylogenetic tree

Layout

The phylogenetic tree can be displayed as a left side panel (default) or over the entire width of the page (you can activate the "**FULL**" layout mode in **Display** options of the side bar).

The side bar options also allow to choose from 5 different layouts for the phylogenetic tree (Rectangular, Radial, Unrooted, Clock, Scatter).

The side menu presents some tree branche view options as well that include toggles to:

- switch between representation of sample distances as time distance scale and divergence scale (*Branch Lenght*);
- show or hide confidence intervals and branch labels (*Branch Labels*);
- select displayed information in the leaf labels (*Tip Labels*);
- display the tree as a break-out chart based on the associated metadata (*Explode Tree By*).

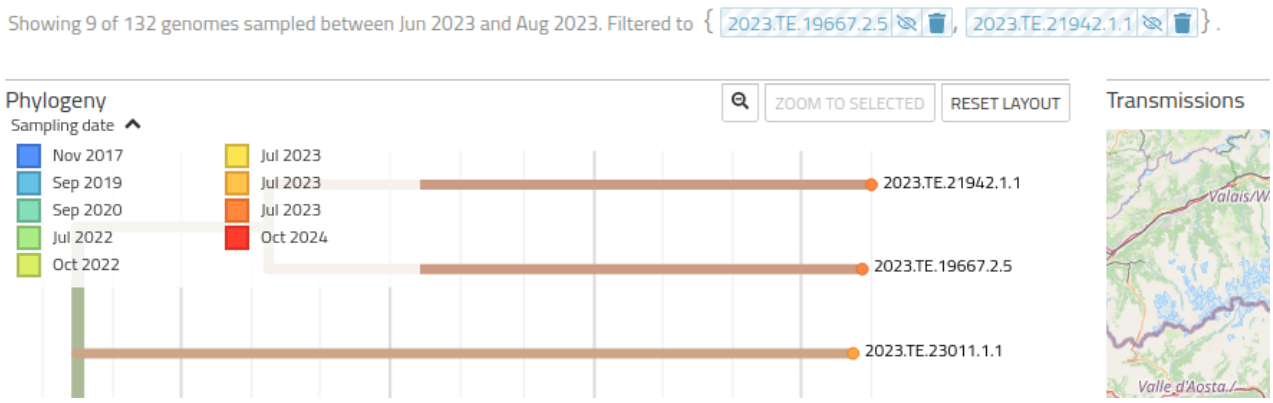
Legend

The **legend** (in the top-left corner of the *Phylogeny* panel) can be displayed or hidden by clicking on the arrow button (^); furthermore, the legend's elements, - which correspond to the sampling dates considered as the "leaves" of the phylogenetic tree -, are highlighted on mouseover.

Zoom and focused samples

Hovering the mouse over a sample in the tree will show a summary of the sample's related information whereas clicking on one of them or on its branches will focus on that area and display the full information sheet of the selected sample. Zoom options can be managed using the 3 buttons in the top dash: .

When clicking on a sample in the graph, the selected sample will be added to a filter used for data visualization; the full list of samples in the filter is displayed the area above the graph:



This interaction empowers users with the ability to hide or display the selected samples individually (hiding all the others in the graph); similarly, it is possible to delete samples from the filter individually. Moreover, samples can also be removed in bulk from the filter using the option in the **Filter Data** section of the side bar which allows to activate/deactivate the visualization of selected samples as well.

Map

Some simple commands are available to navigate the map:

- Use the mouse wheel while holding the `Ctrl` key to adjust zoom;
- Zoom can be also adjusted using the **+/-** buttons in the lower-right corner of the map. Double clicking on a map zone is equivalent to pressing the **+** button;
- Holding the `Shift` key (or `Maiusc` key, depending on keyboard layout) you can create a selection area with the mouse (hold the left mouse button and drag) to zoom inside it;
- Use the button in the upper-right corner of the map to reset the zoom.

Map and graph are tied to each other: a node selection in the graph will also highlight its position on the map.

Map available options in the side bar include the choice to adjust the geographic resolution (depending on the level of accuracy of metadata) and to show or hide transmission lines.

Diversity graph

A third graph, the "*Nucleotide diversity of genome*", located below the map and phylogenetic tree, displays diversity among samples, based on nucleotides or amino acids (the latter are not available in the platform's integration). The graph can show either the "*Entropy*" page or the "*Events*" page (the toggle to switch from one page to the other is in the upper-right corner) and, selecting a range in the graduated line at the bottom of the graph, you can also adjust the zoom.

Filters

Further down the diversity graph, you can find a list of applicable filters, one for each metadata category associated to samples. Filter types can be selected by clicking on the corresponding card to expand it.



Example: to filter Equus caballus samples, open the metadata card "*Filter by Host*" and select Equus caballus.

Animation

Samples can be shown on the map and on the graph through an animation that displays them consecutively, starting from those closer to the tree's root and going towards the leaves. From the side menu, you can choose the animation speed, the activation of loop mode and of cumulative mode (the latter maintains the display of all map nodes after their first appearance rather than displaying them temporarily).

10. SOP

10.1 NGSManagerSOP

This page contains the list of versions of the **NGSManagerSOP** pipeline and the list of analyses (steps) it includes.

> The NGSManagerSOP software is version-controlled on the corresponding Gitea service page (internal to IZSAM).

10.1.1 Versions

NGSManagerSOP - version 2.6

(since Oct. 13, 2026)

It includes the following bioinformatics software:

- fastqc: 0.11.5
- fastp: 0.23.4
- shovill: 1.1.0
- kmerfinder:1.0
- mlst: 2.23.0 - DB updated as of 2026-10-12 - updated MLST schemes for *Listeria monocytogenes* (`listeria_2`) and *Campylobacter* (`campylobacter`)
- chewbbaca: 2.8.5 - added the Oxford cgMLST scheme for *Campylobacter jejuni* and *C. coli*

Corresponding to the Nextflow steps:

- step_1PP_trimming__fastp.nf
- step_2AS_denovo__shovill.nf
- step_2AS_mapping__coverage_estimate.nf
- step_3TX_species__kmerfinder.nf
- step_4TY_MLST__mlst.nf
- step_4TY_flaA__flaA.nf
- step_4TY_cgMLST__chewbbaca.nf

NGSManagerSOP - version 2.5

(since Jan. 24, 2026)

It includes the following bioinformatics software:

- fastqc: 0.11.5
- fastp: 0.23.4
- shovill: 1.1.0
- kmerfinder:1.0 - tool DataBase updated as of 2021-03-30
- mlst: 2.23.0 - Pasteur DB updated as of 2026-01-20
- chewbbaca: 2.8.5

Corresponding to the Nextflow steps:

- step_1PP_trimming__fastp.nf
- step_2AS_denovo__shovill.nf
- step_2AS_mapping__coverage_estimate.nf
- step_3TX_species__kmerfinder.nf
- step_4TY_MLST__mlst.nf
- step_4TY_flaA__flaA.nf
- step_4TY_cgMLST__chewbbaca.nf
- step_4TY_wgMLST__chewbbaca.nf

NGSManagerSOP - version 2.4

(since Jan. 20, 2026)

It includes the following bioinformatics software:

- fastqc: 0.11.5
- fastp: 0.23.4
- shovill: 1.1.0
- kmerfinder:1.0 - tool DataBase updated as of 2021-03-30
- mlst: 2.23.0 - Pasteur DB updated as of 2026-01-20
- chewbbaca: 2.8.5

Corresponding to the Nextflow steps:

- step_1PP_trimming__fastp.nf
- step_2AS_denovo__shovill.nf
- step_3TX_species__kmerfinder.nf
- step_4TY_MLST__mlst.nf
- step_4TY_flaA__flaA.nf
- step_4TY_cgMLST__chewbbaca.nf
- step_4TY_wgMLST__chewbbaca.nf

NGSManagerSOP - version 2.3

(since Jun. 23, 2025)

It includes the following bioinformatics software:

- fastqc: 0.11.5
- fastp: 0.23.4
- shovill: 1.1.0
- mlst: 2.23.0 - Pasteur DB updated as of 2025-06-23
- chewbbaca: 2.8.5

Corresponding to the Nextflow steps:

- step_1PP_trimming__fastp.nf
- step_2AS_denovo__shovill.nf
- step_4TY_MLST__mlst.nf
- step_4TY_flaA__flaA.nf
- step_4TY_cgMLST__chewbbaca.nf
- step_4TY_wgMLST__chewbbaca.nf

NGSManagerSOP - version 2.2

(since Nov. 12, 2024)

It includes the following bioinformatics software:

- fastqc: 0.11.5
- fastp: 0.23.4
- shovill: 1.1.0
- mlst: 2.23.0 - Pasteur DB updated as of 2024-11-06
- chewbbaca: 2.8.5

Corresponding to the Nextflow steps:

- step_1PP_trimming__fastp.nf
- step_2AS_denovo__shovill.nf
- step_4TY_MLST__mlst.nf
- step_4TY_flaA__flaA.nf
- step_4TY_cgMLST__chewbbaca.nf
- step_4TY_wgMLST__chewbbaca.nf

NGSManagerSOP - version 2.1

(since Aug. 28, 2024)

It includes the following bioinformatics software:

- fastqc: 0.11.5
- fastp: 0.23.1
- shovill: 1.1.0
- mlst: 2.23.0 - Pasteur DB updated as of 2024-08-19
- chewbbaca: 2.8.5

Corresponding to the Nextflow steps:

- step_1PP_trimming__fastp.nf
- step_2AS_denovo__shovill.nf
- step_4TY_MLST__mlst.nf
- step_4TY_flaA__flaA.nf
- step_4TY_cgMLST__chewbbaca.nf
- step_4TY_wgMLST__chewbbaca.nf

NGSManagerSOP - version 2.0

(since Sep. 27, 2023)

It includes the following bioinformatics software:

- fastqc: 0.11.5
- fastp: 0.23.1
- shovill: 1.1.0
- mlst: 2.23.0 - Pasteur DB updated as of 2023-09-27
- chewbbaca: 2.8.5

Corresponding to the Nextflow steps:

- step_1PP_trimming__fastp.nf
- step_2AS_denovo__shovill.nf
- step_4TY_MLST__mlst.nf
- step_4TY_flaA__flaA.nf
- step_4TY_cgMLST__chewbbaca.nf
- step_4TY_wgMLST__chewbbaca.nf

NGSManagerSOP - version 1.1

(from May 10, 2023 to September 26, 2023)

The mlst software has been updated to version 2.23.0

NGSManagerSOP - version 1.0

(from February 6, 2021 to May 9, 2023)

It includes the following bioinformatics software:

- fastqc: 0.11.5
- trimmomatic: 0.36
- spades: 3.11.1
- mlst: 2.16.0
- chewbbaca: 2.8.4

Corresponding to the Nextflow steps:

- step_1PP_trimming__trimmomatic.nf
- step_2AS_denovo__spades.nf
- step_4TY_MLST__mlst.nf
- step_4TY_flaA__flaA.nf
- step_4TY_cgMLST__chewbbaca.nf
- step_4TY_wgMLST__chewbbaca.nf

10.1.2 Validation and reproducibility

Datasets

Validation and reproducibility tests are performed on the datasets whose addresses are listed below:

- dataset *Listeria monocytogenes* - izs.te.b912.1.sop001
- dataset *in silico species determination* - izs.te.b3.1.2.sop180
- dataset *Campylobacter* - izs.te.b3.1.3.sop071
- dataset *Salmonella* - izs.te.b3.1.3.sop078

Validation

- Validation NGSManagerSOP pipeline - izs.te.b912.1.sop001 (*Listeria monocytogenes*);
- Validation NGSManagerSOP pipeline - izs.te.b3.1.2.sop180 (*in silico species determination*);
- Validation NGSManagerSOP pipeline - izs.te.b3.1.3.sop071 (*Campylobacter*);
- Validation NGSManagerSOP pipeline - izs.te.b3.1.3.sop078 (*Salmonella*).

Reproducibility

- Reproducibility of the NGSManagerSOP pipeline results - izs.te.b912.1.sop001 (*Listeria monocytogenes*)
- Reproducibility of the NGSManagerSOP pipeline results - izs.te.b3.1.2.sop180 (*in silico species determination*)
- Reproducibility of the NGSManagerSOP pipeline results - izs.te.b3.1.3.sop071 (*Campylobacter*)
- Reproducibility of the NGSManagerSOP pipeline results - izs.te.b3.1.3.sop078 (*Salmonella*)

Control of test outputs

The links listed above lead to the summary pages of the builds for the Listeria and Campylobacter SOPs related to the validation processes (the former) and reproducibility processes (the latter). For all pages, it will be possible to view a graph showing the trend of recent processes launched for validation or reproducibility.

To check the validation or reproducibility status of the pipeline, it is sufficient **to access the links** provided above and **verify that all the colored boxes in the graph are green**.

> The statuses of the individual samples in the validation dataset are available in the .csv file listed at the top of the page. It reports the results as *PASS* (if the results fall within the acceptable limits for validation) or *FAIL* (if the obtained values are not acceptable for validation).

The following video briefly shows how to check the validation status of the NGSmanagerSOP pipeline. The procedure is identical for checking reproducibility, except for the absence of the validation dataset files.

11. News

11.1 2026

11.1.1 September 2026

Version 26.09.1

Version 26.09.1 renews the profile selection at login and the password recovery, introduces an e-mail notification when analyses and uploads finish, and makes everyday work on tables, tags and the file tree more comfortable. The release is completed by improvements to notifications and messages and a number of bug fixes.

This note also collects some improvements released in the past months and not yet communicated: they are marked as *already available*.

The updates in this release cover the following areas:

- login: profile selection and password recovery
- e-mail notification when analyses and uploads finish
- tables: export and preferences
- tags and cart
- file tree of downloads and analysis results
- interface: language and icons
- notifications and messages
- localization improvements and minor bug fixes.

Login: profile selection and password recovery

The login page keeps the functions it had; what changes is how the profile is chosen and the password recovery procedure.

Profile selection

- After login, users with several profiles pick one in a dedicated window listing the available groups; when there are many profiles a search field appears to filter them;
- the last profile used is offered first and remembered for the following logins;
- the same window replaces the *Change group* submenu of the user menu: the profile in use is highlighted and the switch happens immediately, without a confirmation dialog;
- the window can be used with the keyboard: arrow keys to move through the groups and *Enter* to confirm;
- group names and the window interface appear in the user's language, and the login page opens in the language of the user who last signed in from that browser.

Password recovery and management

- The *Forgot password* link under the login form opens a window where the username or e-mail address is entered: an e-mail is then received, in the user's language, with the instructions to set a new password;
- the *Change password* entry of the user menu opens a new window listing the password requirements, which update as the password is typed;
- new users receive, when their account is activated, an e-mail with the instructions to set their own password at the first login.

For more information see the Password management page.

E-mail notification when analyses and uploads finish

On the *Run Analyses* and *Upload* pages, the **E-mail me when finished** checkbox is available next to the run button: when selected, a notification e-mail is sent when the process completes. After the request is sent the checkbox is disabled, confirming that the choice has been recorded.

Tables: export and preferences

- **Export**: the export window of tables gains the *Columns to export* choice, with the options *Visible only* (default) and *All columns*, which also includes hidden columns.
- **Grid preferences**: the save button shows at a glance whether preferences are already saved (check icon and main colour) and the tooltip distinguishes between *Save* and *Update*; the *Update* and *Clear* menu entries have their own icons.
- **Quick filters**: fixed the loading of filters, which in some cases showed those of another table or did not appear at all.
- **XLSX export**: numbers, dates and yes/no values keep their data type in the Excel file, with cell formats and checks, instead of being exported as text. *Already available.*
- **Search**: the table search starts only from three characters; with fewer characters a warning is shown and the table is not reloaded. *Already available.*
- **Advanced filters**: the parameters window shows the name of the saved filter and a summary of the other applied criteria; numeric filters use text labels (*greater than*, *less than*, *equals*) instead of symbols. *Already available.*

Tags and cart

Tag navigation from the sidebar has been made more direct:

- a tag is activated or deactivated by clicking its row; double click still opens the rename and clicking a folder expands it;
- the three-dots icon opens the same menu as the right click;
- in the *Favourites* branch the click correctly activates and deactivates the tag, and the active tag is highlighted with the main colour;
- moving, renaming or starring a tag no longer reloads the page in use, so nothing typed is lost;
- fixed the drag and drop freeze that occurred after dropping a tag on a forbidden target.

For the cart, activation, also via keyboard shortcut, always reloads the table and updates the tag menu. The buttons leading to the samples page (cart menu, cart and tag management, reports) open the page configured for the user's group.

File tree of downloads and analysis results

The file tree, used in the file selection step of the download procedure and when browsing analysis results, offers new tools:

- the *Sort* select reorders the nodes (default, A Z, Z A) while keeping search and selection;
- the *View* button opens HTML results in a new browser tab;
- long names are shortened in the middle and the full name appears on hover (or on long press on touch screens);
- the *Copy* button copies the node name, also available with Ctrl/Cmd+C on the selected node;
- row actions stay visible on narrow screens too.

Interface: language and icons

- The interface language can be changed from the user menu through the entry with the globe icon; the choice is saved in the user preferences. *Already available.*
- The icons of the whole platform have been renewed with a single, more readable set. *Already available.*

Notifications and messages

- In the notification centre the text can be selected and every notification has an icon to copy title and content; a warning appears when new unread notifications arrive. *Already available.*
- When several messages are shown, a **Close all** button with the number of open messages appears below the stack.
- Very long error or warning messages are shortened; the **Read more** link opens the full text.
- Messages no longer cover the page footer.

Localization and minor bug fixes

- EN/IT/FR translations updated for the new interface labels and some French labels corrected;
- more compact table toolbar.

11.1.2 April 2026

Version 26.04.1

The updates in this release cover the following areas:

- addition and editing of metadata, addition of sample pre-acceptance workflows, XLSX import/export;
- tag navigation, tag management and cart, dashboard and custom pages
- views, permissions and backend stability
- UI/localization improvements and minor bug fixes.

Tag Navigation, Tag Management and Cart, Custom Pages

Improvements to the navigation and content management experience have been introduced:

- introduction of tag navigation system: accessible from the sidebar, the menu works like a tree-structured filesystem and allows fast and direct tag management;
- new cart utilities for tags, including removal from tags and fixes for adding, activation and persistence after logout;
- label updates for *Download*, *Upload from NCBI*, and *Report* pages.

Tag Navigation

Views, Permissions and Backend Stability

This release improves reliability and consistency in the management of views, attributes and permissions:

- improved management of views and permissions for greater stability;
- fixed the behavior of required fields in views;
- resolved issues when adding/removing fields in linked tables;
- improved consistency in attribute loading, sorting, and grouping;
- fixed various anomalies in queries and cache management;
- strengthened the reliability of permission restoration;
- improved translation fallback when some localizations are not available.

UI/Localization Improvements and Bug Fixes

The release is completed by additional cross-cutting optimizations:

- centralized fixes for the `datalist` component (input selection/validation and row styling);
- fixes for grid refresh in the absence of the DOM target and for modal closing in the action box;
- improved positioning of deep-link cards and handling of closing non-compatible modals;
- replacement of native alert/confirm with application modals;
- updates to EN/IT/FR localization for new UI keys;
- addition of the admin module toggle in the header and update of the administrator shortcut behavior;
- SASS build fixes and minor autoload corrections.

11.1.3 January 2026

Version 26.01.1

Version 26.01.1 introduces updates to the user experience and layout of *Run Analyses*, *Downloads*, *Upload from NCBI* and *Reports*: the selection of tools is more descriptive, search functionality has been improved and preferences can now be saved. This version also includes optimizations aimed at improving performances.

The updates in this release cover the following areas:

- UI/UX updates
- preferences selection and saving
- performance and optimizations
- localization improvements and minor bug fixes.

UI/UX updates

In this version, the layouts and behaviors of *Run Analyses*, *Downloads*, *Upload from NCBI*, and *Reports* have been updated, featuring a general refresh of the interface and the introduction of more descriptive tools.

The main new features include:

- tool menu with descriptive text and a visible link to the wiki (where available) before selection;
- indexing of descriptions for more effective and precise search;
- correction of search behavior in the tree view to align it with the card view;
- UI improvements, including the use of custom scrollbars and the resolution of issues related to window resizing;
- improvement of translations for certain UI elements.

UI/UX updates

Preferences selection and saving

For *Run Analyses*, *Downloads*, *Upload from NCBI*, and *Reports* pages, it is now possible to manage preferences through a dedicated system, which includes:

- support for user preferences, for favorite tools, based on the user role;
- addition of a *Favorites* category for tools marked with a star.

Preferences selection and saving

Performance and optimizations

Rendering performance for classes and views has been improved with buffering along with smoother animations:

- optimized initial load for classes, views, and join views;
- progressive rendering with smoother animations when elements enter the viewport.

11.2 2025

11.2.1 December 2025

Version 25.12.1

Version 25.12.1 introduces an update regarding the use of certain features, aimed at providing the user with an even more smoother experience on the platform.

The updates in this release concern the following areas:

- help desk
- run analysis and download
- printing functions
- performance improvements and minor bug fixes.

Help desk

This release allows the user to contact the Help Desk directly from the platform, while keeping the possibility to describe the issue, also through images or videos, related to the problem to be solved. The following video shows how to use the *Help Desk* function:

Help Desk update
Further information is available on the dedicated page.

Run analysis and download

Starting from this release, the display of the tools is filtered through the pre-selection of the first available category instead of *All*; moreover, the applied filters will remain stored during the launch session. The following video shows this upgrade:

Category pre-selection

Printing functions

With this release, the print option is restored for:

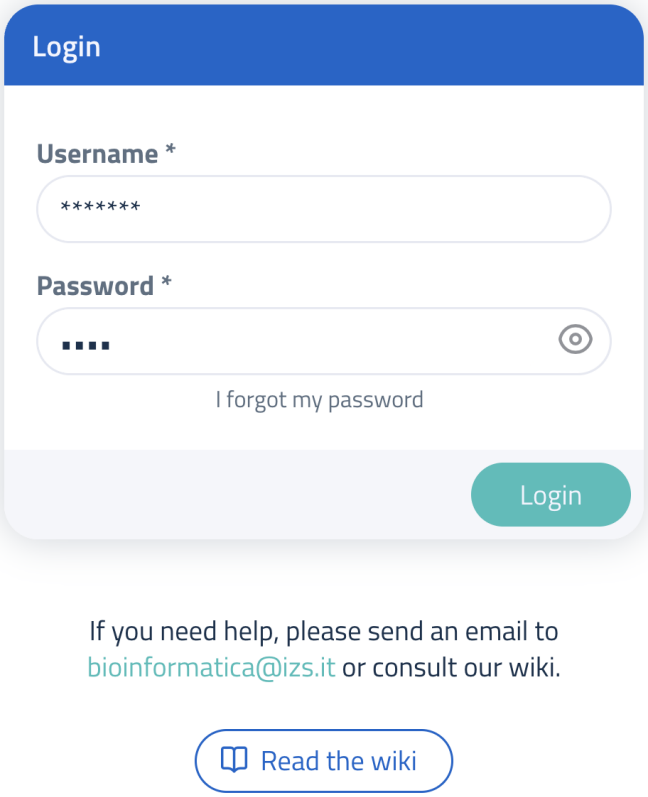
- **single analysis card** using the button , at the top right of the card;
- **a multiple card selection** accessible from the contextual menu via the button .

12. Support

12.1 Password management

The platform has been extended with an **external authentication system**; therefore, for procedures related to password change or reset, you may be redirected to a service outside of it.

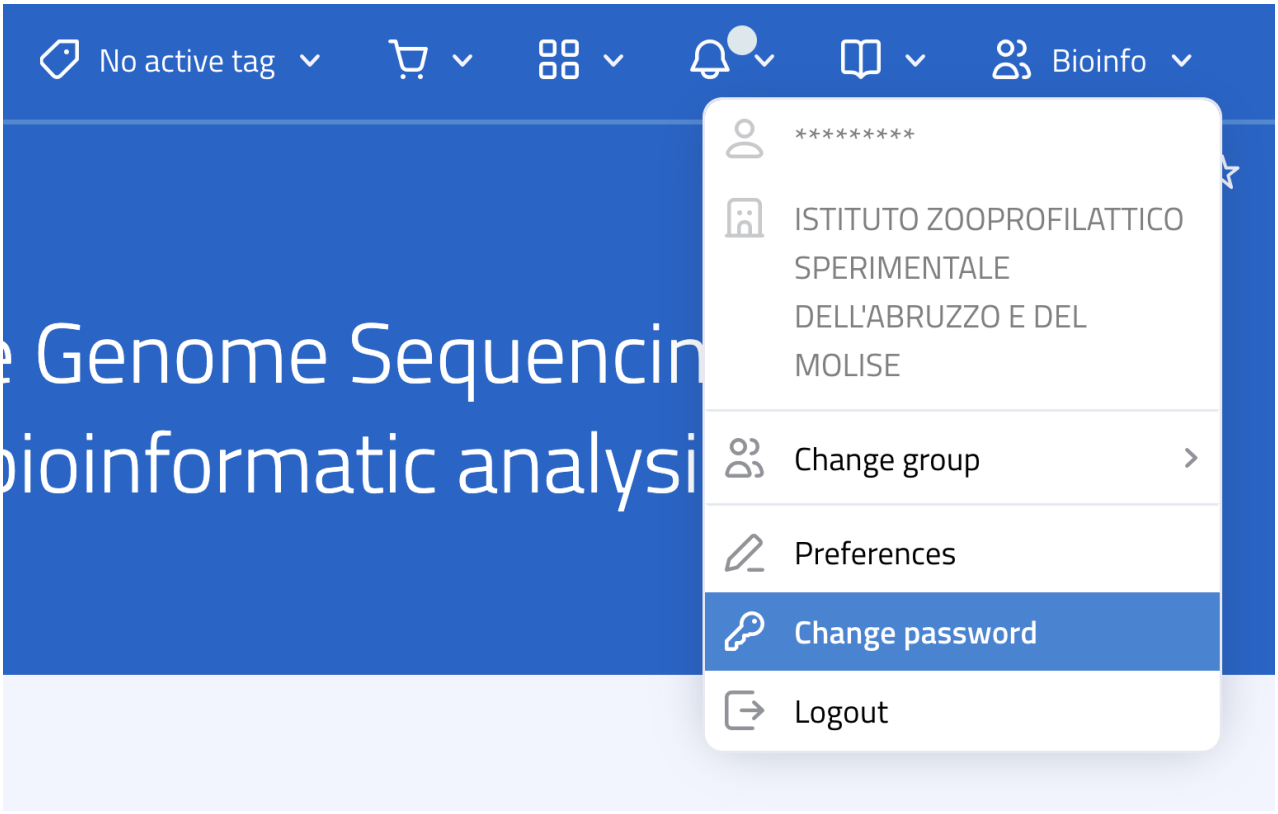
For your password recovery, it is possible to use the **I forgot my password** tip, available below the login form.



Change password

12.1.1 Password change procedure

To change your password, use the **Change password** option, available in the **user menu** at the top right of the platform.



Change password option

By selecting the `Change password` option, a modal window will open briefly explaining the required steps. By clicking `Proceed` *, you will be redirected directly to the external service: https://portale.izs.it/sspr/public/forgottenpassword.



Forgotten Password

If you have forgotten your password, follow the prompts to reset your password.

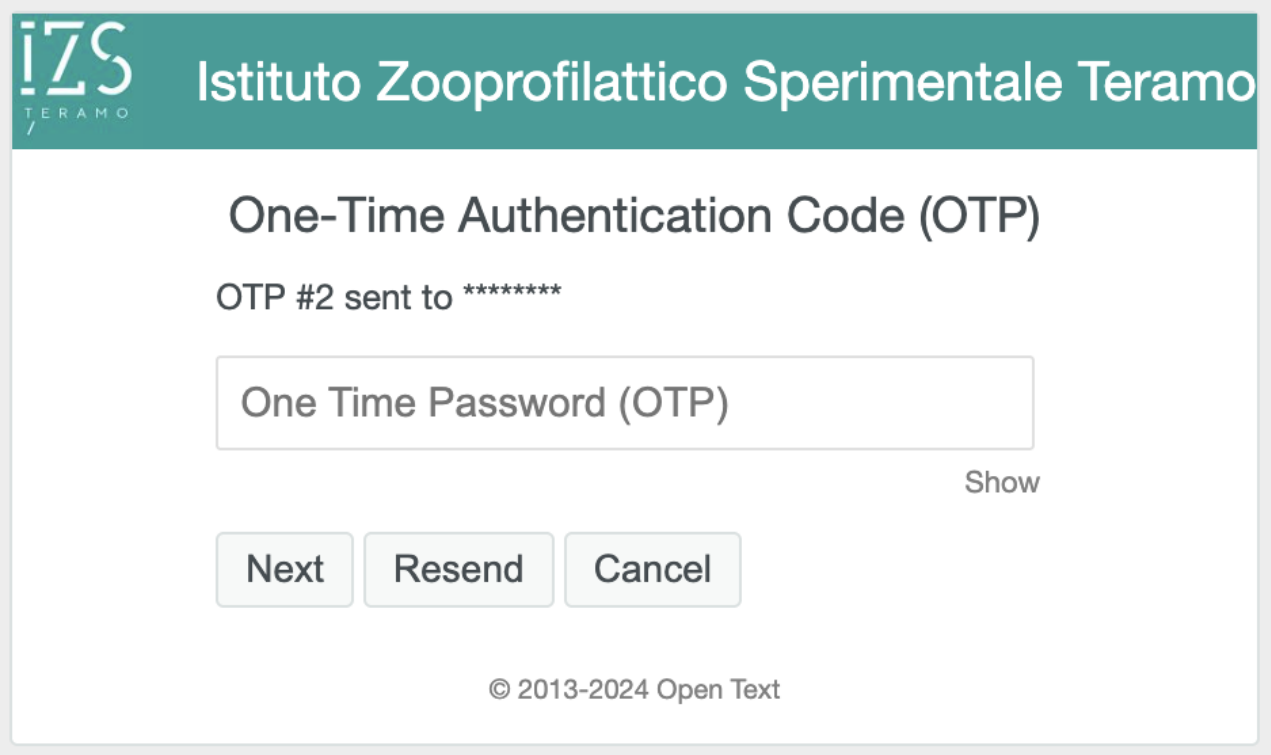
Username*

Search Cancel

Fogotten password option

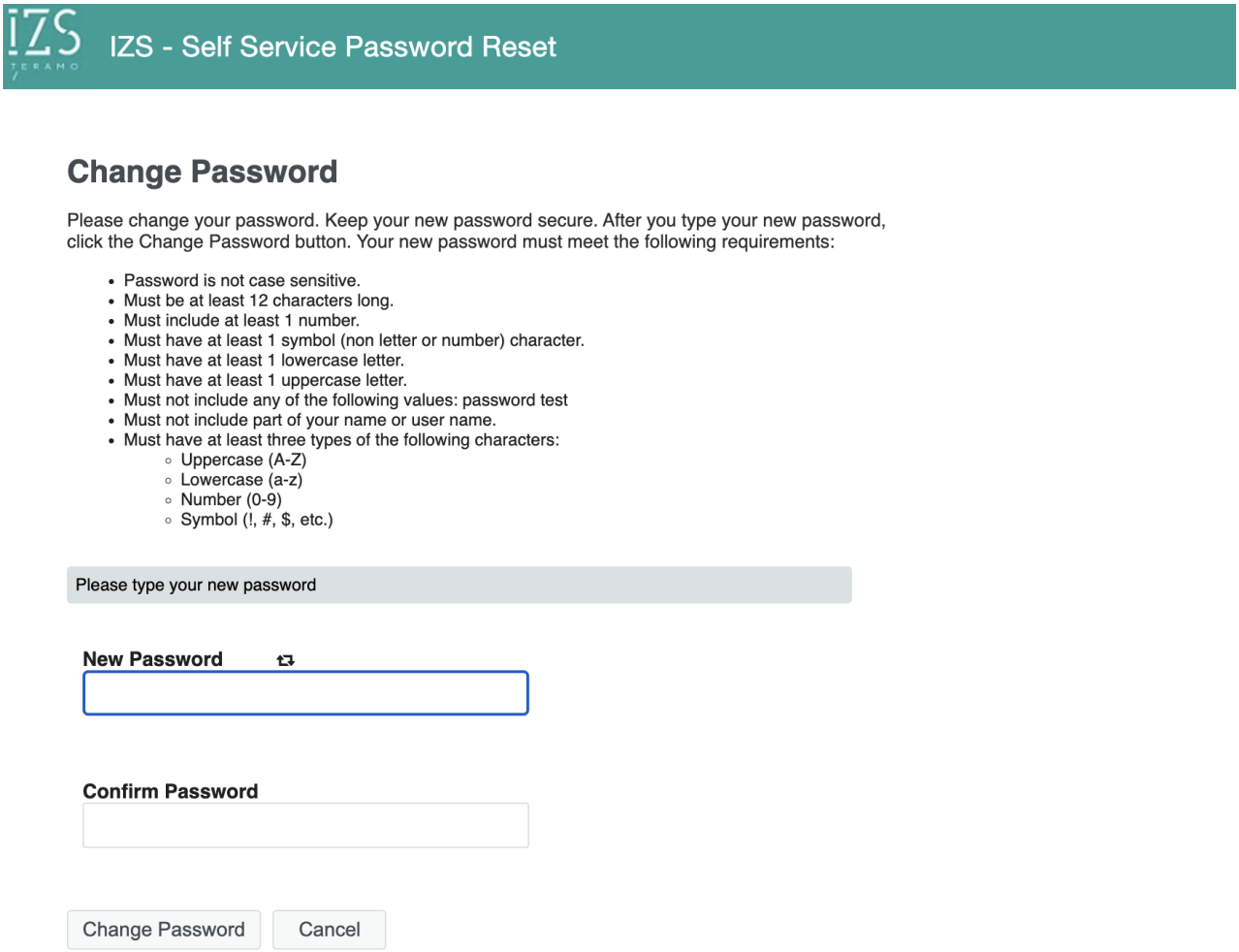
> **Tip** You can change the **language** using the globe icon at the top right. The page title and description may refer to the **forgotten password** procedure, but don't worry as it is the same flow used to change your password as well.

You will be asked to enter your **username** (the same one used to log into the platform). After verification, you will receive an **email**, at the address associated with your account, containing an **OTP code** (*One-Time Password code*).



OTP entry

Once you have entered the OTP code, you will be able to set a **new password**.



Set new password

The reset process may take a few seconds. A **progress bar** will display the status of the operation; once completed, you will be able to go back to the platform.

12.1.2 Forgotten Password

If you have forgotten your password, click **I forgot my password** link, below the login form, to start the recovery procedure.

The steps are identical to those described in the Password Change Procedure section.

12.1.3 Password requirements

To ensure an appropriate level of security, passwords must meet the following requirements:

- **Minimum length**: at least **12 characters** (it is recommended to use complex passwords of 15 or more characters);
- **Uniqueness**: the system does not allow setting a password identical to those used previously;
- **Complexity**: the password must meet **at least three** of the four following criteria:
- uppercase (A–Z);
- lowercase (a–z);
- number (0–9);
- symbol (!, #, $, etc.);
- characters from other languages not listed above;
- **it must NOT** contain parts of your name or username.

Final recommendation: always use long, unique passwords that are not linked to personal data. Consider using a secure password manager to store them.

12.2 Help desk

For support requests, you can use the **Help desk** features, accessible via the wiki's contextual menu located at the top right of the platform (book-shaped icon). Clicking it will open a modal that allows you, without leaving the application, to:

- select the relevant problem area;
- describe any issue, also by attaching an image or a video of up to 25MB (accepted formats: .png, .jpg, .jpeg, .mp4).

The following video shows the procedure:

Help Desk

12.3 Useful videos

In this section the user is provided with useful videos for training or general use of the platform. The videos may refer to events or training activities established within other projects. The digital video library will grow over time and be categorized, with the addition of further videos.

12.3.1 NGS & Bioinformatics - ERFAN

Videos produced within the ERFAN project

Lesson 1 - 22/02/2022

NGS Bioinformatics Lesson - 1

Lesson 2 - 01/03/2022

NGS Bioinformatics Lesson - 2

12.3.2 CIS demo at Matrix WP6

Presentation of COHESIVE Information System at Matrix WP6 (24 February, 2022)

Cohesive demo at matrix wp6

12.4 Keyboard shortcuts

The main functions can be launched via the keyboard, using specific predefined key combinations. Some combinations may work at a general level of the application or only within specific pages. Please find here below a list of the shortcuts defined by default.

> **Important**
> The list is continuously updated, therefore we invite you to visit this page frequently.

12.4.1 Shortcuts list

Function	Combination	Pages
Enable/disable multiple selection	`ctrl+shift+m`	Classes and Views (pages with selectable grid)
Select all	`ctrl+shift+a`	Classes and Views (pages with selectable grid)
Empty cart	`alt+shift+e`	All
Add to cart	`alt+shift+a`	Pages where the *Add to cart* button is present
Replace cart	`alt+shift+r`	Pages where the *Add to cart* button is present
Remove from cart	`alt+shift+d`	Pages where the *Add to cart* button is present
Activate/Deactivate cart	`alt+shift+c`	All
Deactivate tag	`alt+shift+t`	All

12.5 FAQ

This page lists frequently asked questions (FAQ) regarding the use of the platform and their corresponding answers. Before contacting support, please check whether your issue can be solved by reading one of the answers on this page.

12.5.1 Data management

How can I change the character used as a column separator in Excel?

Answer: The column separator can be specified at different levels. Below are guides for the main cases:

- to change the default column separator in Excel, you need to change the character used as a separator in Windows system settings;
- to import into Excel a file with a different separator, you can refer to the various guides available online;
- to split lines in a text file using another character, leveraging Excel built-in features, you can follow this guide: https://it.extendoffice.com/documents/excel/1786-excel-split-text-by-space.html;
- if you want to convert a file, you can use an online tool such as the one available at this link: https://onlinecsvtools.com/change-csv-delimiter.

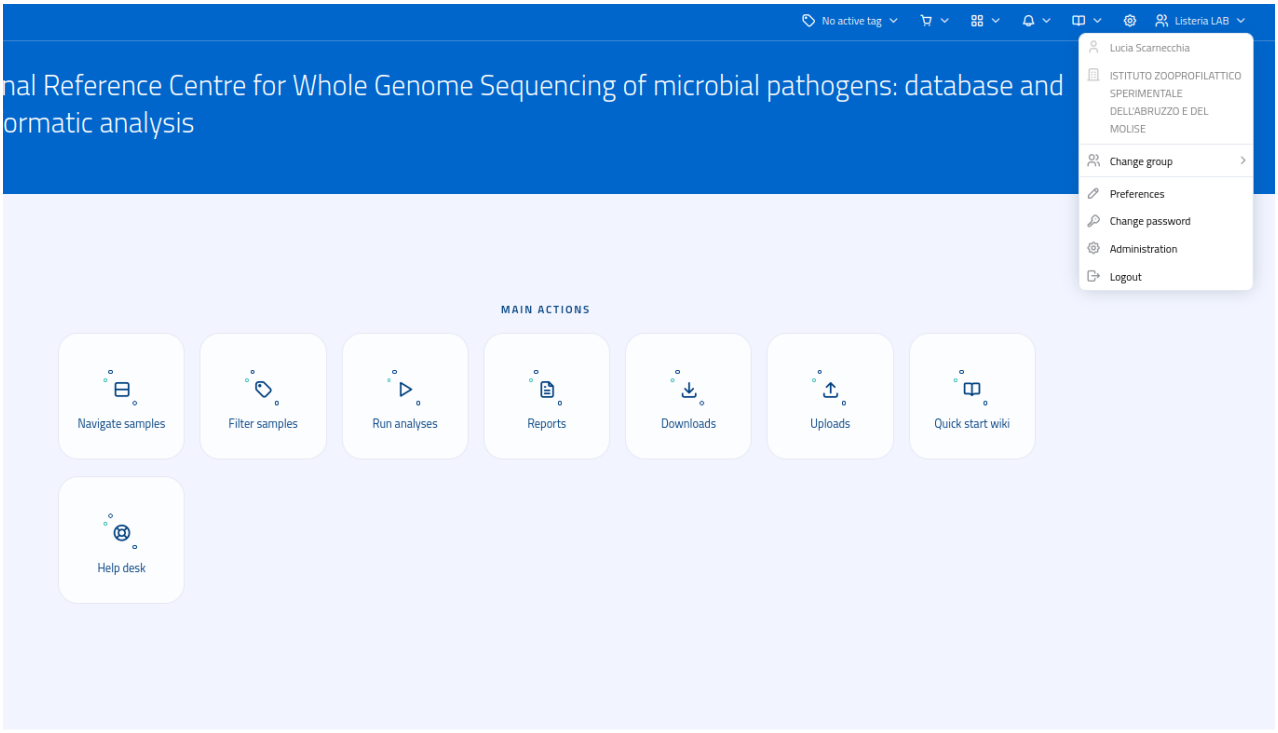
Note: `.csv` files created for the platform must use the semicolon (`;`) as the list separator, whereas the file created for the samplesheet requires a comma (`,`) as the separator.

I can't export a table as CSV.

Answer: CSV export of a platform table is handled through a popup window. Check whether the browser has opened additional windows and, if not, verify that the browser has not blocked popups from the site. Normally, almost all browsers display a small banner at the top of the page, asking the user how to handle the request to display a pop-up.

Can I change the character used as a column separator for CSV export?

Answer: Yes, in the platform it is possible to change the default column separator, used for CSV export, in the **Preferences** section of the user menu: once in *Preferences*, you can change the character in the "**CSV Separator**" field. Once the changes have been saved, from that moment on, when sample lists are exported as CSV files, the selected character will be used as the separator.



12.5.2 Cart and tag

I get an error message when I try to add many samples to the cart.

Answer: The cart can contain a maximum of 2000 samples. If you need to work on a larger number, we suggest to create a tag.

How can I duplicate/clone a tag?

Answer: Activate the tag to duplicate, go to the `Navigate samples` section (from the dropdown menu next to the cart icon), select the ones you are interested in, add them to the cart and create a new tag from the cart.

12.5.3 Interface

How can I display columns with additional information in a sample table?

Answer: Additional columns, including those containing result information (such as calculated species, quality check, N50, PFGE, percentage of called alleles, etc.), can be displayed as described in the Wiki page "Table features".

Additional information, not directly available in the sample table, can be found in the "Analysis Results" page (from the left navigation menu, `Main objects > Results of the analyses`).

12.5.4 Data check and visualization

Where can I find a sample checks?

Answer: In the side *Navigation* menu there are the Check tables: `Report > Detail > Main Quality Check` and `Reports > Detail > Checks/Metrics`. Using a filter, (an active cart or a tag), you can filter the rows of the tables to find the sample(s) of interest. The tables of checks are described in the related page "Tables of checks" of this Wiki.

Where can I find information about the raw reads of a sample?

Answer: Raw reads information can be viewed in the "Check Reads" table (`Report > Detail > Checks/Metrics > Check of the reads`). In addition, as for the analysis results, raw reads can be downloaded via the Download functionality. In this case the button to select is *Osq_rawreads*.

12.5.5 Analyses

Why can't GrapeTree calculate the Minimum Spanning Tree?

Answer: Usually, the reason is usually that there are no allelic differences among the analyzed samples: GrapeTree cannot create an MST on a dataset with multiple samples having 0 differences. Another possible cause is the number of samples: GrapeTree can not calculate a tree if you use only 2 samples or less.

On external samples, can I run GrapeTree or do I need to run ChewBBACA first?

Answer: GrapeTree requires allelic profiles from ChewBBACA (or another tool available for **4TY_cgMLST**) as input; in order for profiles to be usable, ChewBBACA results must have been imported. On external samples, analyses are not run automatically (unlike internal samples) when they are added to the platform. Before running GrapeTree on external samples, ensure that ChewBBACA has been run and its results imported. The table `Main objects > Results of the analyses` allows you to verify that **4TY_cgMLST** has been executed.

What happens if I don't know the host of my sample or I don't want to deplete it?

Answer: If you don't want to perform the host depletion (or you can not do it), you can run mapping directly. Genome reconstruction software will work anyway.

> **WARNING:** This approach is acceptable only if the sample is of excellent quality and you are certain to obtain good results; otherwise the data will be unusable and the platform will accumulate low-quality data, damaging future analyses for users.

What if I don't know the best reference for my viral genome?

Answer: The **3TX_species** analysis, for species identification, can be run using the *vdabricate* software, which identifies the best reference in DB_Virus.

How can I get information about analysis programs/options that were run automatically or by other users in my group?

Answer: If an analysis has been run and imported, results (including report and QC) can be downloaded from the Download page (Download Wiki); moreover, the "Logs / cfg" files are available for each analysis. However, remember that the download system always prepares results of the most recent analysis with the specified features. If you want to investigate the execution of older analyses, you can use the download function via RIS_CD.

I can't see the analyses from other users or the analyses automatically triggered in "Launched pipelines". Am I not able to access them?

Answer: Analysis cards launched by other users cannot be viewed in "Launched pipelines": only administrators can see all analyses launched on the system. The results of those analyses are still available for download, so you still have access to results, reports, and QC, including data displayed in the Check tables.

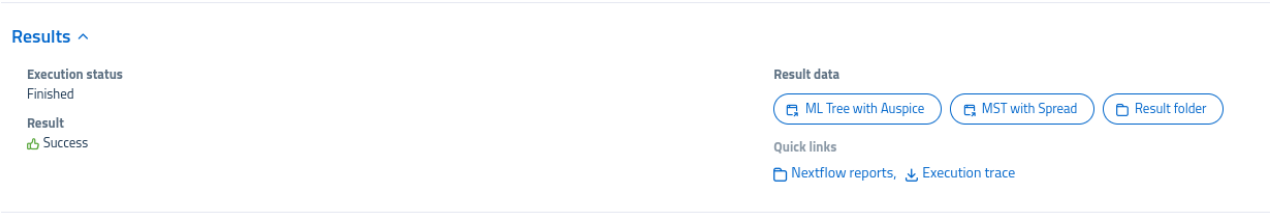
If some samples have been analyzed more than once with the same software, how can I view only the information related to a specific run?

Answer: Each launched analysis has a unique code. In "Results of the analyses" you can filter analyses performed on samples with the same NRG but different DS, using the "Search" field or filters in the "Code" column. If you know the code, it will be sufficient to search for that string, otherwise you can find it based on the execution date (for example, searching or filtering for "210923" returns analyses launched on September 23, 2021; similarly you can specify a time: "240208152038" stands for February 8, 2024 at 15:20:38).

Is it possible to see which programs/options were used to run analyses?

Answer: Certainly. In each analysis folder, logs and a ".cfg" file are preserved, which corresponds to the Bash script executed to run the tool, with all specified parameters. Consequently, programs, their versions, Docker versions, commands and options used for running the analysis are always preserved and can be consulted in different ways:

- For the analyses executed and accessible in the "Launched pipelines" page, you can obtain information by reading the "Execution trace" file, available in the analysis card, "Result data" section, which contains the version of the various tools that were run:



- For analyses imported into the system, i.e. those present in "Results of the analyses", it is possible to download their outputs and select "Logs / cfg" among the download options. By opening the folder containing the downloaded analysis results, the configuration file will be available in the "logs" folder, containing the executed Bash script and its parameters.
- The tools (with their respective versions) that contribute to allele profile calculation are also described in the "NGSManagerSOP" page.

12.5.6 Profile management

How can I change my profile?

Answer: You can change your profile by clicking on the name of the current profile (at the top right) and then on "Change group". The procedure is also illustrated in the video guide "User profiles".

How can I change my password?

Answer: You can find a description of the password change procedures in the Password management page.



GenPat Platform - User manual