COHESIVE Information System

USER MANUAL

EU Emblem

This meeting is part of the European Joint Programme One Health EJP. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 773830.

Table of Contents

1. Welcome 5
1.1 COHESIVE Information System (CIS) 5
1.1.1 Introduction 5
1.1.2 Quick links 5
1.2 About the COHESIVE Project 6
1.2.1 Introduction 6
1.2.2 CIS System Architecture 6
1.2.3 Data Flow 7
1.2.4 ETL Cohesive Information System 8
1.3 Quick start 10
1.3.1 Basic workflow 10
1.3.2 Upload procedure 10
1.3.3 Download procedure 10
1.3.4 User Profiles 11
1.4 Features 12
1.4.1 Main features 12
1.4.2 Table features 14
1.4.3 Data Import / Export through files 16
2. Main objects 17
2.1 Samples 17
2.1.1 "Samples" User Interface 17
2.2 Aliases 19
2.2.1 Alias User Interface 19
2.2.2 Alias Relations 19
2.2.3 Search Codes 20
2.3 Metadata 21
2.3.1 EFSA standard coding 21
2.3.2 Pathogen Species 21
2.3.3 Host Species 22
2.3.4 Material 22
2.3.5 Sampling Point 23
2.4 Exams 24
2.4.1 Exams Results 24
2.4.2 Report errors 24

3. Filter samples 25
3.1 Tags 25
3.1.1 Tags 25
3.1.2 Manage Tags 26
3.2 Cart 28
3.2.1 Introduction 28
3.2.2 Manage Cart 29
3.2.3 Activate Cart 31
4. Analyses 32
4.1 Run analyses 32
4.1.1 Run analysis with cart or active tag 32
4.2 Run analysis steps 33
4.2.1 Step 1: Samples 33
4.2.2 Step 2: Tools 33
4.2.3 Step 3: Inputs 34
4.3 Check analyses 35
4.3.1 Analysis card 35
4.3.2 Import analyses 35
4.3.3 Replay Analysis 36
4.3.4 Lock Analysis 36
4.4 Tool Descriptions 37
4.4.1 Introduction 37
4.4.2 Tools for long reads 42
4.4.3 Single sample 43
4.4.4 Multi sample 97
4.4.5 Pipelines 117
5. Reports 140
5.1 Reports 140
6. Downloads 141
6.1 Downloads 141
6.1.1 Selective download 142
6.1.2 Custom download 142
6.1.3 Preset download 142
6.1.4 Check Downloads 143
6.2 Download methods for Windows 144
6.2.1 Method 1: WSL 144
6.2.2 Method 2: wget.exe 144
6.2.3 Method 3: Browser extension "DownThemAll" 145

7. Upload 146
7.1 Introduction 146
7.1.1 Upload Functionalities 146
7.2 Upload Functionalities 147
7.2.1 Upload from NCBI 147
7.2.2 Manual upload 148
7.3 Prepare upload 149
7.3.1 Add metadata from UI 149
7.3.2 Add metadata from file 149
7.4 Check upload 150
7.4.1 Success 150
7.4.2 Failure 150
7.4.3 Warnings 151
7.5 Zip and metadata rules 152
7.5.1 Metadata 152
7.5.2 Sequence files 153
8. Administration 155
8.1 Authorization and Authentication Rules 155
8.1.1 Access with username 155
8.1.2 Restrictions on sample visibility 155
9. Dashboards 159
9.1 SPREAD, Spatiotemporal Pathogen Relationships and Epidemiological Analysis Dashboard 159
9.1.1 Description 159
9.1.2 Using SPREAD 159
9.1.3 Overview of General Functions 160
9.1.4 Components 160
9.2 Auspice 165
9.2.1 Quick guide 165
9.2.2 Main user interface 165
10. FAQs 169
11. News 170
12. Resources 171
12.1 Keyboard shortcuts 171
12.1.1 Shortcuts list 171
12.2 Video 172

1. Welcome

1.1 COHESIVE Information System (CIS)

1.1.1 Introduction

COHESIVE Information System (CIS) is a data platform to facilitate risk-analysis and outbreak control. The CIS is basically a database with a WEB interface based on the opensource CMDBUILD Project.

CMDBuild is an open source web enterprise environment used to configure custom applications and to manage databases of items. CMDBuild is released under the open source AGPL license: anyone can download it, install it and freely use it.

CMDBuild has native mechanisms to configure custom applications and to manage databases of items and it integrates interoperable software components. Please find more information in the About section of this wiki.

1.1.2 Quick links

- Quick start
- About the COHESIVE Project
- Main Features

1.2 About the COHESIVE Project

1.2.1 Introduction

The "One Health European Joint Programme" (OHEJP) "One Health Structure In Europe" (COHESIVE) aims at improving efficiency of surveillance, risk analysis and outbreak management through a One-Health approach at the Member State level. The achievement of such an objective would require that each Member State integrates data of pathogens from the human, veterinary and food sectors.

The elements under consideration in the OHEJP COHESIVE are:

- **WGS data** of pathogens analyzed by the Member State laboratories,
- **Metadata**, such as the minimum epi-data associated with each pathogen.

Moreover, additional useful metadata sources will be considered.

At the Member State level, each system will be analyzed in order to understand:

- which WGS data is considered (i.e., results of bioinformatics approach/pipelines/tool),
- which metadata is considered (i.e., epi-data),
- which interoperability (data import/export) systhems are implemented,
- how harmonization can be performed and so on.

An Information system, the Prototype COHESIVE information system (CIS), is provided to evaluate the integration of information from pathogens at the Member State level.

The Member State willing to participate can choose to fill the CIS with real data, linked existing data, as well as anonymized or random data, during the evaluation exercise. The Member State can decide to test the CIS system just as a passive archive of information or as an active one (i.e., performing the bioinformatics analysis from integrated data).

1.2.2 CIS System Architecture

The CIS is basically a database with a WEB interface based on the opensource CMDBUILD Project. CMDBuild is an open source web enterprise environment used to configure custom applications and to manage databases of items. CMDBuild is released under the open source AGPL license: anyone can download it, install it and freely use it.

CMDBuild has native mechanisms to model his internal database, to design workflow, to configure reports and dashboards, to build connectors with external systems, to geo-refer items, and to administer the system.

The core code is kept separated from the business logic, to ensure maximum extensibility and to allow the use of CMDBuild as a base system to create custom and configurable vertical applications.

CMDBuild is a web-based system: Java for the server side, Ajax web GUI, WebServices based SOA architecture. CMDBuild exploits the best open source technologies and industry standards.

CMDBuild is exclusively based on open source technologies, and in particular on open source solutions that have been accurately selected for technological features and validity.

Integrated software components or interoperable with CMDBuild are the following:

- Database PostgreSQL, the most advanced and reliable open source database (also includes object oriented features, heavily used in database modeling)
- Tomcat servlet container to run the Java server code
- Reporting engine JasperReports with IReport editor
- Quartzschedulersystem
- OpenLayersGeoServer and GIS features
- Sencha Ext JS library for the desktop client interface based on Ajax technology
- OpenLDAP for accessing external authentication systems

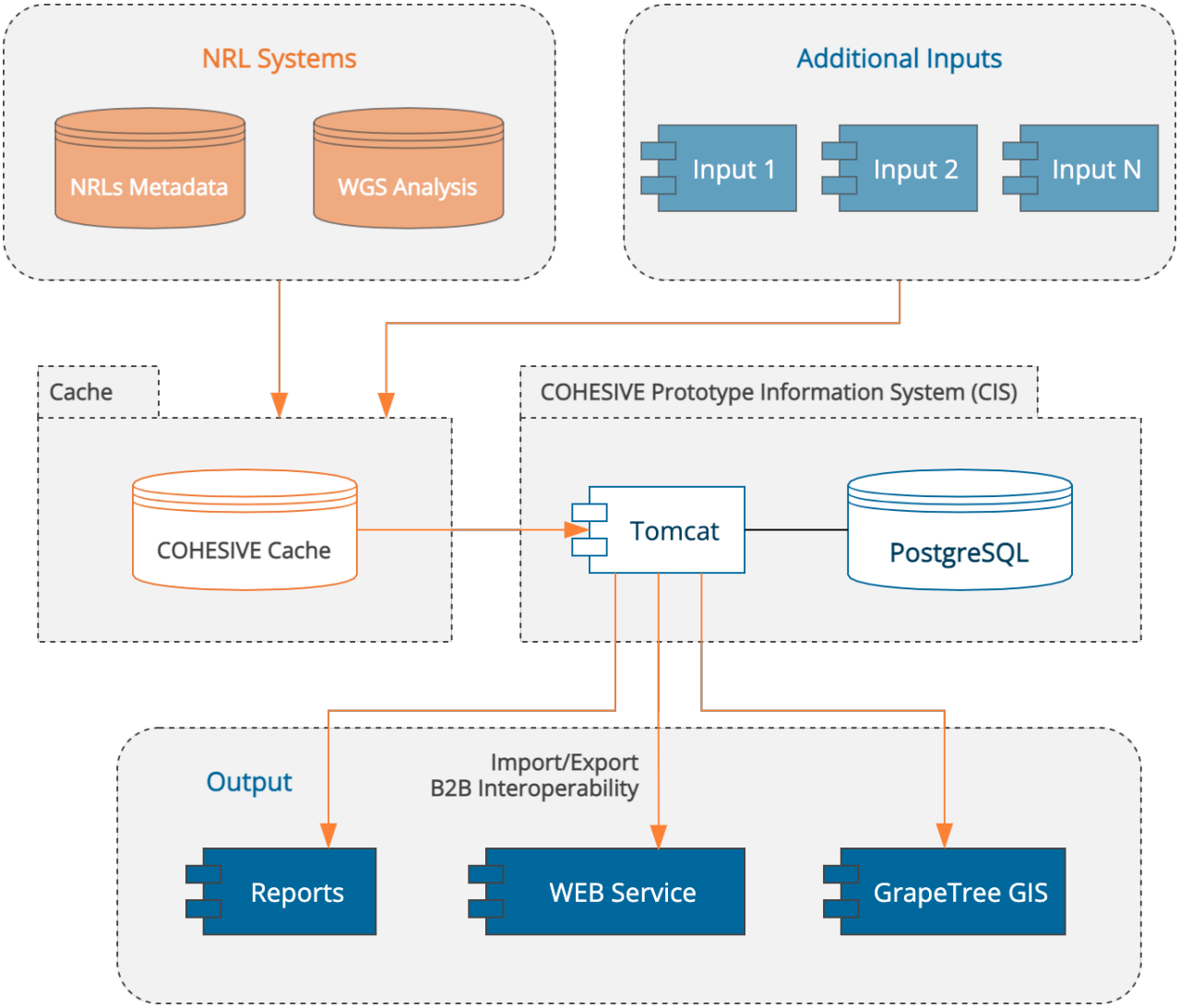
- Additional features can be find here.

Useful references

- Project description and documentation
- Download Sourceforge
- Project Manuals

1.2.3 Data Flow

Elements from different sources (i.e., metadata from Member State, WGS analysis and every kind of relevant input from Member States) are collected into a database (**COHESIVE Cache**) and processed through a cleaning step in order to make them homogeneous (*cleaning*). Subsequently, these elements are processed through a validation step (*validation*), and only at this point they are ready to be transferred to the PostgresSQL database of the CIS.

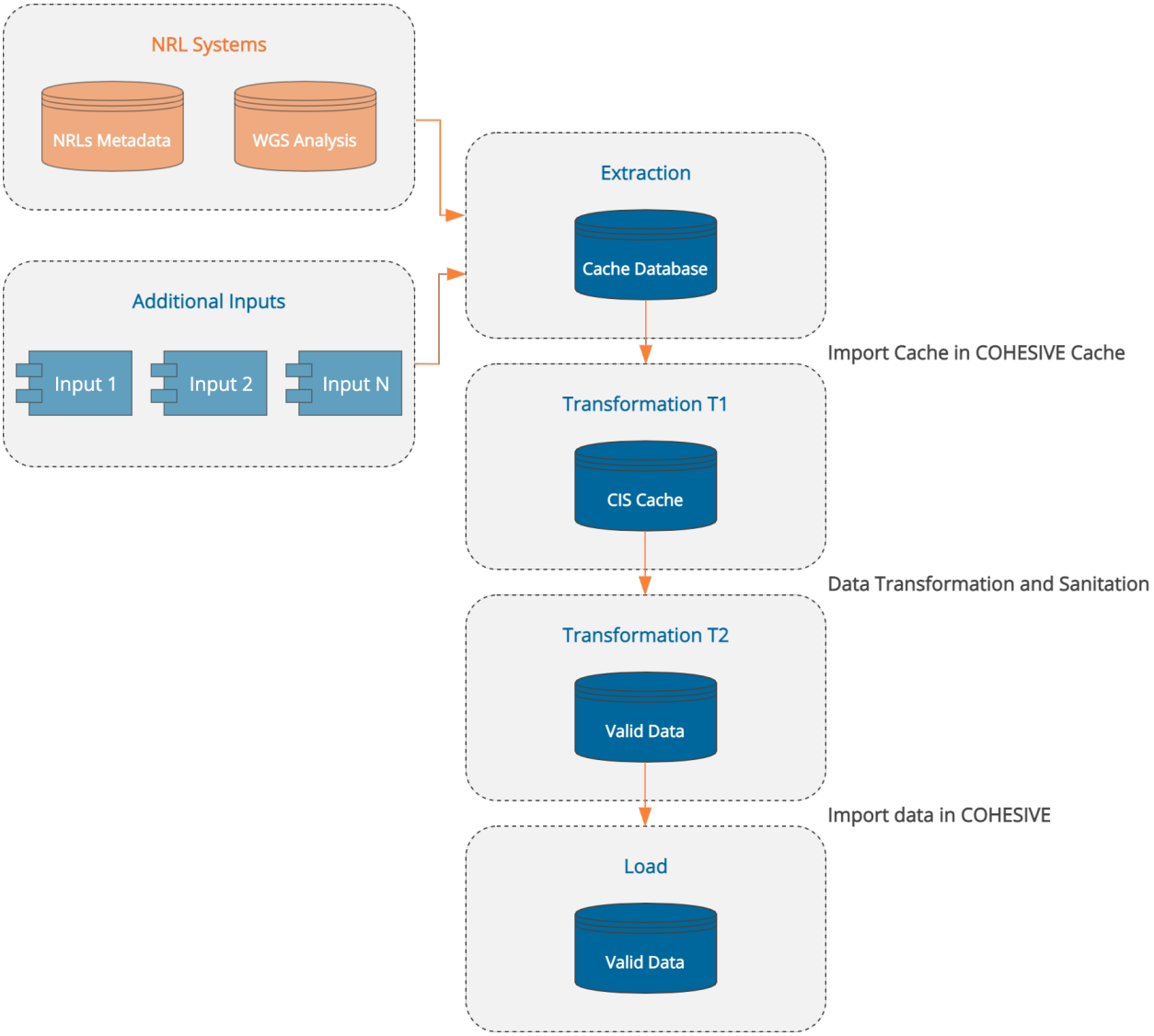


1.2.4 ETL Cohesive Information System

Extract, transform, load (ETL) is the general procedure of copying data from one or more sources into a destination system which represents the original data in a different context than the source(s).

Data extraction involves data extraction from homogeneous or heterogeneous sources (Cache Database); data transformation processes data through a cleaning step and transforms data into a proper storage format/ structure (CIS Cache); finally, data loading describes the insertion of data into the final target database (CIS Database).

A properly designed ETL system extracts data from the source systems, enforces data quality and consistency standards, conforms data so that separate sources can be used together, and finally delivers data in a format ready to be presented (CIS).


NRL Systems
NRLs Metadata
WGS Analysis
Additional Inputs
Input 1
Input 2
Input N
Extraction
Cache Database
Import Cache in COHESIVE Cache
Transformation T1
CIS Cache
Data Transformation and Sanitation
Transformation T2
Valid Data
Import data in COHESIVE
Load
Valid Data

1.3 Quick start

To register an account, send a request to cohesive@izs.it. Then login with the credentials provided to you at:

https://cohesive.izs.it/cmdbuild/ui/#login

1.3.1 Basic workflow

The video below demonstrates a basic workflow to showcase the main activities available within the platform:

- Filtering, selection and management of samples;
- Adding selected samples to the Cart;
- Running analyses;
- Visualization of analysis results through integrated dashboards.

type:video-tag basic-flow.mp4

Please note

- Analyses could take a while. The process runs in background so you can leave the platform and check back later; a built-in notification system will notify you when the analysis is finished.

1.3.2 Upload procedure

The following video shows the basic procedure to upload samples and run analyses on them.

type:video-tag upload-procedure.mp4

Please note

- Analyses could take a while, for instance the run launched in the video took about 3 hours to finish. The process runs in background so you can leave the platform and check back later.
- Please check out the Upload section of this Wiki for more information on preparation of metadata files.

1.3.3 Download procedure

The following video demonstrated the download procedures available in the infromation system.

type:video-tag download-procedure.mp4

- Please refer to the Download section of this Wiki for more information about download methods.

1.3.4 User Profiles

Each user can be associated with different profiles, or as we define them on our platform, each user can be associated with different **groups**.

The user can view or do certain things based on the group to which it belongs. From this point of view, we can assume that groups act as general filters.

In the video, you will see how to switch groups and how this change will affect the view of the main navigation menu and the samples you will be able to see.

type:video-tag group-change-functions.mp4

1.4 Features

1.4.1 Main features

Database schema flexibility

The system has an HIGHLY FLEXIBLE database schema and no software development is needed to perform operations such as creating new tables or new form fields.

As shown in the video you are not only able to define some additional attributes, but you will be able to create an entire custom database from scratch, in terms of:

- "classes", i.e. database tables;
- class "attributes", i.e. table columns (for every type, included lists with closed values, single or multiple, foreign keys, files, formulas) and related layout;
- "domains", i.e. relation types, even with any specific additional attributes;
- "lookup", i.e. lists with closed values that can be associated to a class attribute;
- “inheritance", i.e. the possibility of specializing a class in "subclasses" (even with more levels), by adding specific attributes and domains and sharing the superclass ones;
- "views" based on filters or SQL queries;
- user interface behaviours that are specific for the class: data validation, automatisms at data update, contextual menus, online help, etc.

type:video-tag db-flexibility.mp4

Massive 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, starting from enabling the multiselection.

type:video-tag massive_modification.mp4

Data Import / Export through files

The COHESIVE Information System supports import and export of data through data source in files formats like `CSV`, `XLS` and `XLSX`, as well as relational databases.

The system is based on the configuration of specific templates to allow management of complex operations in a simplified way. Please check the dedicated Data Import/Export section of this Wiki for more in-depth information and video guides on this topic.

Tables

The COHESIVE Information System (CIS) integrates table-based environments to list and interact with entries. Tables support customisation, filters and functions to export, print or store the entries in the Cart. Data relative to each entry can also be accessed and cunsulted by simply clicking on an entry to expand its information card. A detailed guide with videos is available in the tables dedicated section.

Tags

Tags are a feature in CIS that allow grouping of samples. They can act as filters, effectively causing to visualise in the grid only those samples or entries that belong to the active tag.

More details on tags are available in the dedicated section, as well as a demonstration on how to activate a tag.

For a guide on how to use tags, please refer to the Manage Tags subsection, where detailed instructions on how to create, remove or edit tags can be found.

Cart

The cart is a function that allows to temporarily store selections of samples, which can then be used to perform analyses or create tags.

In the introduction to the cart's dedicated page you can find explanations and a video on the cart's basic functions. For more in-depth understanding of the cart functionalities, please visit the Manage Cart page of this Wiki.

Repeat Analyses

Two shortcuts allow the user to re-run an analysis that has been completed successfully on a new set of samples, keeping the old parameters for the new run. The repetition of the analysis is available for a set of samples provided either from cart or from active tag.

More information and a videotutorial are available at the dedicated wiki page.

Dashboards

The COHESIVE Information System (CIS) integrates dashboards that allow direct visualisation of the analysis results. Among those, COHESIVE features an extended version of GrapeTree: SPREAD (you can find more information in the dedicated section of this wiki).

In the following video we show the development over a certain period of time of the Sars-Cov-2 epidemic in the Abruzzo region (Italy). The analysis is based on the metadata related to the samples: dates grouped by year / month of sampling for the tree and their geoJSON data for the map.

type:video-tag GrapeTree-time-lapse.mp4

1.4.2 Table features

Each user can interact with the entries table, customizing it to suit their needs and preferences. The sample table also supports basic and advanced filters, as well as export, download and print options. From the context menu the user can request an analysis or download samples and results of completed analyses.

Table management and customization

The following video demonstrates how the sample table can be managed and modified by changing the columns' sizes and order, as well as choosing the columns to display and the sort order of the entries in each column. The changes made to Cohesive's tables are transient unless they are saved, and they can be reset by clearing the table preferences, as shown in the video.

type:video-tag The-Grid.mp4

Basic filters

Basic filters can be applied to the columns, alone or in combination, to search or narrow down the entries that are displayed in the table.

type:video-tag Filters.mp4

The extended table

By clicking on an entry, the user can expand a sample's card and access more resources and links, among which there are the entry's details and metadata (sampling point, provider, analyses carried out on the sample), as well as related entries, interactive relation graph, previous versions and export options.

type:video-tag Extended-table.mp4

Advanced filters

The table of samples also provides advanced filters that can be applied to the list of entries. Advanced filters will allow for a research by one or more attributes (which can be text strings, date or metadata), with one or more research terms or operators per attribute. Additionally, advanced filters can be edited, copied, saved and set to be applied as default to the grid.

type:video-tag Advanced-Filters.mp4

Context menu and analyses

In addition to filters, the platform's tables also provide a context menu from which the user can edit the view options, enable multi-selection of entries, add samples to the Cart, export the view to a file in table-like format.

type:video-tag Context-Menu.mp4

Hierarchy tree

Tables that contain information about the elements' **hierarchy** will also show a `See hierarchy` button in the context menu. The **See hierarchy** functionality activates an interactive tree visualization of the table's elements, as shown in the video below.

type:video-tag see-hierarchy.mp4

Note: if the table lists more than 2000 elements (lines), the `Expand view` will be automatically disabled.

Export RIS_CD code

Check tables list all analyses which have been run and imported, using univocal codes called "**RIS_CD**".

A RIS_CD (*e.g.*: "210712-10003548-4TY_cgMLST-chewbbaca") identifies analysis based on date and time of the run, together with the corresponding results. **The first 6 digits of the prefix are the execution date** of the analysis ("YYMMDD" format); the terminal part identifies the analysis name (*e.g.*: "4TY_cgMLST-chewbbaca").

In the Check Tables section the selected codes can be exported as CSV or copied, which allows the usage of RIS_CD codes to select samples for running an analysis or for results download.

Note: usage of RIS_CD codes is also the primary method by which older analysis that have been re-run are available for download or as inputs.

The short videotutorial below shows how to use RIS_CD codes to select analysis inputs or files for download.

type:video-tag

1.4.3 Data Import / Export through files

The system offers the possibility of import/export data through data source 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.

In case of importing, you can work in "merge" mode, by editing the already available strings (recognized through a unique key), inserting the new ones and deleting the missing ones (logic cancellation or change of the application status).

There is a complete error handling, shown in the user interface if working in interactive mode and sent per e-mail. Once the errors have been corrected, the file can be recharged without having to change it, since the already imported strings will be ignored.

type:video-tag file-template1.mp4

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

type:video-tag file-template2.mp4

2. Main objects

2.1 Samples

The **Samples** page is one of the system's main areas and is available in the **Main Objects** section of the **Navigation menu**.

2.1.1 "Samples" User Interface

The page's layout consists of a list of samples visible to the current user and related metadata, organized in an integrated interactive table (the table of samples). Please refer to the sample table manual page for an in-depth guide to the **Samples** page usage and to the table's functionalities.

Visibility of samples depends on the user's group and on active tags or Cart.

Table of samples

In the **Samples** page's table, the list can be filtered, samples can be added to the Cart and corresponding metadata can be displayed through the sample's card.

Clicking on a sample will open the detailed sample card. In the first part of the card are listed main metadata and information on analyses run on the samples. In the lower section there are all the relations of the selected samples with other databases, tags and aliases it's associated with. Some sections present quick-links (inline-icon).

The guide to sample cards is available at the extended table section of the manual. The video below demonstrates the actions that can be performed on samples:

type:video-tag Extended-table.mp4

Table functionalities for samples

Main objects organized in the system's tables (Aliases, Metadata and Exams), have export options accessible from the top ribbon. Samples can be managed as shown in the table management section.

Export functions are specific for each type of object: for samples, they are located in the context menu, together with the **Cart**, available only for samples.

Cart

Only samples can be added to the Cart. One or more **selected samples** can be added to the Cart using the corresponding button inline-icon. Actions available to update Cart contents are listed using the arrow on the right side of the button; they are: "Add to Cart", "Replace Cart" and "Remove from Cart".

> **Note:** in order to add to the Cart, **samples must be selected**. Multi selection is active by default and can be turned on or 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 have been selected.

The guide to the Cart and all its functions is available at the Cart Wiki page.

Context menu

The sample table's context menu allows for export of sample codes and reports.

Detailed guides are available in the following sections:

- Context menu;
- Filters and advanced filters;
- Table management.

> **Note:** export functionalities usually require to select samples of interest too. Any attempt at exporting without selecting any sample, will result in export of an empty table or failure.

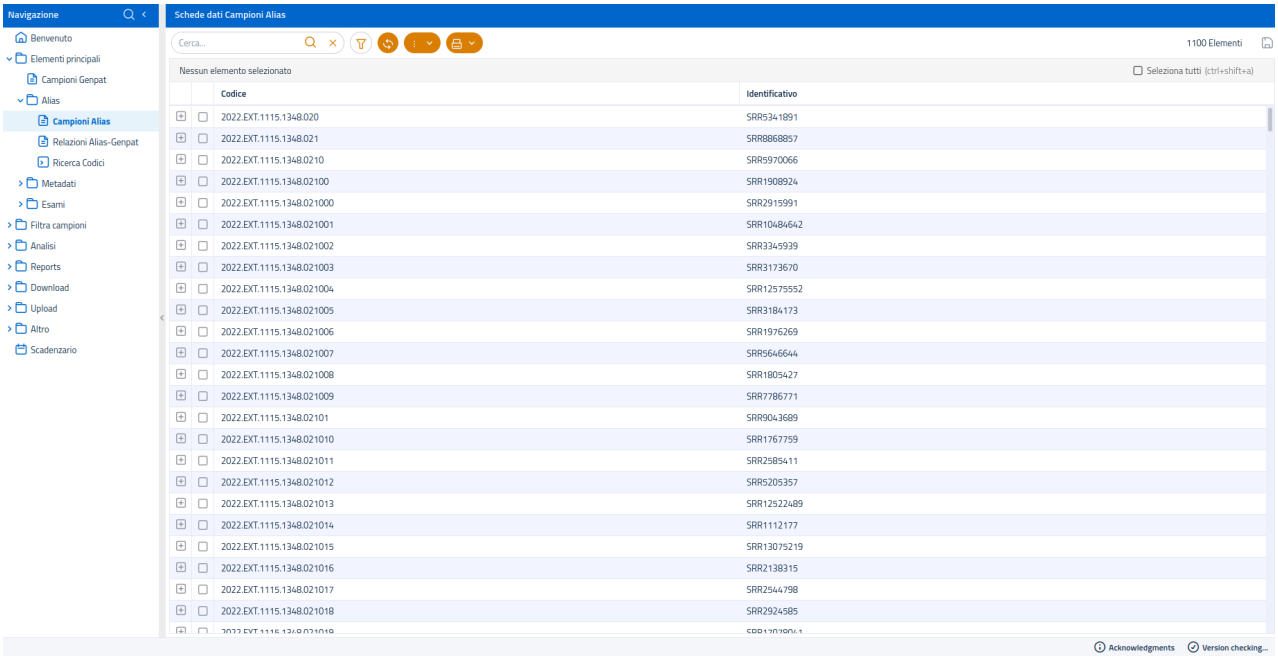
2.2 Aliases

In the **Main Objects** section of the **Navigation menu** is also available the group of **Alias** functionalities, which concern relation among codes referring to the same samples, for example codes from ENA or NCBI databases for samples that have been imported into Cohesive.

2.2.1 Alias User Interface

The **Alias** page lists (in the integrated table), the system's codes and any further corresponding ID (like original codes before import) in the "ID" column.

Alias entries benefit of all functionalities available for **Samples**.

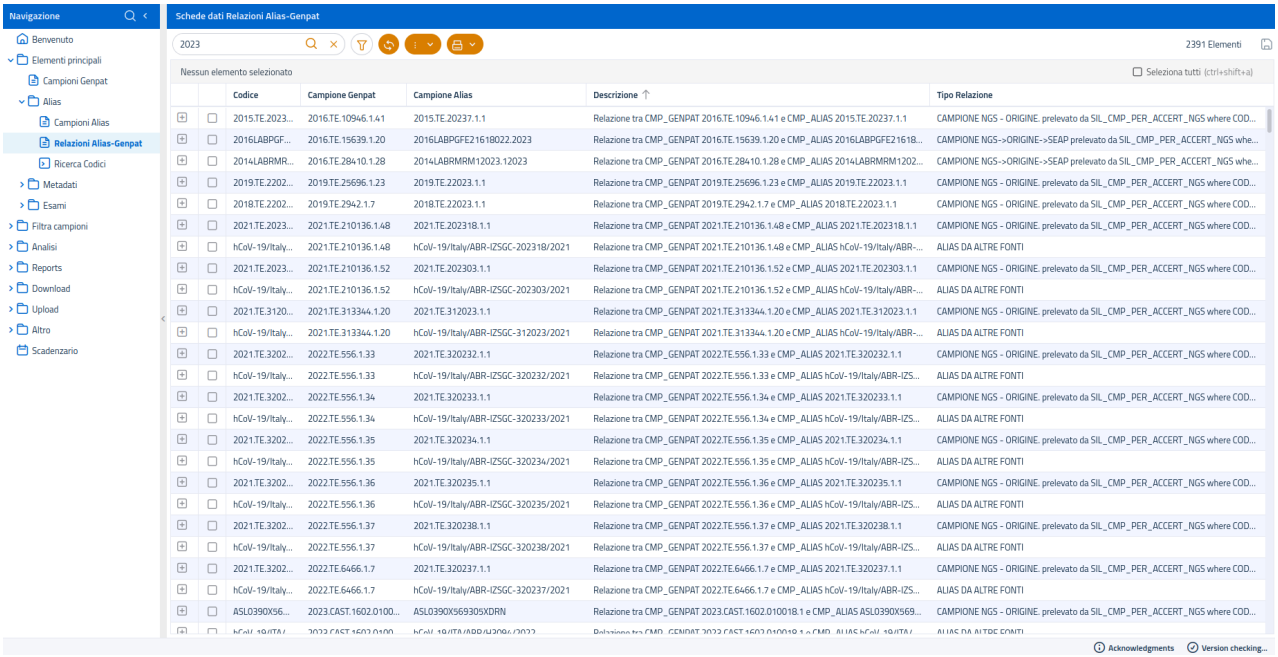


Available columns for Aliases include only those of alternative IDs and other metadata, so they are inferior in number, if compared to those of the Samples page.

2.2.2 Alias Relations

The **Alias Relations** page lists all relations of IDs, together with a description field to allow users to know a sample's origin and track its import into the platoform.

All table features are available in this page.

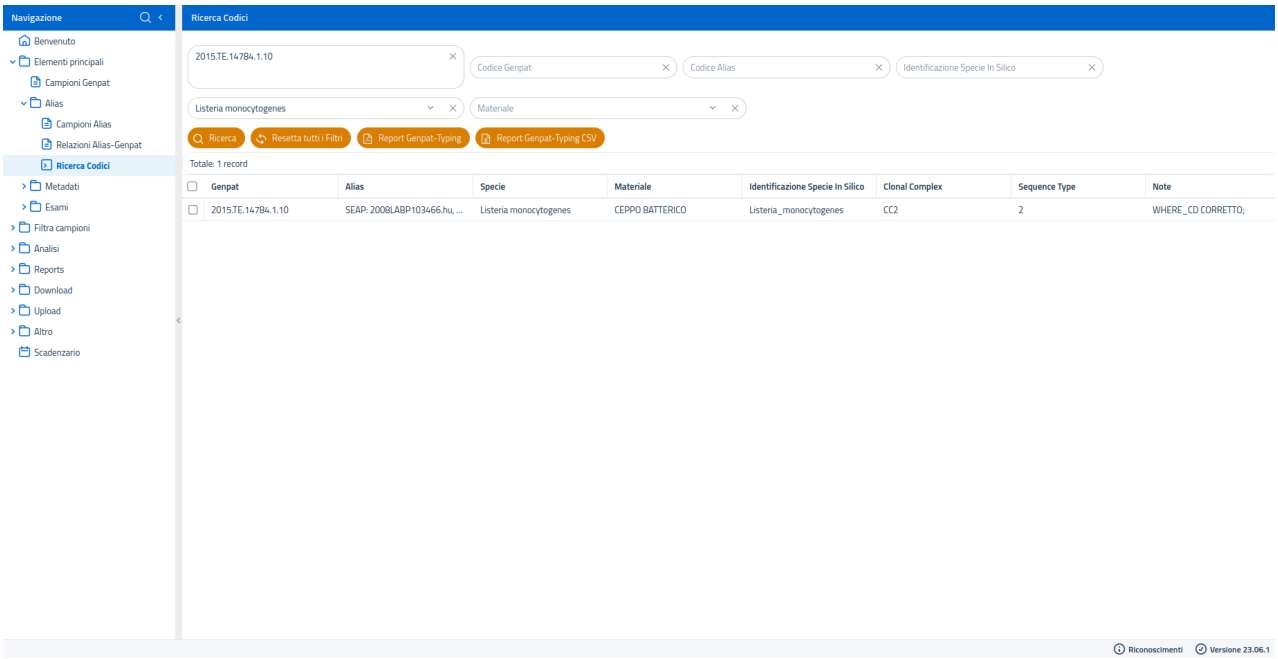


2.2.3 Search Codes

The **Search Codes** page gives access to custom searches for one or more samples. Search results can be thinned out using metadata filters.

It's also possible to export a report as pdf o csv file.

> **Note:** when exporting the report, the downloaded file will have the same columns visualized in the search page.



2.3 Metadata

The "Main Objects" section in the Navigation menu is also available the set of **Metadata** (link pagina Wikipedia) functionalities.

Metadata are organized in the following categories:

- **Species** : list of organisms and their IDs, subdivided into:
  - **Pathogen Species**
  - **Host Species**
- **Material** : list of materials or food matrices from which the sample was originally isolated and corresponding IDs.
- **Sampling Point** : list of geographic places or coordinates of sampling and their IDs.

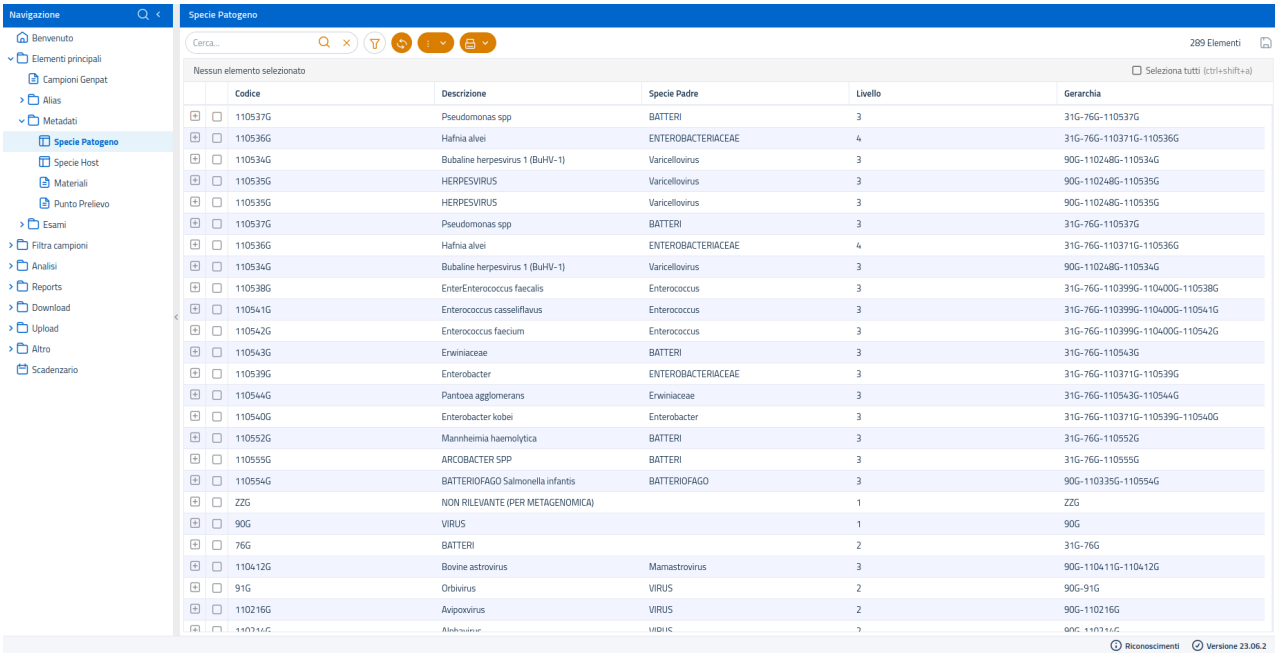
> Metadata are vital to validate uploads.
>
> Metadata have a hierarchy, which can be inspected in "Species Tree" and "Materials Tree".

2.3.1 EFSA standard coding

Sample metadata follow EFSA FOODEX2 specifications.

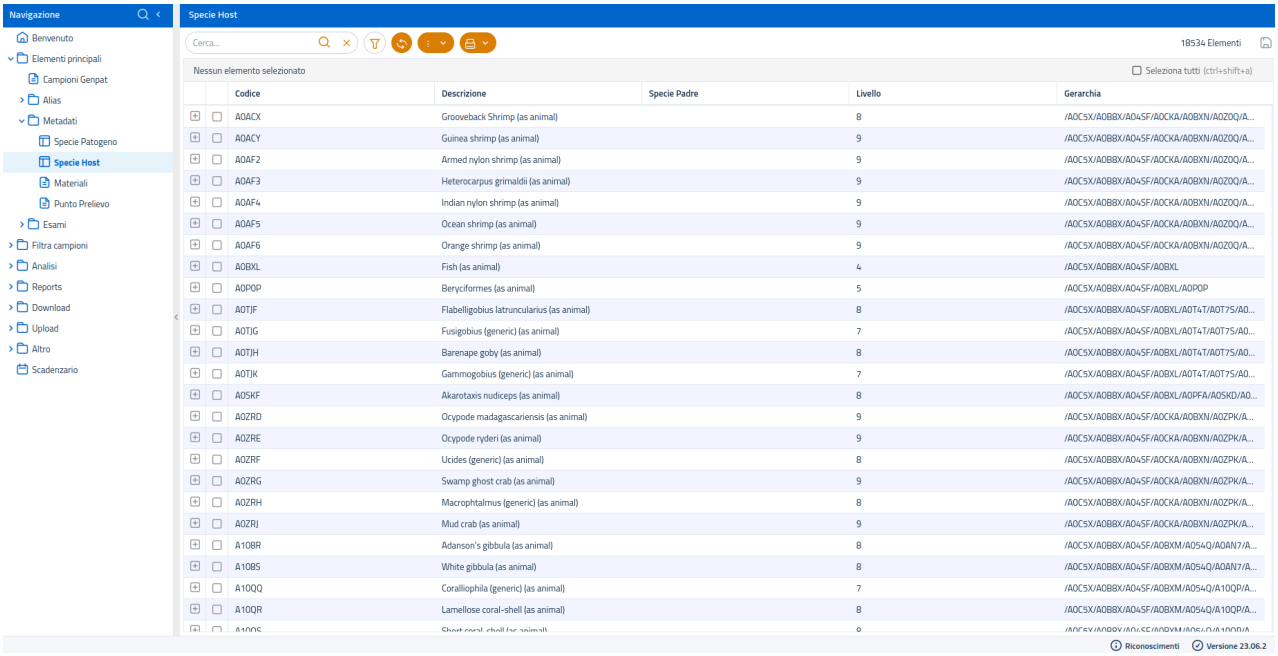
2.3.2 Pathogen Species

The Pathogen Species lists codes for pathogens in the platform's database. Available actions are the same as for table of samples.



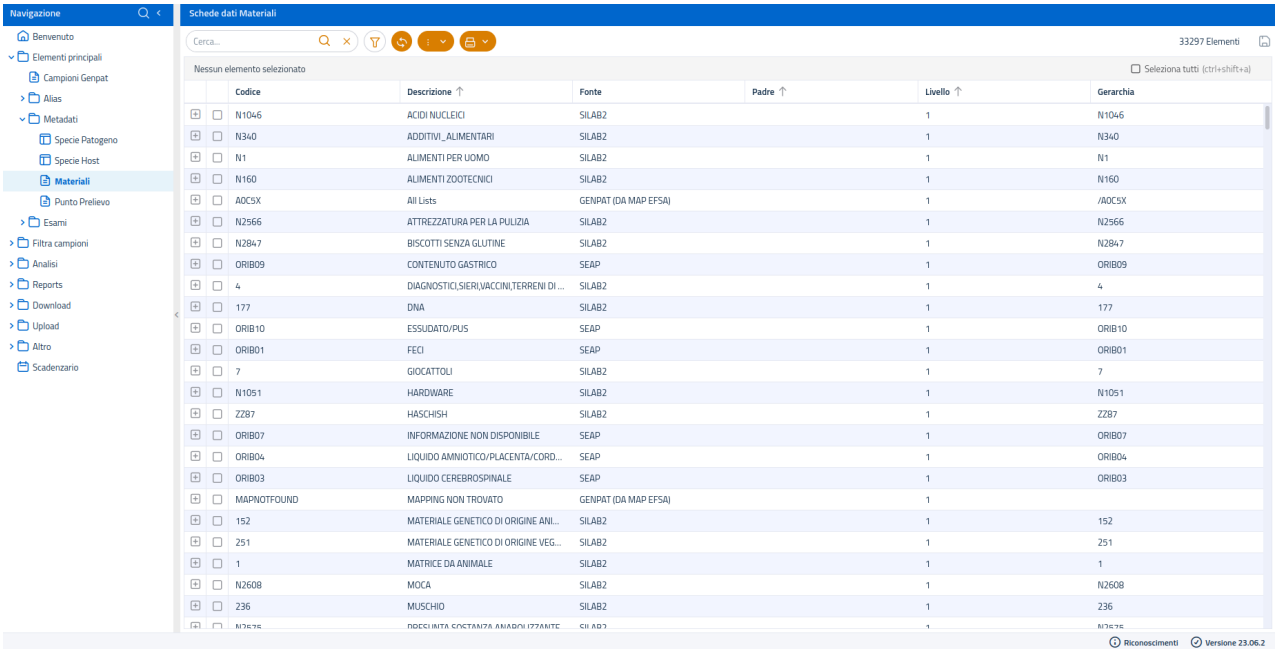
2.3.3 Host Species

The Host Species table lists codes for all organisms whose reference genomes have been added as possible hosts in the platform. Available actions are the same as for table of samples.



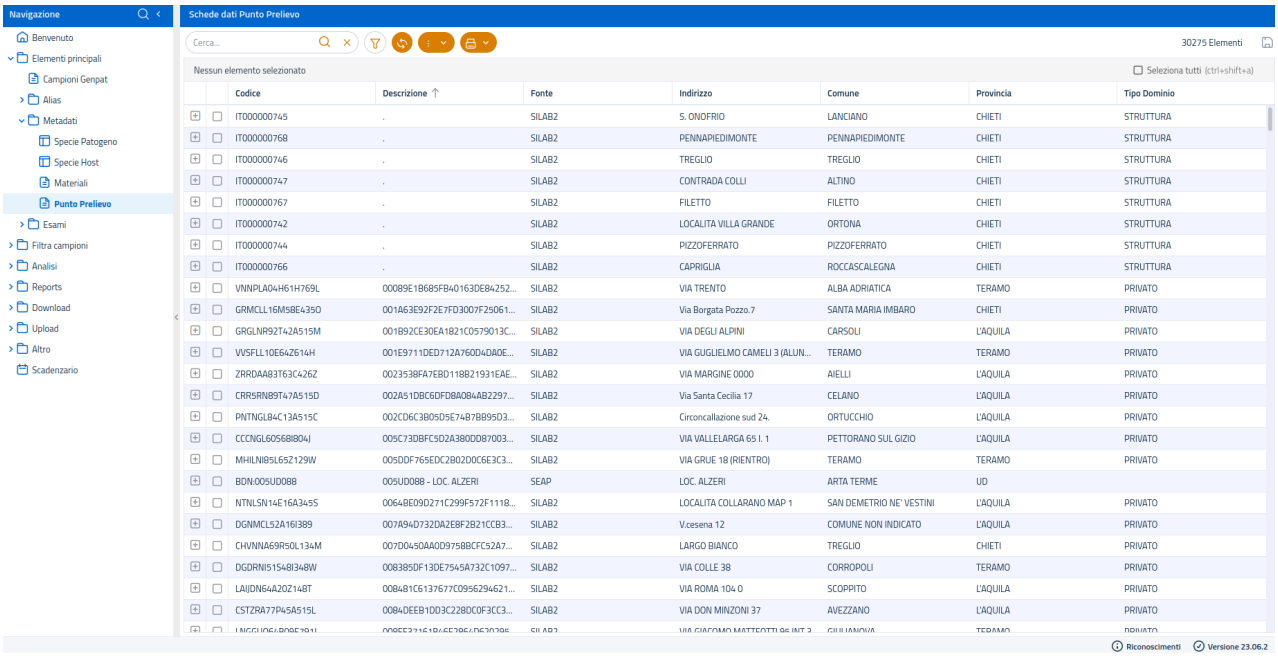
2.3.4 Material

The Material table lists codes corresponding to food matrices, surfaces and other physical or biological structure from which the sample was isolated. Available actions are the same as for table of samples.



2.3.5 Sampling Point

The Sampling Point table lists codes corresponding to places in which the sample was first isolated or the matrix from which it was isolated was first retrieved. Il punto prelievo comprende provincia, comune, eventuale indirizzo e struttura o ente. Available actions are the same as for table of samples.



2.4 Exams

The exams (i.e., analyses) and tools (i.e., methods) available in the CIS are presented in this section.

The page "Results" provides information about the outcomes of analyses launched with samples and related output paths.

2.4.1 Exams Results

The **Exams Results** page lists all results from bioinformatic analyses executed in the system, information on the run and location of results.

Note 1. "Exam" tables are the only ones on which the filter action from activation of the cart has no effect.

Nota 2. In order to run analyses, it is necessary to have all input files from analyses upstream, which can be checked in the "Exam" tables. For more information on how run analyses, check the links to the wiki pages below.

- How to run analyses
- Introduzione ai tools

Available actions are the same as for table of samples.

2.4.2 Report errors

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 an analysis results are not good for usage, to prevent the system to use a specific run as input for other analyses it's possible to mark that analysis as an error-carrying run. To do that, select entries from the table and select from the context menu "**Report error**".

Note: 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:** marking a *de novo* assembly with shovill as an error will result in automatic invalidation of all typing analyses that used that specific assemly as input.

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

type:video-tag mark-error-exam-results.mp4

3. Filter samples

3.1 Tags

3.1.1 Tags

What is a tag

A tag is a useful function which allows grouping of samples with a "reason". The creation of a tag can be helpful in case of specific projects or outbreaks. If specific samples are associated with a tag, only those samples will be visualized in all the tables in the CIS.

> For example, if samples involved in the outbreak XXX have to be analyzed, a tag "outbreak_XXX" can be created only once in order to associate the reated samples, instead of searching and select them each time that new analysis have to be launched. Then, the corresponding tag can be activated to visualize results related to those samples or launch a new pipeline with such samples, by simply specifying the name of the tag.

The following short video shows the process of activating a tag. For in-depth information on tags, please visit the Manage Tags page.

type:video-tag Cohesive-Tags.mp4

3.1.2 Manage Tags

This tutorial will show all the available operations for tags in a series of videos.

The **Manage Tags** page gives access to information about the active tag, if any; it's also possible to disable an active tag or activate an existing one.

The following short video gives an overview of the page: in the **Manage Tags** page you will find 4 expandable sections, or cards, each dedicated to one of the main operations available for tags: **Activate tag**, **Create tag**, **Edit tag** and **Remove tag**.

type:video-tag Manage_Tags.mp4

Activate tag

The first card gives quick access to the process of tag activation: by clicking on the search field or on the "Select tag" button on the right, you can choose what available tag you wish to select from a pop-up window. Once the selection it's done, clicking on the "Activate tag" button will activate the selected tag.

type:video-tag Activate_Tag.mp4

Most used, Recents, Favorites

The cards **Most used**, **Recents** and **Favorites** allow a quick selection of tags to activate. **Most used** and **Recents** are automatically filled with a list of the **5** most used or most recently activated tags, respectively.

The **Favorites** list is populated through direct action from the user and has no upper limit. The following video shows how to add or remove tags using the Favorites list.

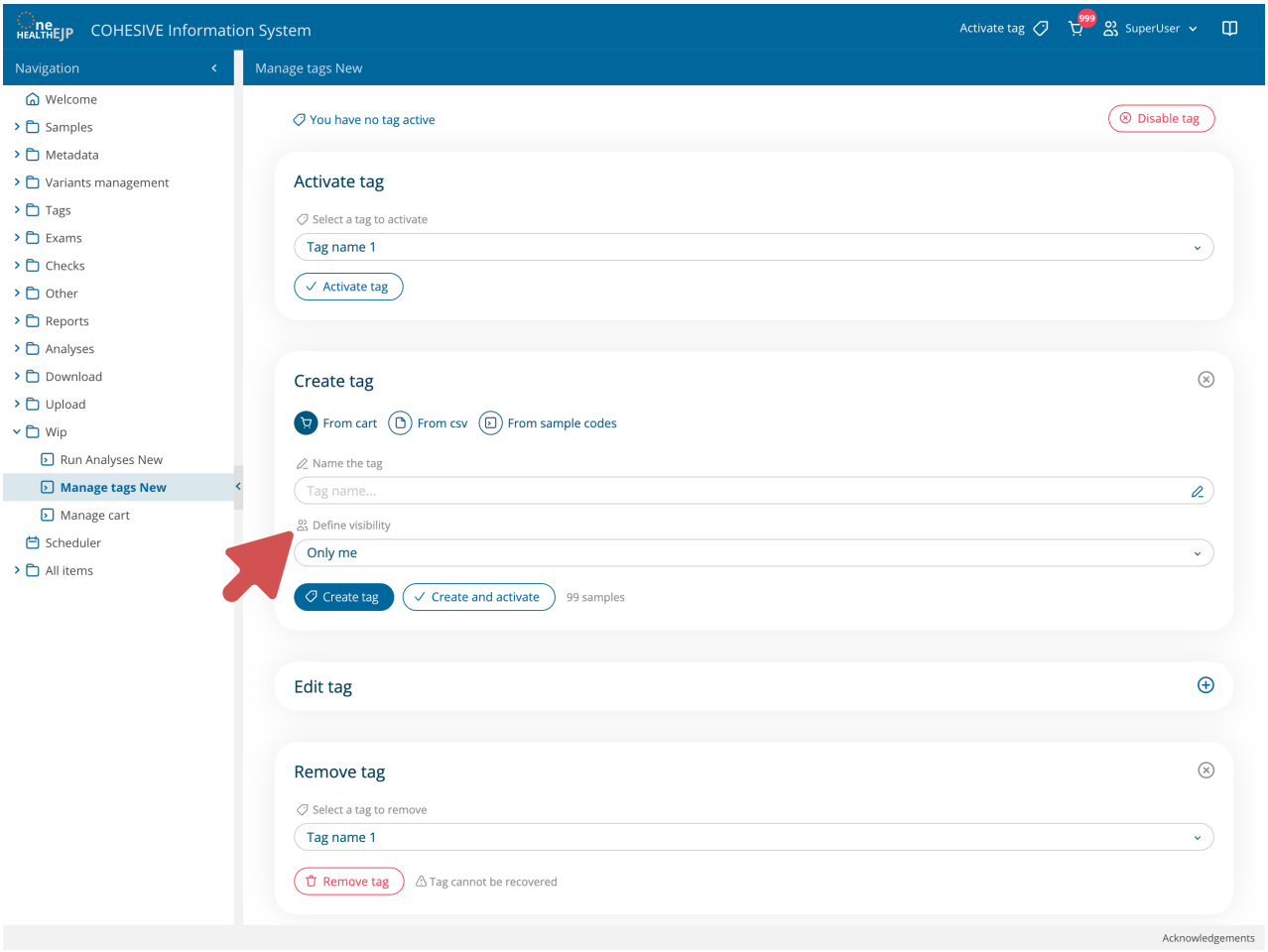
type:video-tag starred-tags.mp4

Create tag

The **Create tag** card offers 3 different ways to create a tag: using a sample collection from the cart, from a csv file or by samples, that is by writing (or pasting) a list of codes, separated by comma, directly in the "Write codes" field. Selecting any of those options will update the card view with the fields available for the desired option.

type:video-tag Create_tag.mp4

A **Define visibility** option will be present for all tag creation method and option. This new functionality allows you to specify if you want to save the tag only for yourself or if you want to make it available also for the group you belong to.



Edit tag

The **Edit tag** card will allow you to edit an existing tag by adding or removing samples: just like in **Create tag**, here too you can choose among cart, csv or list of codes. You will also be able to change the visibility of a selected tag or to remove its contents without deleting the tag itself.

type:video-tag Edit_tag.mp4

Remove tag

The last card serves the purpose of deleting a tag of your choice. Just like for the **Activate tag** card, by clicking on the search field or on the "Select tag" button on the right, you can choose an existing tag; then, to delete it, click on the red "Remove tag" button.

type:video-tag Remove_Tag.mp4

3.2 Cart

3.2.1 Introduction

The cart is highly flexible, useful function in the Cohesive Information System. It closely resembles the cart of an e-commerce system, but with additional features specific to our platform logics and needs.

The cart is used to temporarily store selections of samples that can be stacked or removed from the cart itself, or used to create tags.

For additional information on tags and how to manage them, visit the dedicated Manage Tags page. For detailed information about cart functions, visit the Manage Cart page.

Add Samples to Cart

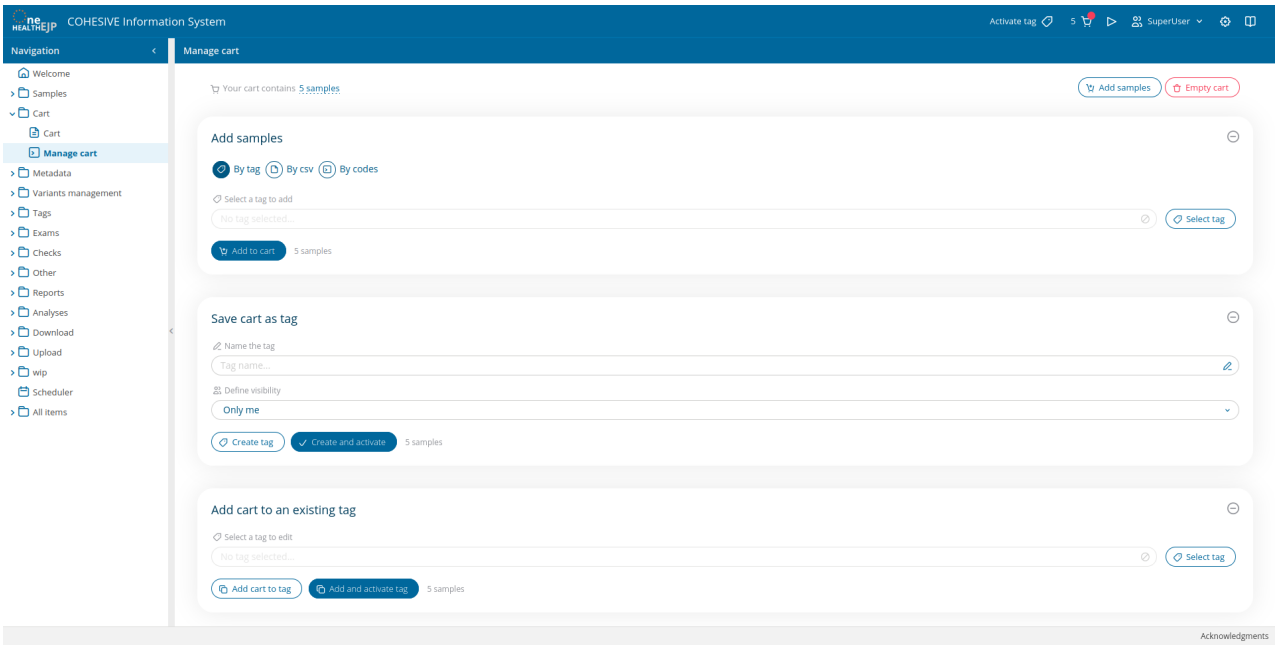
The following short video demonstrates how to add multiple items to the cart.

type:video-tag Cohesive-Cart.mp4

3.2.2 Manage Cart

The cart function behaves similarly to that of an e-commerce system; it can be accessed from the navigation panel on the left, where are available both the cart contents and the cart management functions. The cart can also be accessed from the cart icon on the right side of the top bar, which will open the **Manage cart** page.

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



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

Add samples

From this card it will be possible to edit the cart by adding samples to it in three different ways: from a saved tag, from a csv file or by directly typing/pasting sample codes separated by comma in the "Write codes" field.

The selection of any of the three modes will update the card's view with the options and fields available for the selected mode.

type:video-tag Cart-Add_samples.mp4

Save cart as tag

This card will allow you to create a new tag directly from the samples contained in the cart. To know more about tags please visit the dedicated section or dive into the new functionalities available for tags in the Manage tags page.

The following short video demonstrates how to save a cart as tag for future use.

type:video-tag Cart-Save_as_tag.mp4

Add cart to an existing tag

The last expandable card empowers the user with the ability 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 will show the simple action of adding samples in the cart to a specified tag.

type:video-tag Cart-Add_to_tag.mp4

3.2.3 Activate Cart

The Cart has 2 main function: temporary storage of samples and **Cart activation**.

To activate the Cart means to enable it as a filter: selection of the "Activate Cart" option from the Cart's dropdown menu or from the "**Manage Cart**" page, the current contents of the Cart will be used as a filter list, just like for a tag activation, filtering the whole platform based on the samples in the Cart.

Cart activation (or alternatively, tag activation) is necessary to run an analysis. The cart will be automatically activated if it is selected as source of the sample collection for the analysis.

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

Cart activation differs from tag activation:

- The Cart is transient. It is not saved and can be modified quickly by adding and removing samples;
- With an active Cart, the whole platform is filtered based on the samples in it; this means that all elements in **Check analysis**, **Check download** and **Check import** are filtered too.

The video below shows how to activate and deactivate the Cart and its consequences on the platform.

type:video-tag Cohesive_Activate_cart.mp4

4. Analyses

4.1 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 Analyses** section is available in the side menu and offers options for running a desired analysis on a selection of samples, provided either as a collection currently in the cart or an active tag.

> The **Run Analyses** functions can be reached quickly using the **Run** button ▷ in the top bar (just right of the cart icon).

The following videotutorial will guide through the process of running a new analysis, showing the available options to specify the samples. Usage of the cart or of an active tag does not change the procedure: the only requirement is that the user activated a tag (to run an analysis with tag) or added samples to the cart (to run an analysis with samples in the cart).

4.1.1 Run analysis with cart or active tag

The run analysis process is organized in 3 steps: phase **1** (**Samples**) requires a list of samples to analyze; phase **2** (**Tools**) allows selection of the analysis; phase **3** (**Inputs**) enables selection of parameters and input file source.

type:video-tag Run-analysis_Cohesive.mp4

> **Note 1:** using the Cart as sample source will automatically activate the cart.
>
> **Note 2:** the button in the upper-right corner allows selection of one or more **RIS_CD** codes, which univocally identify a specific result of an analysis run. Using the RIS_CD code for running new analyses allows easy access to a specific run's results and usage of older results of an analysis as inputs.

More information on RIS_CD codes are available in the "Export RIS_CD" section of this wiki.

4.2 Run analysis steps

Cohesive's run analysis system is subdivided into 3 steps:

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

Please consider consulting the "Run analyses" Wiki page for an introduction and quick guide on the whole process.

This Wiki page debates how to go through the 3 steps "Samples", "Tools" and "Inputs".

4.2.1 Step 1: Samples

In order to be able to choose an analysis to perform, the system requires a collection of sample codes. Such collection will be used to request an analysis run on the appropriate input file associated to each of the desired samples.

The user can choose either the **cart** or a **tag** to provide the collection of codes. A tag is an immutable sample code collection, created and saved by the user. The cart is a transient and dynamic code collection system, built by the user, which can be also used as a filter by .

When using the cart to run an analysis, cart activation takes place automatically, overriding any active tag. Cart deactivation willtake back the Information System to the condition prior to cart activation.

If the cart contains no sample codes and there is no active tag, the links in the page will allow to quickly rectify that by leading the user to the appropriate pages.

In the page the user can also find a "Load utilities" function, which allows to perform a quick fill up of the cart, using a list of codes or a CSV file.

The short video below demonstrates the workflow of the sample selection step.

type:video-tag Run_analysis_inputs.mp4

4.2.2 Step 2: Tools

The step "Tools" provides an interface to choose the desired analysis or pipeline and the bioinformatic tool to perform it. All available analyses are listed as buttons with pertinent information and can be filtered by category.

After analysis selection, the user will be able to select the chosen bioinformatic tool among those available. The following short video shows how to perform selection of analysis and tool.

type:video-tag Run_analysis_tools.mp4

4.2.3 Step 3: Inputs

To complete the run analysis process, it's necessary to specify what files, associated with the sample codes, is expected as input for the selected bioinformatic software. As an example, trimming analysis requires the samples' raw reads, cgMLST performed with chewBBACA uses assembly fasta as input, while GrapeTree needs tabular files from cgMLST.

> This step requires that all the necessary files have been produced and imported, (for example, it's not possible to run GrapeTree if chewBBACA has not been executed on the requested samples).

Usually at this stage the fields are pre-filled with the appropriate input and options for the selected tool, so action from the user will have to be taken only in case of multiple suitable inputs or if specific options are available. It is also possible to activate the advanced mode to select inputs.

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

type:video-tag Run_analysis_inputs.mp4

Inputs: advanced mode

In some cases, inputs have to be managed in a way that allows to use files coming from different sources, for the same kind of analysis. To do that, the last run analysis step provides a switch to toggle between the "**Base mode**" and the "**Advanced mode**" for input selection, the latter of which allows to choose multiple input types using checkboxes.

The following short video demonstrates the "**Advanced mode**" for input selection for one of the tools for which such mode is available: the CFSAN pipeline.

type:video-tag Inputs-Advanced-Mode.mp4

Selection of multiple references

Some analyses (notably mapping with **Ivar** or pipelines that include it), can use more than one reference file. Reference selection can be performed by opening the reference genomes pop-up window and selecting any of them through the coresponding checkboxes. Repeating the process will not overwrite references selected previously: the new references will be concatenated to the end of the list.

The list of selected references can be cleaned, ordered, or single references can be removed one by one.

> **Nota:** reference order impacts on the final outcome of the multi-reference mapping. Please order the main references first.

The following videotutorial demonstrates how to interact with options for selection of multiple references.

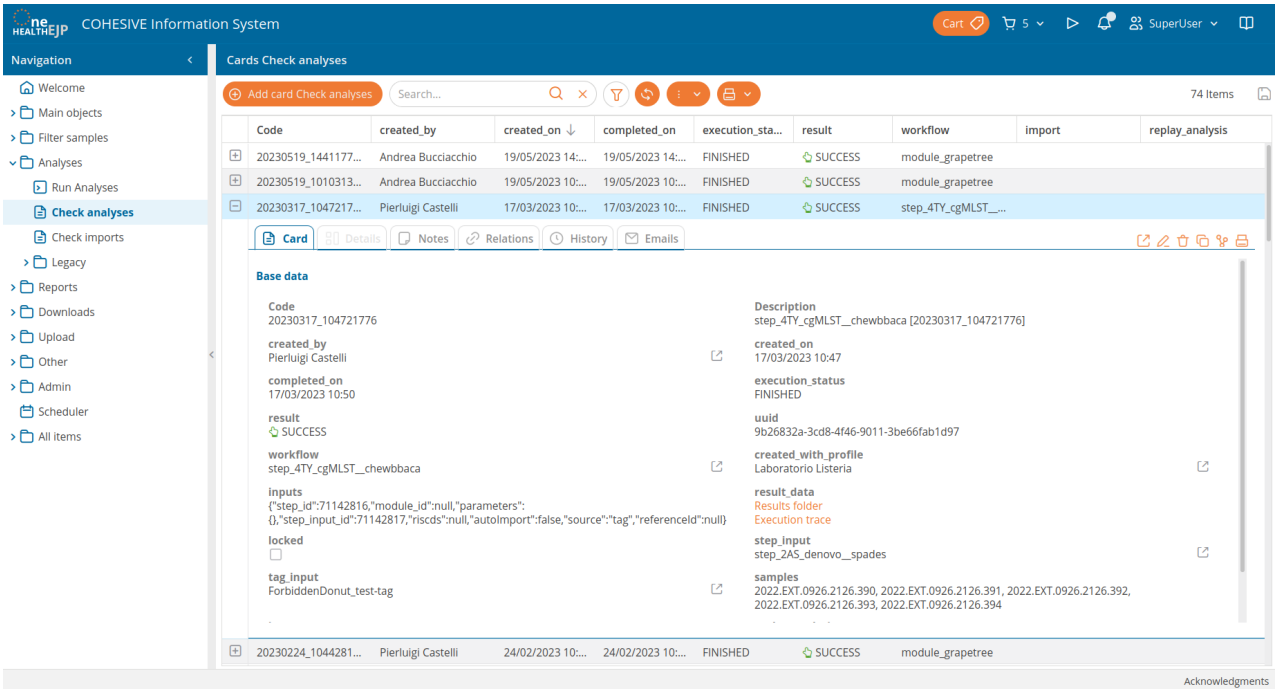
type:video-tag multireference.mp4

4.3 Check analyses

Requested analyses are listed in the **Check analysis** page, available under the **Analyses** section of the navigation menu. Each analysis card can be expanded to display additional information.

4.3.1 Analysis card

The analysis' card holds information about the processed run. If the process concluded successfully, the card will display technical information about the process, used inputs, the "Repeat analysis" function and the import results function. Additionally, in the card will also appear a link to the **Results folder**.



4.3.2 Import analyses

Once an analysis completed the run successfully, output files are stored in a temporary directory for 45 days. The directory will be accessible by users and download its contents, but after the 45 days period has expired, the temporary directory will be removed from the servers. **In order to permanently save the analysis' results, they must be imported in the permanent directory**.

> Note also that usually **to run an analysis on inputs that are produced by another analysis, such files must be imported**.

To import, it's sufficient to click the "Import analysis" button in the analysis card. The Information System will notify the user after import completion. The following short videotutorial shows how to import an analysis.

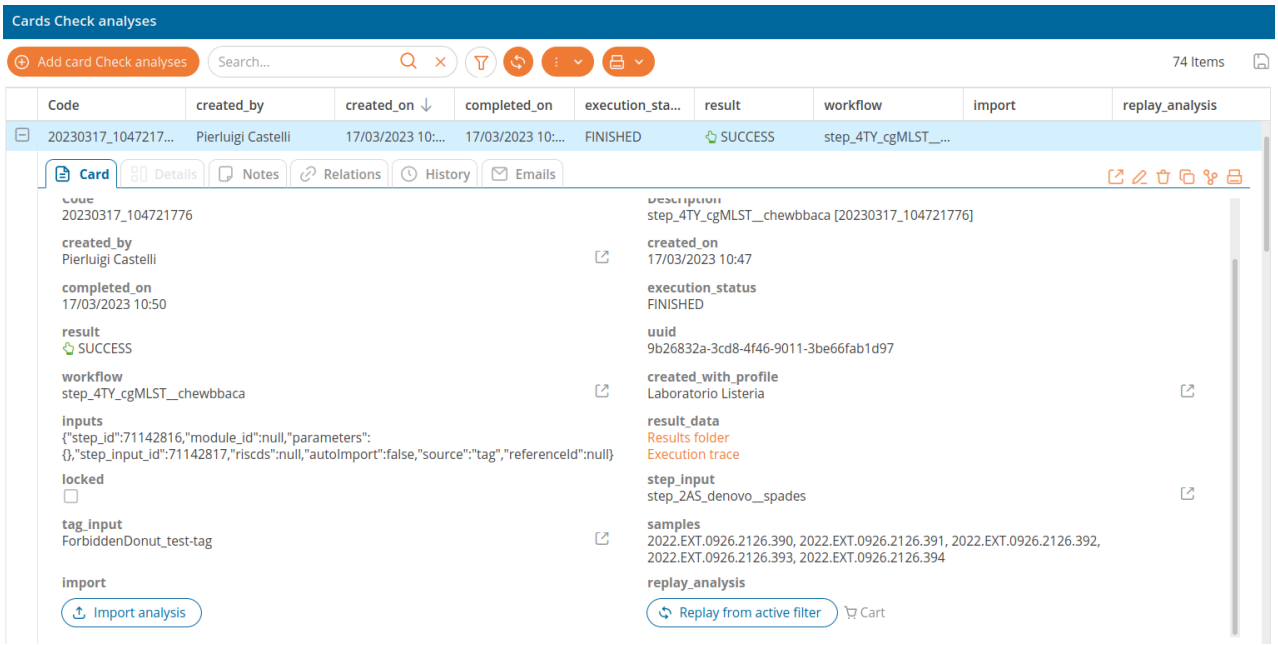
type:video-tag Import-analysis.mp4

> Note that not all analyses can be imported.

A list of import processes and their progression is available in the **Check imports** page (**Analyses** section), of the navigation menu.

4.3.3 Replay Analysis

The option "**Replay from active filter**" (in an analysis summary card) grants the ability to repeat the analysis on a different set of samples, using the same parameters and options. The new set of samples can be provided from cart or active tag (the cart takes precedence on active tag), furthermore, the analysis can be replicated only if it finished successfully the run.



The following videotutorial shows how to replay an analysis, demonstrating both the usage of a tag and of the cart.

type:video-tag repeat_analysis.mp4

4.3.4 Lock Analysis

The analysis cards contain the "**Lock Analysis**" option; if activated, a non-importable analysis will be kept indefinitely, avoiding data removal after the time limit for storage is reached.

type:video-tag lock-analysis.mp4

4.4 Tool Descriptions

4.4.1 Introduction

Analyses on bacterial or viral isolates available in the Choesive Information System are organized into categories. Such groups are assigned names following a specific nomenclature system, which describes the analysis type and execution level inside pipelines.

Nomenclature system

Prefixes

Analysis type and execution level are summarized by a short prefix code. The table below lists such prefixes:

1PP preprocessing analyses
2AS assembly tools
2MG metagenomics analyses
3TX taxonomical classification
4TY *in silico* typing
4AN genome annotation

The first element of the prefix is a number, indicating the usual execution level: preprocessing is usually performed before any other analysis, thus is assigned level 1; taxonomical analyses require files from assembly as input and as such they are assigned level 3, after preprocessing and assembly.

An additional class is represented by the code "**0SQ**", which identifies **S**equence **Q**uality checks performed automatically on all new reads in Cohesive.

Analysis names

The analysis name follows the prefix and describes the kind of data handling performed by available bioinformatic tools.

For example, the "trimming" analysis' full name is "1PP_trimming", since it's classified as preprocessing. Similarly, *de novo* assembly will be called "2AS_denovo" and such name will be maintained independently of the bioinformatic tool used to perform it.

The sections below list all analyses and their groups, together with brief descriptions and links to the appropriate Wiki pages.

Suffixes

Many of the available analyses can be executed with multiple softwares, also called "bioinformatic tools" or "methods".

An analysis name can thus be completed by appending 2 underscore characters ("_") and the name of the selected tool; *e.g.* 2AS_denovo__spades and 2AS_denovo__unicycler, both of which execute the *de novo* assembly, but with different softwares.

Analyses that can be executed with multiple softwares will allow tool selection through a dropdown menu, available in the second step ("Tools") of the Run Analysis wizard.

Available tools for each analysis are listed in the respective analysis Wiki page.

NOTE: some analyses and pipelines are not yet available in the Cohesive Information System Demo.

Single Sample Analysis

Prefix	Analysis Name	Description	Tools
1PP	trimming	removal of low quality nucleotide calls from raw reads	trimmomatic
	hostdepl	depletion of host sequences: reads are mapped against the selected host genome to remove contaminant sequences	bowtie
	downsampling	reduction of the number of sequences in genomic regions with excessive and uninformative vertical coverage	bbnorm
	denovo	de novo assembly: builds genome scaffolds from the pool of contigs	SPAdes unicycler
2AS	mapping	sequence mapping with a reference genome	bowtie ivar snippy
2MG	denovo	de novo assembly for metagenomics: the metaSPAdes software builds a de Bruijn graph for all reads with SPAdes, which is then transformed into an assembly graph, finding paths corresponding to genome fragments in a metagenome	metaSPAdes
3TX	class	taxonomic classification and contamination check of the organisms the reads belong to	kraken kraken2 confindr
	species	closest bacterial or viral species identification / identification of the best viral reference	kmerfinder blast vdabricate
	MLST	Multi-Locus Sequence Typing in silico: it uses schemas of 7 conserved genes specific for each bacterium to assign Sequence Type and Clonal Complex	mlst
	cgMLST	core genome Multi-Locus Sequence Typing in silico phylogenetic analysis: allele calling species-specific core genome allele schemas	chewBBACA mentalist blastMLST
4TY	flaA	analysis specific for Campylobacter. In silico identification of flaA locus variant (MLST for flaA)	flaA
	lineage	lineage assignment for SARS-CoV2 lineage assignment for West Nile Virus	pangolin westnile
	wgMLST	whole genome MLST	chewBBACA
	genes	functional genome annotation through ORF (Open Reading Frame) search in a genome and identification of possible coded proteins	prokka
4AN	AMR	prediction of antibiotic resistance-associated genes presence	abricate blast staramr filtering

Multi Sample Analysis

Analysis type	Analysis name / Tool	Description
Gene-by-gene based clustering	Grapetree	Gene-by-gene based clustering - builds MST (Minimum Spanning Tree) e NJ (Neighbor Joining) tree nwk files
	Reportree	Gene-by-gene based clustering
Pangenome extraction	Panaroo	calculates a presence/absence binary matrix of accessori genes in the sample genomes, starting from gff files from Prokka (genome annotation)
	Snippy-core	executes Snippy to identify mutations (SNPs e indels) in sample reads diverging from a reference haploid genome, followed by Snippy-core to build the core.vcf from Snippy's vcf files. core.vcf contains all core mutations among those listed in Snippy's vcf files
	CFSAN	SNP identification through reference-based phylogenetic analysis
SNP-based clustering	kSNP3	SNP identification through reference-free phylogenetic analysis. It builds a Maximum Likelihood tree graph
	VCF2MST	fast construction of MST graph from a VCF file with no phylogenetic inferences

Pipelines

In addition to single analyses, Cohesive also delivers automatic pipelines, which consist of concatenated single analyses. Pipelines' goal is to make running a workflow simpler and faster, especially if such a workflow is a common one.

The table below lists pipelines and the analyses that are part of them (for more information on softwares used in such analyses, please refer to the corresponding Wiki pages).

Pipeline Name	Description	Analyses
Covid Emergency	SARS-CoV2 fast assembly and lineage assignment	2AS_mapping + 4TY_lineage
Depletion & *de novo*	host depletion from trimmed reads and *de novo* assembly	1PP_hostdepl + 2AS_denovo
Genome Draft	mapping and genome annotation. Mapping is performed with both bowtie and Snippy.	2AS_mapping + 4AN_genes
Enterotoxin *S. aureus* finder	*de novo* assembly and blast to identify enterotoxin gene presence	2AS_denovo + 4AN_AMR
NgsManager	macro-pipeline which, depending on sample type, executes the pipeline "Raw Reads Processing" followed by "Covid Emergency" (for SARS-CoV2), "WGS on Bacteria" and "Typing on Bacteria" (for bacterial isolates) or "Genoma Draft" (for viral isolates or biologic sample)	analyses from the "Raw Reads Processing", "Covid Emergency", "WGS on Bacteria", "Typing on Bacteria" and "Genoma Draft" pipelines
Raw Reads Processing	reads quality check, trimming and virus/bacteria classification	0SQ_rawreads + 1PP_trimming + 3TX_class
Typing on Bacteria	horizontal and vertical coverage calculation, species calculation, gene annotation, identification of antibiotic resistance- and virulence-associated genes, typing	2AS_mapping + 3TX_species + 4AN_genes + 4AN_AMR + 4TY_wgMLST + 4TY_cgMLST + 4TY_MLST + 4TY_flaA
WGS on Bacteria	assembly of bacterial isolate	2AS_denovo
WNV - lineage calculation and mapping	lineage calculation and mapping against the calculated lineage reference for *West Nile Virus* samples	4TY_lineage + 2AS_mapping

4.4.2 Tools for long reads

In this page are listed all analyses and bioinformatic tools available for long reads in CIS. Long reads can be processed **only** with tools that are specifically designed to manage them.

Analyses for long reads from Oxford Nanopore

Analysis	Tool	Notes
1PP_filtering__minimap2	minimap2	only for reads from Nanopore apparatus
1PP_hostdepl__minimap2	minimap2	only for reads from Nanopore apparatus
1PP_trimming__chopper	chopper	only for reads from Nanopore apparatus
2AS_denovo__flye	Flye	only for reads from Nanopore apparatus
2AS_hybrid__unicycler	Unicycler	for hybrid assembly of Illumina short reads and Nanopore long reads
2AS_mapping__medaka	Medaka	only for reads from Nanopore apparatus
3TX_class__centrifuge	Centrifuge	only for reads from Nanopore apparatus

Analyses for long reads from Thermofisher IonTorrent

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

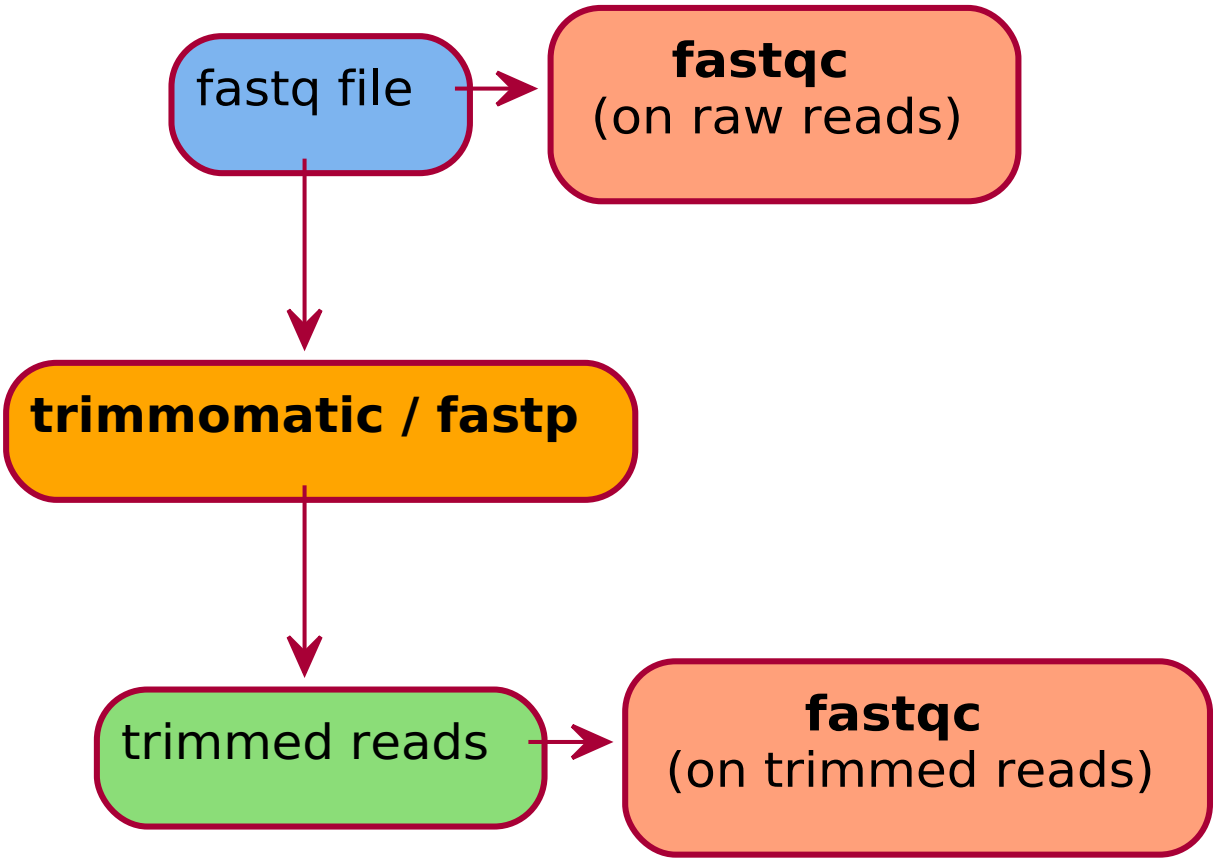
4.4.3 Single sample

1 Pre-processing

1PP_trimming

Introduction

The trimming step removes low quality nucleotide residues from reads produced by the sequencer. The **1PP_trimming** analysis in Cohesive includes the execution of **trimmomatic** and **fastqc**. The latter gives quality metrics about the reads, *i.e.* raw reads (before trimming) and produced trimmed reads.



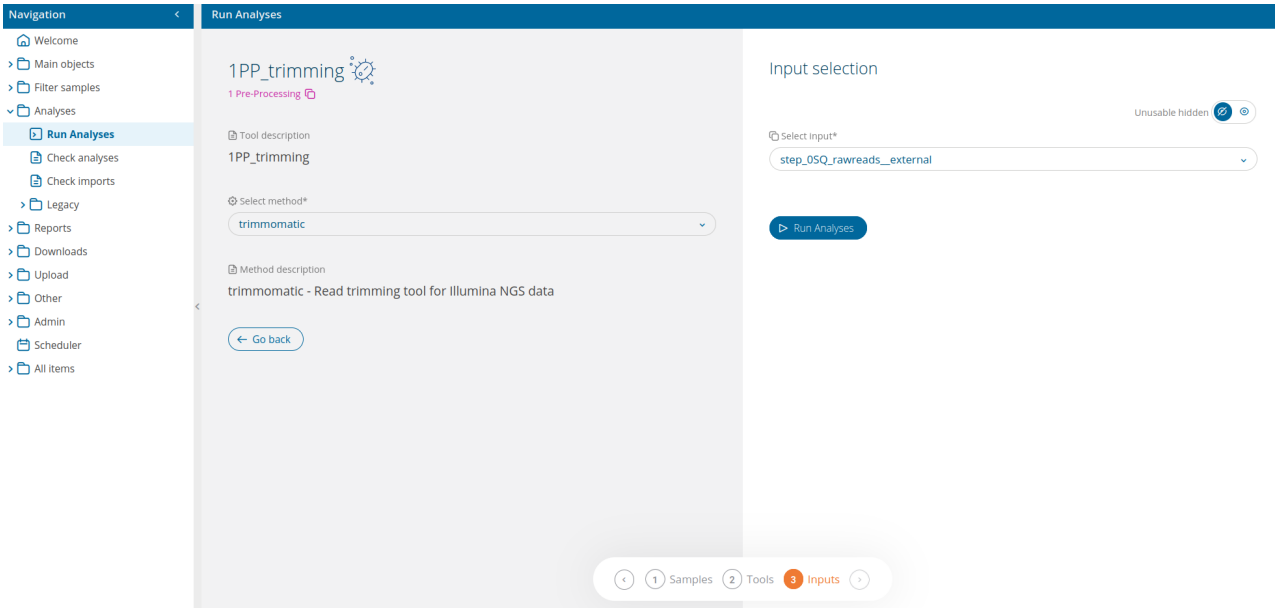
Tool
Input
Result
Quality Check tool

Run Analysis 1PP_trimming

Once the analysis **1PP_trimming** has been selected from the run analyses interface, the user will be able to select which bioinformatic tool to use. The available tool for **1PP_trimming** is:

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

Input is selected in the wizard's last section: "step_0SQ_rawreads__fastq" is for internal fastq files from sequencers, (code 20XX.TE.XXXX.X.X), while "step_0SQ_rawreads__external", is for imported fastq files (code 20XX.EXT.XXXX.X.X). The input selection UI also delivers an advanced input selection mode, to allow selection of all types of supported input files at once.



A link to Check analysis will be created after launching the requested analysis. The system will notify the user after a succesful analysis launch and once execution has ended.

Output directory

> Please refer to Cohesive's specific Wiki page for information on file download.

The output directory is available at the link in the download page or at the link presente in the analysis' summary card, and will have the following structure: `results > YEAR > ID > 1PP_trimming > DSXXXXXXXX-DTXXXXXX_trimmomatic`. At that path there will be 3 directories:

- **meta**: ("metadata") contains log and configuration files.
- **qc**: ("quality check") it contains 2 directories (meta and result). In this case quality check is performed with **fastqc**.
- **result**: contains the analysis' output files.

The table below lists files available in the output directory structure, alongside some useful information.

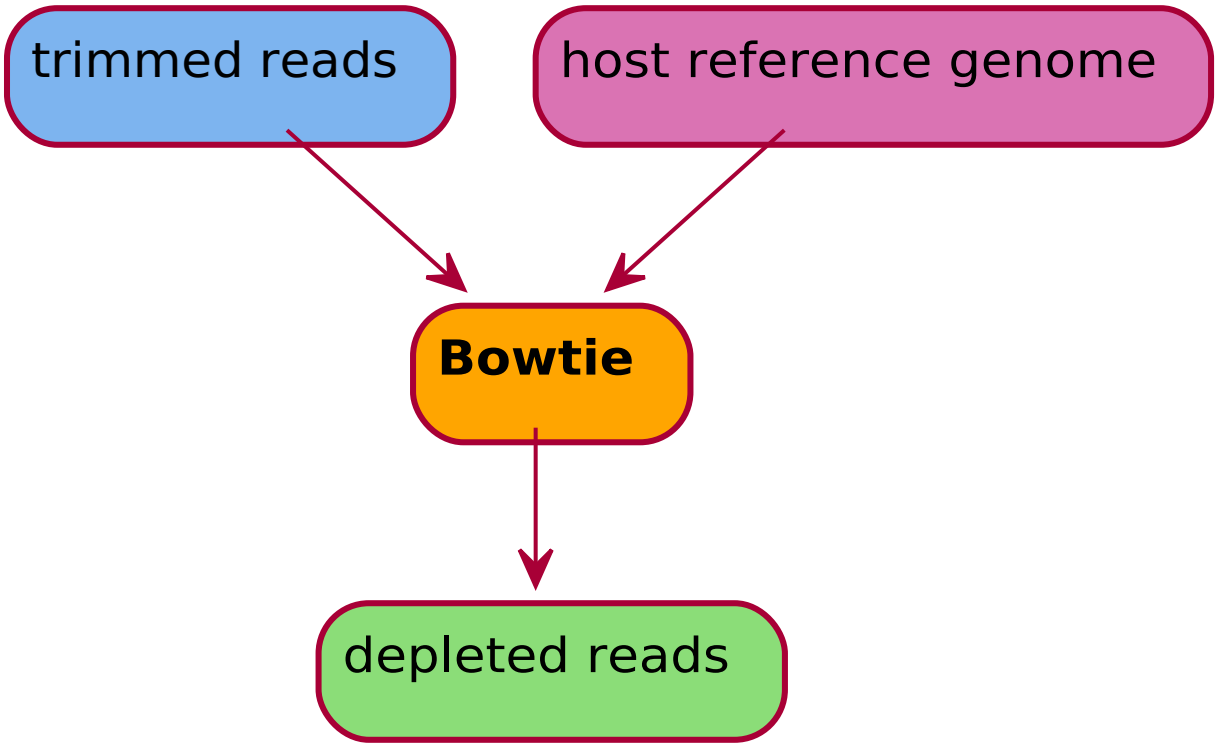
File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_R1_trimmomatic.fastq	trimmed read 1 (R1)	result directory
DSXXXXXXXX-DTXXXXXX_ID_R2_trimmomatic.fastq	trimmed read 2 (R2)	result directory
DSXXXXXXXX-DTXXXXXX_ID_unpaired_trimmomatic.fastq	trimmed unpaired reads	result directory
DSXXXXXXXX-DTXXXXXX_ID_R1_trimmomatic_fastqc.html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

1PP_hostdepl

Introduction

The 1PP_hostdepl analysis maps input reads against a selected host's reference genome and removes contaminant reads originating from the host from the fastq files, a process also called "host depletion".

Output files are new fastq with clean reads.



Tool
Input
Result
Reference

Run Analysis 1PP_hostdepl

Once the analysis **1PP_hostdepl** has been selected from the run analyses interface, the user will be able to select which bioinformatic tool to use. The available tools for this analysis are listed below.

For all kinds of reads:

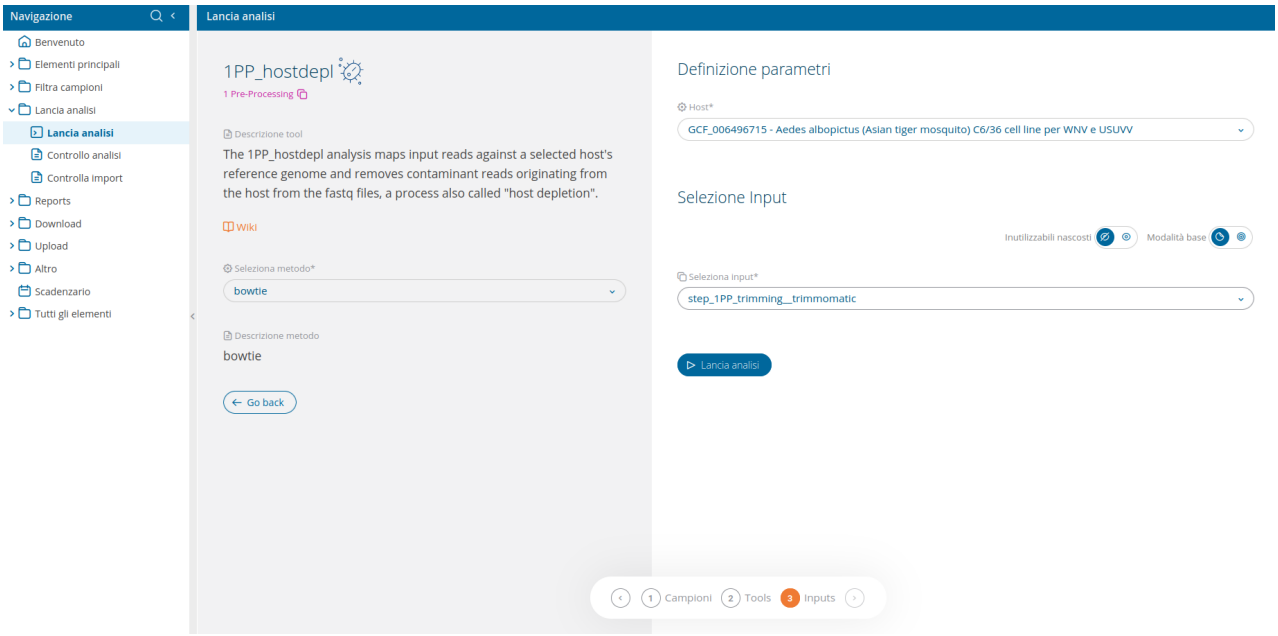
- **Bowtie** - An ultrafast, memory-efficient short read aligner

Only for long reads from nanopore technology apparatus:

The input selection wizard delivers an advanced input selection mode, to allow selection of all types of supported input files at once.

Inputs for Bowtie:

- step_1PP_trimming
- step_1PP_hostdepl
- step_1PP_downsampling



A link to Check analysis will be created after launching the requested analysis. The system will notify the user after a succesful analysis launch and once execution has ended.

Output directory

> Please refer to Cohesive's specific Wiki page for information on file download.

The output directory is available at the link in the download page or at the link presente in the analysis' summary card, and will have the following structure: `results > YEAR > ID > 1PP_hostdepl > DSXXXXXXXX-DTXXXXXX_bowtie`. The last directory's suffix will be replaced with the name of the chosen tool. At that path there will be 2 directories:

- **meta**: ("metadata") contains log and configuration files.
- **result**: contains the analysis' output files.

The tables below list files available in the output directory structure, alongside some useful information.

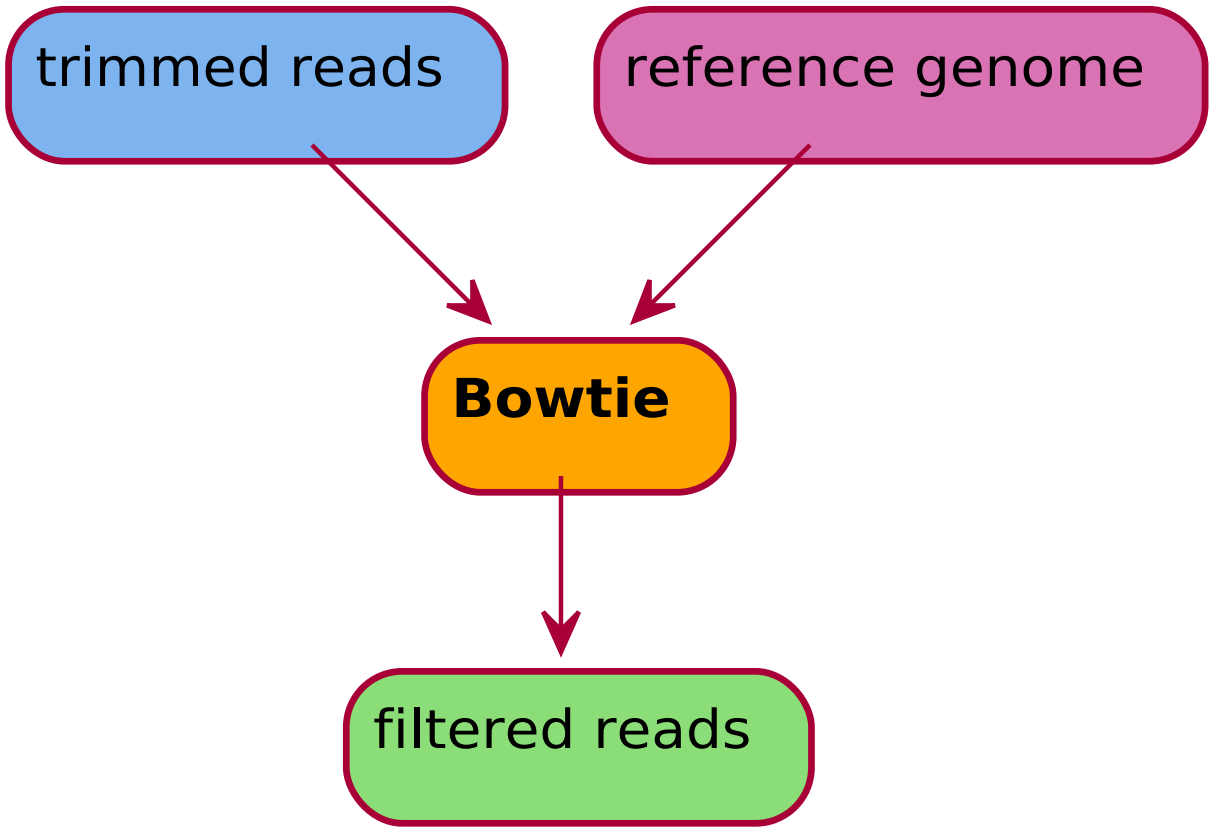
Bowtie:

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

1PP_filtering

Introduction

The 1PP_filtering maps reads against a selected reference genome and returns only those that mapped successfully, removing (filtering) reads that did not map, like reads from the host.



Tool
Input
Result
Reference

Run Analysis 1PP_filtering

Once the analysis **1PP_filtering** has been selected from the run analyses interface, the user will be able to select which bioinformatic tool to use. The available tools for this analysis are listed below.

For all kinds of reads:

- **Bowtie** - An ultrafast, memory-efficient short read aligner

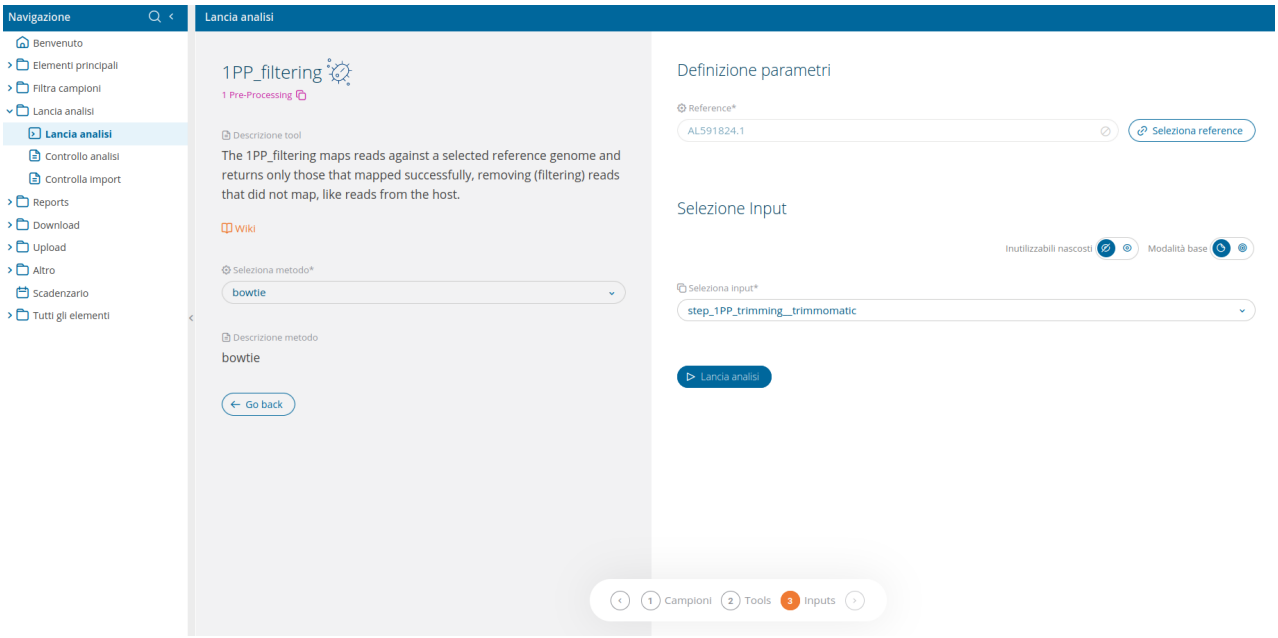
Only for long reads from nanopore technology apparatus:

The wizard for input selection will require input reads and a reference genome for mapping.

The input selection wizard delivers an advanced input selection mode, to allow selection of all types of supported input files at once.

Inputs for Bowtie:

- step_1PP_trimming
- step_1PP_hostdepl
- step_1PP_downsampling
- step_1PP_filtering



A link to Check analysis will be created after launching the requested analysis. The system will notify the user after a succesful analysis launch and once execution has ended.

Output directory

Please refer to Cohesive's specific Wiki page for information on file download.

The output directory is available at the link in the download page or at the link presente in the analysis' summary card, and will have the following structure: `results > YEAR > ID > 1PP_filtering > DSXXXXXXXX-DTXXXXXX_bowtie`. The last directory's suffix will be replaced with the name of the chosen tool. At that path there will be 2 directories:

- **meta**: ("metadata") contains log and configuration files.
- **result**: contains the analysis' output files.

The tables below list files available in the output directory structure, alongside some useful information.

Bowtie:

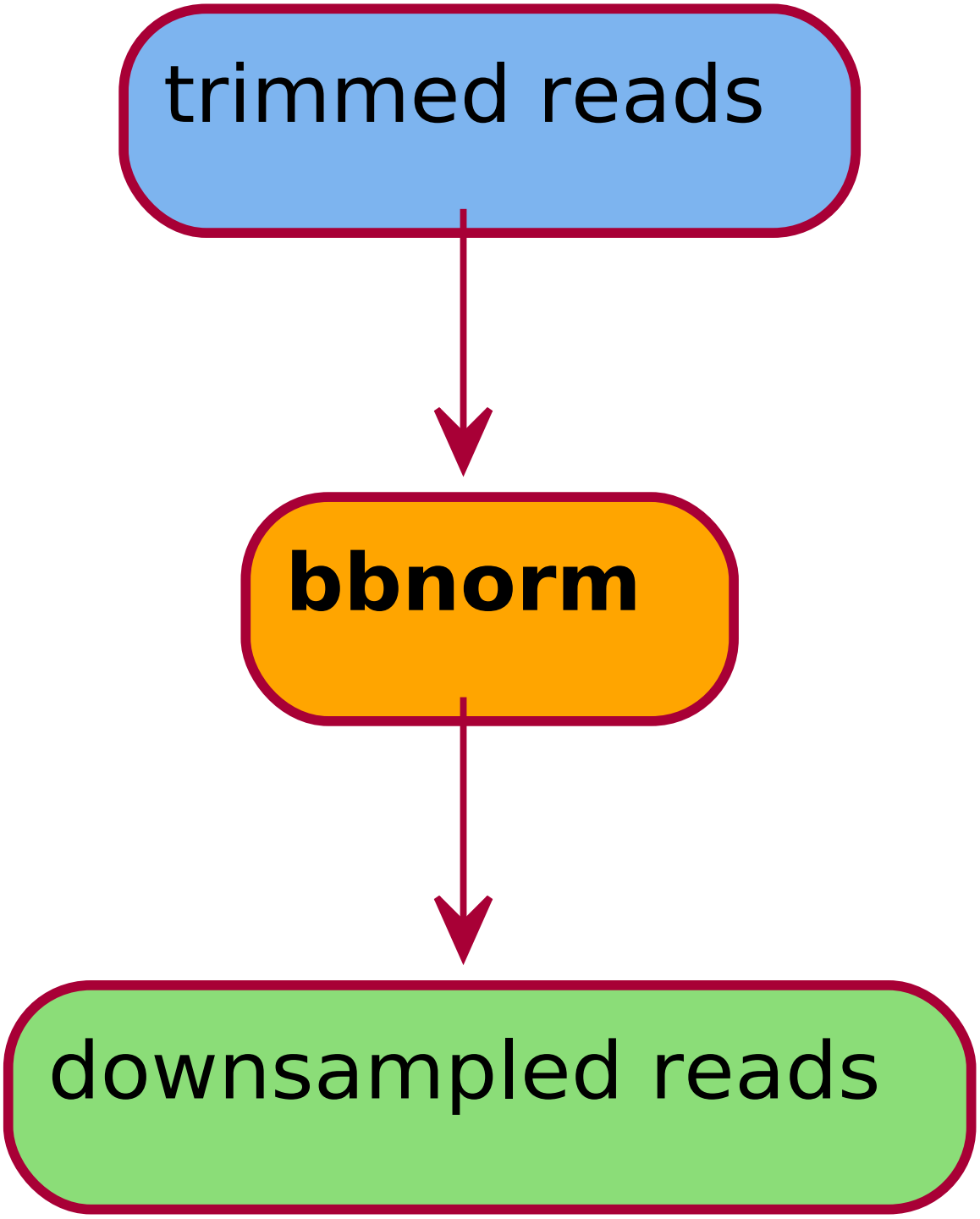
File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_R1_bowtie_REFID.fastq.gz	filtered read 1 (R1) (compressed file)	result directory
DSXXXXXXXX-DTXXXXXX_ID_R2_bowtie_REFID.fastq.gz	filtered read 2 (R2) (compressed file)	result directory

1PP_downsampling

Introduction

Downsampling is defined as a process of reduction in read depth (*vertical coverage*), at specific positions or regions of the genome.

Sequencing protocols can cause stacking of reads for a specific region, leading to excessive data that slows down execution of downstream analyses, while providing no additional information. Downsampling reads prevents extension of calculation times by discarding read pairs until a defined threshold for desired vertical coverage.



Tool
Input
Result

Run Analysis 1PP_downsampling

Once the analysis **1PP_downsampling** has been selected from the run analyses interface, the user will be able to select which bioinformatic tool to use. The only available tools for this analysis is **bbnorm**.

The wizard will require input reads and additional parameters to define the target vertical coverage.

The "kmer length" parameter is specific for used samples, while the "Target" parameter is dependent on kmer length and species. As a result, the second parameter has to be determined empirically.

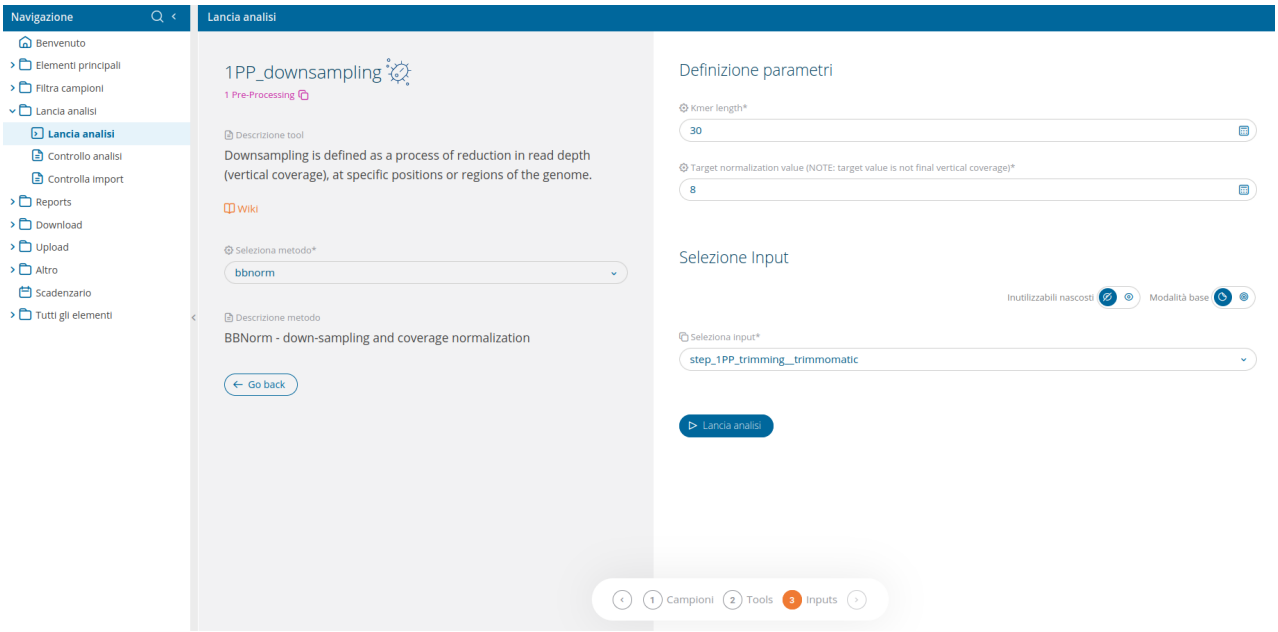
> **Warning:** the second parameter **does not** correspond to the final verticale coverage.

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

The input selection wizard delivers an advanced input selection mode, to allow selection of all types of supported input files at once.

- step_1PP_trimming
- step_1PP_hostdepl
- step_1PP_downsampling



A link to Check analysis will be created after launching the requested analysis. The system will notify the user after a succesful analysis launch and once execution has ended.

Output directory

> Please refer to Cohesive's specific Wiki page for information on file download.

The output directory is available at the link in the download page or at the link presente in the analysis' summary card, and will have the following structure: `results > YEAR > ID > 1PP_downsampling > DSXXXXXXXX-DTXXXXXX_bbnorm`. At that path there will be 2 directories:

- **meta**: ("metadata") contains log and configuration files.
- **result**: contains the analysis' output files.

The table below lists files available in the output directory structure, alongside some useful information.

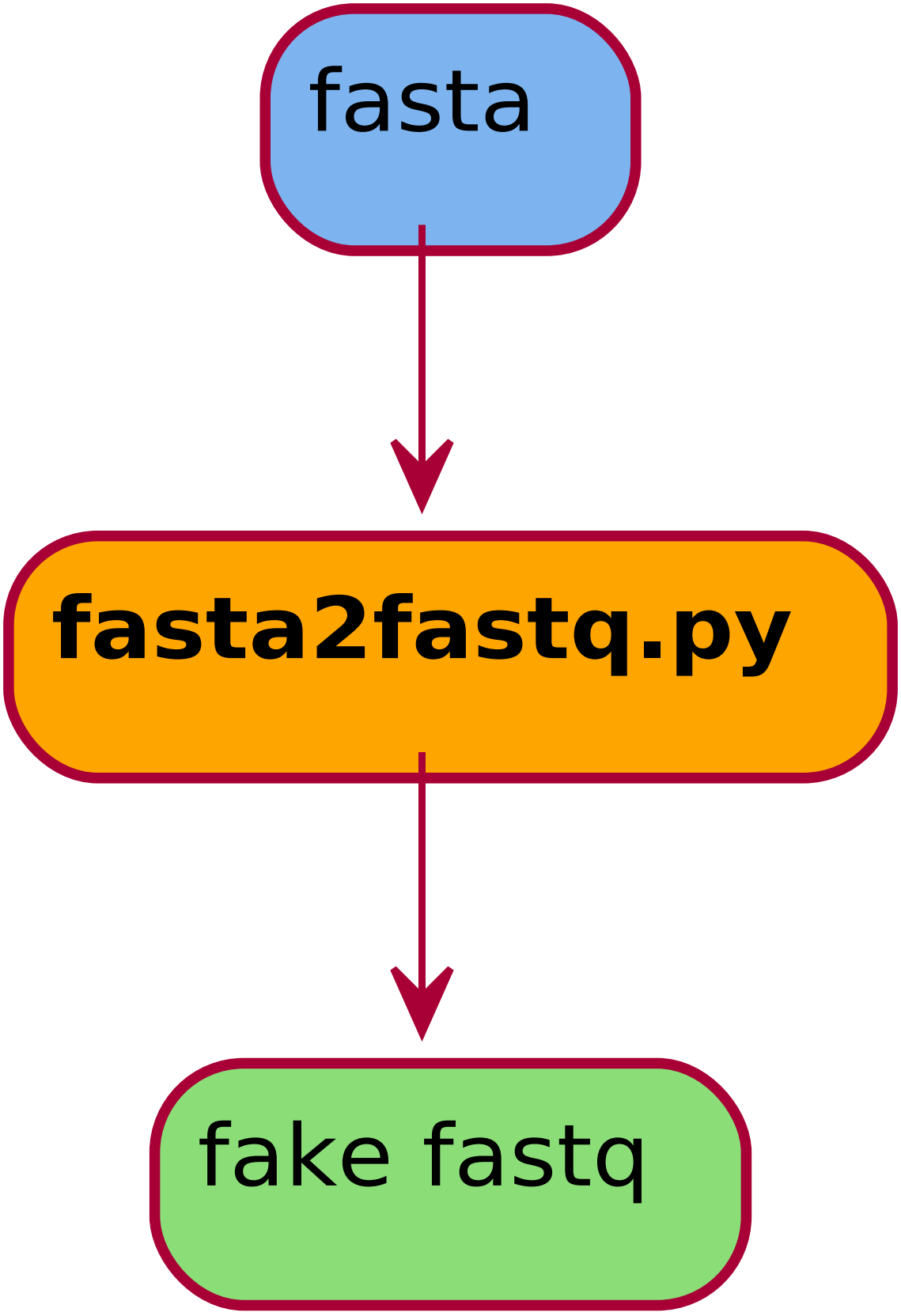
File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_bbnorm_kXX_tX_R1.fastq.gz	downsampled read 1 (R1)	result directory
DSXXXXXXXX-DTXXXXXX_ID_bbnorm_kXX_tX_R2.fastq.gz	downsampled read 2 (R2)	result directory

1PP_generated

Introduction

The 1PP_generated builds a fastq file from a fasta file. To perform fastq generation, CIS uses a custom script (**fasta2fastq.py**), which creates artificial quality lines and modifies headers to make them compliant to fastq standards. Files for both antiparallel DNA strands (R1 e R2).

Warning: Since generated quality line are artificial, this analysis is only aimed at allowing to run softwares that use fastq files on sequences only available as fasta. Output files are not to be used as replacement for proper fastq sequences in workflows for which read quality evaluation matters.



Tool

Run Analysis 1PP_generated

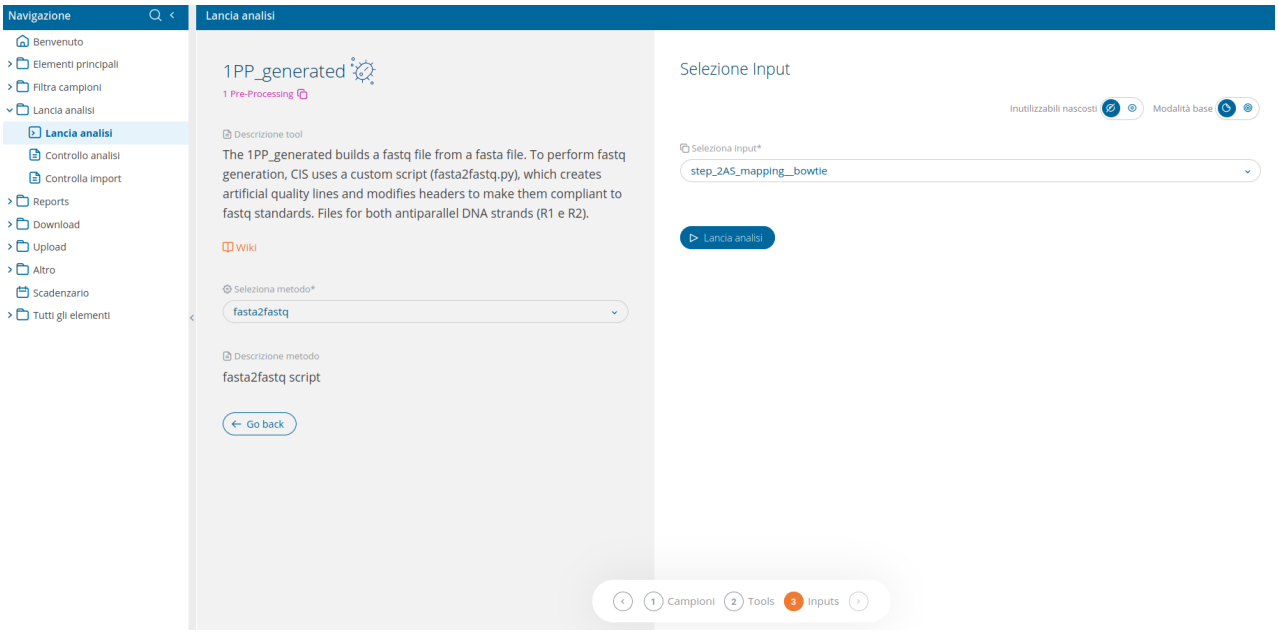
Once the analysis **1PP_generated** has been selected from the run analyses interface, the user will be able to select which bioinformatic tool to use. The only available tool for this analysis is **fasta2fastq**.

The input selection wizard delivers an advanced input selection mode, to allow selection of all types of supported input files at once.

Available inputs for 1PP_generated are:

- step_2AS_mapping
- step_2AS_denovo
- step_2AS_import

The wizard requires input sequences, usually from *de novo* assembly or mapping and the reference genome that has been used for mapping, in case such input is selected.



A link to Check analysis will be created after launching the requested analysis. The system will notify the user after a succesful analysis launch and once execution has ended.

Output directory

> Please refer to Cohesive's specific Wiki page for information on file download.

The output directory is available at the link in the download page or at the link presente in the analysis' summary card, and will have the following structure: `results > YEAR > ID > 1PP_generated > DSXXXXXXXX-DTXXXXXX_fasta2fastq`. At that path there will be 2 directories:

- **meta**: ("metadata") contains log and configuration files.
- **result**: contains the analysis' output files.

The table below lists available fasta2fastq.py output files, alongside some useful information.

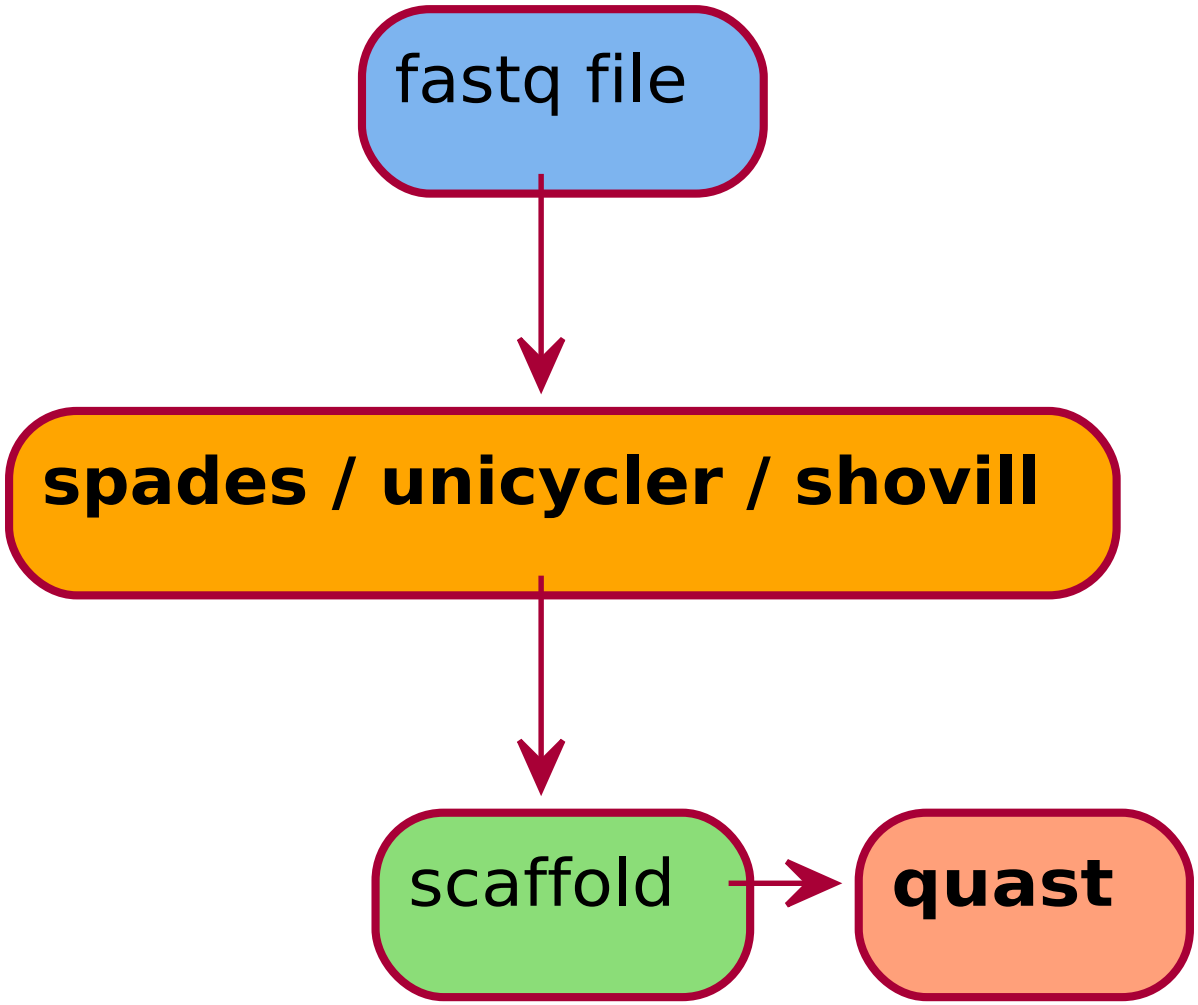
File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_fasta2fastq_R1.fastq.gz	artificial fastq of + strand (compressed file: .gz)	results directory
DSXXXXXXXX-DTXXXXXX_ID_fasta2fastq_R2.fastq.gz	artificial fastq of - strand (compressed file: .gz)	results directory

2 Assembly

2AS_denovo

Introduction

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



Tool
Input
Result
Quality Check tool

Run Analysis 2AS_denovo

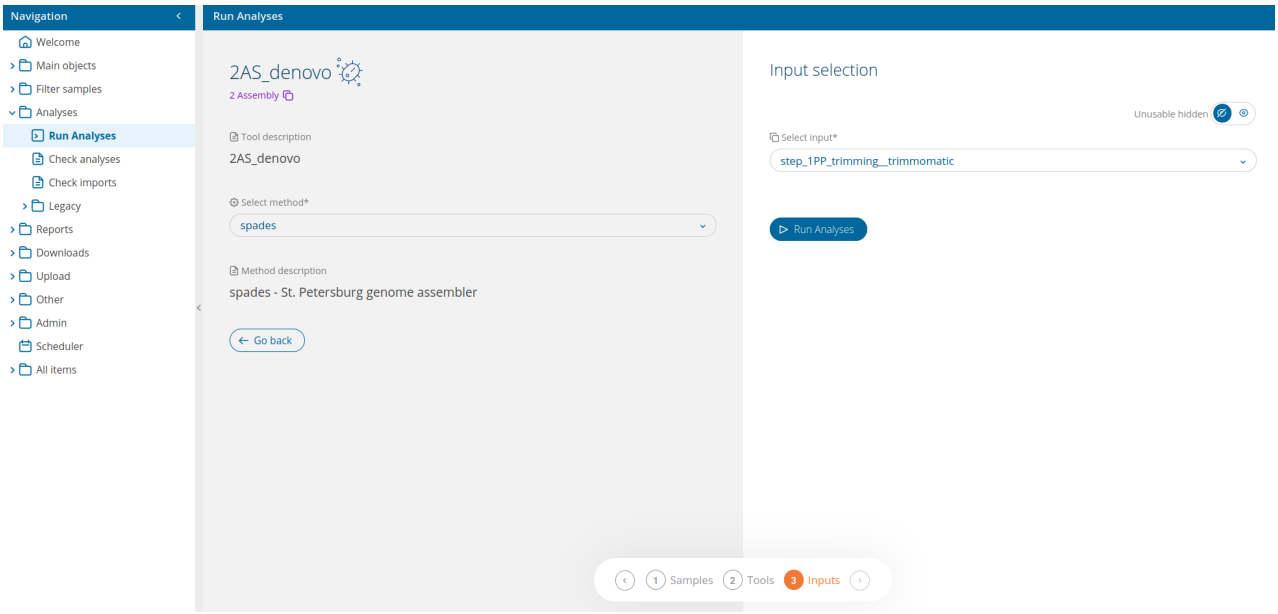
Once the analysis **2AS_denovo** has been selected from the run analyses interface, the user will be able to select which bioinformatic tool to use. The available tool is:

- **spades** - St. Petersburg genome assembler

The input selection wizard delivers an advanced input selection mode, to allow selection of all types of supported input files at once.

Accepted inputs can be from:

- step_1PP_trimming
- step_3TX_class
- step_4AN_AMR



A link to Check analysis will be created after launching the requested analysis. The system will notify the user after a succesful analysis launch and once execution has ended.

Output directory

> Please refer to Cohesive's specific Wiki page for information on file download.

The output directory is available at the link in the download page or at the link presente in the analysis' summary card, and will have the following structure: `results > YEAR > ID > 2AS_denovo > DSXXXXXXXX-DTXXXXXX_spades`. The last directory's suffix will be replaced with the name of the chosen tool. At that path there will be 3 directories:

- **meta**: ("metadata") contains log and configuration files.
- **qc**: ("quality check") it contains 2 directories (meta and result). In this case quality check is performed with **quast**.
- **result**: contains the analysis' output files.

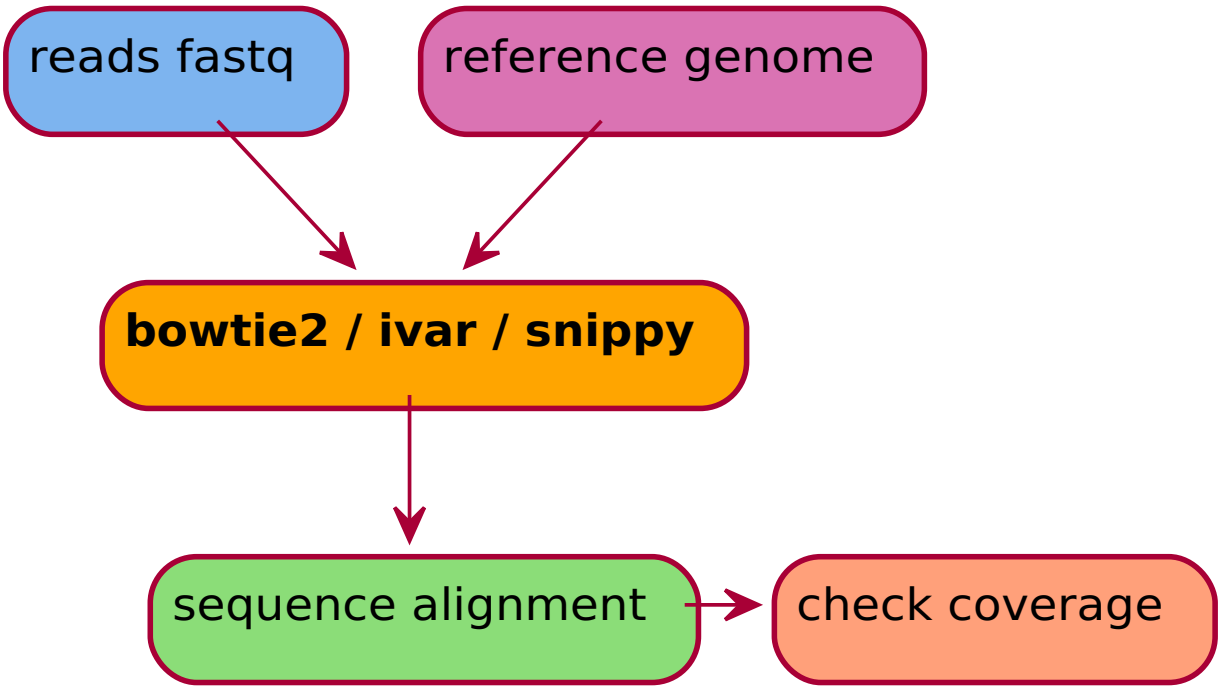
The table below lists files available in the output directory structure, alongside some useful information.

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_spades_contigs.fasta	*de novo* assembly's contigs file	result directory
DSXXXXXXXX-DTXXXXXX_ID_spades_scaffolds.fasta	scaffolds file	result directory
DSXXXXXXXX-DTXXXXXX_ID_spades_scaffolds_L200.fasta	file with scaffolds longer than 200bp	result directory
DSXXXXXXXX-DTXXXXXX_ID_quast.csv	file with assembly's quality metrics	qc directory > result

2AS_mapping

Introduction

2AS_mapping performs the **assembly with reference** (*i.e.* **mapping**) of nucleotide sequences.



Tool
Input
Result
Reference
Quality check

Run Analysis 2AS_mapping

Once the analysis **2AS_mapping** has been selected from the run analyses interface, the user will be able to select which bioinformatic tool to use. The available tools are:

- **bowtie2** - Fast and sensitive read alignment
- **ivar** - computational package for viral amplicon-based sequencing

Note: In case the software "snippy" is chosen, samples will need to have an associated 4AN_genes output files from "Prokka".

To select the mandatory reference genome, use the "Select reference" button in the input selection wizard. This will open a pop-up table, listing 2 different kind of sequences, both usable as reference:

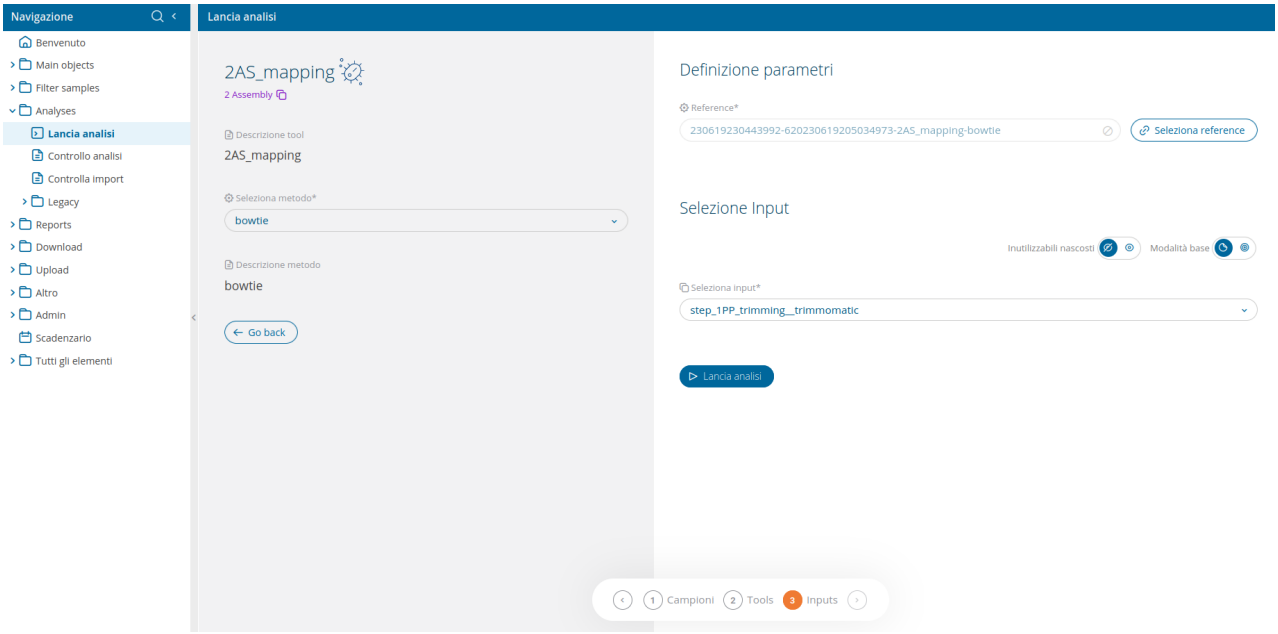
1. **reference fasta files**;
2. consensus sequence or *de novo* assembly of **a sample available in Cohesive**.

With the tool "Ivar" it's possible to select multiple references (please consult the "Multiple references" section).

The input selection UI delivers an advanced input selection mode, to allow selection of all types of supported input files at once.

Accepted inputs can be from:

- step_1PP_trimming
- step_1PP_hostdepl



A link to Check analysis will be created after launching the requested analysis. The system will notify the user after a succesful analysis launch and once execution has ended.

Multiple references

If 2AS_mapping is performed with **Ivar**, it will be possible to select more than one reference, as shown in the image below. A guide to the multiple reference selection system is available at the corresponding section of the run analysis Wiki.

Output directory

Please refer to Cohesive's specific Wiki page for information on file download.

The output directory is available at the link in the download page or at the link presente in the analysis' summary card, and will have the following structure: `results > YEAR > ID > 2AS_mapping > DSXXXXXXXX-DTXXXXXX_bowtie2`. The last directory's suffix will be replaced with the name of the chosen tool. At that path there will be 2 directories:

- **meta**: ("metadata") contains log and configuration files.
- **result**: contains the analysis' output files.

The tables below list files produced by 2AS_mapping's available tools.

Bowtie2

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_bowtie_REFID.fasta	consensus file	result directory
DSXXXXXXXX-DTXXXXXX_ID_bowtie_REFID.sorted.bam	bam (Binary Alignment Map) format alignment file	result directory
DSXXXXXXXX-DTXXXXXX_ID_bowtie_REFID.sorted.bam.bai	bai (bam file's index) file	result directory
DSXXXXXXXX-DTXXXXXX_ID_bowtie_REFID.var.flt.vcf	vcf (variant calling format) file with identified varaints	result directory
DSXXXXXXXX-DTXXXXXX_ID_bowtie_REFID_coverage_plot.png	coverage distribution graph	result directory

Ivar

Note: Ivar's execution consists of "Snippy", "Samtools" and "Ivar" tools execution.

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_ivar_REFID.fasta	consensus sequence from Ivar	result directory
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.aligned.fa	reference with `-` in positions with `sequencing depth = 0` and `depth`'s `N` between 0 and the minimum number of reads considered for site coverage (**no variants in this file**)	result directory
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.bam	bam (Binary Alignment Map) format alignment file	result directory
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.bam.bai	bai (bam file's index) file	result directory
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.bed	bed (Browser Extensible Data) file	result directory
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.consensus.fa	reference genome with representation of all variants	result directory
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.consensus.subs.fa	reference genome with representation of substitution variants	result directory
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.txt	snippy run summary	result directory
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.tab	variant table in tsv format	result directory
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.csv	variant table in csv format	result directory
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.filt.vcf	variants filtered by Freebayes	result directory
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.raw.vcf	variants not filtered by Freebayes	result directory
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.subs.vcf	table of substitution variants in vcf format	result directory
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.gff	variants in GFF3 format	result directory
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.html	html version of the .tab table of variants	result directory
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.vcf	Snippy's output file with identified variants in vcf format	result directory
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.vcf.gz	snippy's vcf output file (archive)	result directory
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID.vcf.gz.csi	bcftools index of vcf.gz file	result directory
DSXXXXXXXX-DTXXXXXX_ID_vdsnippy_REFID_coverage_plot.png	coverage distribution graph	result directory

For further information about Snippy's output files, file formats and contents, please refer to Snippy's official manual.

Data visualization

Alignment of reads on the reference genome can be visualized with specific softwares (*i.e.* Tablet, BioEdit e uGene), which are able to read bam e bam.bai files:

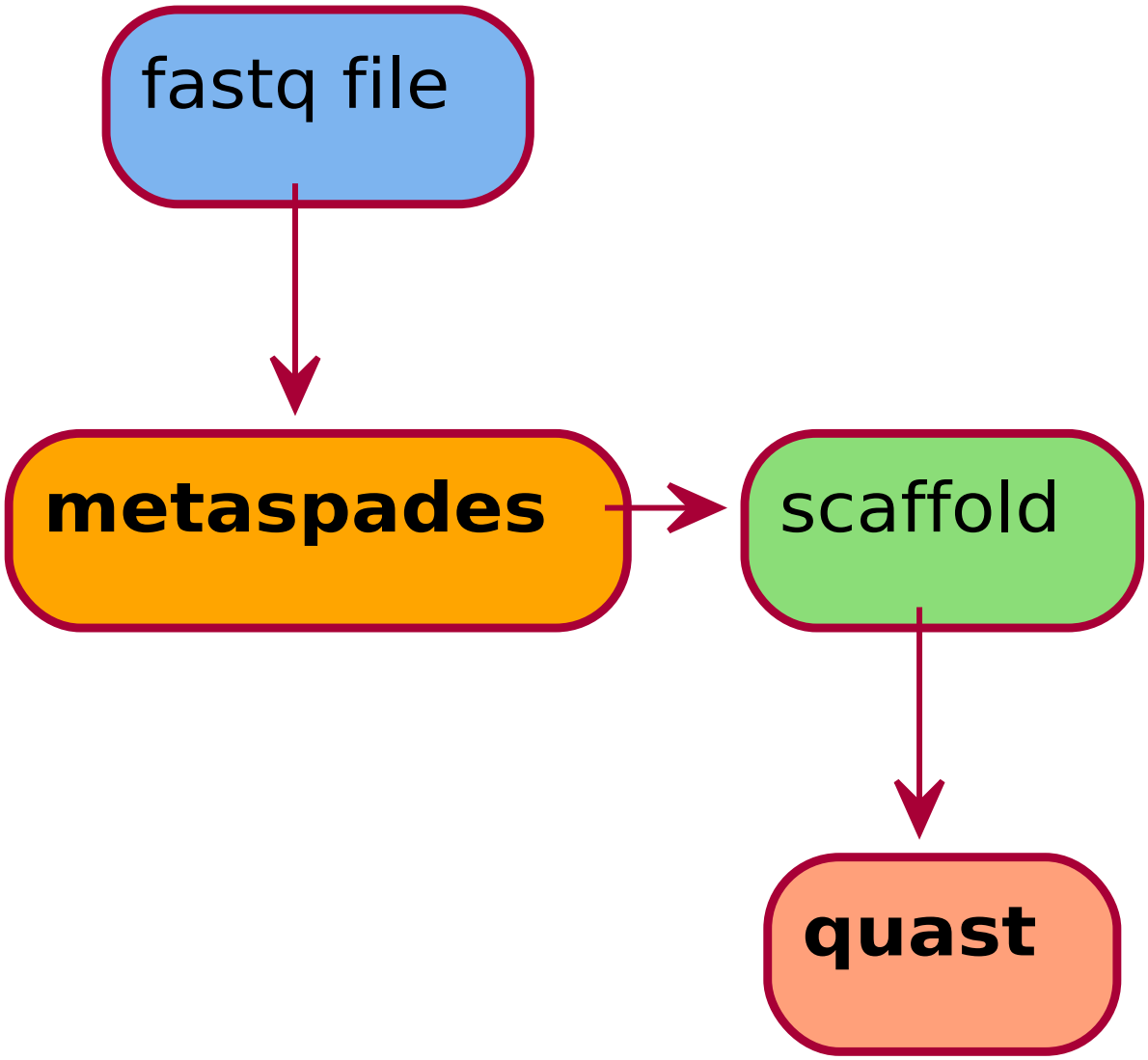
- Tablet (GNU/Linux, macOS, Windows): https://ics.hutton.ac.uk/tablet/download-tablet/;
- BioEdit (Windows): https://thalljiscience.github.io/;
- uGene (GNU/Linux, macOS, Windows): http://ugene.net/ugene/.

2 Metagenomics

2MG_denovo

Introduction

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



Tool
Input
Result
Quality Check tool

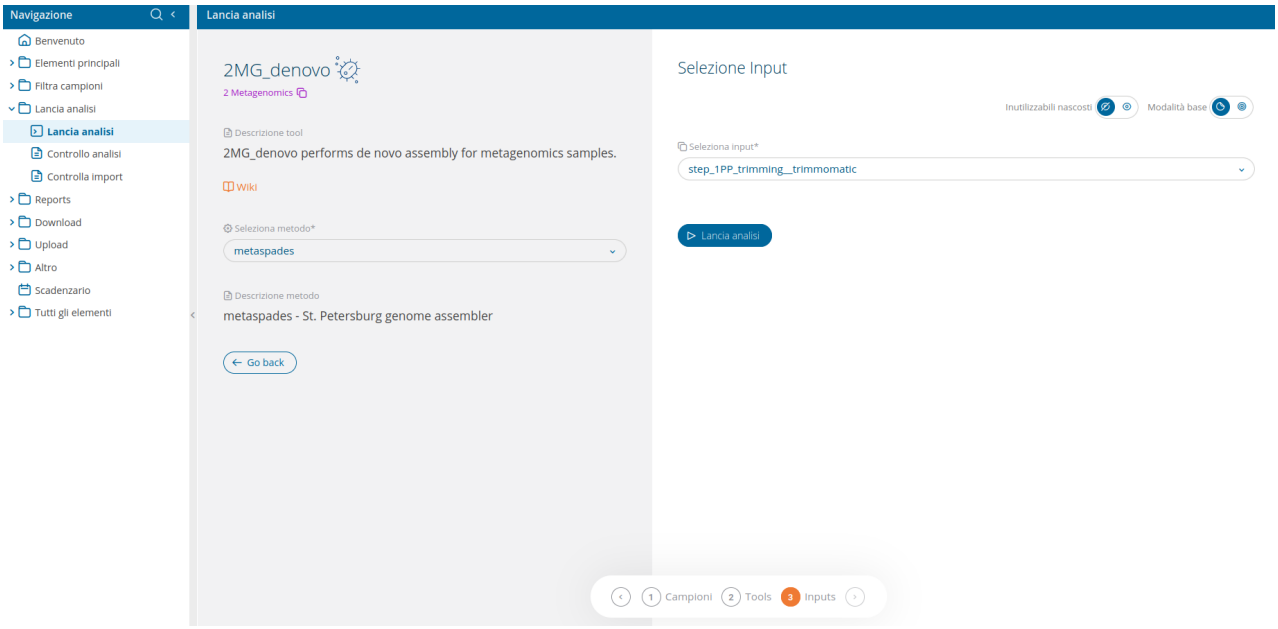
Run Analysis 2MG_denovo

Once the analysis **2MG_denovo**has been selected from the run analyses interface, the user will be able to select which bioinformatic tool to use. The tool used for this analysis is **metaspades** - *St. Petersburg genome assembler*.

-The input selection wizard delivers an advanced input selection mode, to allow selection of all types of supported input files at once.

Accepted inputs can be from:

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



A link to Check analysis will be created after launching the requested analysis. The system will notify the user after a succesful analysis launch and once execution has ended.

Output directory

> Please refer to Cohesive's specific Wiki page for information on file download.

The output directory is available at the link in the download page or at the link presente in the analysis' summary card, and will have the following structure: `results > YEAR > ID > 1PP_generated > DSXXXXXXXX-DTXXXXXX_metaspades`. At that path there will be 3 directories:

- **meta**: ("metadata") contains log and configuration files.
- **qc**: ("quality check") it contains 2 directories (meta and result). In this case quality check is performed with **quast**.
- **result**: contains the analysis' output files.

The table below lists files available produced by metaspades, alongside some useful information.

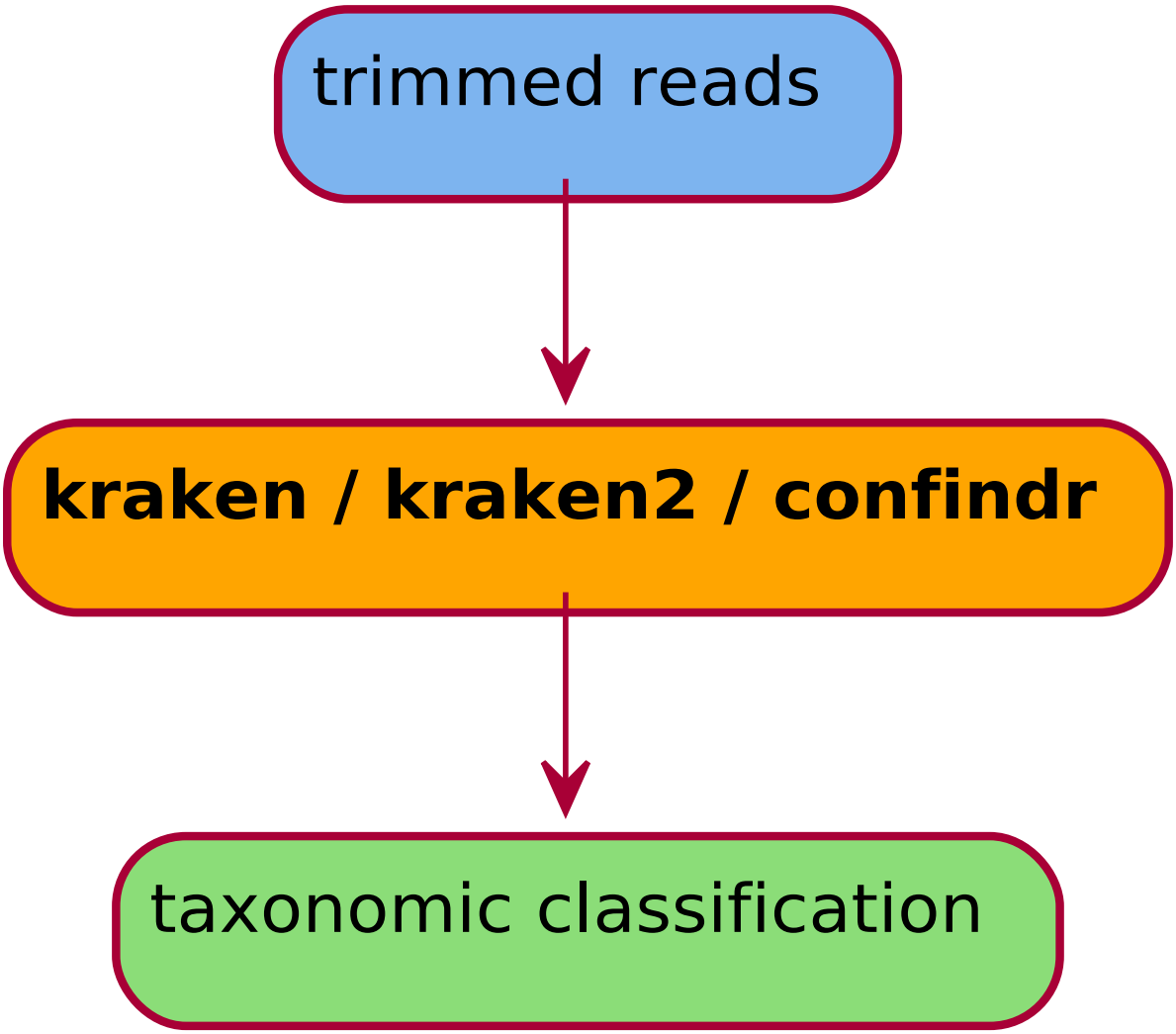
File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_spades_contigs.fasta	file with contigs of the *de novo* assembly	result directory
DSXXXXXXXX-DTXXXXXX_ID_spades_scaffolds.fasta	scaffolds	result directory
DSXXXXXXXX-DTXXXXXX_ID_spades_scaffolds_L200.fasta	file with scaffolds longer than 200bp	result directory
DSXXXXXXXX-DTXXXXXX_ID_quast.csv	file with assembly quality metrics	qc > result directory

3 Taxonomy

3TX_class

Introduction

3TX_class performs taxonomic classification of organisms and checks for contamination.



Tool
Input
Result

Run Analysis 3TX_class

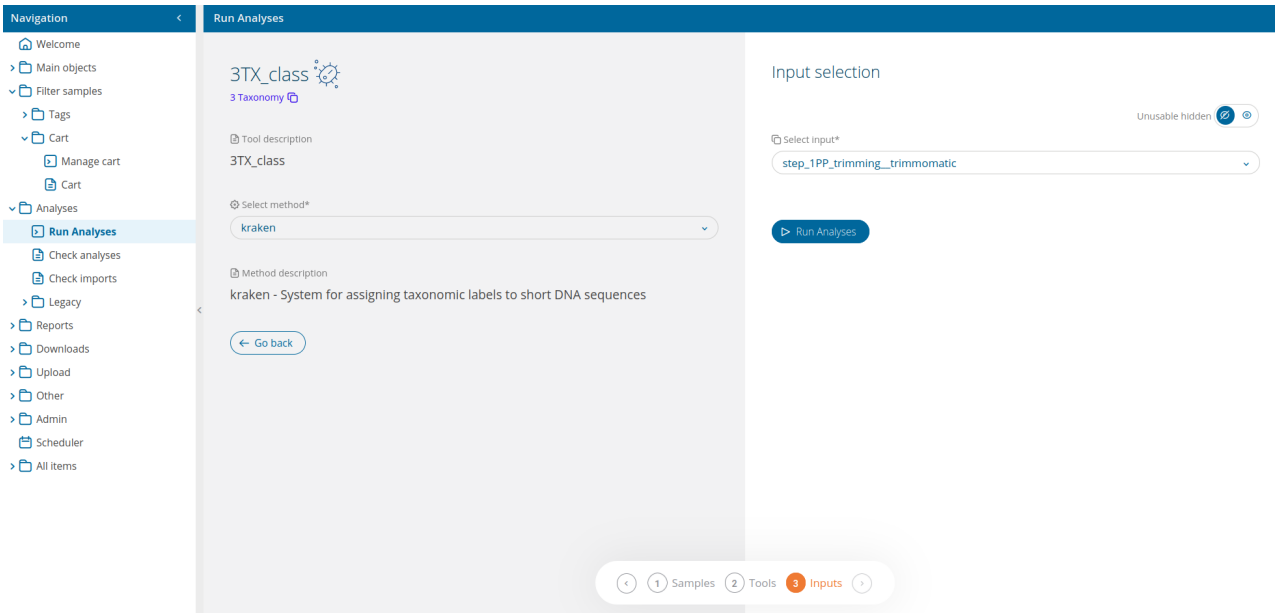
Once the analysis **3TX_class** has been selected from the run analyses interface, the user will be able to select which bioinformatic tool to use. The available tools are listed below.

- **kraken** - System for assigning taxonomic labels to short DNA sequences (version 1)

The input selection UI delivers an advanced input selection mode, to allow selection of all types of supported input files at once.

Accepted inputs can be from:

- step_1PP_trimming
- step_1PP_hostdepl



A link to Check analysis will be created after launching the requested analysis. The system will notify the user after a succesful analysis launch and once execution has ended.

Output directory

> Please refer to Cohesive's specific Wiki page for information on file download.

The output directory is available at the link in the download page or at the link presente in the analysis' summary card, and will have the following structure: `results > YEAR > ID > 3TX_class > DSXXXXXXXX-DTXXXXXX_kraken`. The last directory's suffix will be replaced with the name of the chosen tool. At that path there will be 2 directories:

- **meta**: ("metadata") contains log and configuration files.
- **result**: contains the analysis' output files.

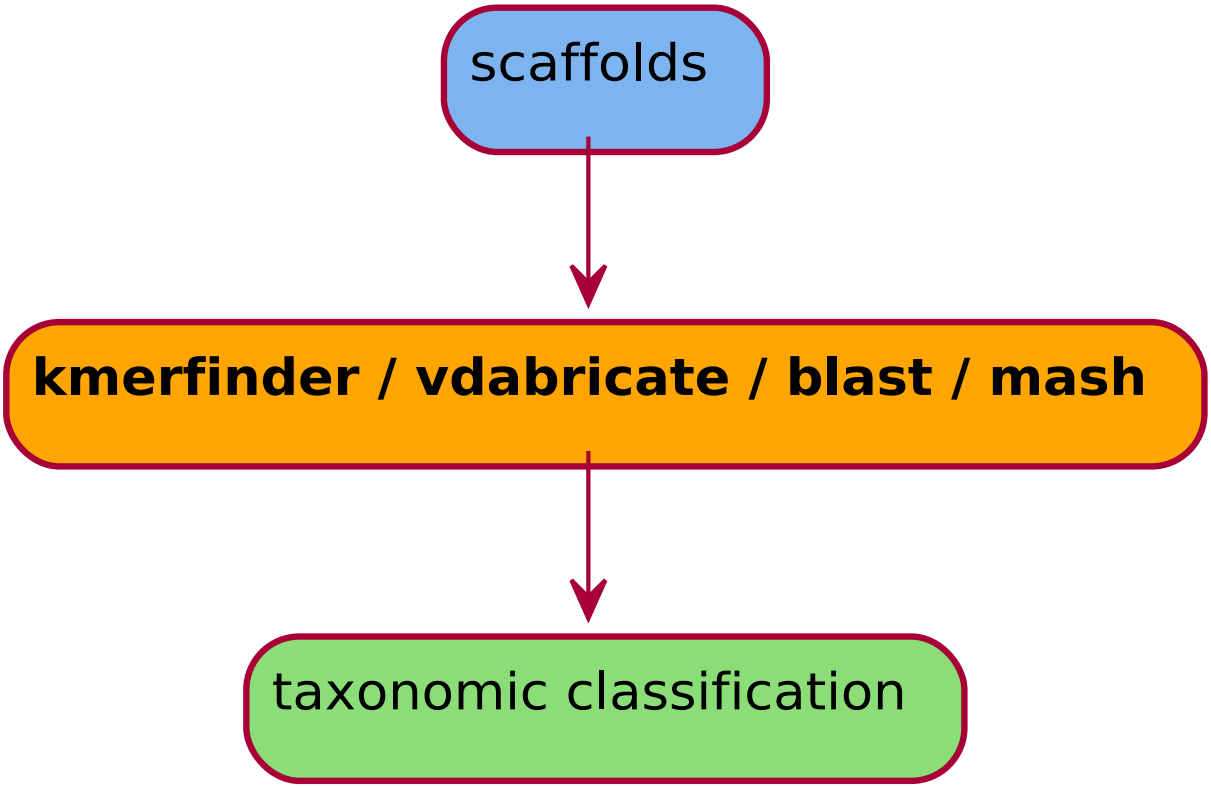
The tables below list files produced by available tools. For more details on Kraken and Kraken2's output files, please refer to kraken's official manual and Kraken2's help sheet.

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_kraken.tsv	kraken's taxonomic classification file	cartella result

3TX_species

Introduction

3TX_species assigns the closest bacterial or viral species to the input reads.



Tool
Input
Result

Run Analysis 3TX_species

Once the analysis **3TX_species** has been selected from the run analyses interface, the user will be able to select which bioinformatic tool to use. The available tool is:

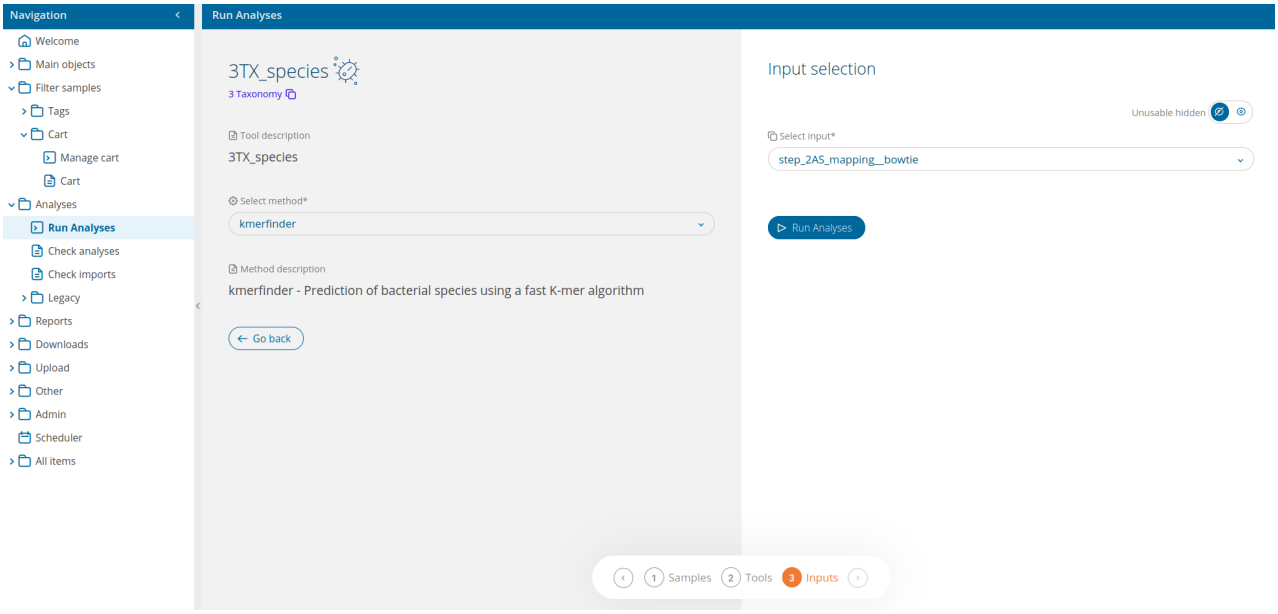
- **kmerfinder** - Prediction of bacterial species using a fast K-mer algorithm (**assegnazione di specie batteriche**)

3TX_species requires input sequences from *de novo* assembly or mapping; if the latter are provided, the reference genome that has been used for mapping will also be required.

The input selection UI delivers an advanced input selection mode, to allow selection of all types of supported input files at once.

Accepted inputs can be from:

- step_2AS_mapping
- step_2AS_denovo



A link to Check analysis will be created after launching the requested analysis. The system will notify the user after a succesful analysis launch and once execution has ended.

Output directory

> Please refer to Cohesive's specific Wiki page for information on file download.

The output directory is available at the link in the download page or at the link presente in the analysis' summary card, and will have the following structure: `results > YEAR > ID > 3TX_species > DSXXXXXXXX-DTXXXXXX_kmerfinder`. The last directory's suffix will be replaced with the name of the chosen tool. At that path there will be 2 directories:

- **meta**: ("metadata") contains log and configuration files.
- **result**: contains the analysis' output files.

The table below list files produced by tools available for 3TX_species.

kmerfinder

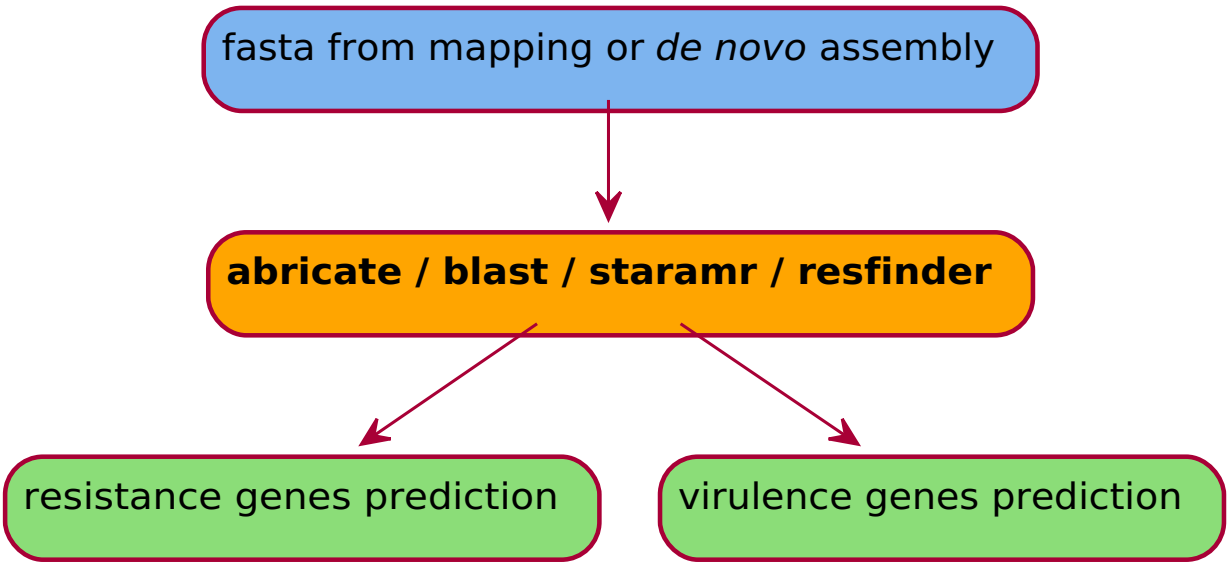
File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_kmerfinder_bacterial.tsv	file with the assigned batterial species	results directory
DSXXXXXXXX-DTXXXXXX_ID_kmerfinder_viral.tsv	file with the assigned viral species	results directory

4 Genome annotation

4AN_AMR

Introduction

4AN_AMR (Anti-Microbial Resistance) performs prediction of antibiotic resistace- and virulence-associated gene presence in the reconstructed sequence of the microorganism of interest.



Tool
Input
Result

Run Analysis 4AN_AMR

Once the analysis **4AN_AMR** has been selected from the run analyses interface, the user will be able to select which bioinformatic tool to use. The available tools are listed below.

For all microorganisms:

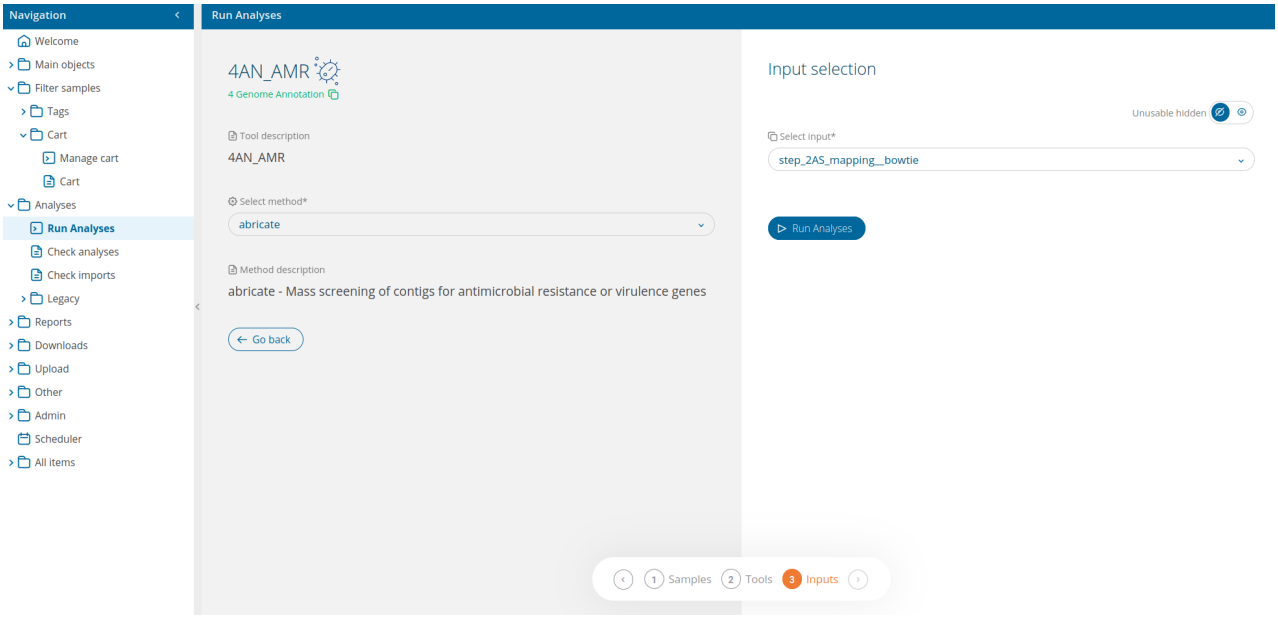
- **abricate** - Mass screening of contigs for antimicrobial resistance or virulence genes

4AN_AMR requires input sequences from *de novo* assembly or mapping; if the latter are provided, the reference genome that has been used for mapping will also be required.

The input selection UI delivers an advanced input selection mode, to allow selection of all types of supported input files at once.

Accepted inputs can be from:

- step_2AS_mapping
- step_2AS_denovo



A link to Check analysis will be created after launching the requested analysis. The system will notify the user after a succesful analysis launch and once execution has ended.

Output directory

> Please refer to Cohesive's specific Wiki page for information on file download.

The output directory is available at the link in the download page or at the link presente in the analysis' summary card, and will have the following structure: `results > YEAR > ID > 4AN_AMR > DSXXXXXXXX-DTXXXXXX_abricate`. The last directory's suffix will be replaced with the name of the chosen tool. At that path there will be 2 directories:

- **meta**: ("metadata") contains log and configuration files.
- **result**: contains the analysis' output files.

The table below list files produced by tools available for 3TX_species.

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

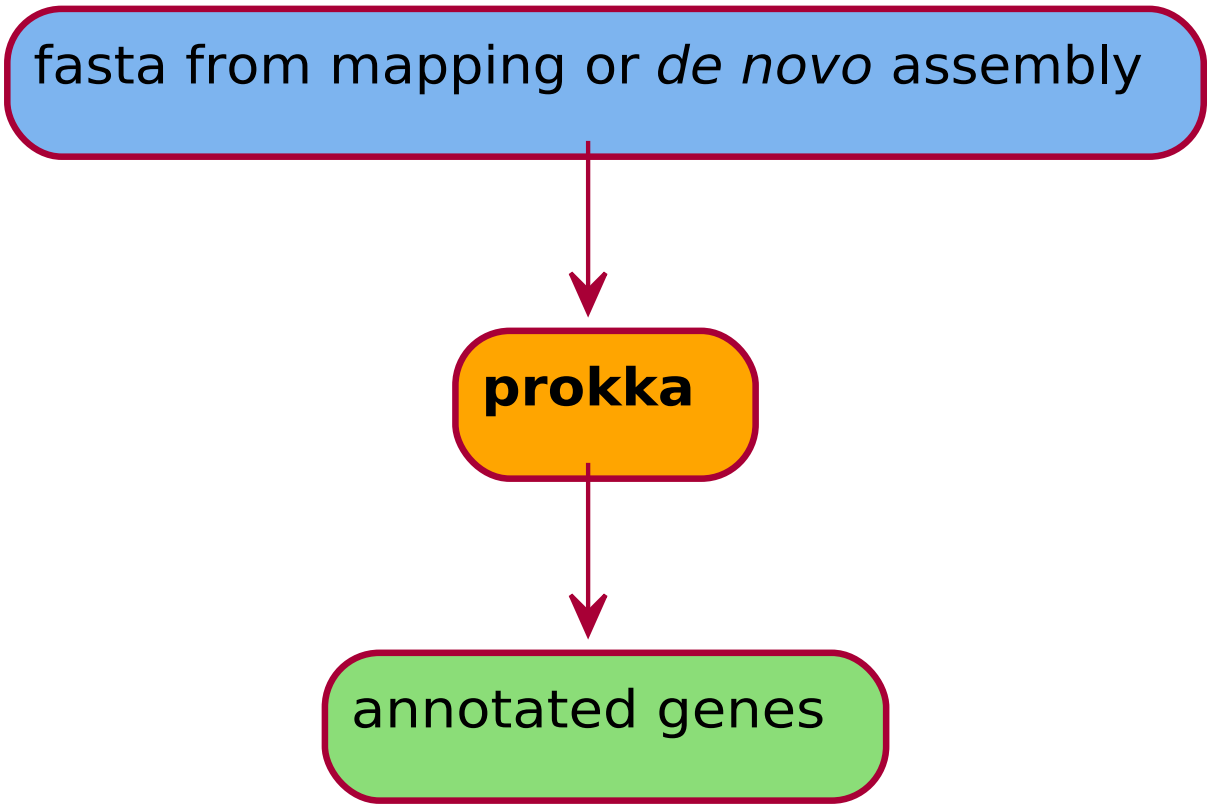
> In the .summary file, for each sample there will also be a number listed in the genes column, which is the value of the horizontal coverage of the associated gene. In case of fragmented genes, multiple semicolon-separated values will be listed, each associated to one of the contigs from *de novo* assembly that matched the gene.

For more details on abricate output files, please refer to abricate's official manual.

4AN_genes

Introduction

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



Tool
Input
Result

Run Analysis 4AN_genes

Once the analysis **4AN_genes** has been selected from the run analyses interface, the user will be able to select which bioinformatic tool to use. The only available tool for this analysis is **Prokka** - *Tool to annotate bacterial, archaeal and viral genomes*.

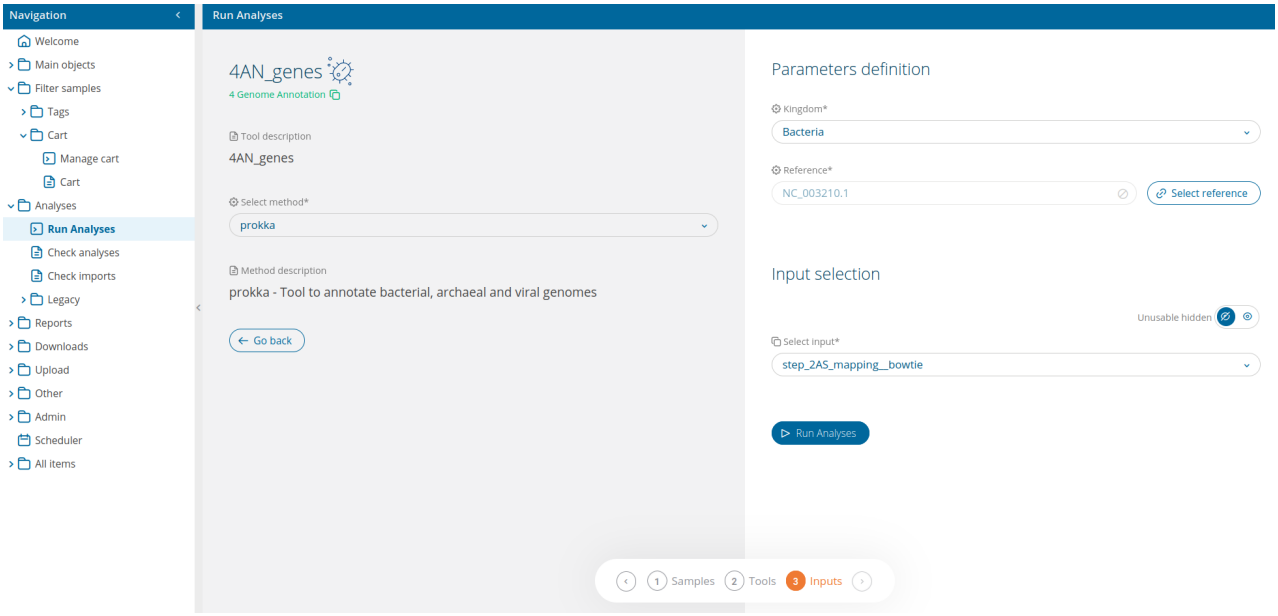
The input selection UI delivers an advanced input selection mode, to allow selection of all types of supported input files at once.

The first required parameter is the kingdom (*i.e.* virus or bacteria, plus "host", an artificial group which includes possible host organisms). The second parameter is a reference genome.

Accepted inputs can be from:

- step_2AS_mapping
- step_2AS_denovo

If output from mapping is provided, the reference genome that has been used for mapping will also be required.



A link to Check analysis will be created after launching the requested analysis. The system will notify the user after a succesful analysis launch and once execution has ended.

Output directory

> Please refer to Cohesive's specific Wiki page for information on file download.

The output directory is available at the link in the download page or at the link presente in the analysis' summary card, and will have the following structure: `results > YEAR > ID > 4AN_genes > DSXXXXXXXX-DTXXXXXX_prokka`. At that path there will be 2 directories:

- **meta**: ("metadata") contains log and configuration files.
- **result**: contains the analysis' output files.

The following table lists Prokka's output files.

File	Description	Location
log_errore_controlli_esami.log	run's warning and error log	main directory
metadata_samples.tsv	samples' metadata summary table in tsv format	main directory
results.csv	summary table separated by semicolon (";") containing sample IDs and information	main directory
DSXXXXXXXX-DTXXXXXX_ID_prokka_REFID_result.err	text report file with run's errors	results directory
DSXXXXXXXX-DTXXXXXX_ID_prokka_REFID_result.faa	amminoacidic sequences from translation of identified coding genes (faa format - fasta aminoacid)	results directory
DSXXXXXXXX-DTXXXXXX_ID_prokka_REFID_result.ffn	nucleotidic sequences of identified coding genes (fnn format - fasta nucleotide)	results directory
DSXXXXXXXX-DTXXXXXX_ID_prokka_REFID_result.fna	nucleotidic sequences of identified coding genes (fna format)	results directory
DSXXXXXXXX-DTXXXXXX_ID_prokka_REFID_result.fsa	sequences in fsa format (fragment analysis data file)	results directory
DSXXXXXXXX-DTXXXXXX_ID_prokka_REFID_result.gbk	output file in GenBank format	results directory
DSXXXXXXXX-DTXXXXXX_ID_prokka_REFID_result.gff	output file in gff format (General Feature Format)	results directory
DSXXXXXXXX-DTXXXXXX_ID_prokka_REFID_result.log	Prokka's run log	results directory
DSXXXXXXXX-DTXXXXXX_ID_prokka_REFID_result.sqn	file per la sottomissione a GenBank in formato Sequin	results directory
DSXXXXXXXX-DTXXXXXX_ID_prokka_REFID_result.tbl	text file with information on sequence and *loci*	results directory
DSXXXXXXXX-DTXXXXXX_ID_prokka_REFID_result.tsv	tsv list of *loci* and proteins from mapped coding genes	results directory
DSXXXXXXXX-DTXXXXXX_ID_prokka_REFID_result.txt	metrics on identified CDS	results directory
proteins.faa	protein sequnces in faa format	results directory

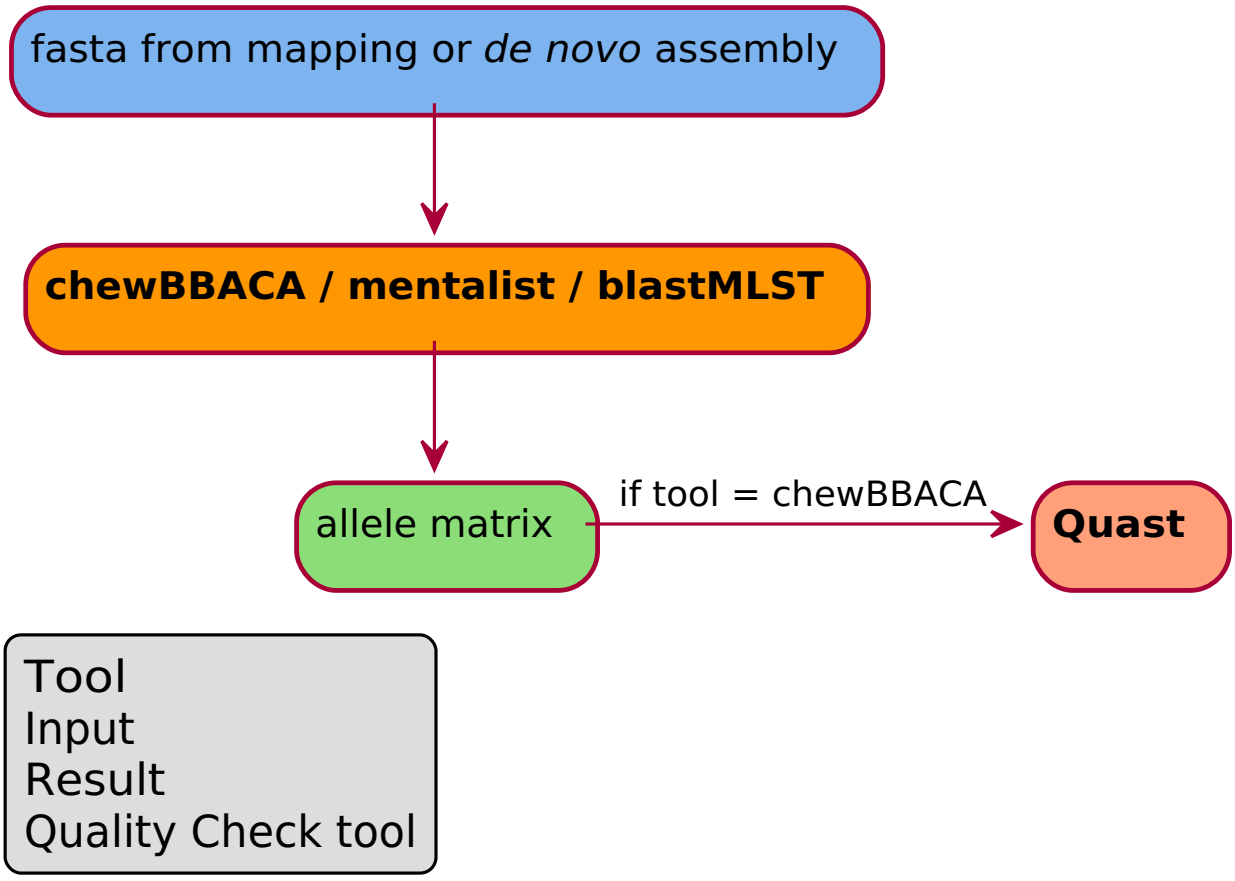
For more details on Prokka's output files, please refer to Prokka's official manual.

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 Analysis 4TY_cgMLST

Once the analysis **4TY_cgMLST** has been selected from the run analyses interface, the user will be able to select which bioinformatic tool to use. The available tool is:

- **chewBBACA** - BSR-Based Allele Calling Algorithm

Tools for cgMLST possess schemas for specific bacteria. Schemas available for chewBBACA are listed in the table below.

Tool	Available schemas
chewBBACA	*Listeria monocytogenes*, *Campylobacter jejuni*, *Campylobacter coli*, *Staphylococcus aureus*, *Brucella*, *Brucella melitensis*, *Klebsiella pneumoniae*.

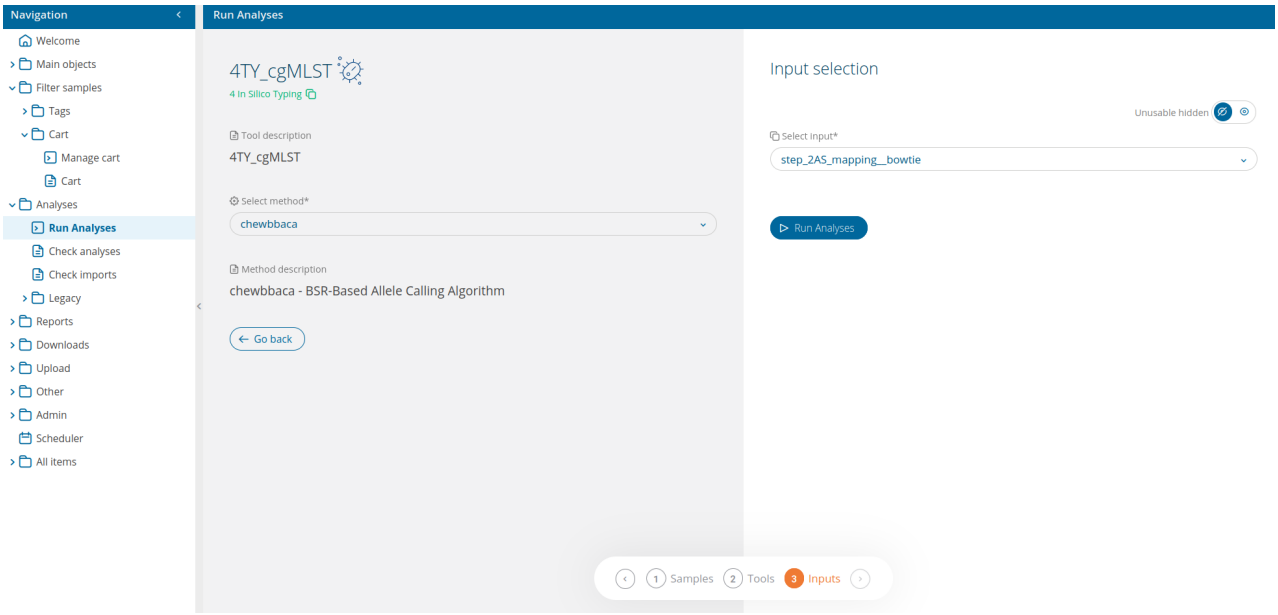
Note 1: Running 4TY_cgMLST on a microorganism, for which there is no corresponding cgMLST schema will cause the run to fail.

The input selection UI delivers an advanced input selection mode, to allow selection of all types of supported input files at once.

Accepted inputs can be from:

- step_2AS_mapping
- step_2AS_denovo

4TY_cgMLST requires input sequences from *de novo* assembly or mapping; if the latter are provided, the reference genome that has been used for mapping will also be required.



A link to Check analysis will be created after launching the requested analysis. The system will notify the user after a succesful analysis launch and once execution has ended.

Output directory

Please refer to Cohesive's specific Wiki page for information on file download.

The output directory is guida ufficiale di available at the link in the download page or at the link presente in the analysis' summary card, and will have the following structure: `results > YEAR > ID > 4TY_cgMLST > DSXXXXXXXX-DTXXXXXX_chewbbaca`. At that path there will be 3 directories:

- **meta**: ("metadata") contains log and configuration files.
- **result**: contains the analysis' output files.
- **qc**: ("quality check") it contains 2 directories (meta and result). In this case quality check is performed with **Quast**.

Output files from allele call with chewBBACA are available with 3 different encoding:

- **IZS encoding:** each allele is identified with a progressive numeric ID. ID assignment considers all *loci*, thus it **DOES NOT** restart from 1 at each new *locus*.
- **Pasteur encoding:** the code for identified alleles consists of a numeric value. Progression restarts from 1 at each *locus*. For each execution the analysis restarts from the unmodified, downloaded database. Used schema displays download date.
- **MD5 encoding:** each allele is identified with an alphanumeric code of 16 characters (MD5 code), obtained through a "hash" applied to the allele's sequence.

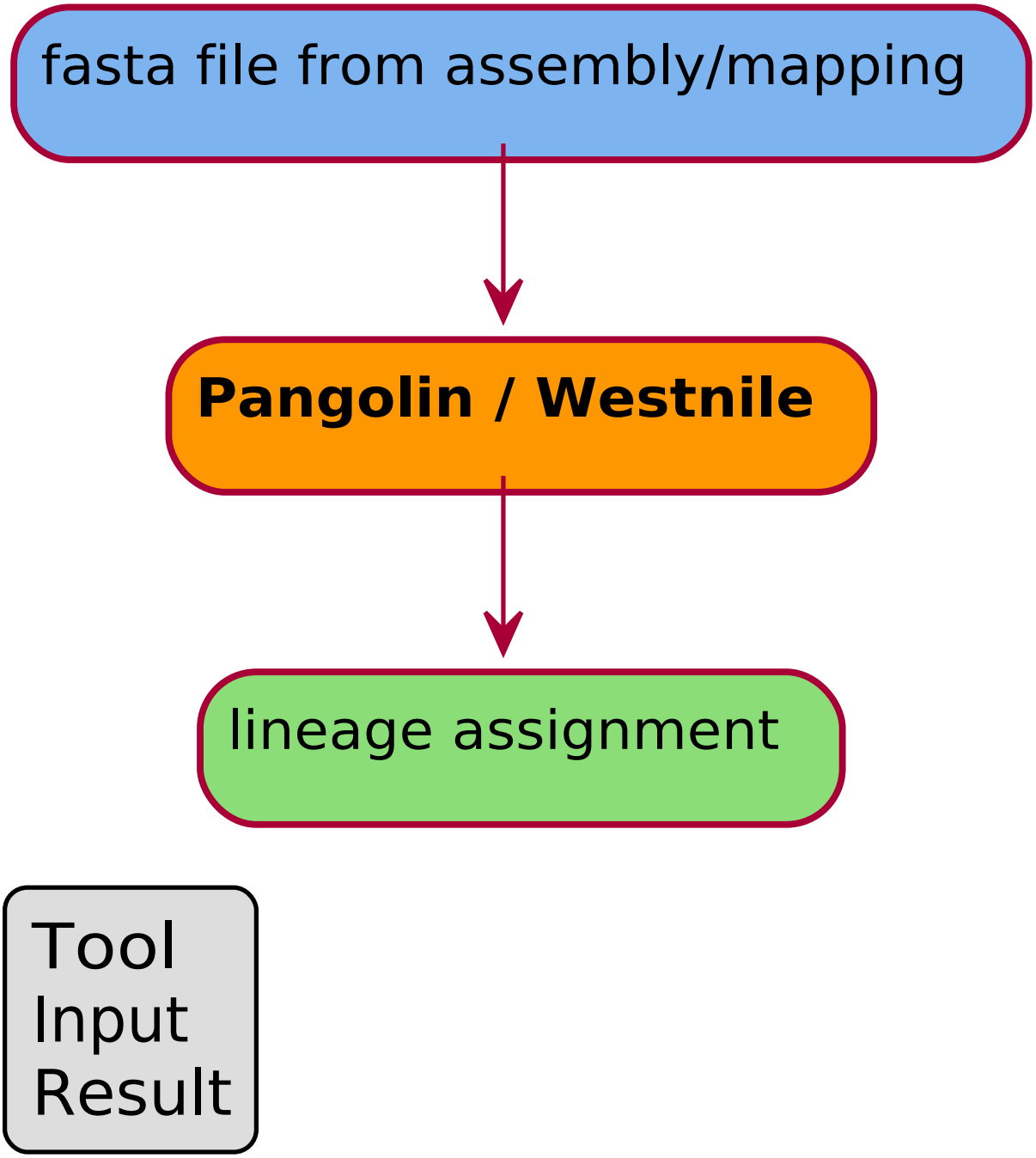
File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_chewbbaca_new_alleles.txt	sequences of newly-identified alleles	result directory
DSXXXXXXXX-DTXXXXXX_ID_chewbbaca_results_alleles.tsv	allele call with Pasteur encoding in csv format	result directory
DSXXXXXXXX-DTXXXXXX_ID_chewbbaca_results_contigsInfo.tsv	info about the contig mapped on each *locus*	result directory
DSXXXXXXXX-DTXXXXXX_ID_chewbbaca_results_izsam.csv	allele call with IZS encoding	result directory
DSXXXXXXXX-DTXXXXXX_ID_chewbbaca_results_md5.csv	allele call with md5 encoding	result directory
DSXXXXXXXX-DTXXXXXX_ID_chewbbaca_results_pasteur_2021-05-28.csv	allele call with Pasteur encoding in tsv format	result directory
DSXXXXXXXX-DTXXXXXX_ID_chewbbaca_results_statistics.tsv	metrics on *loci* encoded as EXC, INF, LNF, PLOT, NIPH, ALM, ASM	result directory
DSXXXXXXXX-ID_import_chewbbaca_check.csv	quality check with info on calledPerc, calledNum, annotated, new, notFound, discarded	qc > result directory

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

4TY_lineage

Introduction

4TY_lineage assigns the lineage of SARS-CoV2 and West Nile Virus. This analysis is specific for such virus.



Run Analysis 4TY_lineage

Once the analysis **4TY_lineage** has been selected from the run analyses interface, the user will be able to select which bioinformatic tool to use. The available tools are:

- **Pangolin** - Phylogenetic Assignment of Named Global Outbreak LINeages
- **Westnile** - West Nile virus lineage calculation

> Pangolin is exclusively for SARS-CoV2 lineage assignment, while Westnile is specific for West Nile Virus.

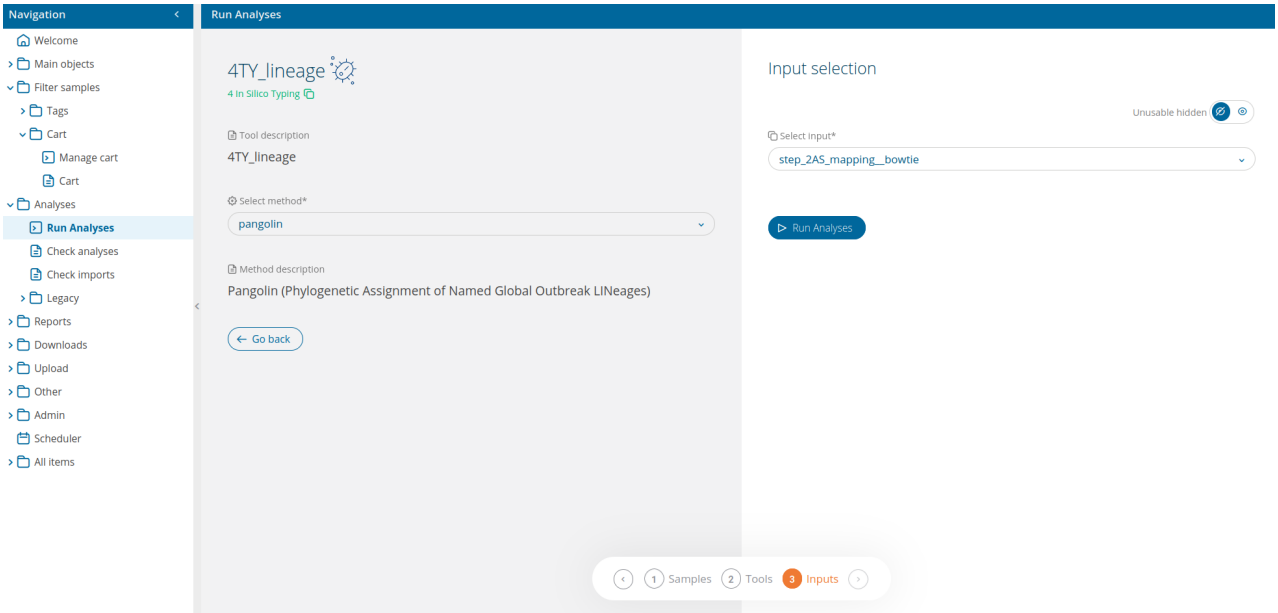
The input selection UI delivers an advanced input selection mode, to allow selection of all types of supported input files at once.

Accepted inputs can be from:

- step_2AS_mapping
- step_2AS_denovo
- step_1PP_trimming

> The tool **Westnile** only accepts fasta file from 2AS_mapping as input.

4TY_lineage requires input sequences from *de novo* assembly or mapping; if the latter are provided, the reference genome that has been used for mapping will also be required.



A link to Check analysis will be created after launching the requested analysis. The system will notify the user after a succesful analysis launch and once execution has ended.

Output directory

> Please refer to Cohesive's specific Wiki page for information on file download.

The output directory is guida ufficiale di available at the link in the download page or at the link presente in the analysis' summary card, and will have the following structure: `results > YEAR > ID > 4TY_lineage > DSXXXXXXXX-DTXXXXXX_pangolin`. At that path there will be 2 directories:

- **meta**: ("metadata") contains log and configuration files.
- **result**: contains the analysis' output files.

Pangolin

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_pangolin_lineage_report.csv	csv file with assigned lineage	result directory

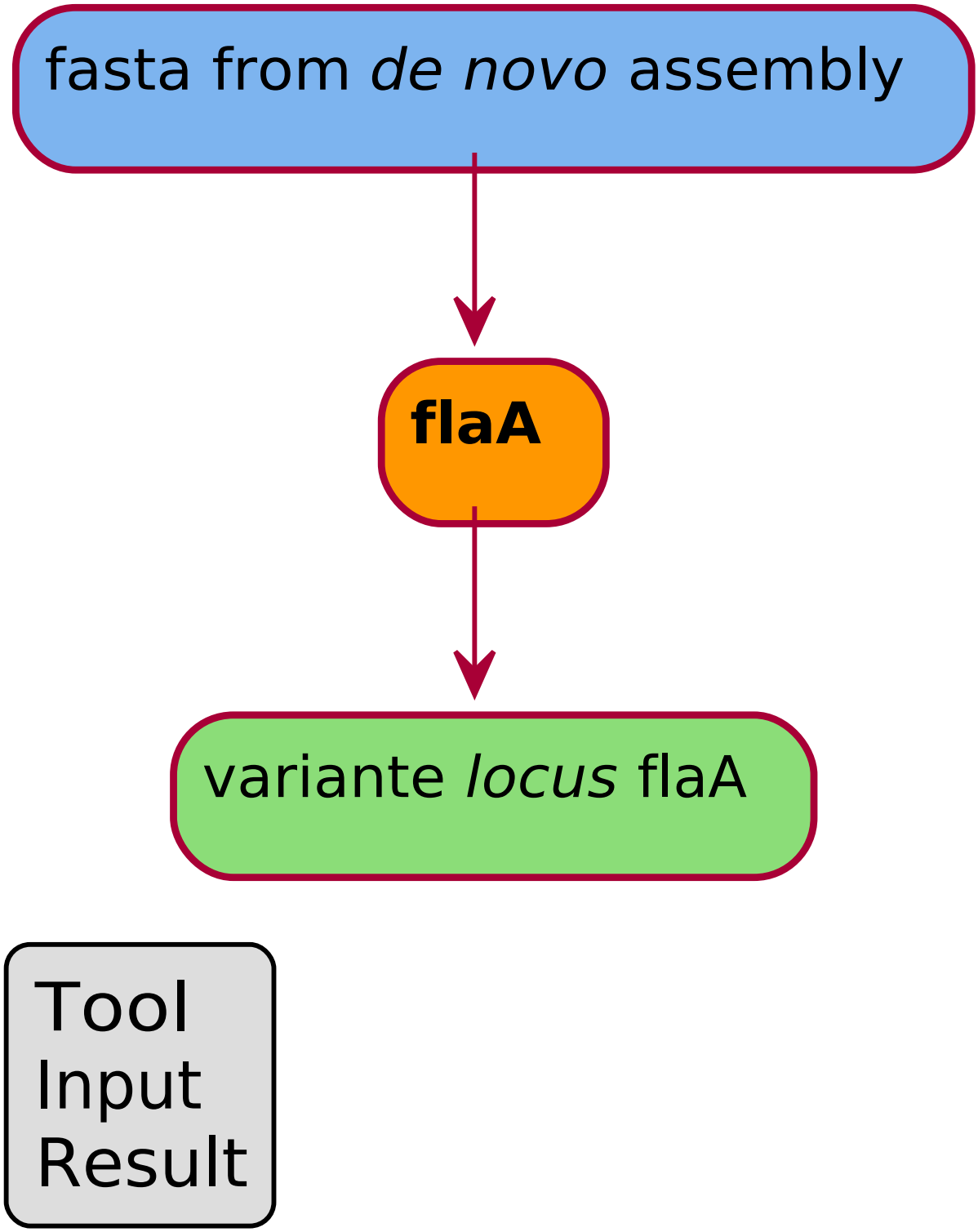
Westnile

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_westnile_lineage.csv	csv file with assigned lineage	result directory
DSXXXXXXXX-DTXXXXXX_ID_westnile_lineage_summary.csv	csv summary table	result directory

4TY_flaA

Introduction

The 4TY_flaA analysis infers *in silico* the flaA *locus* variant for *Campylobacter*.



Run Analysis 4TY_flaA

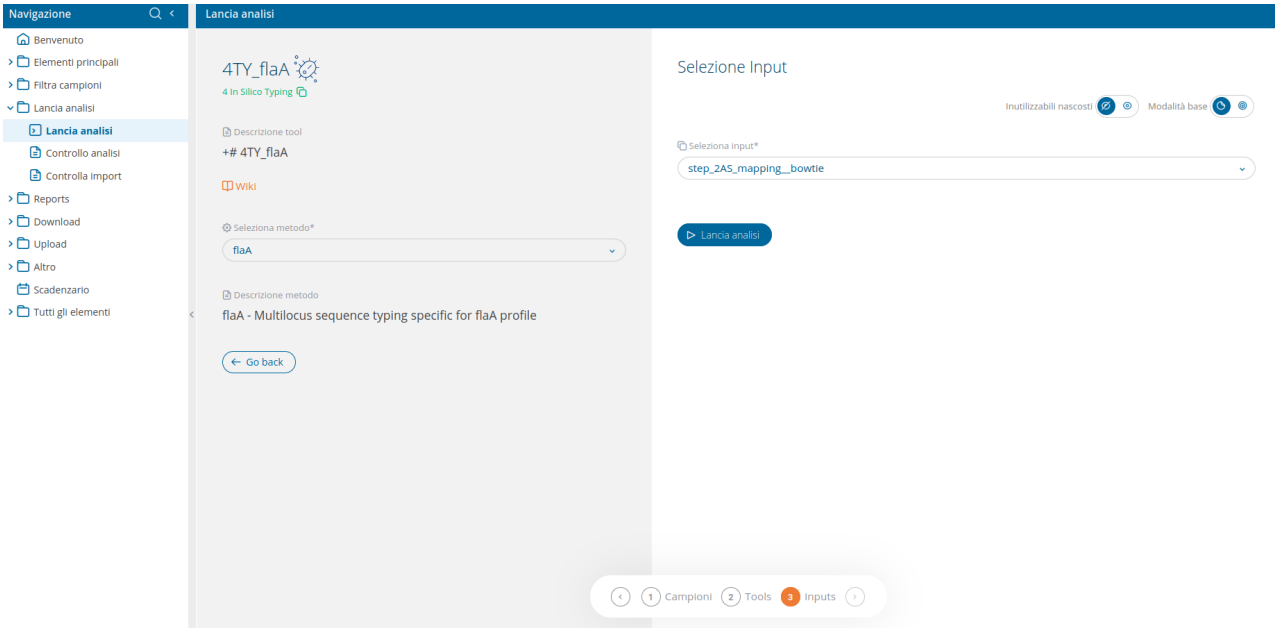
Once the analysis **4TY_flaA** has been selected from the run analyses interface, the user will be able to select which bioinformatic tool to use. The available software for this analysis is flaA - Multilocus sequence typing specific for flaA profile.

The input selection UI delivers an advanced input selection mode, to allow selection of all types of supported input files at once.

Accepted inputs can be from:

- step_2MG_denovo
- step_2AS_denovo
- step_2AS_mapping
- step_2AS_import

The software requires the input sequences from *de novo* assembly or from mapping and, in case the latter is selected, the reference genome used for mapping.



A link to Check analysis will be created after launching the requested analysis. The system will notify the user after a succesful analysis launch and once execution has ended.

Output directory

Please refer to Cohesive's specific Wiki page for information on file download.

The output directory is guida ufficiale di available at the link in the download page or at the link presente in the analysis' summary card, and will have the following structure: `results > YEAR > ID > 4TY_flaA > DSXXXXXXXX-DTXXXXXX_flaA`. At that path there will be 2 directories:

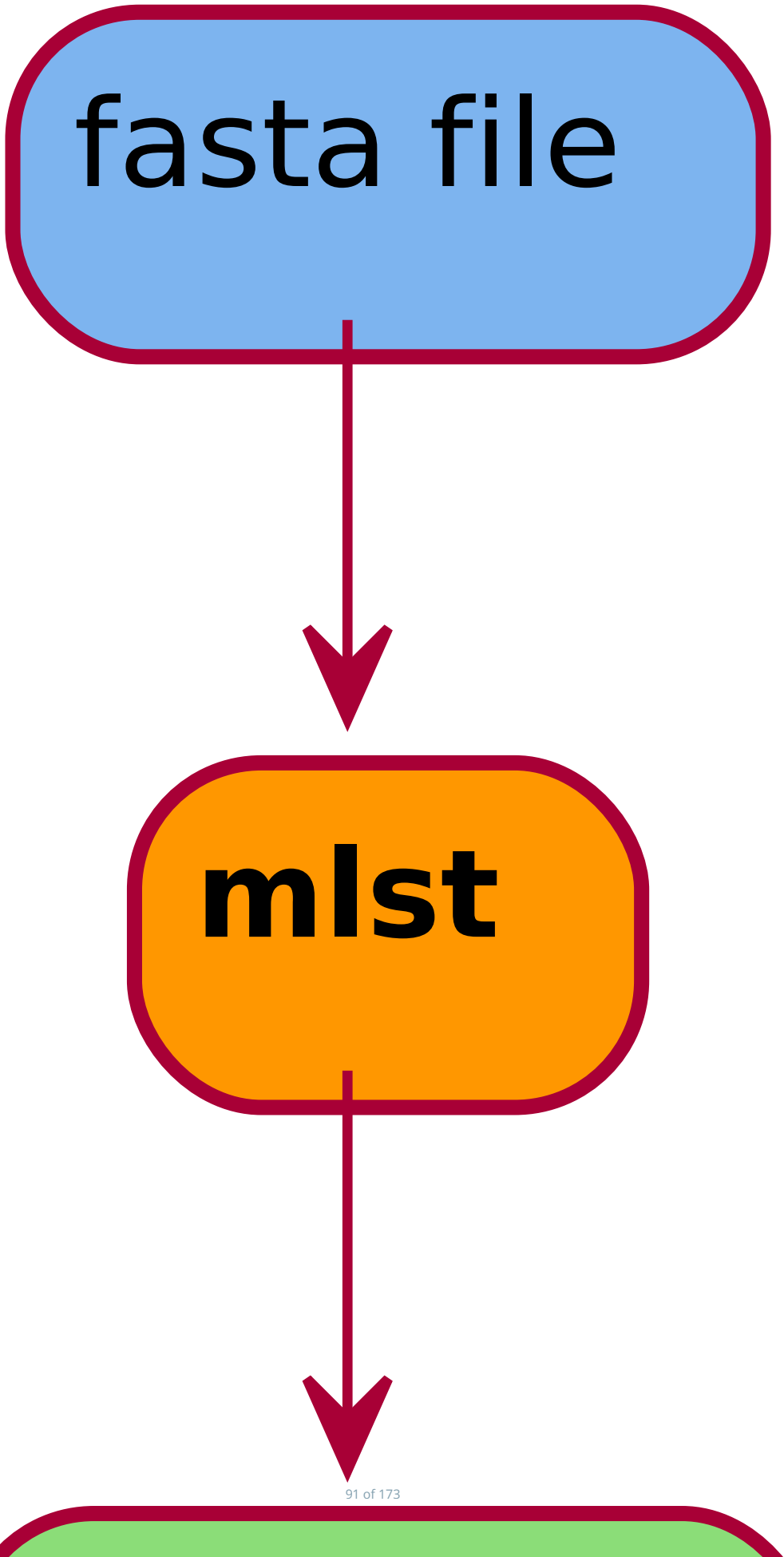
- **meta**: ("metadata") contains log and configuration files.
- **result**: contains the analysis' output files.

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_flaA.tsv	allele of the flaA *locus*	result directory

4TY_MLST

Introduction

4TY_MLST performs the *in silico* Multi Locus Sequence Typing, which assigns Sequence Type (ST) and Clonal Complex (CC), where applicable.


fasta file
mlst

Run Analysis 4TY_MLST

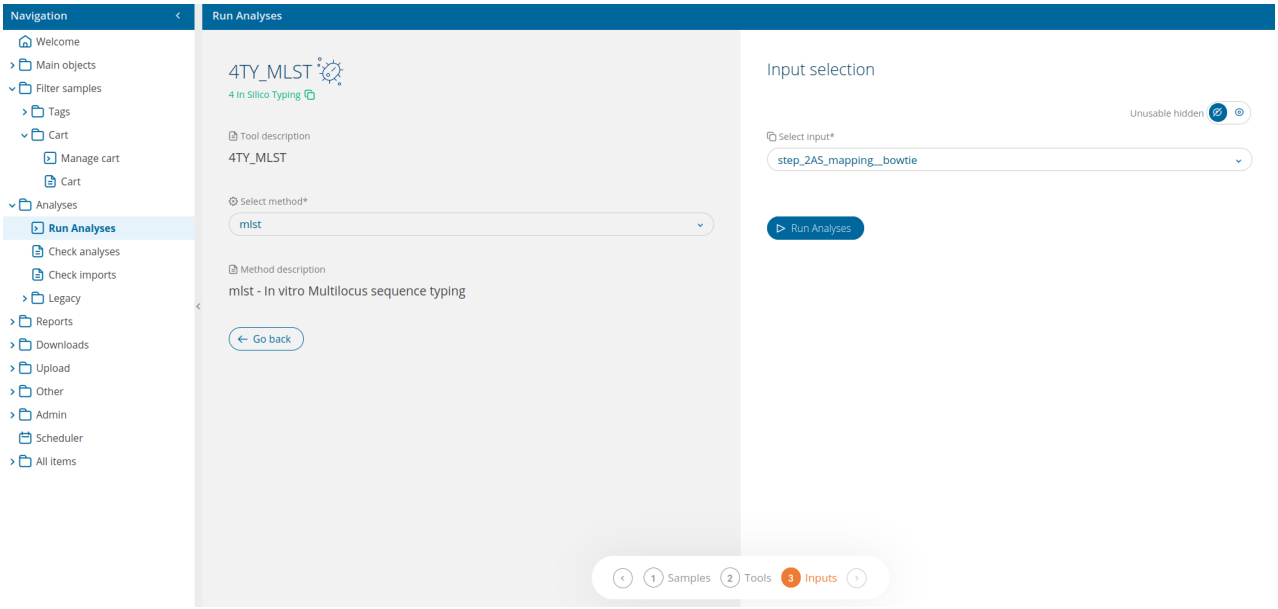
Once the analysis **4TY_MLST** has been selected from the run analyses interface, the user will be able to select which bioinformatic tool to use. The tool available for this analysis is **mlst** - *In vitro* Multi-Locus Sequence Typing.

The input selection UI delivers an advanced input selection mode, to allow selection of all types of supported input files at once.

Accepted inputs can be from:

- step_2AS_mapping
- step_2AS_denovo

4TY_MLST requires input sequences from *de novo* assembly or mapping; if the latter are provided, the reference genome that has been used for mapping will also be required.



A link to Check analysis will be created after launching the requested analysis. The system will notify the user after a succesful analysis launch and once execution has ended.

Output directory

> Please refer to Cohesive's specific Wiki page for information on file download.

The output directory is guida ufficiale di available at the link in the download page or at the link presente in the analysis' summary card, and will have the following structure: `results > YEAR > ID > 4TY_MLST > DSXXXXXXXX-DTXXXXXX_mlst`. At that path there will be 2 directories:

- **meta**: ("metadata") contains log and configuration files.
- **result**: contains the analysis' output files.

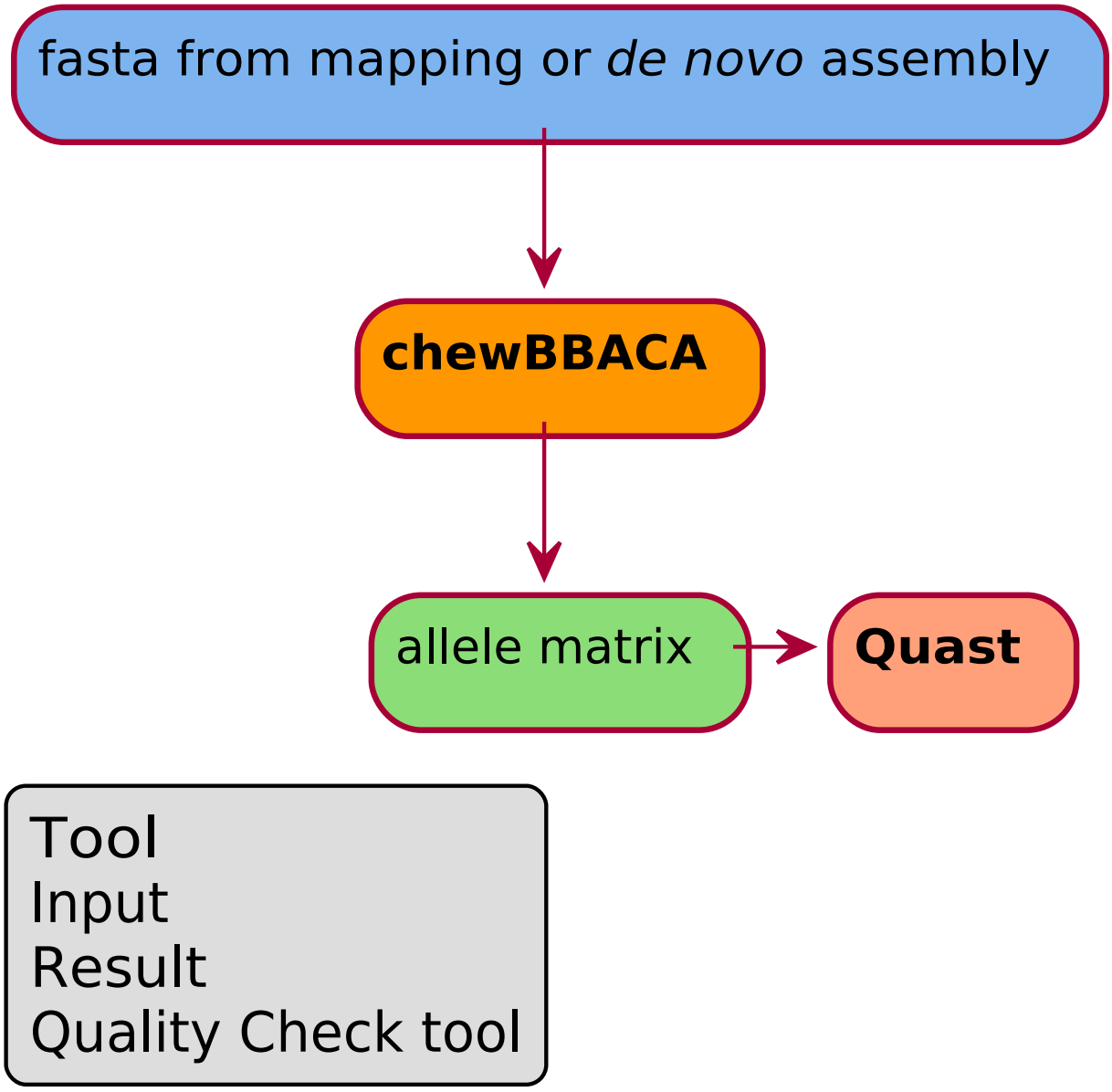
The table below lists output files and some useful information.

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_mlst.tsv	results of *locus* calls for MLST	result directory
DSXXXXXXXX-DTXXXXXX_ID_mlst_cc.csv	results with assigned CC	result directory

4TY_wgMLST

Introduction

4TY_wgMLST performs the "whole genome Multi-Locus Sequence Typing" (wgMLST) *in silico* for the *Campylobacter jejuni* and *Campylobacter coli* bacteria. Differently than cgMLST (core genome MLST), characterization takes into account the *whole genome*.



Run Analysis 4TY_wgMLST

Once the analysis **4TY_wgMLST** has been selected from the run analyses interface, the user will be able to select which bioinformatic tool to use. The available tool is **chewBBACA** - BSR-Based Allele Calling Algorithm.

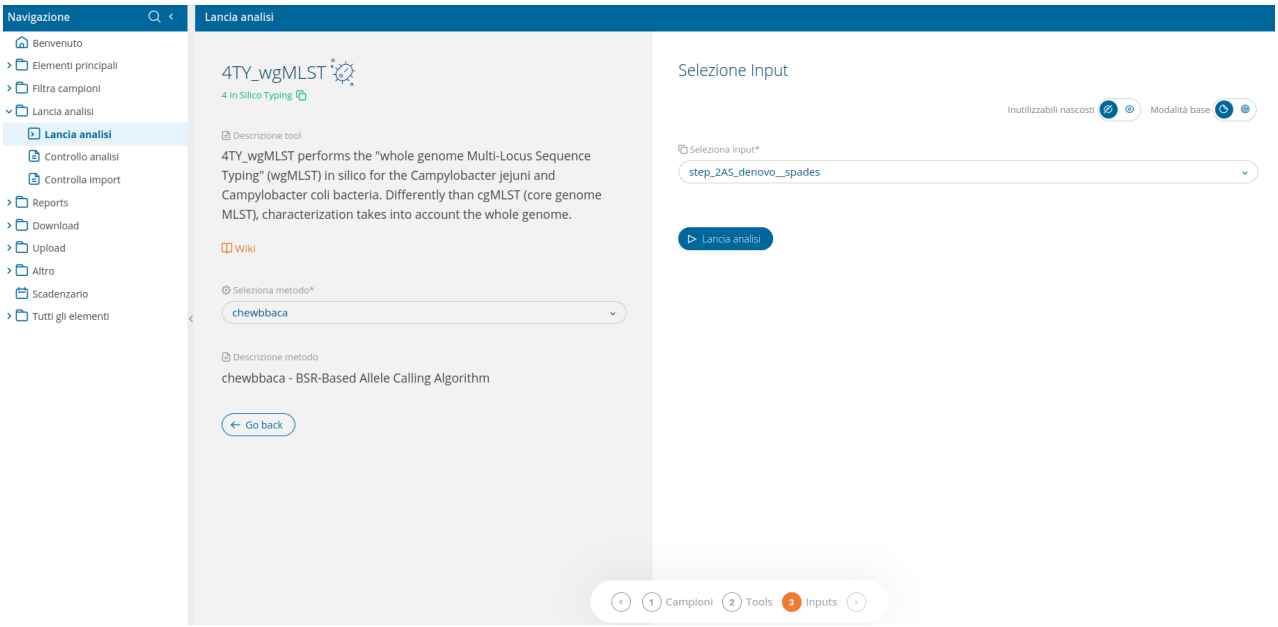
Note: Only *Campylobacter jejuni* and *Campylobacter coli* wgMLST schemas are available. Running the analysis on a different microorganism wiil cause the run to fail.

The input selection UI delivers an advanced input selection mode, to allow selection of all types of supported input files at once.

Accepted inputs can be from:

- step_2MG_denovo
- step_2AS_denovo
- step_2AS_mapping
- step_2AS_import

The software requires the input sequences from *de novo* assembly or from mapping and, in case the latter is selected, the reference genome used for mapping.



A link to Check analysis will be created after launching the requested analysis. The system will notify the user after a succesful analysis launch and once execution has ended.

Output directory

Please refer to Cohesive's specific Wiki page for information on file download.

The output directory is guida ufficiale di available at the link in the download page or at the link presente in the analysis' summary card, and will have the following structure: `results > YEAR > ID > 4TY_wgMLST > DSXXXXXXXX-DTXXXXXX_chewbbaca`. At that path there will be 3 directories:

- **meta**: ("metadata") contains log and configuration files.
- **result**: contains the analysis' output files.
- **qc**: ("quality check") it contains 2 directories (meta and result). In this case quality check is performed with **Quast**.

Output files from allele call with chewBBACA are available with 3 different encoding:

- **IZS encoding:** each allele is identified with a progressive numeric ID. ID assignment considers all *loci*, thus it **DOES NOT** restart from 1 at each new *locus*.
- **Pasteur encoding:** the code for identified alleles consists of a numeric value. Progression restarts from 1 at each *locus*. For each execution the analysis restarts from the unmodified, downloaded database. Used schema displays download date.
- **MD5 encoding:** each allele is identified with an alphanumeric code of 16 characters (MD5 code), obtained through a "hash" applied to the allele's sequence.

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

File	Description	Location
DSXXXXXXXX-DTXXXXXX_ID_chewbbaca_new_alleles.txt	sequence of newly-identified alleles	result directory
DSXXXXXXXX-DTXXXXXX_ID_chewbbaca_results_alleles.tsv	allele call with Pasteur encoding in csv format	result directory
DSXXXXXXXX-DTXXXXXX_ID_chewbbaca_results_ccoli_curated_210722.csv	allele csv table. Rows: samples, columns: *loci*	result directory
DSXXXXXXXX-DTXXXXXX_ID_chewbbaca_results_contigsInfo.tsv	info about the contig mapped on each *locus*	result directory
DSXXXXXXXX-DTXXXXXX_ID_chewbbaca_results_crc32.csv	allele csv table encoded in CRC32. Rows: samples, columns: *loci*	result directory
DSXXXXXXXX-DTXXXXXX_ID_chewbbaca_results_izsam.csv	allele call with IZS encoding	result directory
DSXXXXXXXX-DTXXXXXX_ID_chewbbaca_results_md5.csv	allele call with md5 encoding	result directory
DSXXXXXXXX-DTXXXXXX_ID_chewbbaca_results_statistics.tsv	metrics on *loci* encoded as EXC, INF, LNF, PLOT, NIPH, ALM, ASM	result directory
DSXXXXXXXX-DTXXXXXX_ID_import_chewbbaca_check.csv	quality check with info on calledPerc, calledNum, annotated, new, notFound, discarded	qc > result directory

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

4.4.4 Multi sample

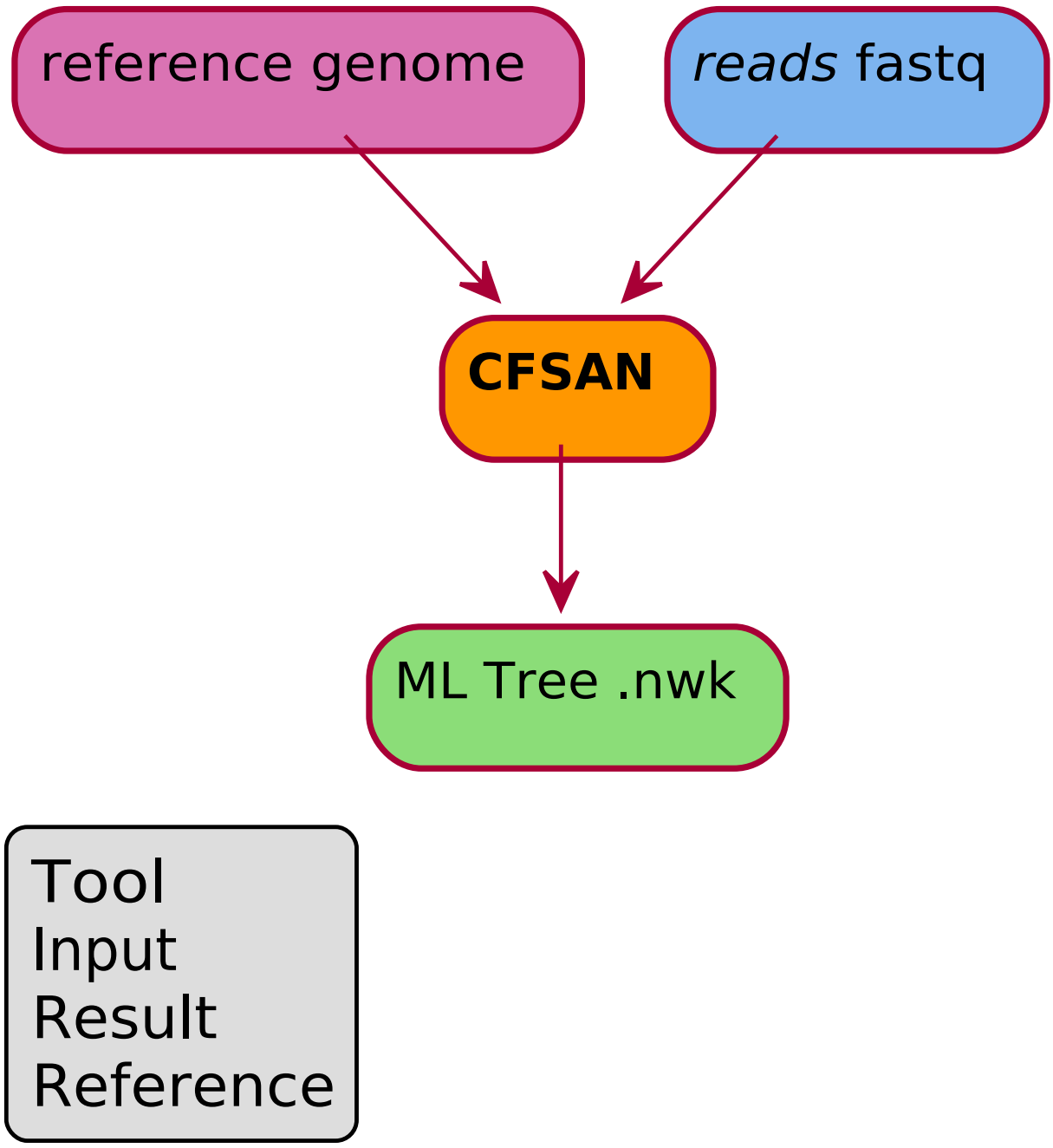
SNP-based clustering

CFSAN

Introduction

CFSAN performs a reference-based phylogenetic analysis and builds a tree graph of distance between microorganisms, using the Maximum Likelihood (ML) algorithm, starting from a Single-Nucleotide Polimorphisms (SNPs) matrix.

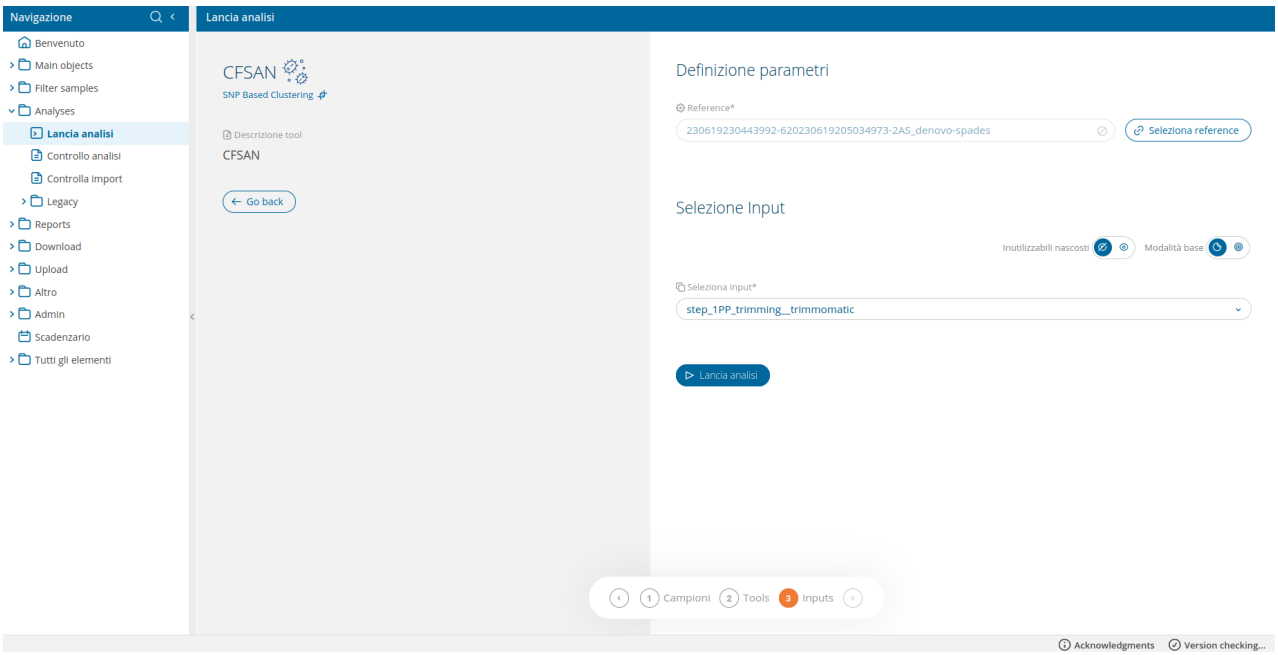
More information about the software are available at CFSAN's official guide.



Run Analysis CFSAN

The analysis **CFSAN** can be selected from the run analyses page.

CFSAN requires a reference genome, which can be selected in the parameter section. The input selection UI delivers an advanced input selection mode, to allow selection of all types of supported input files at once. Accepted inputs are always raw reads in fastq format (either internal, *i.e.* **step_0SQ_rawreads__fastq** or imported, *i.e.* **step_0SQ_rawreads__external**).



A link to Check analysis will be created after launching the requested analysis. The system will notify the user after a succesful analysis launch and once execution has ended.

Output directory

> Please refer to Cohesive's specific Wiki page for information on file download.

The output directory is available at the link in the download page or at the link in the analysis' summary card. The `results` directory is located directly in the root directory. Inside `results` there are 2 subdirectories:

- **meta**: ("metadata") contains log and configuration files.
- **result**: contains the analysis' output files.

The following table lists output files stored in `results`.

File	Description	Location
metrics.tsv	tsv summary table with sample metrics and SNPs	results directory
referenceSNP.fasta	fasta file with bases from the reference sequence for each SNP position	results directory
snp_distance_matrix.tsv	tsv table containing the samples' distance matrix, based on SNP differences	results directory
snp_distance_pairwise.tsv	tsv table containing the samples' *pairwise distances* based on SNP differences	results directory
snpma.fasta.nwk	nwk (newick) treefile obtained with the ML algorithm	results directory
snpma.vcf	list of SNPs and their relative positions for each sample, in VCFv4.2 format	results directory

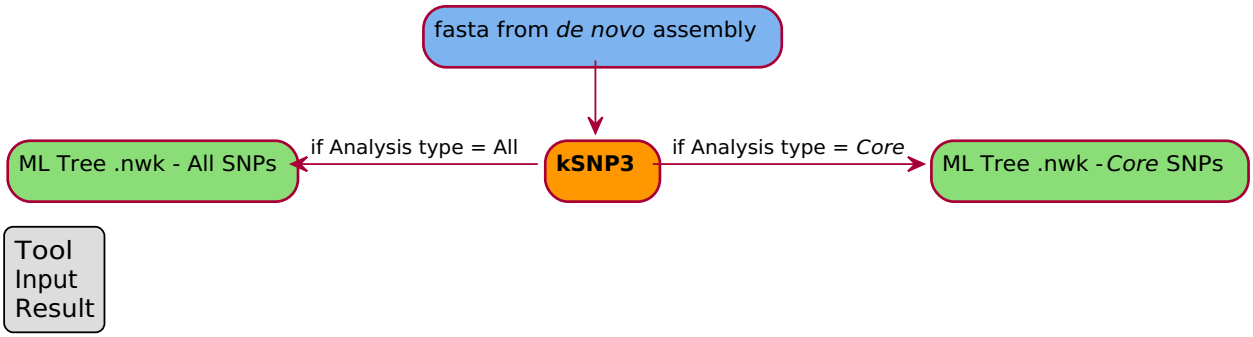
> For more information on CFSAN's outputs, please refer to CFSAN's official manual.

kSNP3

Introduction

kSNP3 identifies Single Nucleotide Polymorphisms (SNPs) in the samples, it performs a reference-free phylogenetic analysis (without alignment) and builds a phylogenetic tree with Maximum Likelihood (ML) algorithm, starting from the samples' SNP matrix.

The complete user guide is available at the kSNP3 guide's dedicated page.



Run Analysis kSNP3

The analysis **kSNP3** can be selected from the run analyses page.

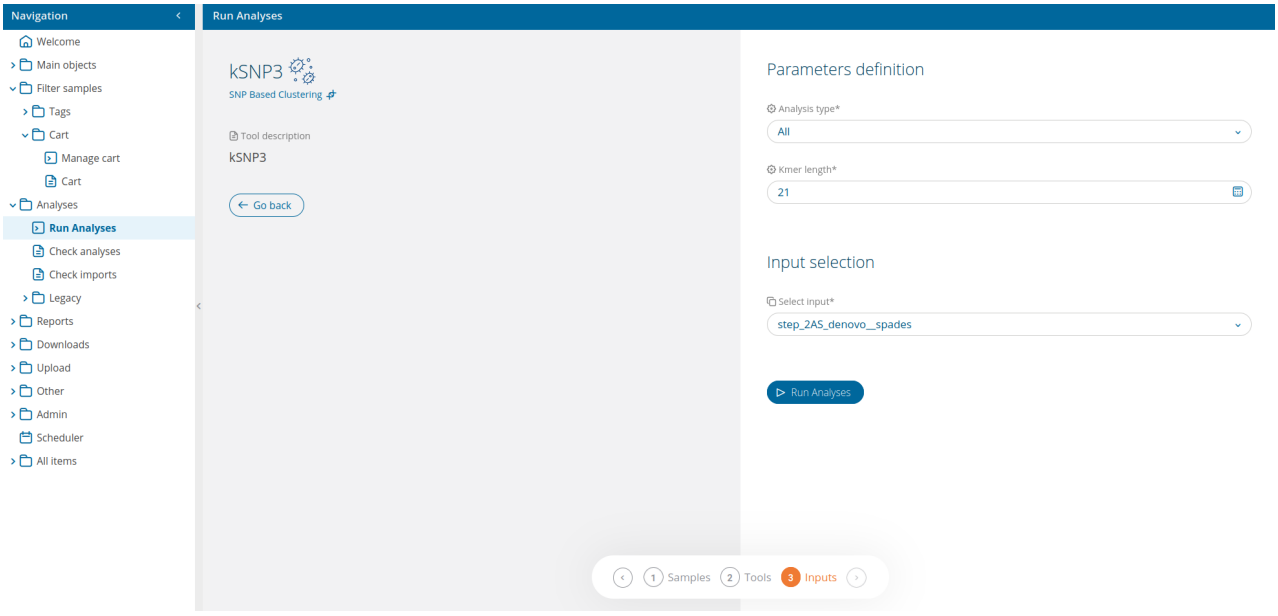
The parameters section will allow to choose:

1. to run the phylogenetic analysis using all the SNPs or only *core* SNPs ("**Analysis type**" option);
2. to change kmer length ("**kmer length**" option, default = 21).

The input selection UI delivers an advanced input selection mode, to allow selection of all types of supported input files at once.

Accepted inputs are:

- step_2AS_mapping
- step_2AS_denovo



A link to Check analysis will be created after launching the requested analysis. The system will notify the user after a succesful analysis launch and once execution has ended.

The analysis summary lists the output directory and additional options, such as nwk tree visualization, access to some of the output files, log files and metadata.

Output directory

> Please refer to Cohesive's specific Wiki page for information on file download.

The output directory is available at the link in the download page or at the link in the analysis' summary card. The `results` directory is located directly in the root directory. Inside `results` there are 2 subdirectories:

- **meta**: ("metadata") contains log and configuration files.
- **result**: contains the analysis' output files.

The following table lists output files stored in `results`.

> Please bear in mind that output file names will change depending on the option selected for the **Analysis type** parameter:
>
> - SNPs_all_matrix (if "All" has been selected);
> - core_SNPs_matrix (if "Core" has been selected).
>
> In the table below the listed file names are those in case of execution on all SNPs.

File	Description	Location
SNPs_all_matrix.fasta	multifasta file with *headers* bearing the sample name and the line with the sequence of the concatenated SNPs. "N" indicates absence of a SNP in the corresponding sample	results directory
SNPs_all_matrix.fasta.nwk	nwk (newick) treefile obtained with ML algorithm	results directory

> For additional information about formatting of kSNP3's outputs, please refer to kSNP3's user guide.

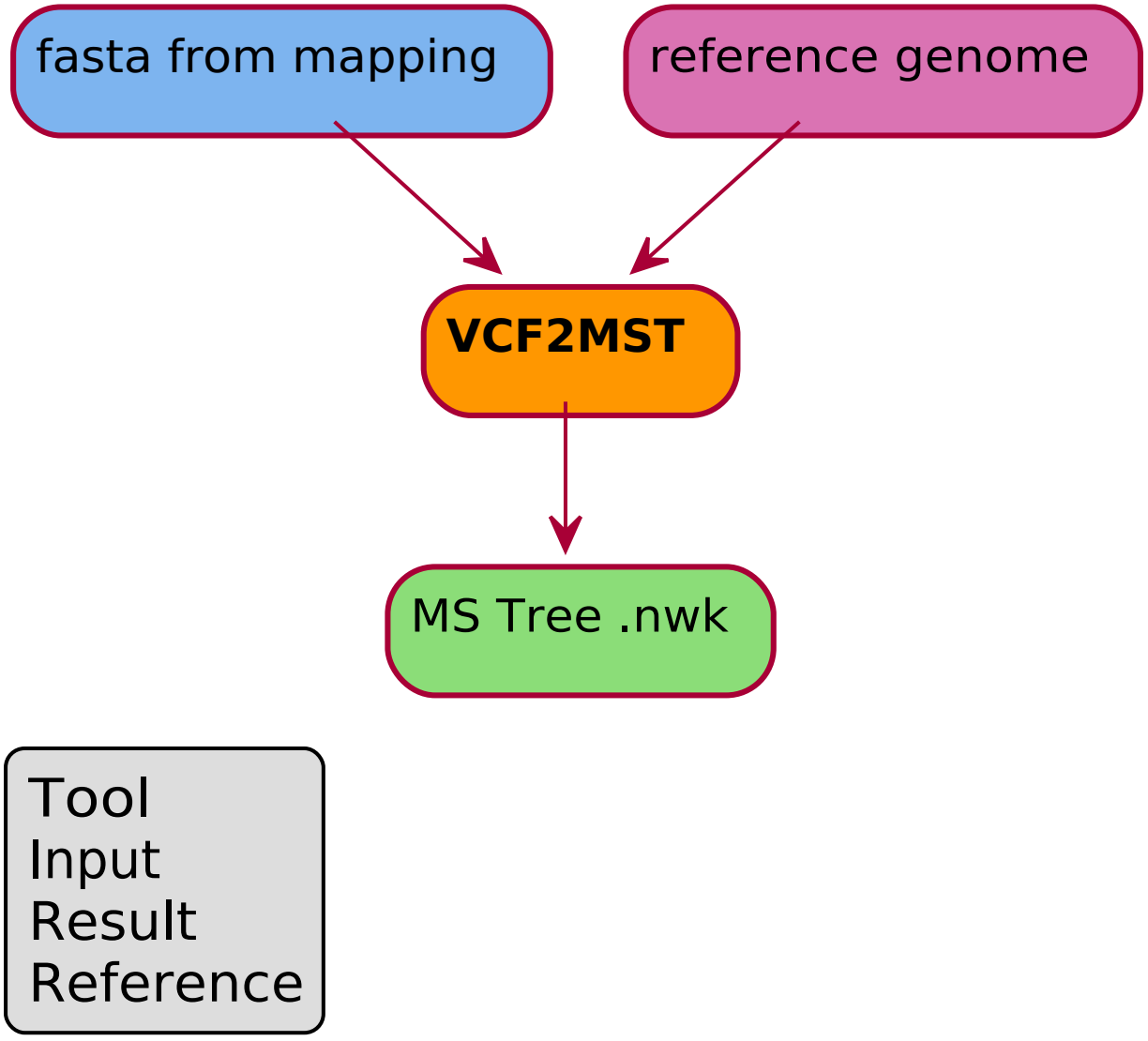
VCF2MST

Introduction

VCF2MST identifies SARS-CoV-2's PANGO lineages and builds a phylogenetic tree graph through the Minimum Spanning Tree (MST) algorithm, starting from VCF files of Single Nucleotide Polymorphysms (SNPs), without alignment-based phylogenomic inference.

The software builds a Minimum Spanning Tree based on Hamming distances, which measure the number of substitutions necessary for a sequence (string) to be transformed into another string.

More information on the software is available at VCF2MST's official GitHub page ("VCF2MST - Hamming Distance based Minimum Spanning Tree from Samples vcf using graptree") or at the corresponding research article.

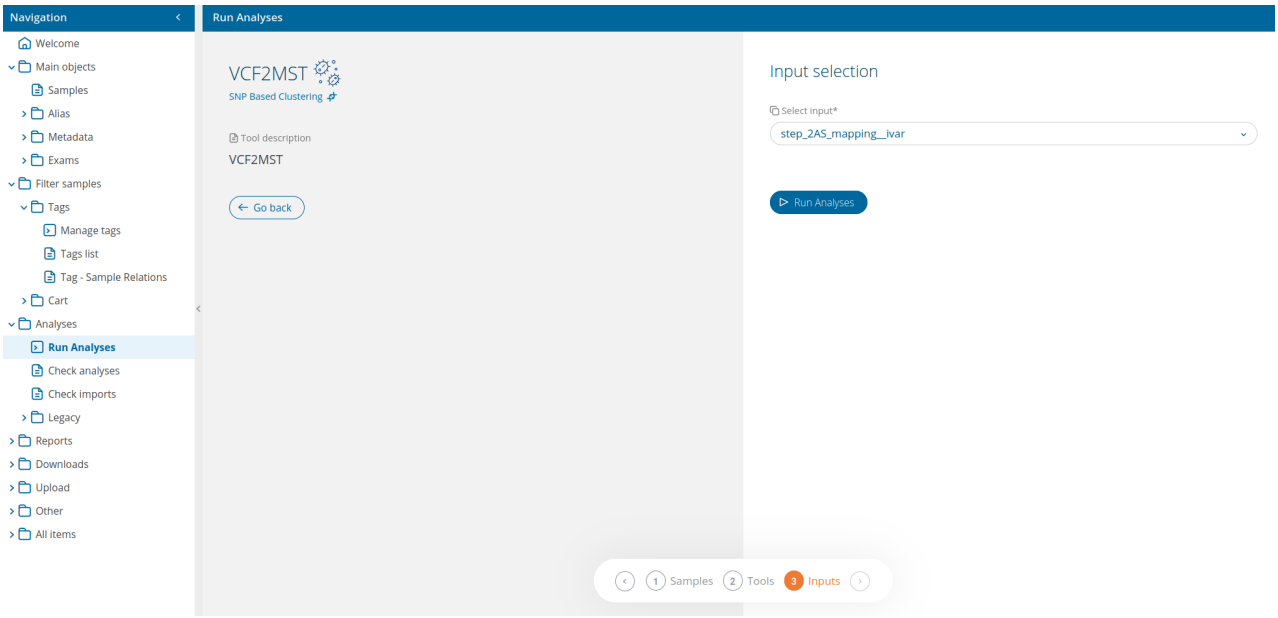


Run Analysis VCF2MST

The analysis **VCF2MST** can be selected from the run analyses page.

The input selection UI delivers an advanced input selection mode, to allow selection of all types of supported input files at once.

Accepted inputs for VCF2MST are from step_2AS_mapping. In the input slection interface there will also be a field to provide the reference genome (auto-filled).



A link to Check analysis will be created after launching the requested analysis. The system will notify the user after a succesful analysis launch and once execution has ended.

The analysis summary lists the output directory and additional options, such as access to some of the metadata, log and output files and direct graph visualization thanks to GrapeTree's integration in Cohesive.

Output directory

> Please refer to Cohesive's specific Wiki page for information on file download.

The output directory is available at the link in the download page or at the link in the analysis' summary card. The `results` directory is located directly in the root directory. Inside `results` there are 2 subdirectories:

- **meta**: ("metadata") contains log and configuration files.
- **result**: contains the analysis' output files.

The following table lists output files stored in `results`.

File	Description	Location
tree.nwk	MST treefile in nwk (newick) format	results directory

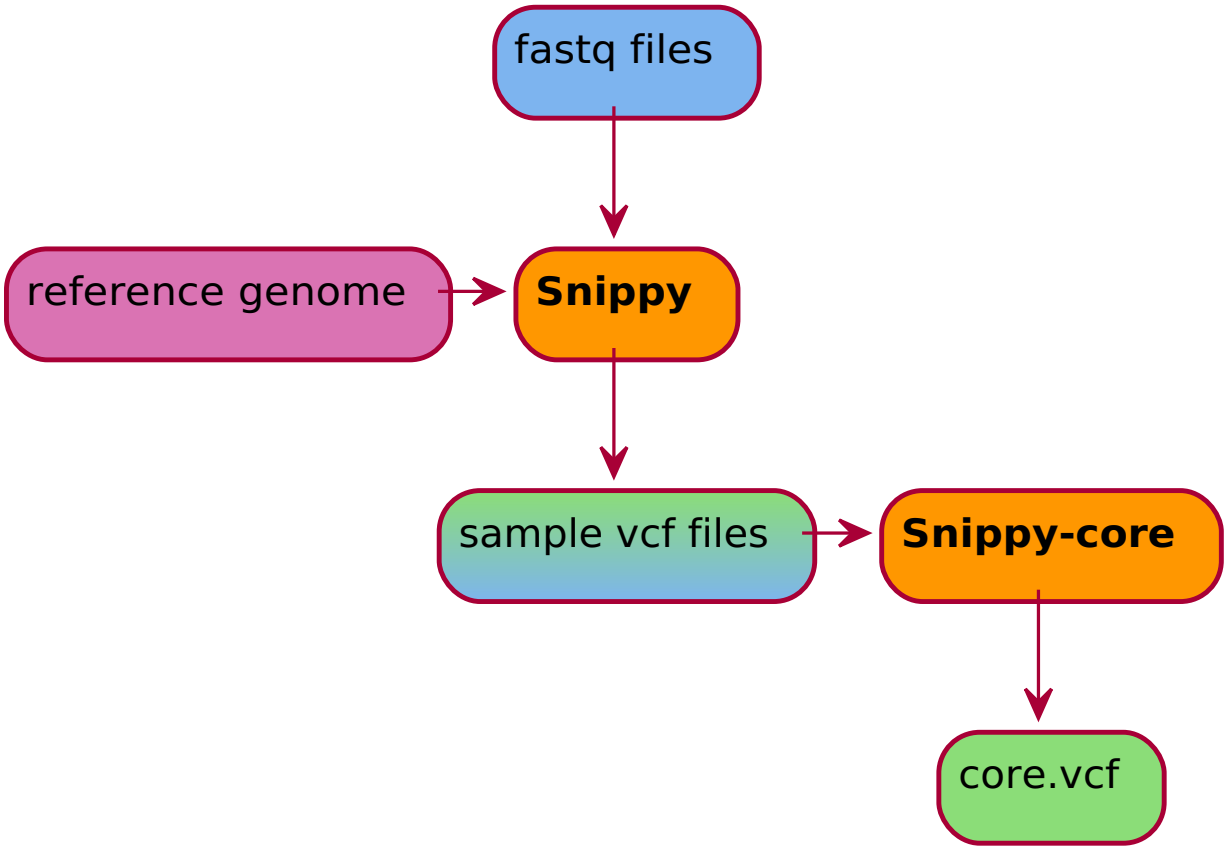
Pangenome extraction

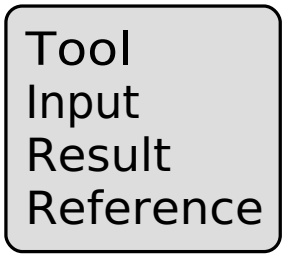
Snippy-core

Introduction

The Snippy-core analysis includes a Snippy run (*Rapid haploid variant calling and core genome alignment*), which aligns the selected samples to a reference in .gb format, and Snippy-core (hence the name) to merge Snippy's VCFs into a `core.vcf` file, finding core variants (SNPs e indels common to all samples).

Snippy & Snippy-core GitHub Page: https://github.com/tseemann/snippy





Run Analysis Snippy-core

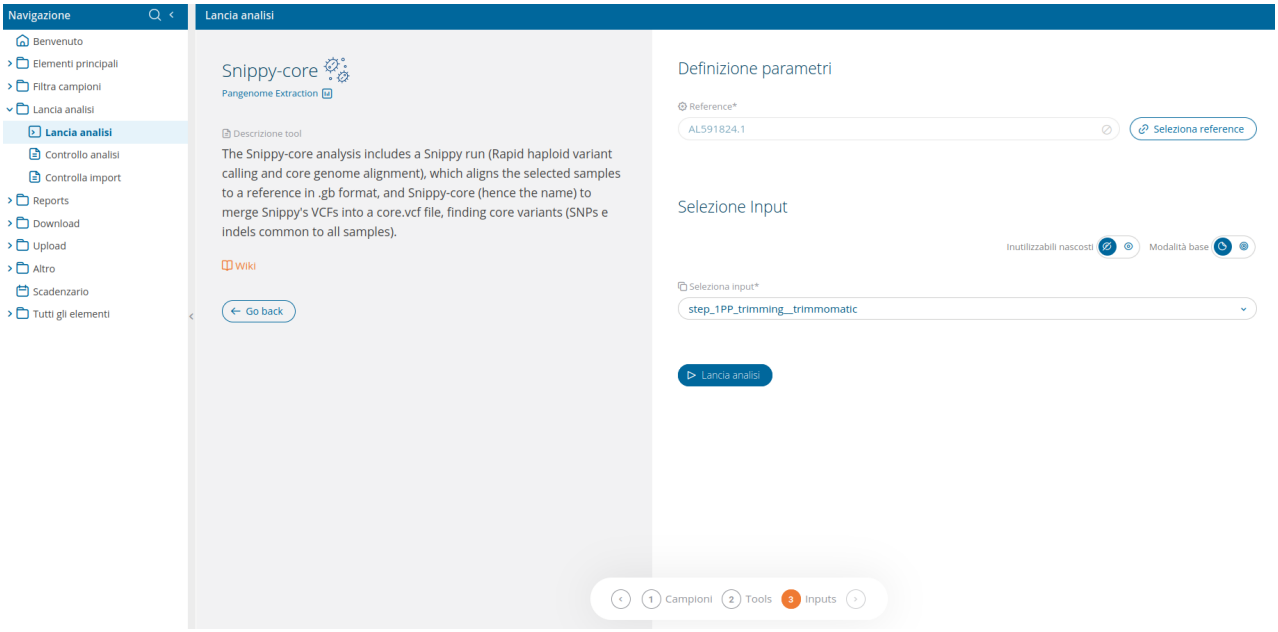
Once the analysis **Snippy-core** has been selected from the run analyses interface, the wizard will present a confirmation UI. The analysis is specific for Snippy+Snippy-core and there are no other available tools.

The input selection wizard will require selection of a `.gb` reference for Snippy-core execution, after Snippy.

The input selection wizard delivers an advanced input selection mode, to allow selection of all types of supported input files at once.

Accepted inputs can be from:

- step_1PP_trimming
- step_1PP_generated
- step_1PP_hostdepl
- step_1PP_downsampling



A link to Check analysis will be created after launching the requested analysis. The system will notify the user after a succesful analysis launch and once execution has ended.

Output directory

Please refer to Cohesive's specific Wiki page for information on file download.

The output directory can be reached from the link of the download page or from the link in the analysis summary. The `results` directory is located directly in the root directory and it contains the following 2 directories:

- **meta**: ("metadata") contains log and configuration files.
- **result**: contains the analysis' output files.

The table below lists files available in the output directory structure, alongside some useful information.

File	Description	Location
core.aln	fasta file with sequences from samples and from reference	result directory
core.full.aln	fasta file with samples' sequences aligned to the reference's	result directory
core.ref.fa	fasta file with the reference genome's nucleotide sequence	result directory
core.tab	tsv variants table. Columns: Chromosome, position, nucleotide in reference, nucleotide in sample 1, nucleotide in sample 2...	result directory
core.txt	text file with summary table of sample and reference sequence features	result directory
core.vcf	VCFv4.2 variants file. Includes an informative header and the table listing variant type and vairant presence/absence binary matrix for samples	result directory

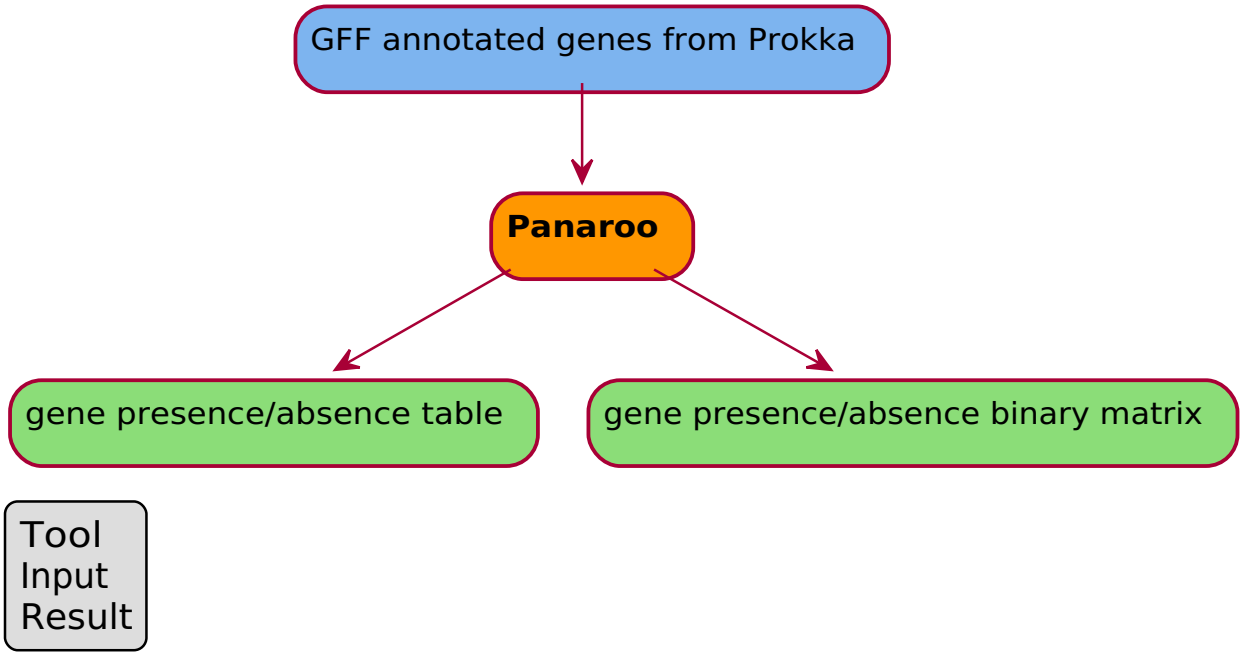
Panaroo

Introduction

Panaroo elaborates the pan-genome and builds presence/absence matrices of samples' annotated genes, starting from Prokka's genome annotation.

For in depth information about Panaroo and its operation, please refer to Panaroo's official guide.

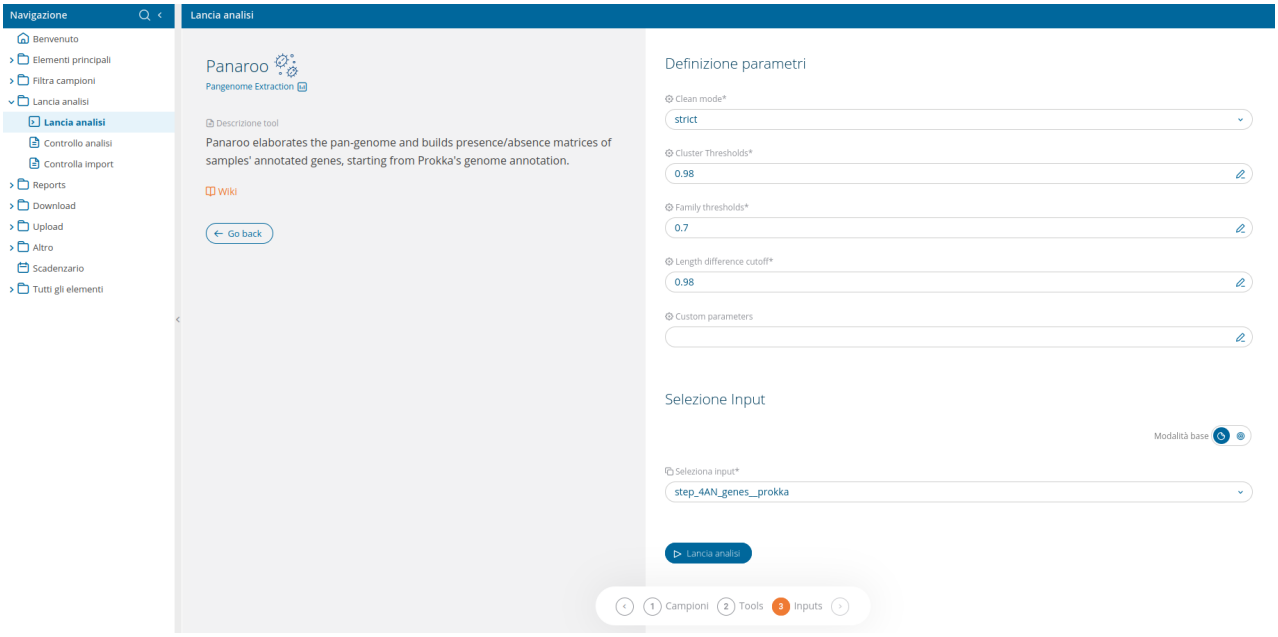
Panaroo's GitHub Page: https://github.com/gtonkinhill/panaroo



Run Analysis Panaroo

Once the analysis **Panaroo** has been selected from the run analyses interface, the wizard will present a confirmation UI: there is no need for tool selection, since the only tool available is "Panaroo - An updated pipeline for pangenome investigation".

The input selection wizard will allow to confirm the input for Panaroo, which is deisgned to work on Prokka's output (thus the annotation files from 4AN_genes will be the input). Fields are pre-filled and no further selection by the user is required.



A link to Check analysis will be created after launching the requested analysis. The system will notify the user after a succesful analysis launch and once execution has ended.

Output directory

> Please refer to Cohesive's specific Wiki page for information on file download.

The output directory can be reached from the link of the download page or from the link in the analysis summary. The `results` directory is located directly in the root directory and it contains the following 2 directories:

- **meta**: ("metadata") contains log and configuration files.
- **result**: contains the analysis' output files.

The following table lists files created by Panaroo, alongside some useful information. More information on Panaroo's output files are available at the official guide to Panaroo's outputs.

File	Description	Location
combined_DNA_CDS.fasta	fasta file of nucleotide sequences from annotated genes and genes identified by Panaroo	results directory
combined_protein_CDS.fasta	fasta file of aminoacid sequences from annotated genes and genes identified by Panaroo	results directory
combined_protein_GFF3cdhit_out.txt	log of Panaroo's CD-HIT phase	results directory
combined_protein_cdhit_out.txt.clstr	CD-HIT cluster info	results directory
core_alignment_header.embl		results directory
core_gene_alignment.aln	alignment file	results directory
final_graph.gml	pan-genome graph	results directory
gene_data.csv	csv table with sequences of annotated genes and corresponding Panaroo internal codes	results directory
gene_presence_absence.Rtab	gene presence/absence binary matrix	results directory
gene_presence_absence.csv	csv file of gene presence in samples	results directory
gene_presence_absence_roary.csv	csv file of gene presence in samples (Roary model)	results directory
pan_genome_reference.fa	pan-genome reference fasta for genes in the dataset	results directory
pre_filt_graph.gml	raw pan-genome graph	results directory
struct_presence_absence.Rtab	presence/absence binary matrix for gene rearrangement events	results directory
summary_statistics.txt	metrics summary file	results directory

Panaroo's authors suggest Cytoscape for graph visualization. more information on pan-genome graph visualization are available at Panaroo's official documentation page.

Gene-by-gene based clustering

GrapeTree

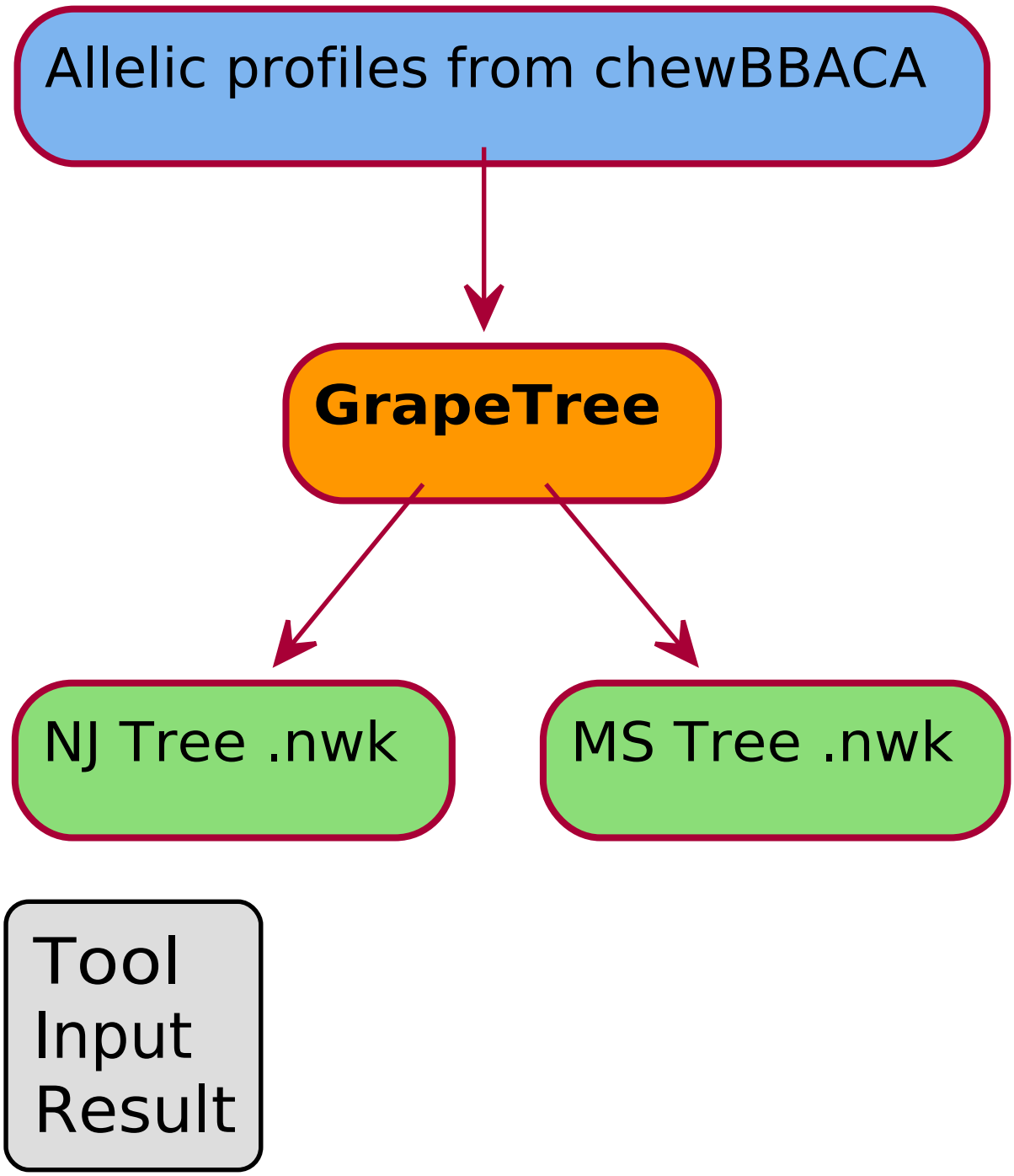
Introduction

GrapeTree builds graphs through the "Minimum Spanning Tree" (MSTree) or "Neighbor Joining" (NJ) algorithms, starting from the samples' allelic profiles, obtained from cgMLST.

Cohesive enables both tree calculatation and visualization, thanks to the integration of the standalone version of GrapeTree's interacitve visualization system.

A full guide to GrapeTree's visualization system interface and its usage is available at the GrapeTree dashboard guide page.

For more information on the software, please refer to the research article and to the official documentation of the EnteroBase system, to which GrapeTree belongs.



Run Analysis GrapeTree

The analysis **GrapeTree** can be selected from the run analyses page.

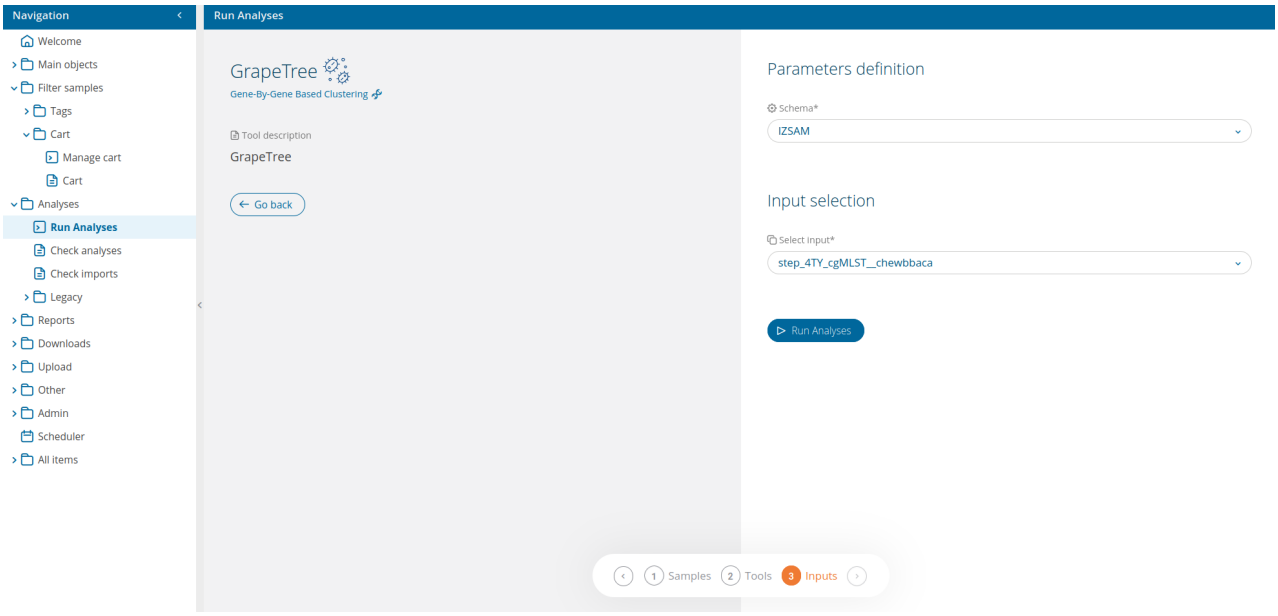
The input selection UI delivers an advanced input selection mode, to allow selection of all types of supported input files at once.

Accepted inputs are:

- step_4TY_wgMLST
- step_4TY_cgMLST

> For all possible inputs, is imperative that the program used to calculate allelic profiles is **chewBBACA**, so please mind the selected software, when launching 4TY_cgMLST or 4TY_wgMLST, if you wish to run GrapeTree later on.

It will be necessary to specify the 4TY_cgMLST__chewbbaca schema for the appropriate species (selection available in the parameter section), if the selected input consists of cgMLST allelic profiles. The field is pre-filled with the most appropriate schema for the microorganism matching the codes of samples in the cart or tag used for requesting the analysis.



A link to Check analysis will be created after launching the requested analysis. The system will notify the user after a succesful analysis launch and once execution has ended.

The analysis summary lists the output directory and additional options, such as access to some of the output files and direct graph visualization (both MSTree and NJ graphs) thanks to GrapeTree's integration in Cohesive.

Output directory

> Please refer to Cohesive's specific Wiki page for information on file download.

The output directory is available at the link in the download page or at the link in the analysis' summary card. The `results` directory is located directly in the root directory. Inside `results` there are 2 subdirectories:

- **meta**: ("metadata") contains log and configuration files.
- **result**: contains the analysis' output files.

The following table lists output files stored in `results`.

File	Description	Location
cgMLST.nwk	nwk (newick) treefile obtained with MSTree algorithm	results directory
cgMLST.tsv	tsv table with alleles identified by chewBBACA during cgMLST. Rows: samples; Columns: *loci*.	results directory
cgMLST_NJ.nwk	nwk (newick) treefile obtained with NJ algorithm	results directory
missing_loci.tsv	list of *loci* not shared between samples, thus discarded because uninformative for tree construction	results directory

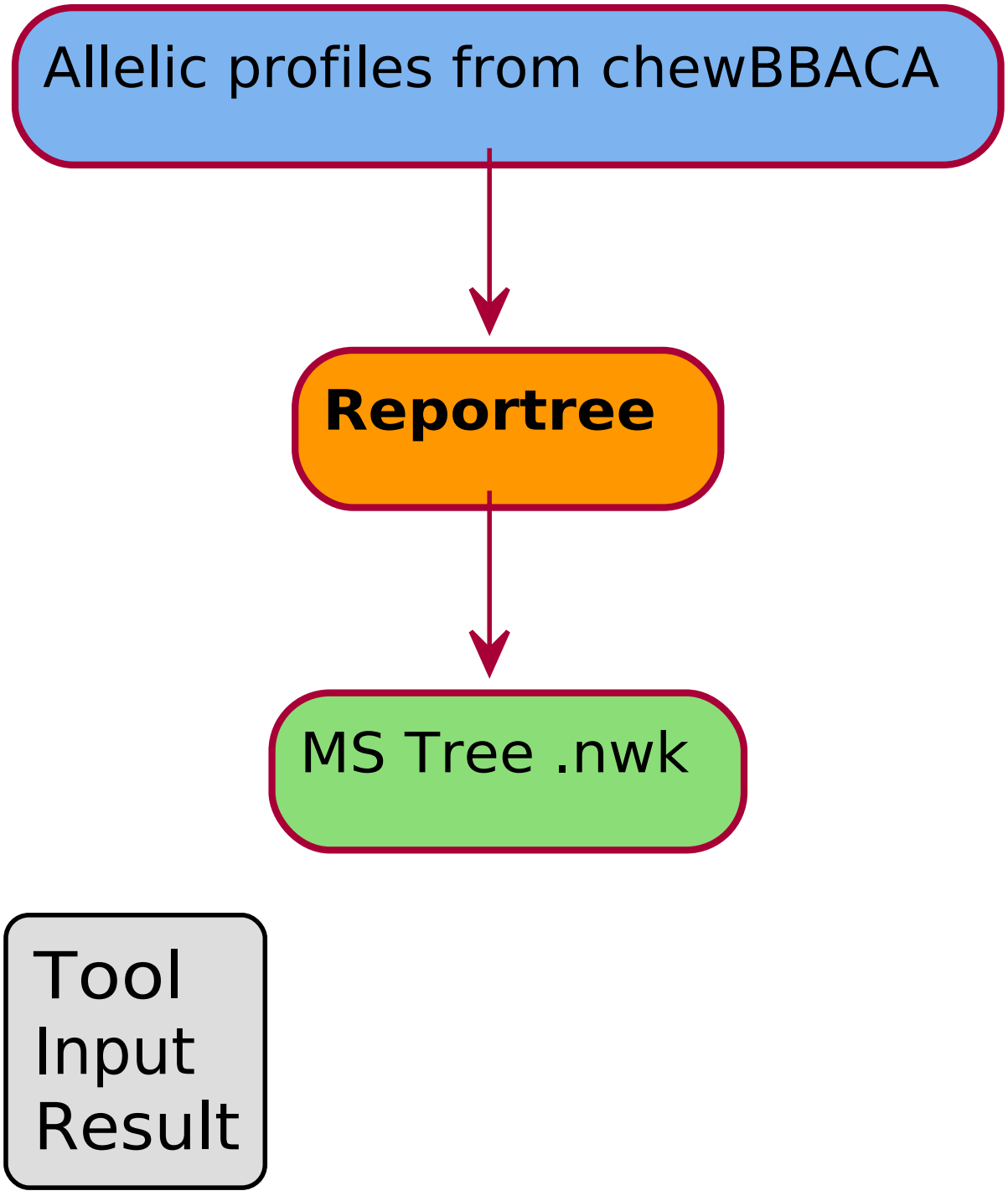
nwk files contain all of the tree's data and can be visualized through GrapeTree or an external tree-graph visualization software.

Reportree

Introduction

Reportree, similarly to GrapeTree, buidls tree-graphs of distance for microorganisms, starting from allelic profiles of the samples. The tree is built using the "Minimum Spanning Tree" (MSTree or MST) algorithm.

Once the tree has been calculated, it can be visualized directly in Cohesive (using GrapeTree) through the integrated system for tree visualization, available in the completed analysis' summary.



Run Analysis Reportree

The analysis **Reportree** can be selected from the run analyses page.

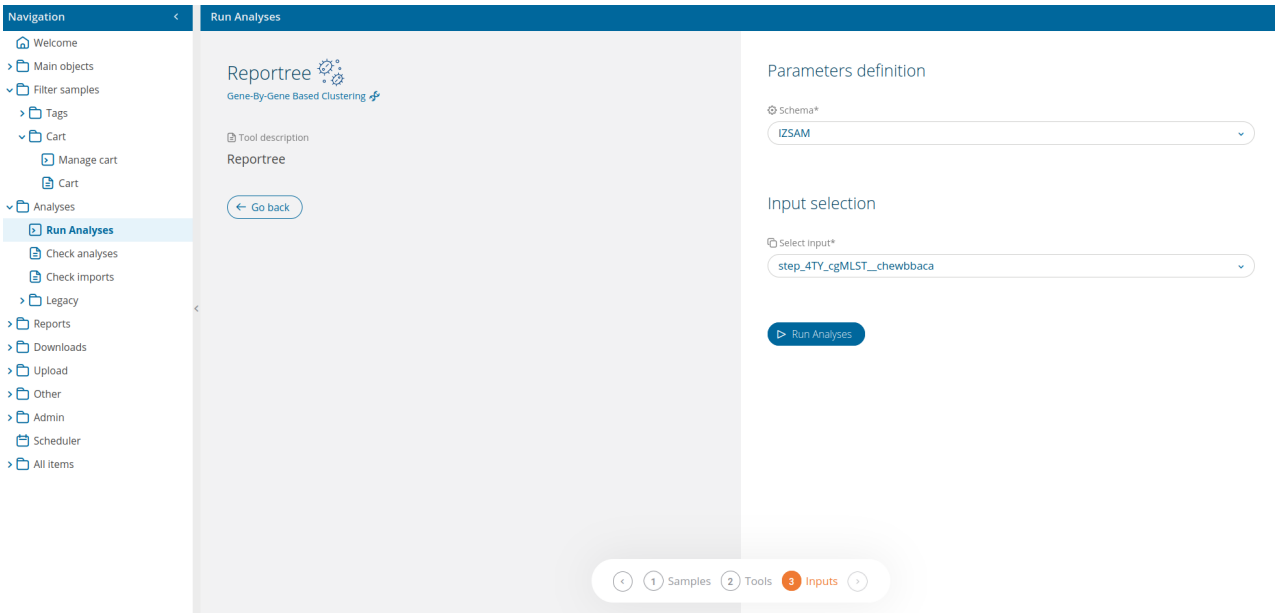
The input selection UI delivers an advanced input selection mode, to allow selection of all types of supported input files at once.

Accepted inputs are:

- step_4TY_wgMLST
- step_4TY_cgMLST

> For all possible inputs, is imperative that the program used to calulate allelic profiles is **chewBBACA**, so please mind the selected software, when launching 4TY_cgMLST or 4TY_wgMLST, if you wish to run GrapeTree later on.

It will be necessary to specify the 4TY_cgMLST__chewbbaca schema for the appropriate species (selection available in the parameter section), if the selected input consists of cgMLST allelic profiles. The field is pre-filled with the most appropriate schema for the microorganism matching the codes of samples in the cart or tag used for requesting the analysis.



A link to Check analysis will be created after launching the requested analysis. The system will notify the user after a succesful analysis launch and once execution has ended.

The analysis summary lists the output directory and additional options, such as access to some of the metadata, log and output files and direct graph visualization (both MSTree and NJ graphs) thanks to GrapeTree's integration in Cohesive.

Output directory

> Please refer to Cohesive's specific Wiki page for information on file download.

The output directory is available at the link in the download page or at the link in the analysis' summary card. The `results` directory is located directly in the root directory. Inside `results` there are 2 subdirectories:

- **meta**: ("metadata") contains log and configuration files.
- **result**: contains the analysis' output files.

The following table lists output files stored in `results`.

File	Description	Location
cgMLST.tsv	tsv table with alleles identified by chewBBACA during cgMLST. Rows: samples; Columns: *loci*.	cartella result
missing_loci.tsv	list of *loci* not shared between samples, thus discarded because uninformative for tree construction	cartella result
reportree.dist	summary text file with Reportree's caculated distances	cartella result
reportree.mx	table with samples' relative distance matrix	cartella result
reportree.nw	nwk (newick) treefile obtained with MSTree algorithm	cartella result
reportree_ST_iso_year_count_matrix.tsv	tsv table with Sequence Type (ST) quantities by year	cartella result
reportree_ST_iso_year_freq_matrix.tsv	tsv table with Sequence Type (ST) frequencies by year	cartella result
reportree_ST_summary.tsv	summary table of STs and metadata	cartella result
reportree_clusterComposition.tsv	tsv table of info on sample clusters	cartella result
reportree_iso_year_summary.tsv	summary table of ST percentages and origin by year	cartella result
reportree_metadata.tsv	sample metadata tsv table	cartella result
reportree_metadata_w_partitions.tsv	sample metadata tsv table with cluster info	cartella result
reportree_metrics.tsv	Reportree cluster metrics	cartella result
reportree_partitions.tsv	tsv table of cluster subdivision	cartella result
reportree_partitions_summary.tsv	tsv summary table of STs with metadata and cluster info	cartella result
reportree_puntoprelievo_summary.tsv	tsv summary table of STs by origin	cartella result
reportree_stableRegions.tsv	Wallace coefficient *Stability Regions*	cartella result

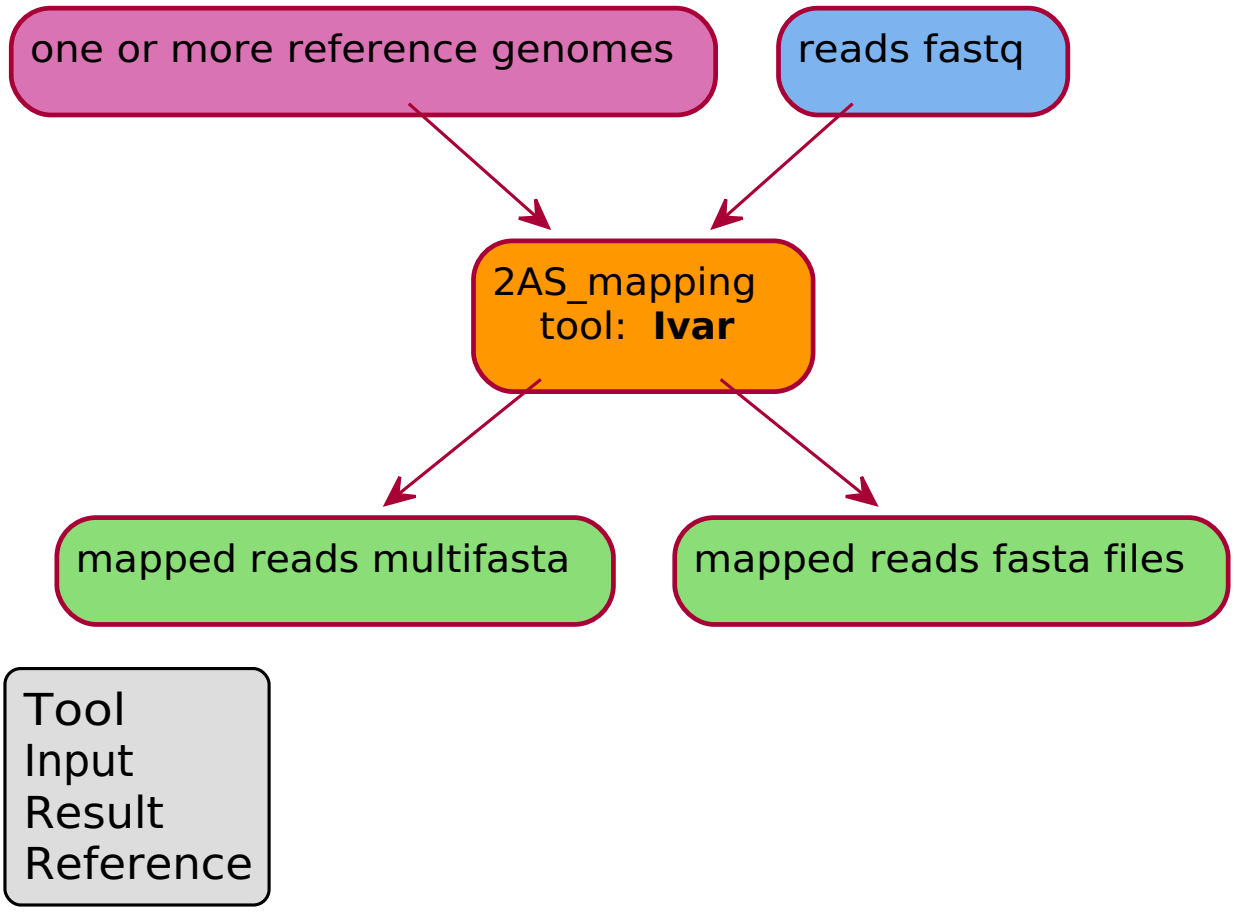
nwk files contain all of the tree's data and can be visualized through GrapeTree or an external tree-graph visualization software.

4.4.5 Pipelines

Mapping for Segmented Viruses Pipeline

Introduction

The pipeline "Mapping for Segmented Viruses" performs mapping of genome fragments from segmented viruses using multiple references, which allows mapping of each segment to the most appropriate reference. This pipeline outputs both alignment files and the whole multifasta file.



Run Mapping for Segmented Viruses Pipeline

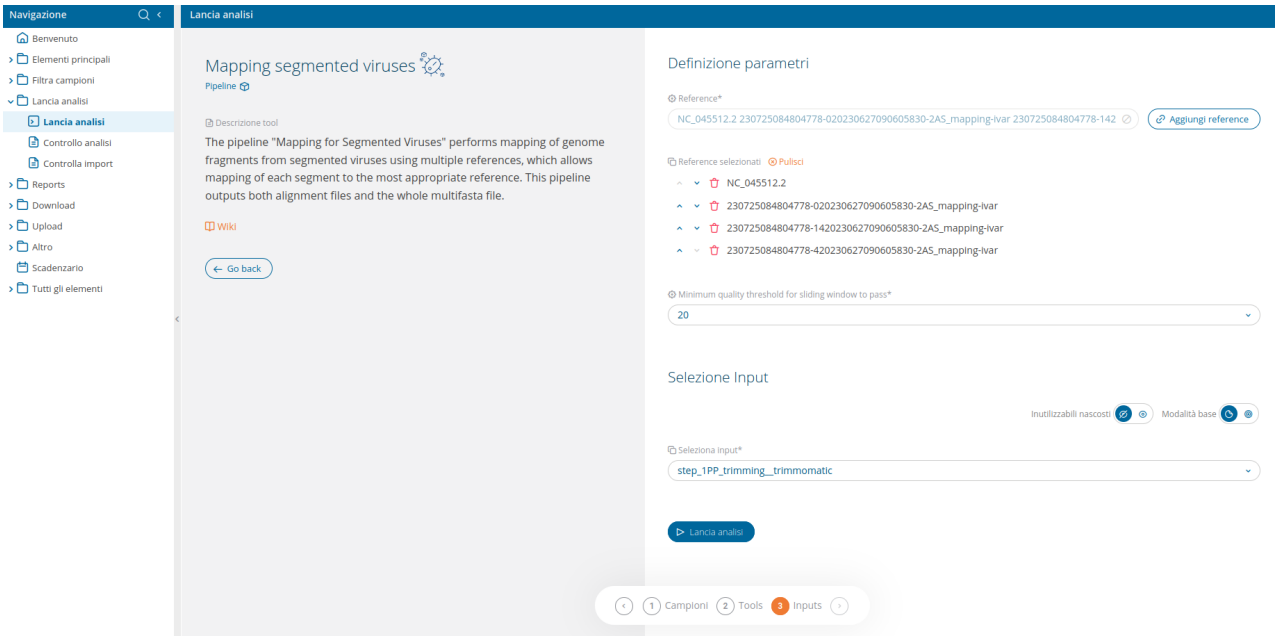
The filter at the top of the run analyses page allows to display only the pipelines. Once the **Mapping for Segmented Viruses** pipeline has been selected, a confirmation interface will be displayed.

The **Mapping for Segmented Viruses** pipeline performs mapping with **Ivar** (2AS_mapping__ivar) for all genome fragments from segmented viruses provided by the user.

Accepted inputs can be 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 UI delivers an advanced input selection mode, to allow selection of all types of supported input files at once.



A link to Check analysis will be created after launching the requested pipeline. The system will notify the user after a succesful pipeline launch and once execution has ended.

Multiple References

All required reference files will have to be provided by the user, through the dedicated reference selection wizard (please refer to the corresponding Wiki section on the run analysis topic).

Results

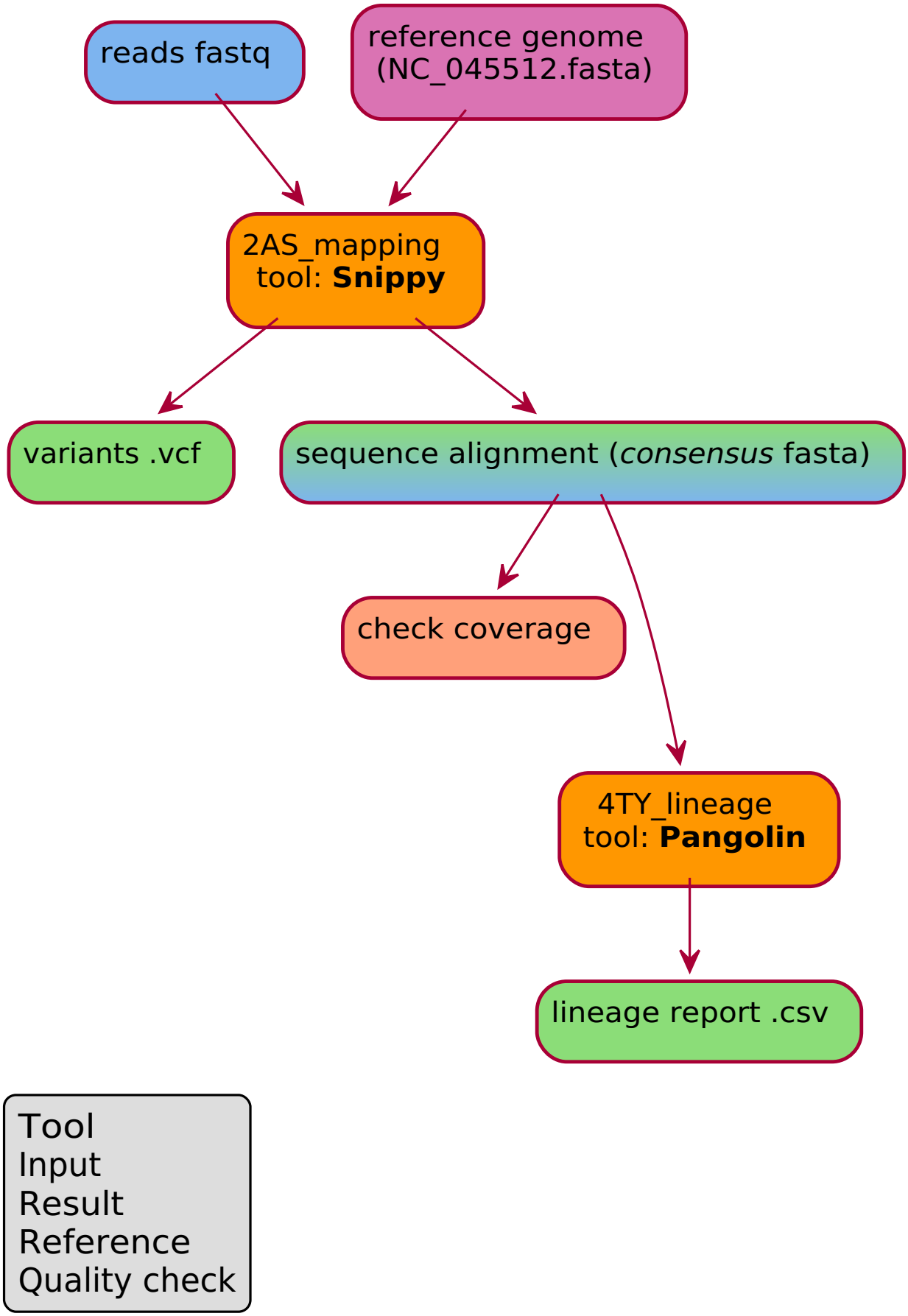
Output files from the "Mapping for Segmented Viruses" pipeline are the same as those produced by the single analyses included in it, in the same directory structure. Please refer to the corresponding 2AS_mapping Wiki section. On top of those files, the pipelina also outputs a multifasta file (more details in the table below).

File	Description	Location
	multifasta file containing all the fragment sequences mapped to the appropriate references	
warnings.log	warning text file with information about incorrect references	main directory

Covid Emergency Pipeline

Introduction

The **Covid Emergency** pipeline allows fast execution of assembly (step_2AS_mapping) and lineage assignment (4TY_lineage) of SARS-CoV2 samples. The reference genome used dutring the mapping step is NC_045512 (Wuhan-Hu-1). Since the tool used for mapping is Snippy, the pipeline also produces the variants' VCF file.


reads fastq
reference genome (NC_045512.fasta)
2AS_mapping tool: **Snippy**
variants .vcf
sequence alignment (*consensus* fasta)
check coverage
4TY_lineage tool: **Pangolin**
lineage report .csv
Tool
Input
Result
Reference
Quality check

Run Covid Emergency Pipeline

The filter at the top of the run analyses 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:

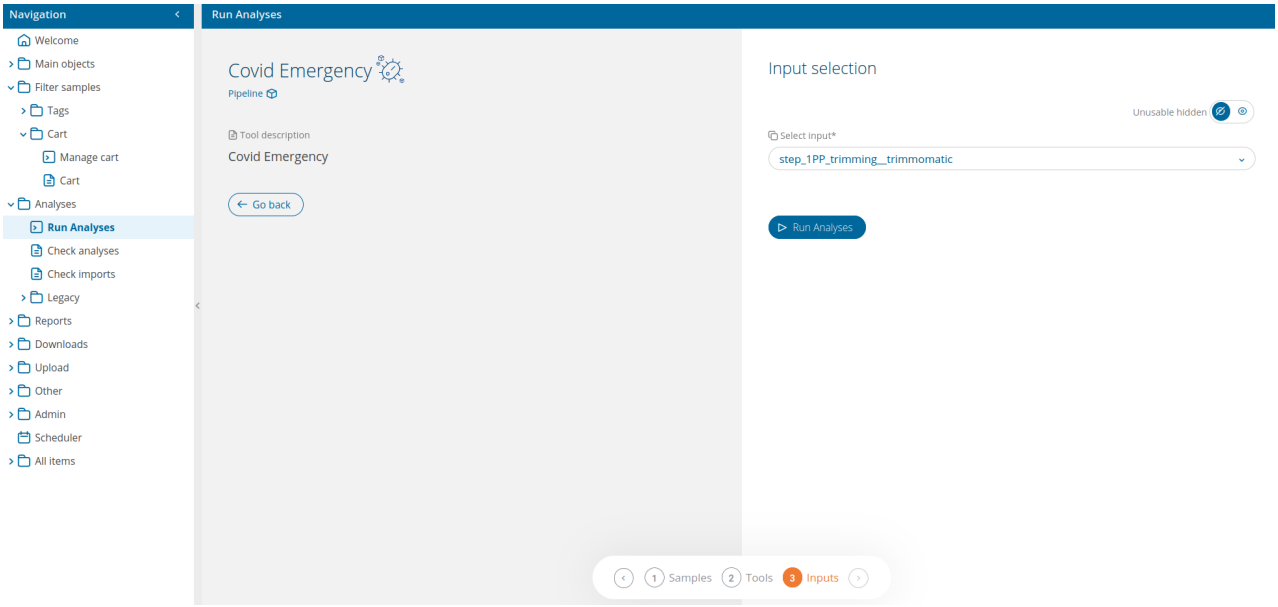
1. Mapping with **Snippy** (2AS_mapping__snippy);
2. Lineage assignment with **Pangolin** (4TY_lineage__pangolin).

Accepted inputs are the same as for mapping with Snippy, *i.e.* reads in fastq format:

- step_1PP_trimming

The input selection UI delivers an advanced input selection mode, to allow selection of all types of supported input files at once.

In the parameter dedicated section it's possible to choose a minimum quality value (20 o 14) for predicted bases, through the field for quality threshold selection.



A link to Check analysis will be created after launching the requested pipeline. The system will notify the user after a succesful pipeline launch and once execution has ended.

Results

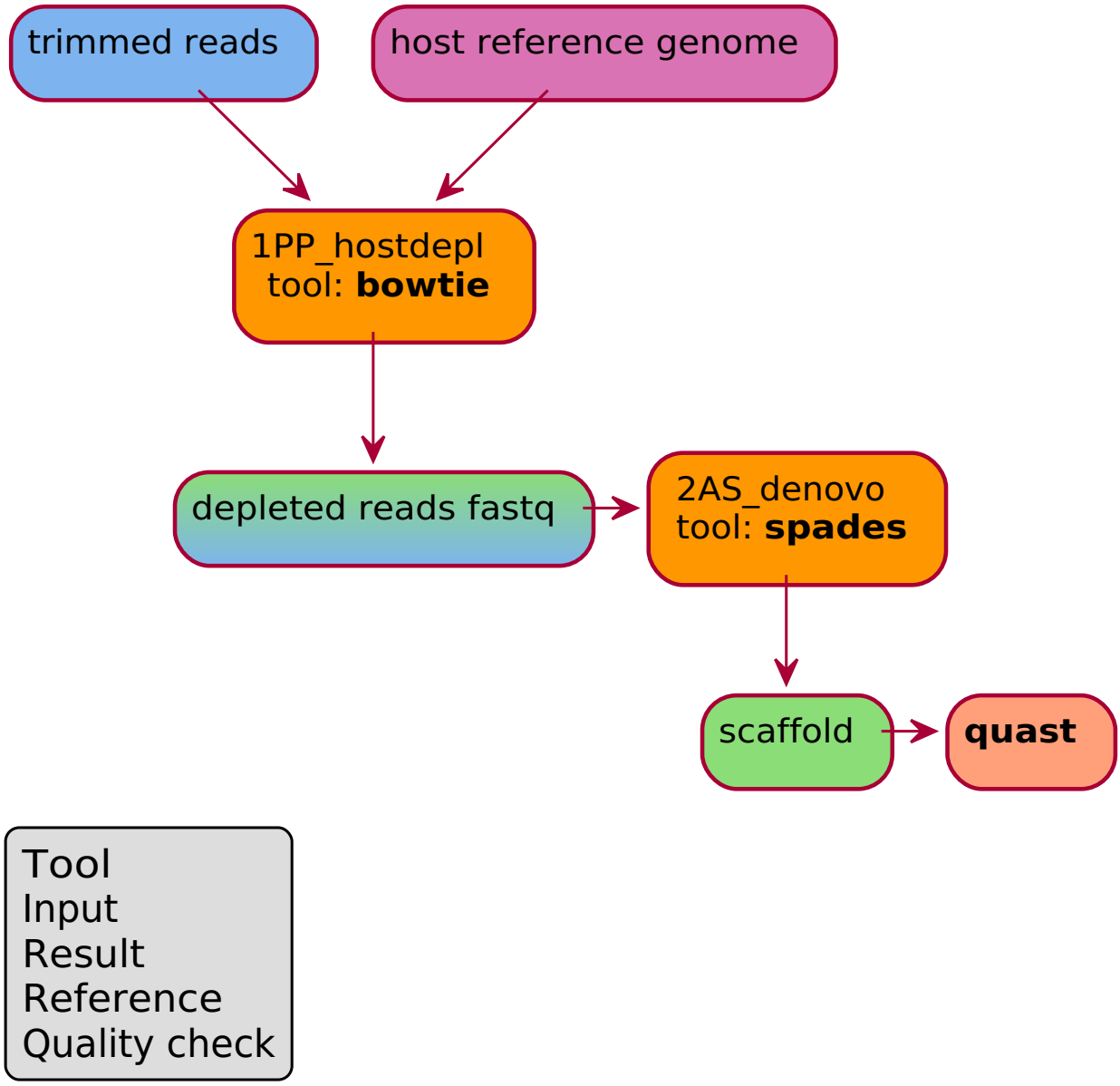
Output files from the "Covid Emergency" pipeline are the same as those produced by the single analyses included in it, in the same directory structure. For details on the pipeline's output files, please refer to the output sections of the appropriate analyses:

- Output of mapping with Snippy
- Output of lineage assignment with Pangolin

Depletion & *de novo* Pipeline

Introduction

The **Depletion & *de novo*** pipeline performs consecutively host depletion (1PP_hostdepl) and *de novo* assembly (2AS_denovo).



Run Depletion & *de novo* Pipeline

The filter at the top of the run analyses page allows to display only the pipelines. Once the **Depletion & *de novo*** pipeline has been selected, a confirmation interface will be displayed.

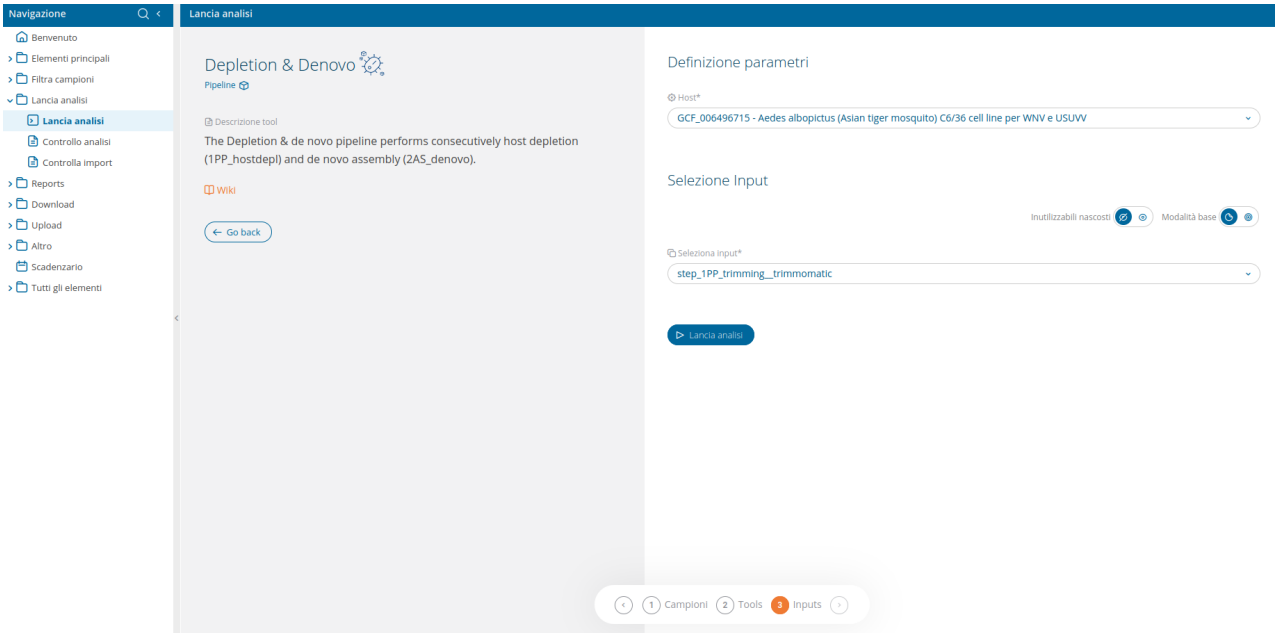
The pipeline performs the following 2 analyses:

1. Depletion of host sequences with **bowtie** (1PP_hostdepl__bowtie);
2. *de novo* assembly with **Spades** (2AS_denovo__spades) e controllo qualità con **Quast**.

Inputs are the same as those available for host depletion, *i.e.* the host's reference genome and the fastq reads:

- step_1PP_hostdepl
- step_1PP_downsampling
- step_1PP_trimming

The input selection UI delivers an advanced input selection mode, to allow selection of all types of supported input files at once.



A link to Check analysis will be created after launching the requested pipeline. The system will notify the user after a succesful pipeline launch and once execution has ended.

Results

Output files from the "Deplezione & *de novo*" pipeline are the same as those produced by the single analyses included in it, in the same directory structure:

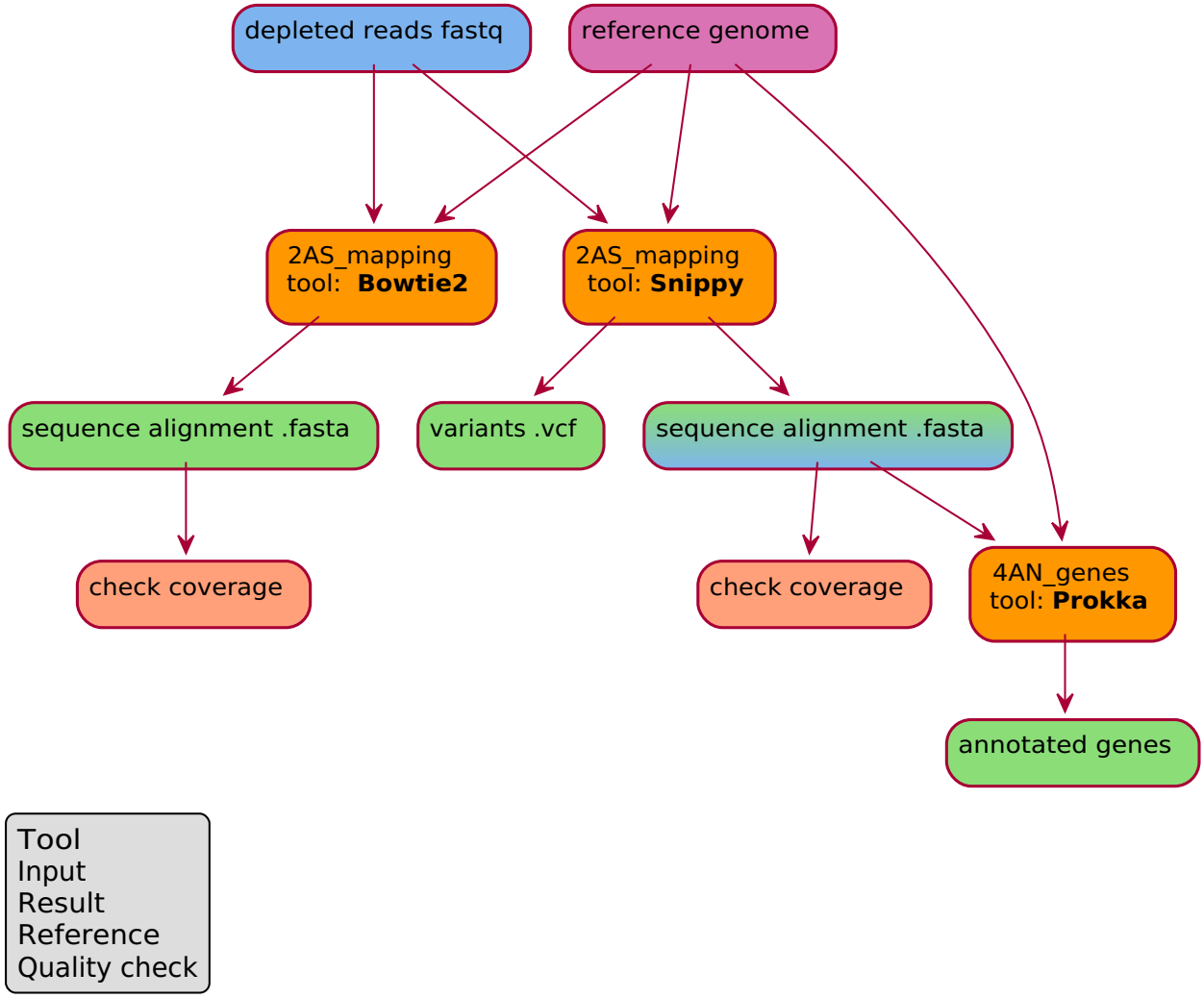
- Output of ost depletion with bowtie
- Output of *de novo* assembly with SPAdes and of Quast's quality check

Genome Draft Pipeline

Introduction

The **Genome Draft** pipeline is part of the **Vdraft** pipeline. **Genome Draft** performs mapping (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 the reference used has a genebank (.gb or .gbk) file format.



Run Genome Draft Pipeline

The filter at the top of the run analyses 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).

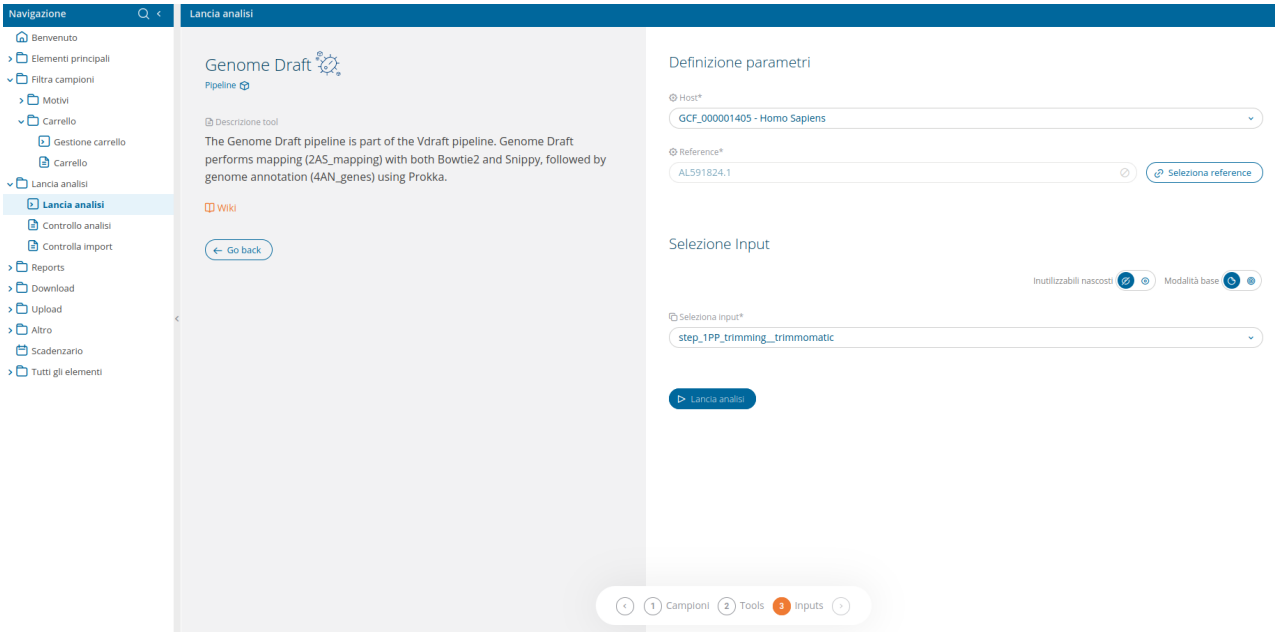
Both mapping executions include computing of coverage.

Accepted inputs can be 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 UI delivers an advanced input selection mode, to allow selection of all types of supported input files at once.

The pipeline also requires to specify, through the dedicated wizard, the host's genome and a reference genome, necessary for mapping and for functional genome annotation.



A link to Check analysis will be created after launching the requested pipeline. The system will notify the user after a succesful pipeline launch and once execution has ended.

Results

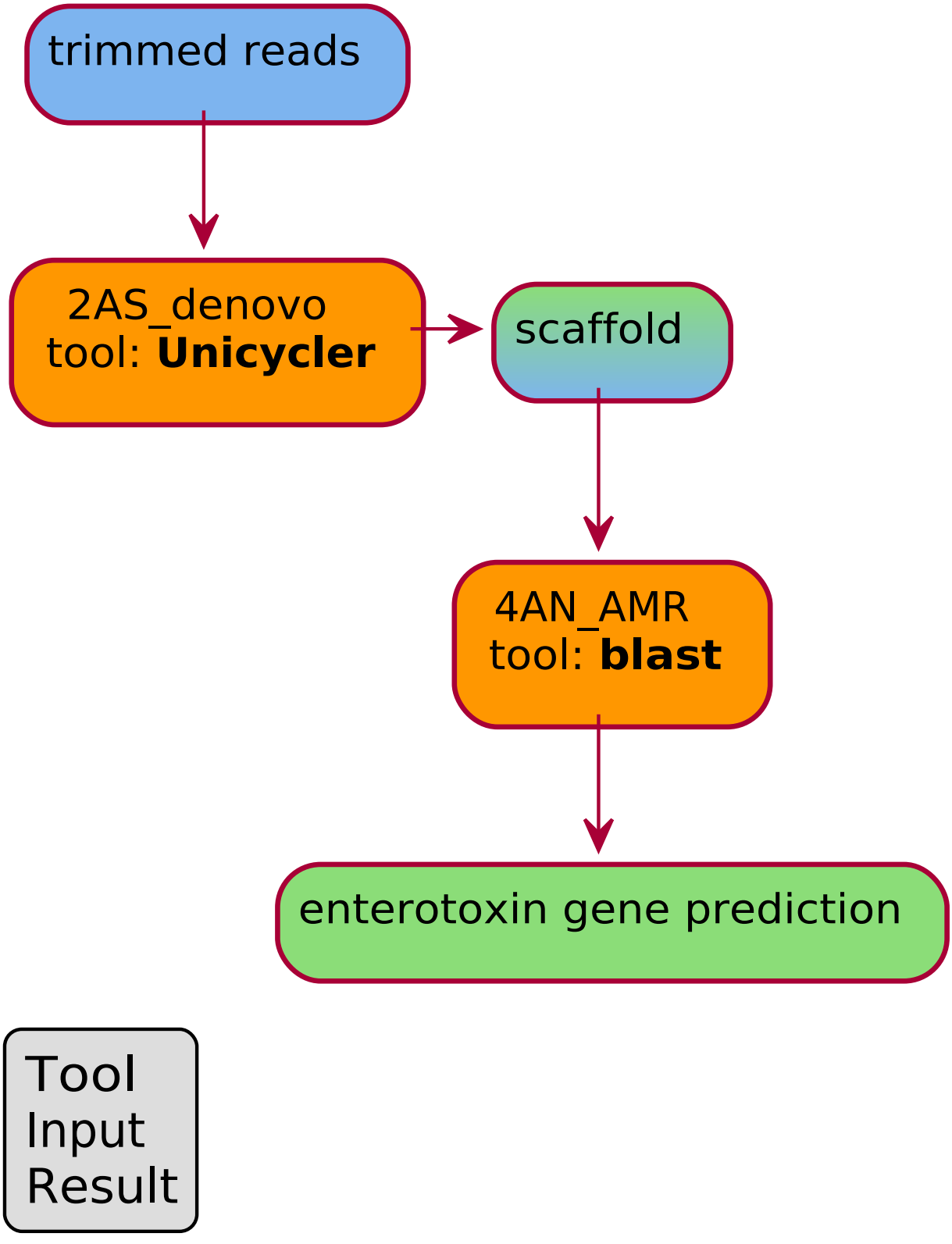
Output files from the "Genome Draft" pipeline are the same as those produced by the single analyses included in it, in the same directory structure:

- Output of mapping
- Output genome annotation

Enterotoxin *S. aureus* Finder Pipeline

Introduction

The **Enterotoxin *Staphylococcus aureus* Finder** pipeline detects the gene coding for enterotoxin in the genome of *S. aureus*, running first the *de novo* assembly (2AS_denovo), followed by typing 4AN_AMR.



Run Enterotoxin *S. aureus* Finder Pipeline

The filter at the top of the run analyses page allows to display only the pipelines. Once the **Enterotoxin *S. aureus* Finder** pipeline has been selected, a confirmation interface will be displayed.

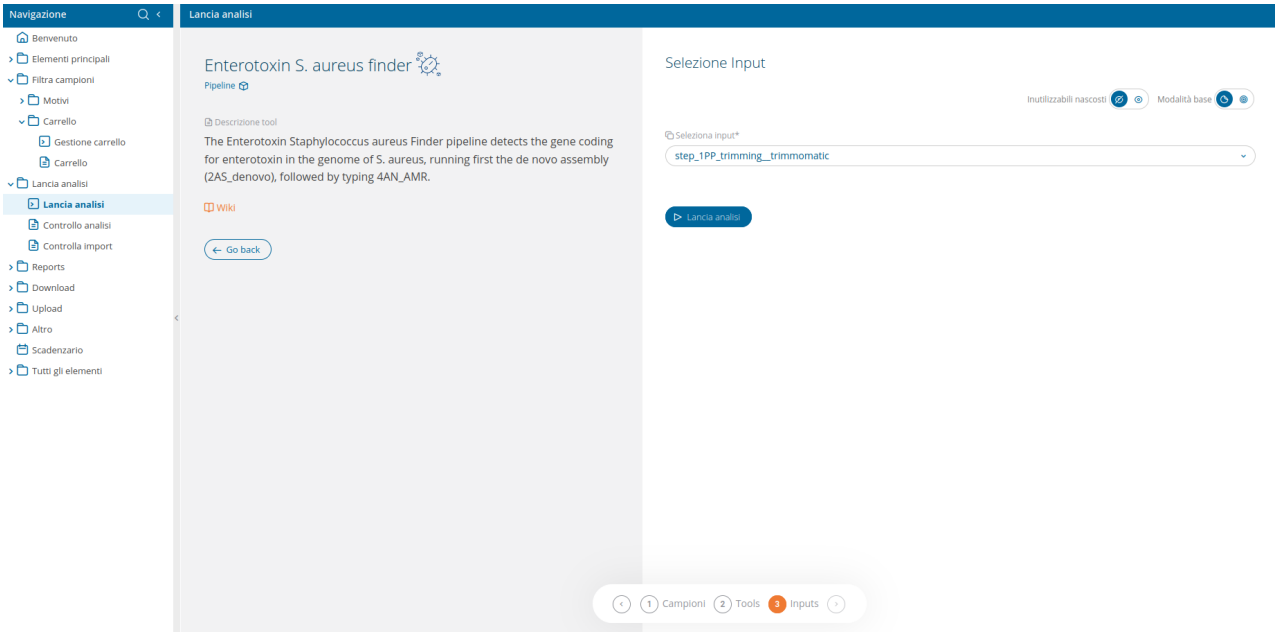
The pipeline performs the following 2 analyses:

1. *de novo* assembly with **Unicycler** (2AS_denovo__unicycler)
2. detection of enterotoxin-coding gene with **blast** (4AN_AMR__blast)

Accepted inputs can be 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 UI delivers an advanced input selection mode, to allow selection of all types of supported input files at once.



A link to Check analysis will be created after launching the requested pipeline. The system will notify the user after a succesful pipeline launch and once execution has ended.

Results

Output files from the "Enterotoxin *S. aureus* Finder" pipeline are the same as those produced by the single analyses included in it, in the same directory structure:

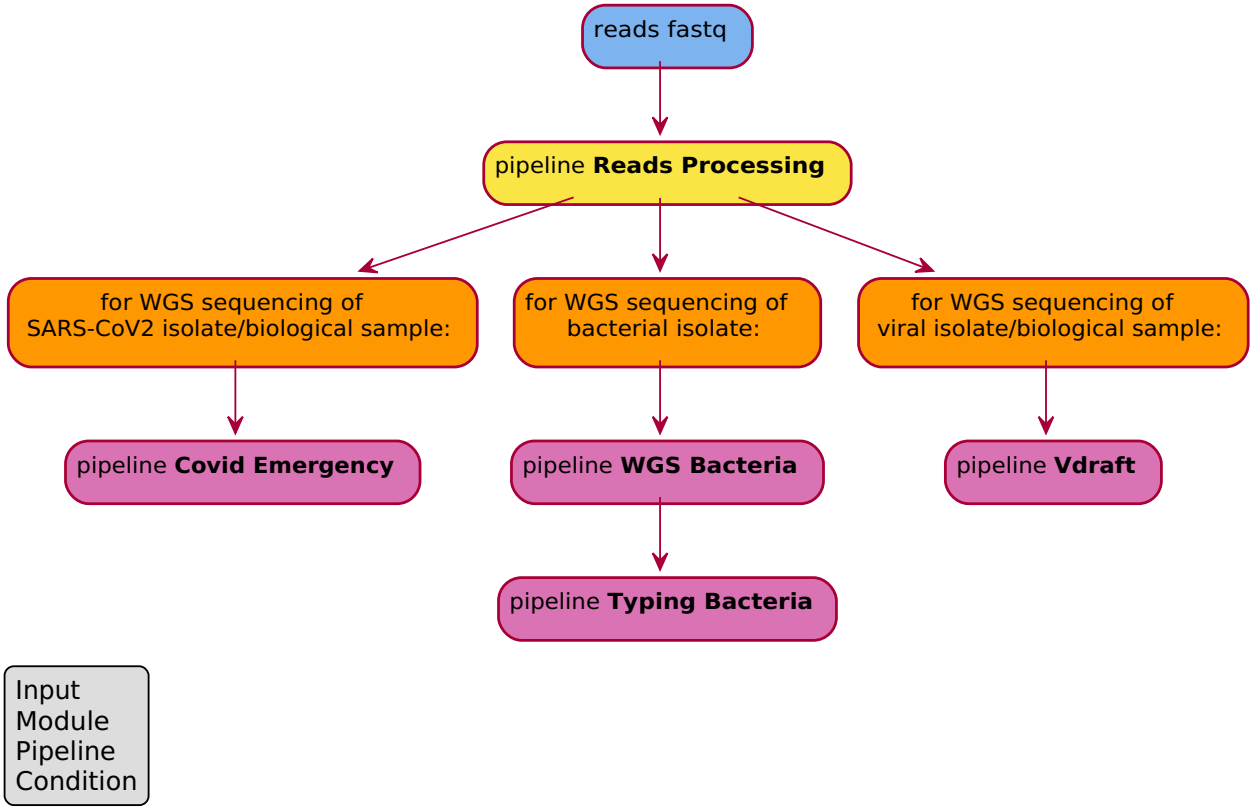
- Output of *de novo* assembly with Unicycler
- Output of enterotoxin-coding gene detection

NgsManager Pipeline

NOTE: some of the NgsManager pipeline's components are not yet implemented in Cohesive Demo's current version.

Introduction

The NgsManager pipeline manages reads and executes subordinate pipelines depending on the sample type.



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.

Depending on sample type, **NgsManager** can execute the following pipelines:

- Raw Reads Processing
- For bacteria:
  - WGS on Bacteria
  - Typing on Bacteria
- For viruses:
  - Vdraft pipeline
- For SARS-CoV2:
  - Covid Emergency

(For more information on single analyses, please refer to the corresponding pages, available from the list above).

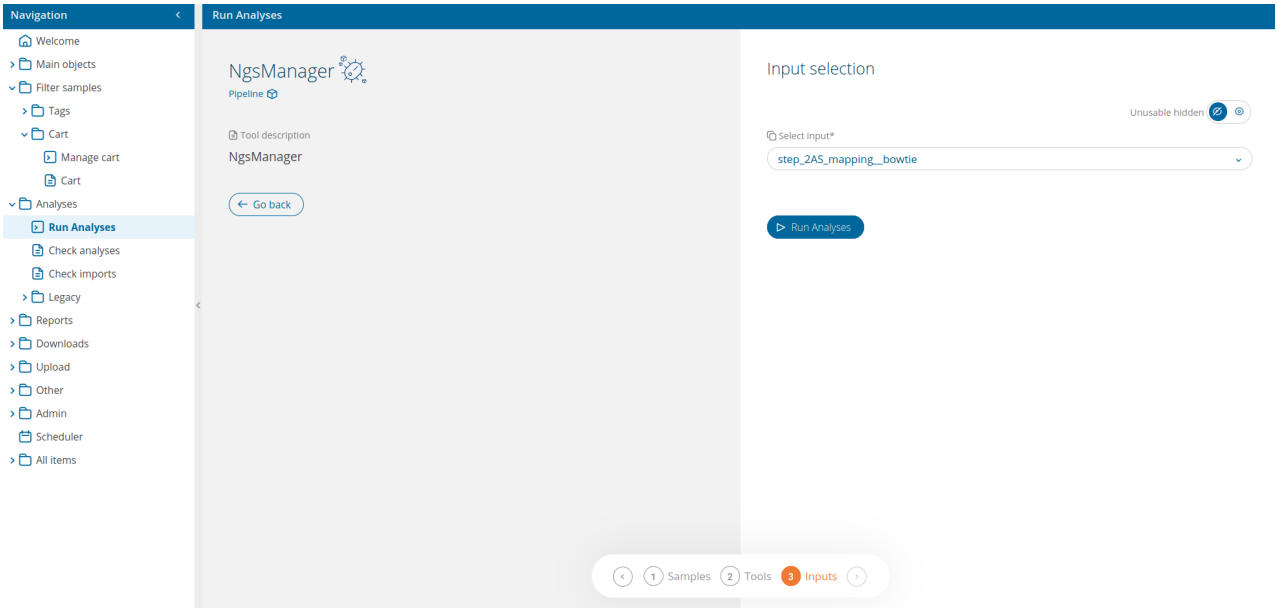
NgsManager's pipeline selection is controlled by choosing the appropriate sample type among the 3 available in the Parameters section.

Accepted inputs are:

- raw reads
- fasta from mapping (step_2AS_mapping)

In case 2AS_mapping is selected as input, NgsManager will also require the reference genome that has been used for mapping. The Raw Reads Processing pipeline will be executed on any sample type, provided that raw reads are used as input.

The input selection UI delivers an advanced input selection mode, to allow selection of all types of supported input files at once.



A link to Check analysis will be created after launching the requested pipeline. The system will notify the user after a succesful pipeline launch and once execution has ended.

Results

Output files from the "NgsManager" pipeline depend on the constituent pipelines that are executed, and are organized in the same directory structure. For details on the pipeline's output files, please refer to the output sections of the appropriate analyses:

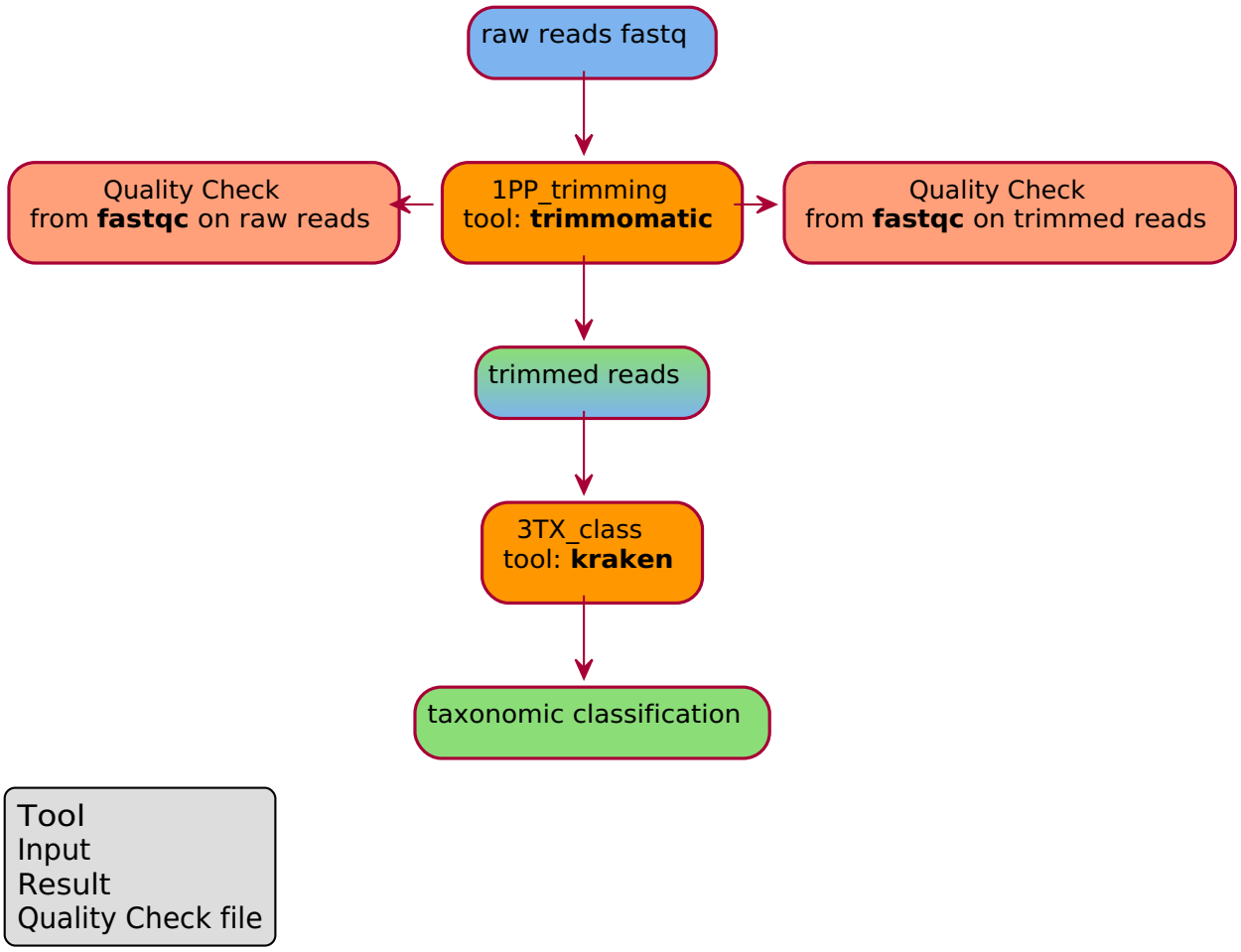
- Results of the pipeline for bacteria:
  - Results of the pipeline "WGS su Batteri"
  - Results of the pipeline "Typing su batteri"
- Results of the pipeline "Vdraft" for viruses
- Results of the pipeline "Covid Emergency" for SARS-CoV2

Raw Reads Processing Pipeline

Introduction

The **Raw Reads Processing** pipeline performs reads preprocessing (**1PP**) followed by taxonomic classification. Reads from the sequencer undergo a quality check with fastQC and trimming with Trimmomatic (1PP_trimming); subsequent taxonomic analysis is executed with Kraken (3TX_class).

> Note that in 1PP_trimming quality check with **fastQC** is performed both on raw reads and the obtained trimmed reads.



Run Raw Reads Processing Pipeline

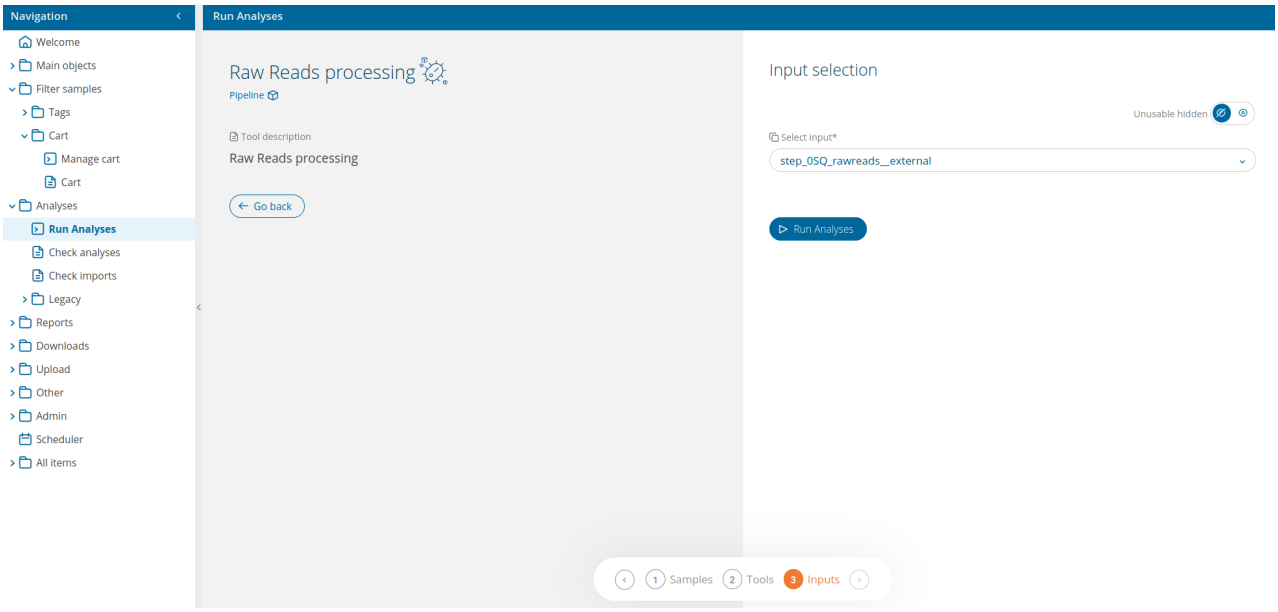
The filter at the top of the run analyses page allows to display only the pipelines. Once the **Raw Reads Processing** pipeline has been selected, a confirmation interface will be displayed.

The pipeline consists of 2 steps:

1. Trimming of the raw reads to remove low-quality bases with **trimmomatic** (1PP_trimming__trimmomatic) and **fastQC**;
2. Taxonomic classification of identified bacteria or viruses with **Kraken** (3TX_class__kraken).

Accepetd inputs are raw reads in fastq format.

The input selection UI delivers an advanced input selection mode, to allow selection of all types of supported input files at once.



A link to Check analysis will be created after launching the requested pipeline. The system will notify the user after a succesful pipeline launch and once execution has ended.

Results

Output files from the "Raw Reads Processing" pipeline are the same as those produced by the single analyses included in it, in the same directory structure. For details on the pipeline's output files, please refer to the output sections of the appropriate analyses:

- Output del trimming delle *reads* eseguito con trimmomatic e *quality check* di fastQC
- Output della classificazione tassonomica eseguita con Kraken e del *quality check* di Quast

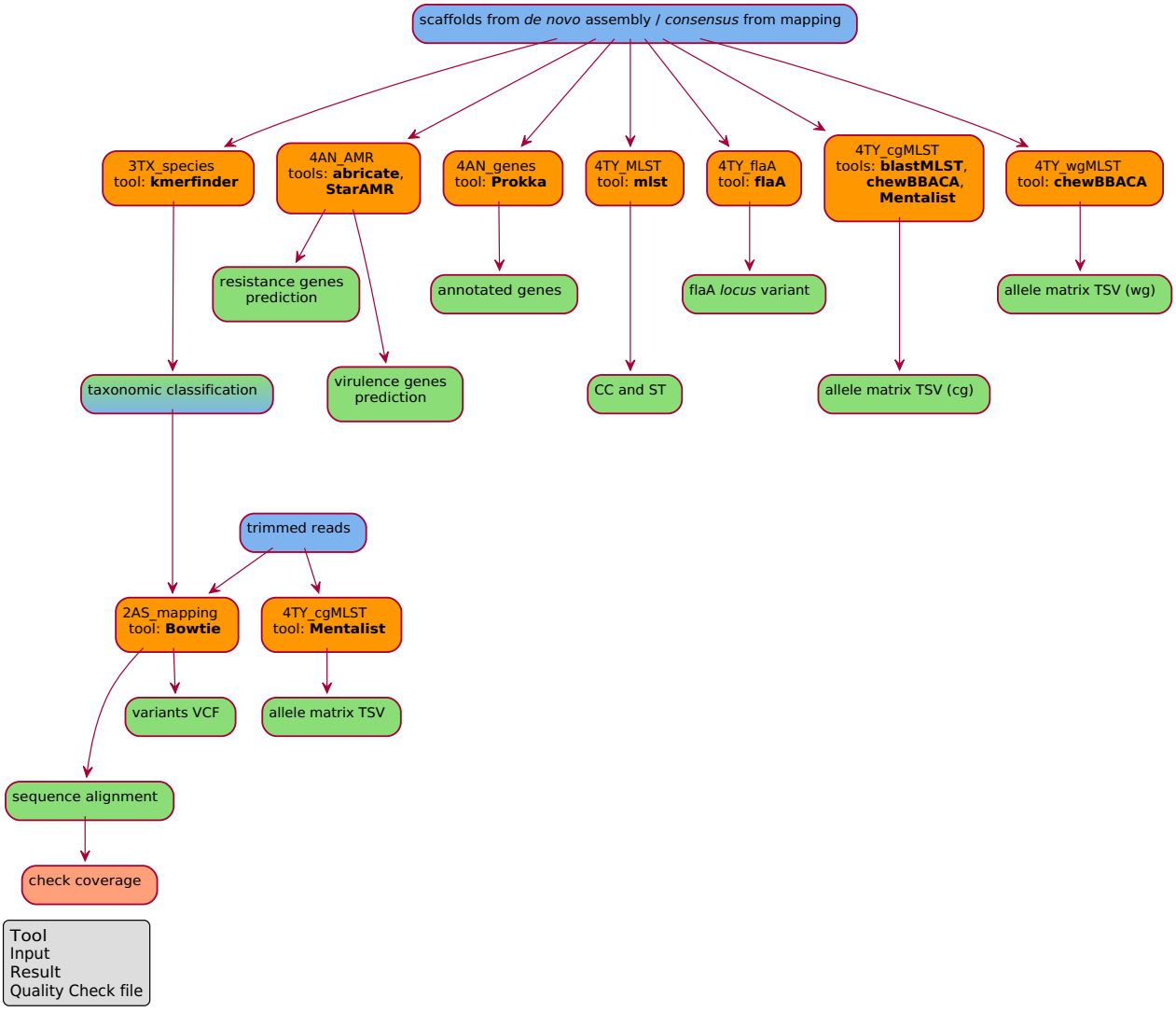
Pipeline Typing on Bacteria

NOTE: some of the Typing on Bacteria pipeline's components are not yet implemented in Cohesive Demo's current version.

Introduction

The pipeline **Typing on Bacteria** executes all typing analyses for bacteria, including:

- calculation of species ;
- identification of antibiotic resistance- and virulence-related genes;
- identification of flaA *locus* variant (*Campylobacter* only);
- MLST;
- core-genome MLST;
- whole-genome MLST;
- genome annotation;
- calculation of horizontal and vertical coverage.



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 pipeline consists of the following analyses:

1. 3TX_species
2. 4AN_AMR
3. 4TY_flaA
4. 4TY_MLST
5. 4TY_cgMLST
6. 4TY_wgMLST
7. 4AN_genes

Sui risultati di kmerfinder (3TX_species) vengono inoltre eseguite:

- 4TY_cgMLST
- 2AS_mapping

The software used for **4TY_cgMLST** and **4TY_wgMLST** is selected based on the schemas for microorganisms that are available for each tool, as explained in the appropriate Wiki pages:

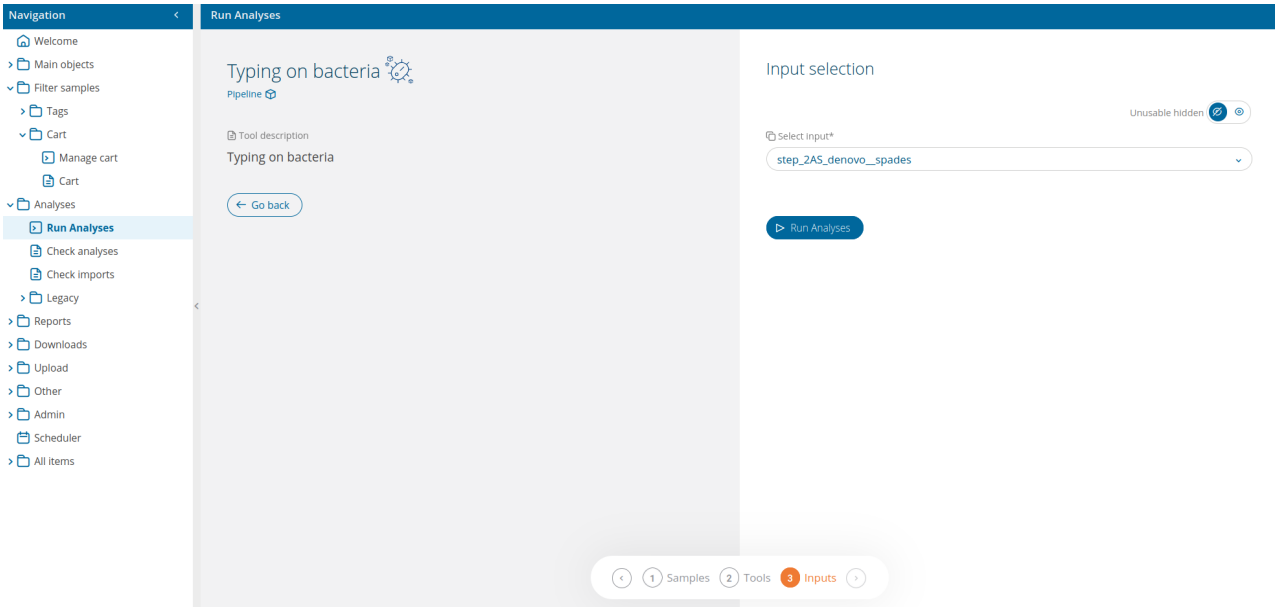
- Available schemas for cgMLST
- Available schemas for wgMLST

Similarly, **4TY_MLST_mlst** is executed only if the appropriate microorganism schema is available (the full list of schemas is accessible at the official mlst guide on GitHub), **4TY_flaA** is only executed on *Campylobacter* samples and **4TY_cgMLST_mentalist** on trimmed reads is only executed in case of *Listeria monocytogenes* samples.

Accepetd inputs are *consensus* sequences (.fasta format) obtained from mapping (2AS_mapping) or scaffolds from *de novo* assembly (2AS_denovo).

In case 2AS_mapping is selected as input, the pipeline will also require a reference genome among those available.

The input selection UI delivers an advanced input selection mode, to allow selection of all types of supported input files at once.



A link to Check analysis will be created after launching the requested pipeline. The system will notify the user after a succesful pipeline launch and once execution has ended.

Results

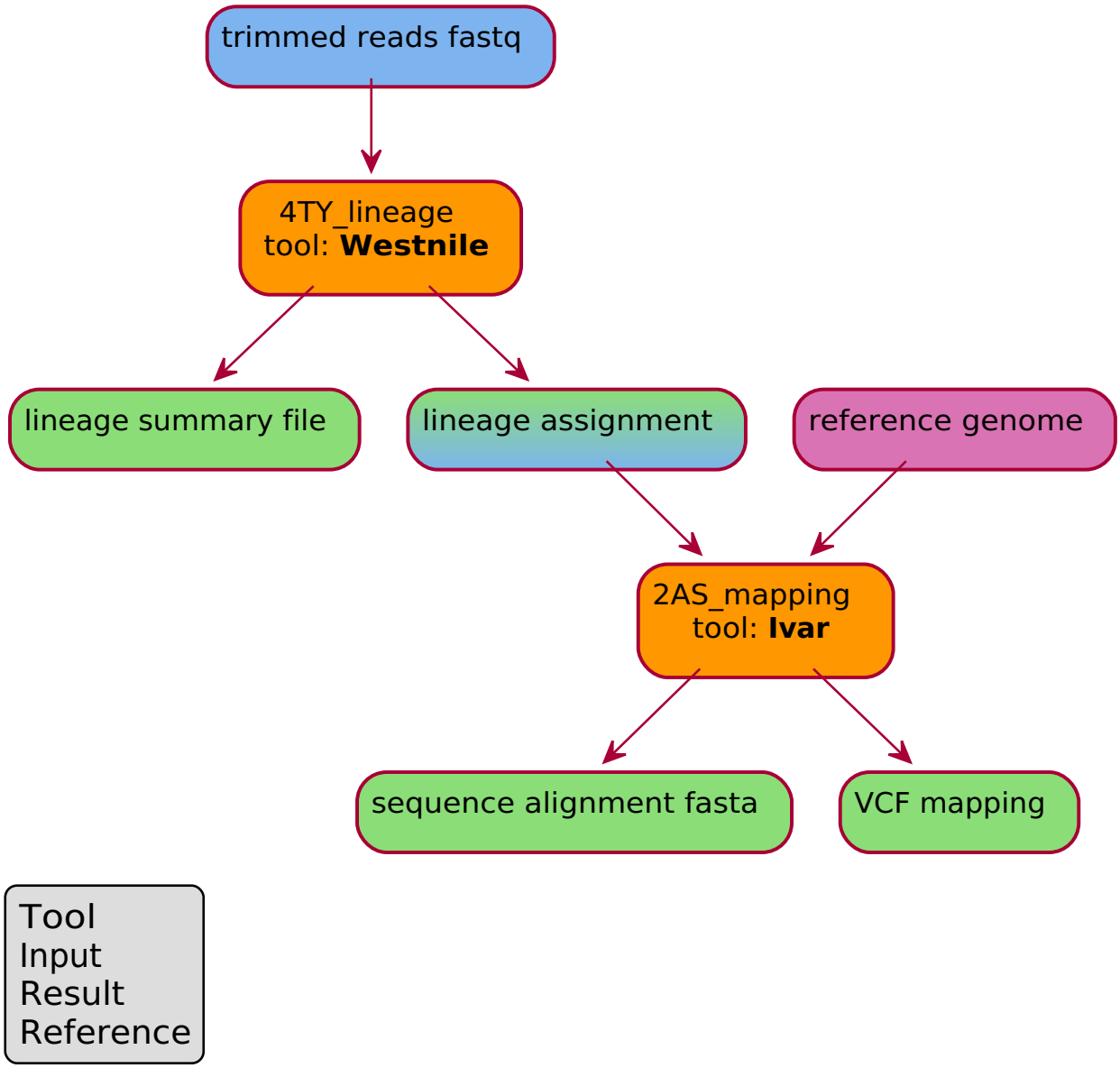
Output files from the "Typing on Bacteria" pipeline are the same as those produced by the single analyses included in it, in the same directory structure. For details on the pipeline's output files, please refer to the output sections of the appropriate analyses:

- Output of species assignment
- Output of antibiotic-resistance and virulence prediction
- Output per la determinazione della variante del *locus* flaA
- 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 computes lineage for West Nile Virus maps the sample to the appropriate 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.

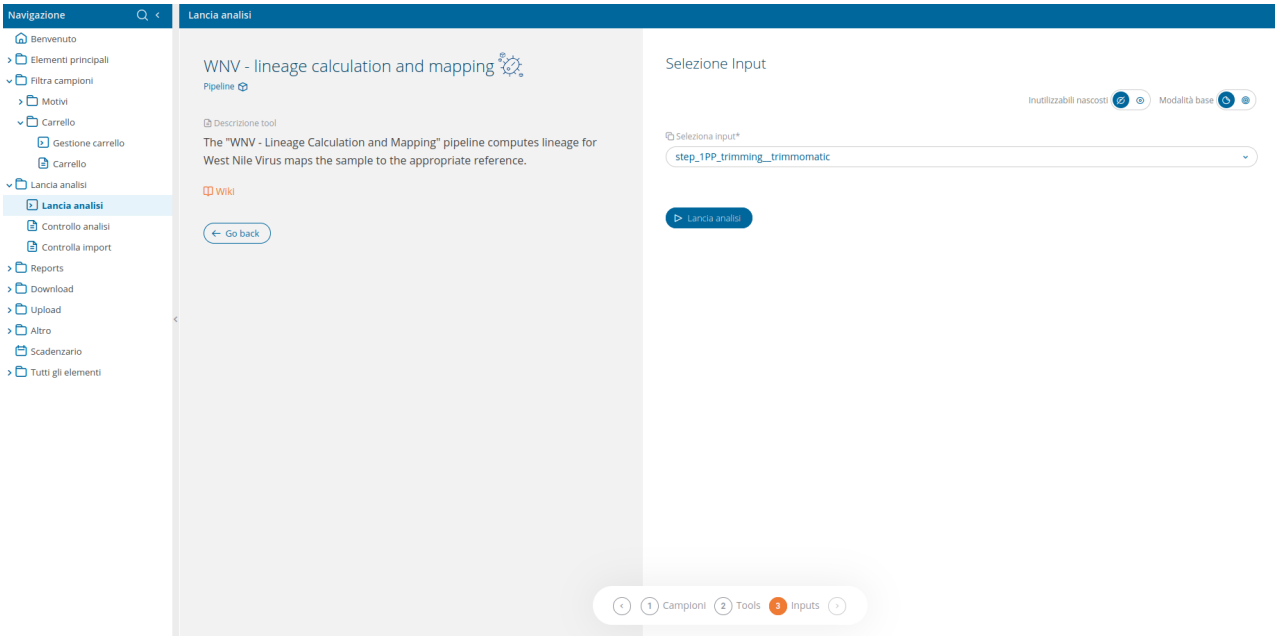
The **WNV** pipeline consists of 2 steps:

1. West Nile Virus lineage computation with **Westnile** (4TY_lineage__westnile);
2. Mapping of all the provided genome fragments with **Ivar** (2AS_mapping__ivar), using multiple *references*.

Accepted inputs can be 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 UI delivers an advanced input selection mode, to allow selection of all types of supported input files at once.



The reference genome used for mapping will be the base genome corresponding to the lineage assigned by the **Westnile** tool, following the relationships 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

A link to Check analysis will be created after launching the requested pipeline. The system will notify the user after a succesful pipeline launch and once execution has ended.

Results

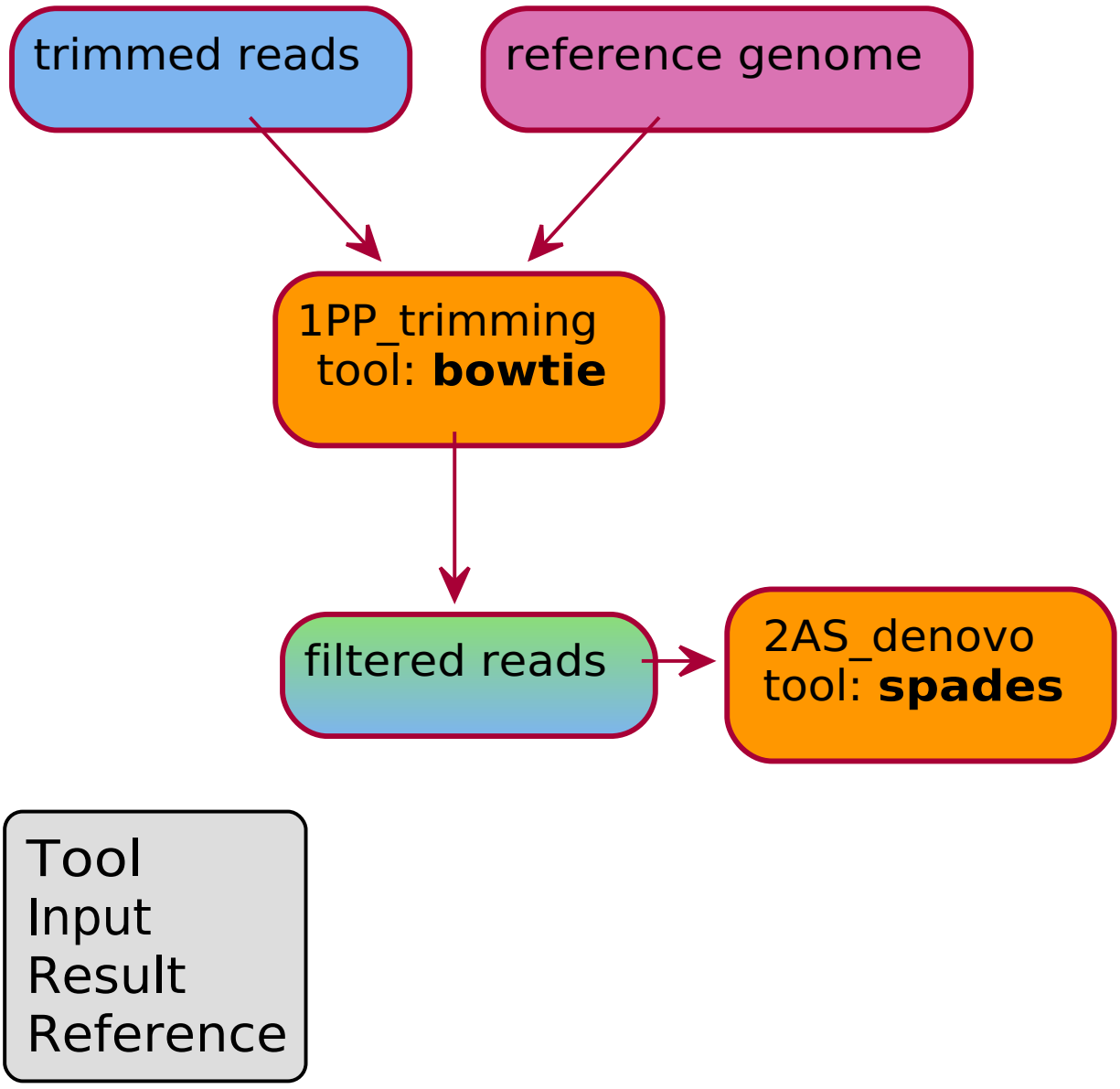
Output files from the "WNV" pipeline are the same as those produced by the single analyses included in it, in the same directory structure. For details on the pipeline's output files, please refer to the output sections of the appropriate analyses:

- 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 on the selected reference (1PP_filtering) and performs *de novo* assembly (2AS_denovo).



Run Filtering & *de novo* Pipeline

The filter at the top of the run analyses 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:

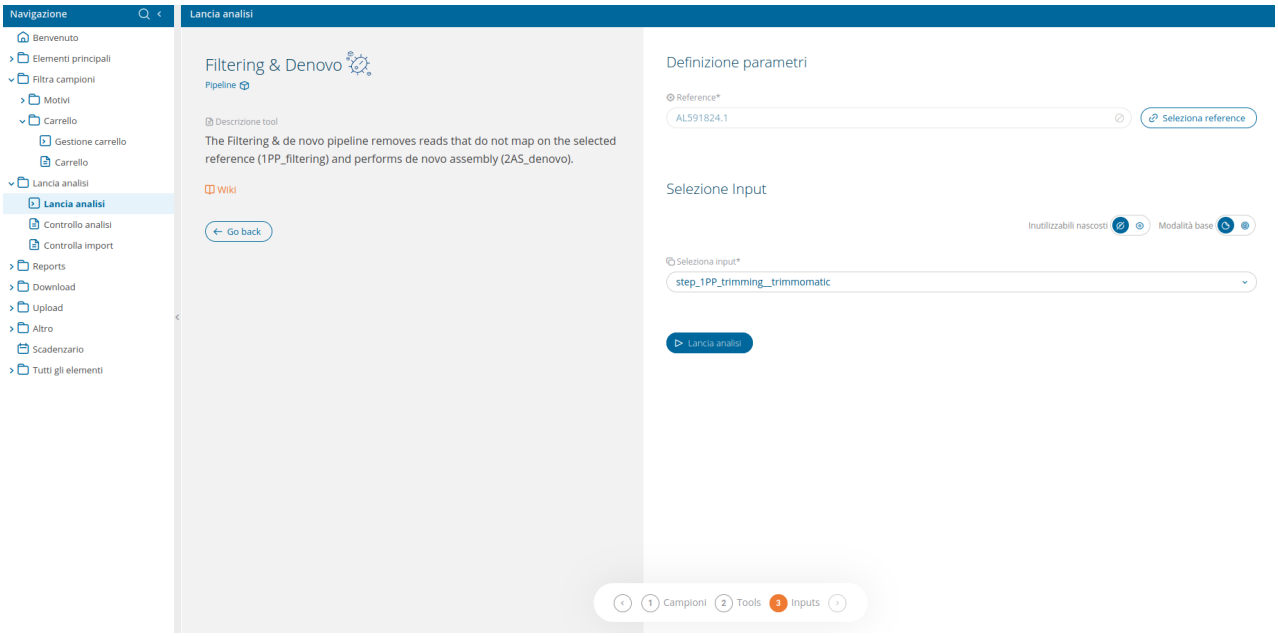
1. removal of reads that do not map to the reference with **Bowtie** (1PP_filtering__bowtie)
2. *de novo* assembly with **SPAdes** (2AS_denovo__spades)

Accepted inputs can be 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 UI delivers an advanced input selection mode, to allow selection of all types of supported input files at once.

The pipeline also requires to specify, through the dedicated wizard, a reference genome to use for the first analysis of the pipeline: *filtering* (1PP_filtering).



A link to Check analysis will be created after launching the requested pipeline. The system will notify the user after a succesful pipeline launch and once execution has ended.

Results

Output files from the "Filtering & *de novo*" pipeline are the same as those produced by the single analyses included in it, in the same directory structure:

- Output of reads filtering using Bowtie
- Output of *de novo* assembly with SPAdes

5. Reports

5.1 Reports

From the side menu are available tools to see reports. The standard reports method requires either a collection of IDs in the cart or an active tag to specify the entries on which to produce 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.

The following videotutorial shows the procedure on how to see reports by using an active tag.

See reports with active tag

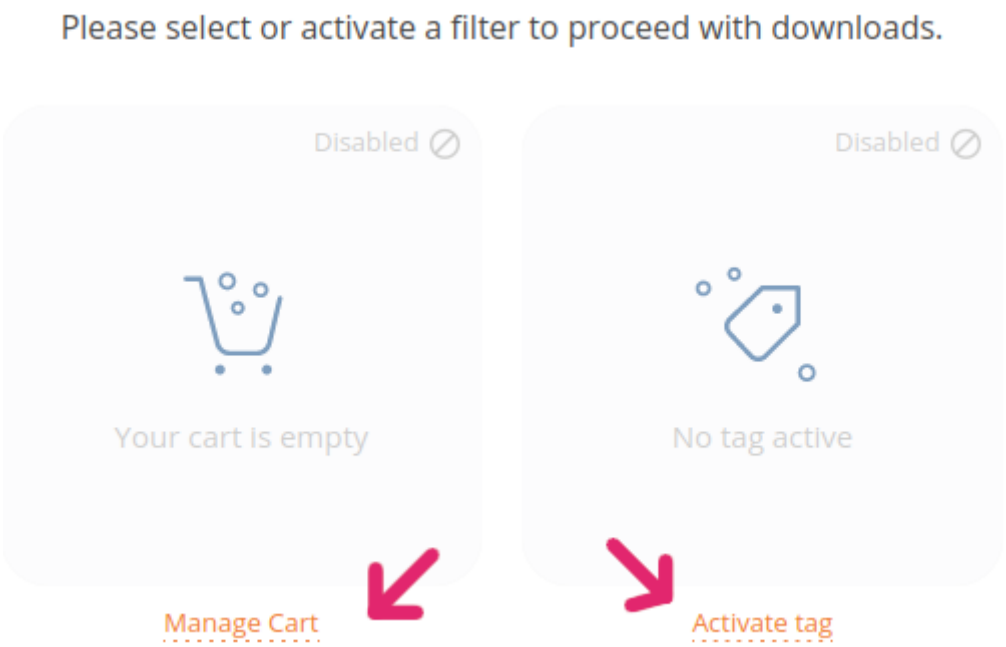
type:video-tag Reports_Video.mp4

6. Downloads

6.1 Downloads

The user can access download tools from the **Downloads** section in the left-side menu. The standard download method requires either a collection of IDs in the Cart or an active tag to specify the entries to download.

The quick links below the Tag and Cart buttons allow to edit samples in the Cart or change active tag on the fly.



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.

After selection of samples, users will be able to filter and select the kind of file(s) they wish to download, using the buttons in the top-left corner. Two main tools are available to download files:

- **Custom download**, which allows users to select freely results from any combination of analysis and tool available;
- **Preset download**, which provide shortcuts to the most downloaded analysis results.

Note: analysis results will be available for download only if previously **imported**.

6.1.1 Selective download

After selection of the analysis results to download, users can use the available checkboxes to choose which files, produced by the bioinformatic tool, they wish to download (more on the topic in the videotutorial below):

Thanks to the selective download functionality, both the generated `wget` command and the link to the directory will gather only the selected files, thus avoiding large downloads.

6.1.2 Custom download

The videotutorial below demonstrates the download workflow using the active Cart or tag to select samples. The Custom download is the download method which grants a greater level of personalization.

type:video-tag Genpat-download.mp4

Note 1: just as for running analyses, the Cart is automatically activated in order to carry on with the download.

Note 2: the **Check download** section in the navigation menu lists detail cards for all download processes that have been launched.

Nota 3: in the top-right corner, next to the **Load utilities** button, there is the **Load RIS_CD** functionality. **Load RIS_CD** allows usage of codes of analysis that are finished and imported (pasting codes in the text field or uploading a csv file). This functionalities grant the ability to directly download older analyses. More information on "RIS_CD" usage to run analyses or more download features are available in the "Export RIS_CD" section of the Wiki.

6.1.3 Preset download

The **Preset** download method differs from the Custom download in 2 characteristics:

- it avoids manual specification of all necessary parameters;
- is only available for a subset of the possible analyses (the most downloaded).

The **Preset** download is designed to facilitate the workflow of users who routinely download specific files: all parameters are preset, to quickly reach the download, reducing the number of clicks from the user.

Files available as **Preset** downloads include:

- Raw reads
- Trimmed reads from `trimmomatic`
- Trimmed reads from `fastp`
- *De novo* assembly from `shovill`
- *De novo* assembly from `spades`
- Virus *consensus* from `ivar`
- Antibiotic-resistance profiles from `*AMR` (starAMR)
- FlaA profiles from Multilocus sequence typing with the `flaA` software
- wgMLST allelic profiles from `chewBBACA`
- cgMLST allelic profiles from `chewBBACA`

Note: running an analysis creates a link to the process' card; for some analyses it will be possible to access some files, visualization and download options directly from the process card. More information on the generic workflow are available at the following Wiki pages:

- Run analysis steps;
- Check analysis.

6.1.4 Check Downloads

The **Check Downloads** page stores the user's download history. Each record contains useful information on the process, such as IDs, links and the `wget` command to repeat the download.

Like the elements in other tables, each card can be expanded to show additional details.

6.2 Download methods for Windows

Download features mainly rely on `wget` commands generated automatically by the download system; such commands cannot be used natively in Microsoft Windows. In this page are listed a few methods to allow Windows users to perform downloads of multiple or large files.

6.2.1 Method 1: WSL

Starting from Windows 10, Microsoft supports usage of the GNU/Linux Terminal through the Windows Subsystem for Linux (WSL), a tool for virtualization of the Linux kernel in a Windows environment.

WSL does not require to use additional softwares, since it is a Windows 10/11 functionality that only needs download and activation. Official guides to WSL and instructions on how to install it are available on Microsoft's own websites:

- https://learn.microsoft.com/en-us/windows/wsl/about (what is WLS)
- https://learn.microsoft.com/en-us/windows/wsl/install (installation guide)

After successful installation and activation of WLS on a Windows 10/11 machine, open a GNU/Linux Terminal. It will be possible to run Bash commands from it, so paste the generated `wget` command to download the desired files.

6.2.2 Method 2: wget.exe

A similar result can be obtained using an external program to execute the UNIX command `wget` directly from Windows's command line.

1. Download the "wget.exe" program through this external link;
2. Save or move the wget.exe file in the "Download" folder;
3. Open the Windows command line (**cmd** or **Command Prompt**, available in the Start Menu), then move inside the "Download" folder using the command below:

   ```
   cd Download
   ```

4. Copy the `wget` command generated by the download page in the Command Prompt and run it;
5. After execution, files will be located in the "Download" folder, inside the subfolder named as specified in the `wget` command (*e.g.*: "f0086d8f-948d-4753-ad85-5171f1a5dda8").

6.2.3 Method 3: Browser extension "DownThemAll"

The web browser extension "DownThemAll!" can be added to Firefox, Chrome or Edge to download multiple files from a web page.

1. "DownThemAll!" can be obtained from the following external link: https://www.downthemall.org/;

2. In the "Get it!" section, select the browser and version and install the software:

1. Once the extension has been installed, when exploring a results folder it will be possible to use the "DownloadThemAll!" add-on to download all files listed in the page (the extension icon is usually located in the top-right corner of the browser's function bar):

1. The pop-up window enables filtering and selection of the files before launching the download:

7. Upload

7.1 Introduction

The Cohesive system includes 2 main upload features (`Upload > Upload page` from the Navigation Menu): **Upload from NCBI** and **Manual upload**.

The upload wizard guides users in a 2-steps procedure to upload files to the information system:

1. **Source**: guides through the selection of the upload method, "Upload from NCBI" and "Manual upload";
2. **Upload**: supports in the selection of the type of file to upload.

Two more ancillary areas are available:

- The "**Prepare upload**" page (`Upload > Prepare upload` from the Navigation Menu) is dedicated to preparation and modification of sequence files' metadata. Metadata are a requirement for manual upload; (in the main Upload page is also available the quick link "Manage metadata" to the **Prepare upload** page).
- The "**Check upload**" (`Upload > Check upload` from the Navigation Menu) shows the user's upload history.

7.1.1 Upload Functionalities

- **Upload from NCBI Wiki page**
- **Manual upload Wiki page**

Upload from NCBI grants import of sequence files directly from NCBI's databases. This upload functionality only requires a list of NCBI Accession Numbers.

Manual upload is the main tool to upload sequence files from a user's device. This upload functionality requires:

1. compilation and validation of metadata, to be submitted in the **Prepare upload** page;
2. the sequence file(s) or a `.zip` archive containing the sequence i files corresponding to the validated metadata.

7.2 Upload Functionalities

7.2.1 Upload from NCBI

The **Upload from NCBI** features allow upload of one or more sequence files from NCBI (National Center for Biotechnology Information) databases to Cohesive. Cohesive can query 3 NCBI databases through 3 different upload options:

- **Assembly fasta**, to import fasta files from NCBI's "Assembly" database;
- **Nucleotide fasta**, to import fasta files from NCBI's "Nucleotide" database;
- **Reads fastq SRA**, to import fastq files from NCBI's "Sequence Read Archive - SRA" database.

Upload of files from NCBI requires only a list of NCBI Accession Numbers: users will choose from which databes the file should be imported, select the organism's species and paste the Accession Numbers in the text field.

Note: multiple sequences can be uploaded all at once only if they belong to the same **species**; organisms of different species will require a separate upload request.

The following video shoes how to upload files from NCBI. All 3 upload types (Assembly, Nucleotide and SRA) share the same requirements and workflow.

type:video-tag CIS-upload_NCBI.mp4

7.2.2 Manual upload

The **Manual upload** services are for submission of sequence files from the user's device to Cohesive, contingent upon prior compilation of metadata. Accepted sequence files belong to one of the following upload categories:

- **Fasta**, for direct upload of **fasta** sequences;
- **Raw reads Illumina**, to upload **fastq** raw reads from Illumina apparatus (**short, paired-end reads**);
- **Raw reads Iontorrent**, to upload **fastq** raw reads from Iontorrent apparatus (**long, single-end reads**);
- **Raw reads Nanopore**, to upload **fastq** raw reads from Nanopore apparatus (**long, single-end reads**).

Access to Manual upload can be granted only if **valid metadata** for the files to upload have been submitted in the **Prepare upload** section. Users will then be able to select the file type and the technology the reads were produced with (in case of raw reads), to finally choose the sequence file(s) or the `.zip` archive containing the files to upload (through drag-and-drop or by clicking to open a file explorer).

Users can also opt to automatically run a pipeline tailored to the specific sample type upon completion of the upload process.

> **Note:** please look up the Prepare upload Wiki page for further details on how to submit files and prepare corresponding metadata.

The following video shows the available Manual upload functionalities. The common prerequisite is metadata submission.

type:video-tag CIS-manuale_upload.mp4

> **Note:** uploads may require some time for execution; closing the web browser tab will stop the upload process. Update of the Check upload page may also not be immediate. Users are notified of process completion through the integrated notification system.

7.3 Prepare upload

Metadata associated to samples to submit are necessary for the manual upload process. In the "**Prepare upload**" page (`Upload > Prepare upload` from the Navigation Menu), users can add cards with sample metadata, by means of 2 available methods:

- Add from UI (User Interface)
- Add from file

Metadata cards associated to samples will have to be validated by Cohesive, which will ensure absence of mistakes and presence of minimum mandatory data. After the validation request ("Validate" button), the validation state will be displayed in the hononymous column.

7.3.1 Add metadata from UI

The "Add data" button empowers users to quickly create a metadata card for a single sample. The card will have to be filled manually with at least the minimum mandatory metadata for the sample.

The following video shows how to fill a metadata card and some other useful features of metadata card management.

type:video-tag CIS-upload_metadata_card.mp4

UI tools can also be used to **modify** one or more existing metadata cards simultaneously, regardless of how those cards were added to the list; this feature also allows spot corrections of invalid sample metadata, as shown in the following video:

type:video-tag CIS-upload_metadata_card_2.mp4

7.3.2 Add metadata from file

The "Add from file" button allows users to submit a tabular file containing the list of samples to upload and corresponding metadata. Depending on the upload type (*e.g.*: "upload of official samples" as opposed to "default upload"), the upload wizard will generate a link to download the corresponding template.

The following video shows how to submit a metadata file.

type:video-tag CIS-upload_metadata_file.mp4

Metadati table templates

Templates of default metadata tables are available for download at the following links:

- Template_default.tsv
- Template_default.csv

7.4 Check upload

The "**Check upload**" page (`Upload > Check upload` from the Navigation Menu) hosts the user's upload history.

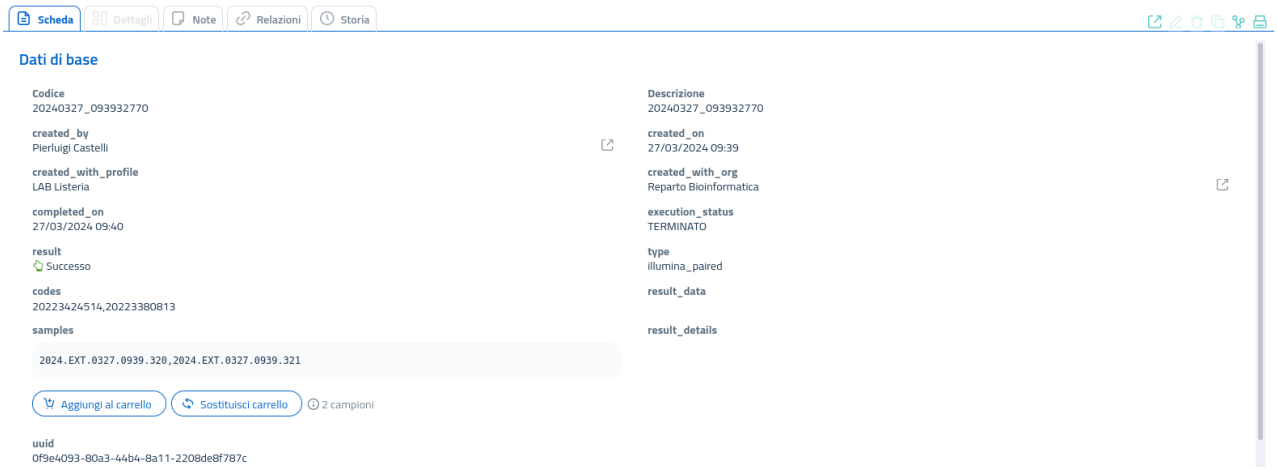
All elements in the **Check upload** table can be interacted with in the same way as for all other Cohesive tables.

Uploads can conclude with 3 possible exit statuses: **Success**, **Failure** or **Warnings**.

7.4.1 Success

The exit status is "**success**" when all submitted files caricati fit the corresponding metadata. When the upload exits with "success" the uploaded sample(s) metadata cards are removed from the "**Prepare upload**" page and paired to the corresponding files.

Cards of successful uploads bear the sequnces' newly generated internal IDs (field "samples") and their orignal che IDs (field "codes"). Both fields carry the "Add to Cart" button.



7.4.2 Failure

The exit status is "**failure**" when all of the submitted files raise an error. When an upload fails, metadata cards are not removed from the Prepare upload page, so that they remain available for revision (as shown in this video).

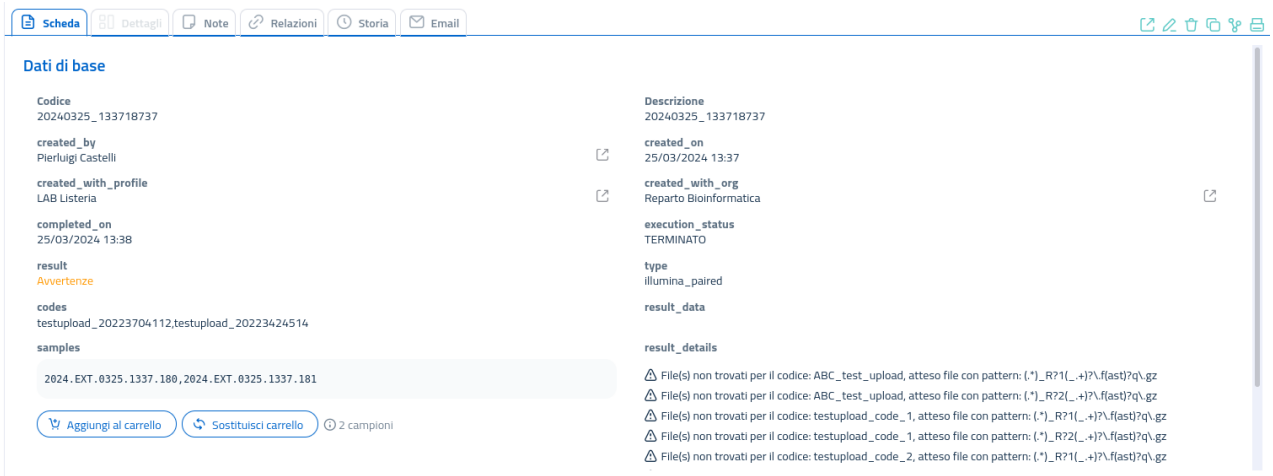
The card of a failed upload gives information regarding the failure cause(s), listed in the "result_details" field. In the image below, for example, the process execution failed because the names of all files were either wrong or not corresponding to those listed in the validated metadata:

7.4.3 Warnings

Uploads fail only if all submitted files raise errors: if even just one file is valid, it will be uploaded and the exit status of the process will be "**warnings**".

Cards of uploads with "warnings" exit status list codes of samples that were successfully uploaded (which can be added to the Cart), but also error reports for failed uploads (field "result_details"). Only cards corresponding to metadata of successfully uploaded samples are paired to the new accession numbers and removed from the Prepare upload page's table, while all others are qued for manual revision or deletion.

In the image below, some files in the uploaded zip archive had invalid filenames, while 2 files were valid. The 2 valid files were uploaded, the others generated warning messages.



7.5 Zip and metadata rules

Metadata files and manually-created metadata cards have to follow some rules in order to be correctly validated. This section of the Wiki lists and explains such rules.

> Please check the **FAQ section** in case of doubt.

7.5.1 Metadata

Add data

If metadata are submitted using the "**Add data**" button, or when metadata cards are modified manually, users are guided by a wizard to compile metadata. In this case only 2 conditions need to be met to ensure a successful metadata validation:

1. Submit the correct **Code** (*i.e.* **filename**);
2. Fill in at least the **minimum required metadata**.

1 - Code

The "Code" corresponds to the name of the fasta or fastq file to which the metadata refer.

- The code should include the actual file name **without extension**;
- The code should not contain blank spaces;
- In case of fastq of paired-end reads, the code should not include the read forward or read reverse identifier (*i.e.* "_R1" or "_1") and it should be **unique** (the metadata card refers to both reads in a pair).

2 - Minimum required metadata

Mandatory fields for metadata may vary depending on sample type; as an example, a public sample will required less mandatory metadata if compared to a sample marked as an official submission.

Minimum required metadata for a generic sample are:

- The filename or "Code";
- Sample type (depends on available metadata templates);
- Species.

> The sample type can be selected from a dropdown menu, while the species is selected from the pop-up table of species available in the information system's database.

Add file

Templates for metadata tables are available as `.tsv` and `.csv` files and can be downloaded from the **Prepare Upload page**, or directly from the "Add file" pop-up, as shown in the following video:

type:video-tag CIS-metadata_template.mp4

Each of the file's columns corresponds to a field available in the metadata card.

Depending on user profile, one or more templates may be available, one for each sample type that user is allowed to upload to Cohesive.

Note: always use the metadata template corresponding to the sample type to upload.

Templates include the header and an example line to demonstrate formatting and expected content of each field.

In case the selected template requires to fill in metadata about "species", "material", "host", "matrix" or "sampling point", such fields **should not** be filled with text or taxonomy: they require **the corresponding metadata code of Cohesive's database**.

Tables of metadata codes can be consulted in `Main elements > Metadata` **from the Navigation Menu.**

- Species: `Main elements > Metadata > Pathogen Species`
- Material: `Main elements > Metadata > Material`
- Host: `Main elements > Metadata > Host Species`
- Matrix: `Main elements > Metadata > Matrix`
- Sampling point: `Main elements > Metadata > Sampling Point`

Please refer to the **Metadata** section of this Wiki for more information on the topic.

7.5.2 Sequence files

Accepted extensions for submitted fasta files are: `.fasta` , `.fa` .

Accepted extensions for submitted fastq files are: `.fastq.gz` , `.fq.gz` .

Note: compression of fastq files in `.gz` archives is mandatory. Fasta files are accepted with or without `.gz` compression.

Zip archive

When uploading multiple sequences as a single compressed file, `.zip` compression has to be used. When loading the `.zip` file, the system will automatically check metadata cards, which will be paired to the corresponding samples. For this reason, the name of each fasta or fastq file needs to meet the rules discussed above.

In case the zip file contains files with names that are unpairable to any validated metadata card, the upload will exit with "failure" or "warnings" status.

Upload with as zip file can only be used if **the zip files contains files of the same type**, because each file type needs to be managed through a different upload procedure (*e.g.* upload of fastas versus upload of paired-end fastqs versus single-end fastqs).

> The zip file may contain folders, since the upload system supports retrieval of compressed files up to 10 levels of sub-folders in the archive.

8. Administration

8.1 Authorization and Authentication Rules

8.1.1 Access with username

- Each user is assigned a username and a password to log into the system.
- Each user is linked to one or more "profiles". Profiles determine data visibility and which features and functionalities can be accessed by the user.

The profile 'External User' has no limitation regarding species visibility, thus can be associated to all users in need to visualize all samples for which they requested upload to the system, with no restriction of species. It will be necessary to specify the latter behavior in the "Applicant-User Relationship" page (accessible by navigating "All Elements" > "Classes" > "Applicant-User Relationship" as Super User).

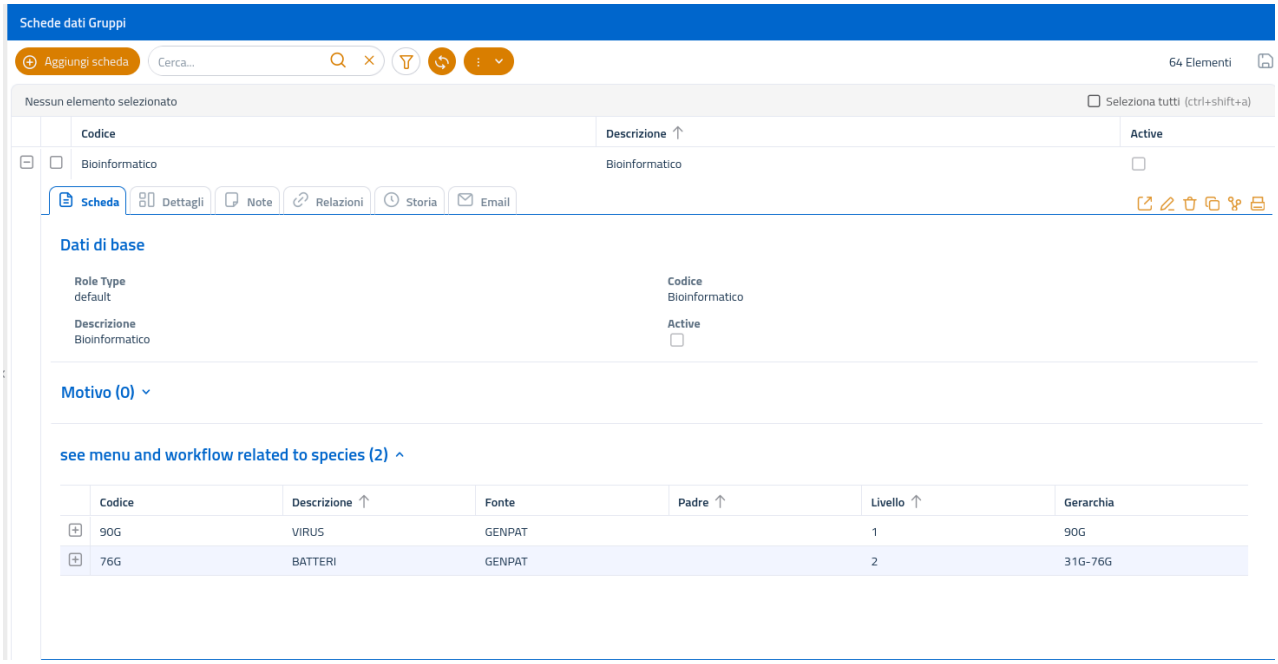
8.1.2 Restrictions on sample visibility

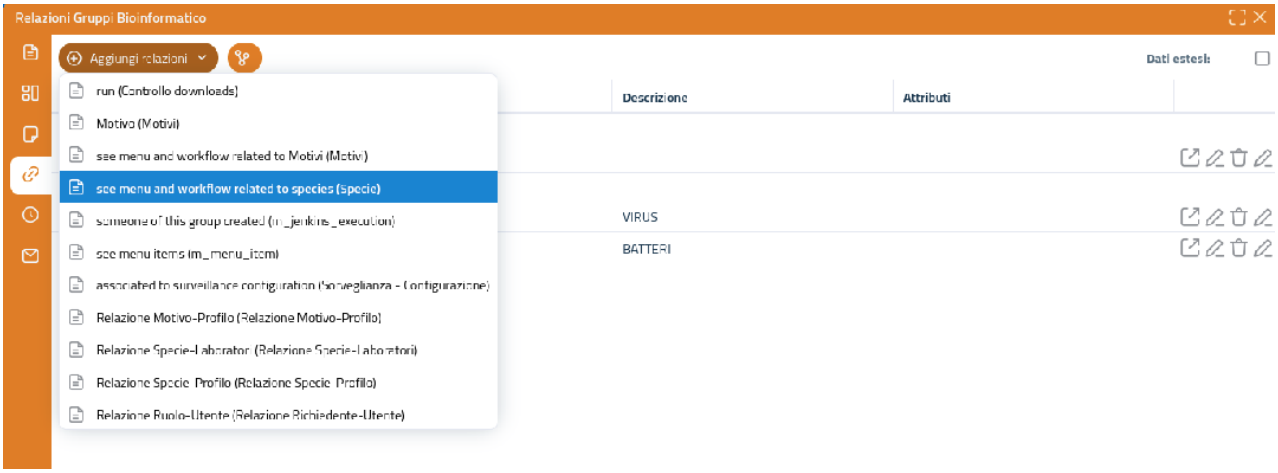
Every time a user logs into GenPat, 3 different levels of filtering are used to manage sample visibility for that user:

- **filter by Species**;
- **filter by Tag**;
- **filter by Applicant**.

Filter by Species

Each user is assigned one or more laboratory profiles. Each laboratory will be able to see only samples belonging to a set species or collection of species. Which species are visible to each laboratory is defined by the Super User through a mapping table (`All Elements > Classes > Groups > Relationships Table`), so each user will be able to see only samples of species linked to the laboratory profile they are logging in with.

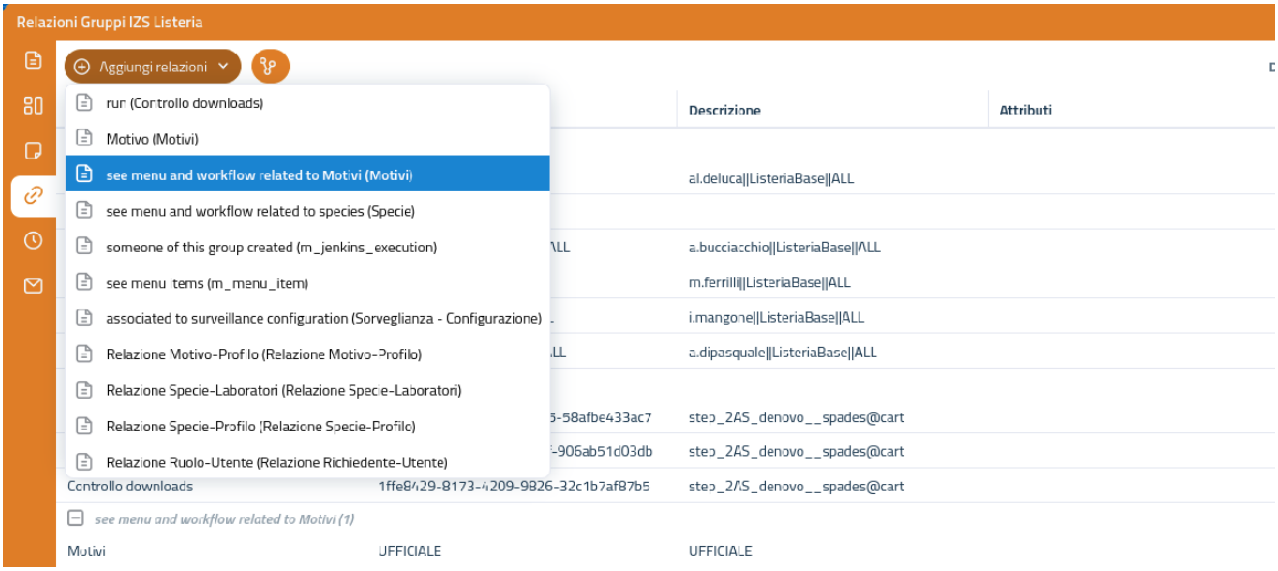




Filter by Tag

A guide to Tags is available in the Tags Wiki page.

Administrators can choose if a laboratory profile should be able to see only **official** samples: this behavior can be set by a Super User using the mapping table for laboratory profiles and Tags (`All Elements > Classes > Groups > Relationships Table`).



In addition, any user can activate their own Tags in the "**Activate Tag**" page (`Filter samples > Tags > Manage tags`). The selected Tag will filter all tables on the system (samples, analyses, checks, *etc.*), restricting the visualization to samples listed in the Tag and related analysis results.

To remove the Tag filter, just click on the "**Deactivate Tag**" button ("Active Tag" combined button on the top bar). If a Tag is active, it will show in the "Activate Tag" page, the home page and the top bar (ribbon).

Users can create new Tags, add samples to a pre-existing Tag or remove samples from it. Such functionalities are available in the Manage Tags page (`Filter samples > Tags > Manage tags`).

An in-depth guide on how to use Tag funtionalities is available at the Manage Tags Wiki page.

Note: Tag profiling

A custom Tag (code: "**P0**"; description: "**Profiles/Genpat Groups**"") is available in the `Tags/Applicants > Tags > Tags` page. Such Tag is used to group together new sub-Tags representing laboratories able to access GenPat.

When a new tag is created in the "**Genpat-Tag Relationship**" page, the new Tag will be associated to the user's corresponding laboratory Profile.

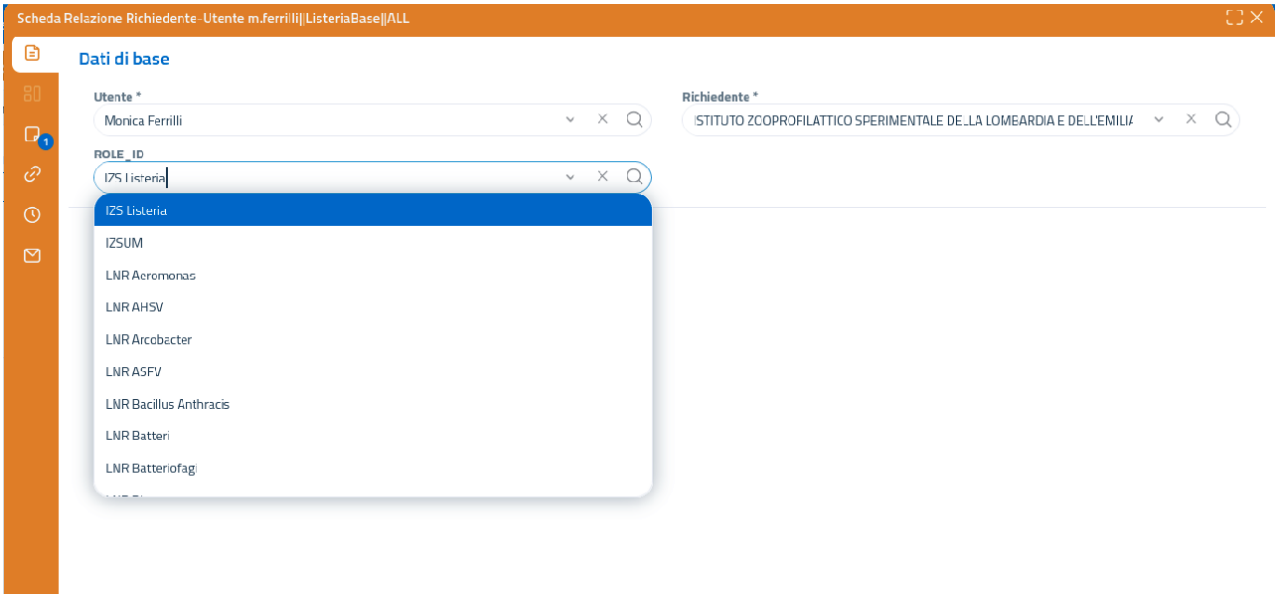
Note: Each laboratory will be able to visualize and activate Tags originated from SILAB and SEAP, in addition to Tags belonging to the laboratory.

Filter by Applicant

The page "**Applicant-User Relationship**" available in the system (`All Elements > Classes > Applicant-User Relationship`) grants the ability to link an Applicant to a specific user-laboratory profile combination; the page is only visible to, and accessible by, the Super User. When a user logs in, the system will automatically check if that user is associated with a specific Applicant; if they are, all data belonging to that Applicant will be displayed. There can be only one Applicant linked to a user-laboratory profile combination. To all user-laboratory profile combinations that have no associated Applicant will be applied a dummy Applicant ("ALL"), able to see all data from all Applicants from that profile.

Applicants taken in consideration for filter management are all those with "GENPAT" source. Such Applicants are organized hierarchycally to manage data accessibility.

E.g.: if an institution consists of 2 laboratories (LAB1 and LAB2, for the sake of this example), it will be "parent" of those laboratories. As a result, a user logged in as LAB1 will be able to access only samples of LAB1, a user logged in as LAB2 will be able to access only samples of LAB2, but a user logged in as the parent company will be able to access samples from both LAB1 and LAB2.



9. Dashboards

9.1 SPREAD, Spatiotemporal Pathogen Relationships and Epidemiological Analysis Dashboard

9.1.1 Description

The dashboard is based on GrapeTree is a fully interactive tree visualisation program within EnteroBase, which supports straightforward manipulations of both tree layout and metadata. It generates GrapeTree figures using the Neighbor-Joining (NJ) algorithm, Minimal Spanning Tree algorithm (MSTree), or an improved Minimal Spanning Tree algorithm called MSTree V2. The original version is an integral part of EnteroBase, and you can refer to its documentation to learn more.

For a formal description, please see the accepted manuscript in Genome Research.

The standalone version of GrapeTree has been integrated into Cohesive Information System (CIS) 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 display them, reflecting the selection on the tree and the map interchangeably. Additionally, it enables users to reproduce a temporal visualization both on the map and the tree.

9.1.2 Using SPREAD

The dashboard is integrated with the CIS 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 is 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.

type:video-tag grapetree-upload.mp4

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 CIS Platform, you will always find yourself in the same configurable workspace. The main components that make up the dashboard along with the **Tree** are:

- **Settings**
- **Map**
- **Legend**
- **Metadata**
- **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 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 information or support regarding this, please contact `bionformatica@izs.it` .

9.1.4 Components

Let's take a closer look at the various parts that make up the dashboard.

Settings

By activating the settings, you will enter a sort of "edit mode" and you will see 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.

Tree layout

The *Tree Layout* section contains tools to modify the appearence and layout of the tree, as well as options to revert the tree to previous state and undo changes and modifications.

- **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. It will ask for confirmation before leaving the page. **Caution**: this function will cause you to lose all the changes you made to the tree.

Node style

Under this category you can find 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 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 dropdown menu just below.
- **Font Size:** the font size for the labels can be chosen either by dragging the slider or by specifying a size in the box.
- **Highlight Label:** use the text box to provide a search term for the labels to highlight the corresponding nodes. 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. The breakdown is based on the category used for the Colour by settings.

Node size

- **Node Size (%):** the overall node size can be increased or decreased with the *Node Size (%)* option, either by using the slider or through 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 Labels:** the *Show Labels* switch will allow you to choose whether to show the branch labels or hide them.
- **Font Size:** options from where the font size for the labels can be chosen by means of the slider or the box.
- **Scaling (%):** the main *Scaling* option allows you to increase or decrease the overall branches' length, by using the slider or by specifying a value in the box below the slider.
- **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 will 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). Use the box to enter a value. For branches longer than the specified value, it's possible to choose whether to display them, hide them, or shorten them:

- **Display:** show long branches as normal (default).
- **Hide:** makes long 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. The branches will appear as dashed lines to indicate those that are affected.

Rendering

This section contains *Rendering Layout* options, which control how nodes are positioned in the tree. Two modes are available: *Static* and *Dynamic*.

- **Static:** in *Static* mode, the tree layout is calculated when the tree is generated and remains static (but interactive). Relative branch length and 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 branch length scaling will be modified. Branches will NOT be to scale when in *Dynamic* mode. Toggle on the **Selected Only** option to apply only to selected nodes.

type:video-tag grapetree-settings.mp4

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, markers are drawn with varying 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 also the selection. You can interact with the map through its **context menu**, which can be accessed by right-clicking on the map or by using the button located at the top-right corner with three vertical dots.

Map tools

Through the options available in this section, you can control the appearance of the map.

- **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 based on the values of the loaded metadata. In the first case, you can define a numerical 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 present 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.

type:video-tag grapetree-map.mp4

Metadata

By activating the **Metadata** toggle, you can view the metadata within an interactive grid. The grid 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 present on 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.

type:video-tag grapetree-metadata.mp4

Legend

This is a key tool for both interacting with the tree and for the temporal visualization/selection of data. Through the legend, you can change the category and color of the tree, customize the colors, or select samples from a specific category. In **video mode**, the legend transforms into a proper timeline.

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.

type:video-tag grapetree-legend.mp4

Video

The video player 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.

type:video-tag grapetree-video.mp4

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 selection in the chart workspace, press and hold the `Shift` key on the keyboard. As long as the cursor 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).

Certainly, more interesting things can 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.

In particular, we will focus on the **save functions** here.

Save SPREAD

Through `Save SPREAD` button you will have access to a modal that allows you to save and/or export SPREAD in various formats:

- **Save as complete json (.json)**: Allows you to save the current work in a complete `.json` file containing all the information about the tree object, 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)**: Exports data in the `.nwk` format.
- **Export metadata (.tsv)**: Exports metadata in the `.tsv` format.
- **Export geoJson (.geoJson)**: Exports the spatial metadata in the `.geojson` format.

All the files produced are compatible to be reloaded later in the dashboard, as mentioned in Using SPREAD.

type:video-tag grapetree-tree.mp4

9.2 Auspice

The Auspice dashboard integrated into Cohesive allows you to view phylogenetic trees, associated metadata, and geographic data produced by Augur.

- Official documentation

Access to the Auspice integrated dashboard is available in all execution cards of **Augur** analyses that finished execution successfully (see the "Check analysis" section).

9.2.1 Quick guide

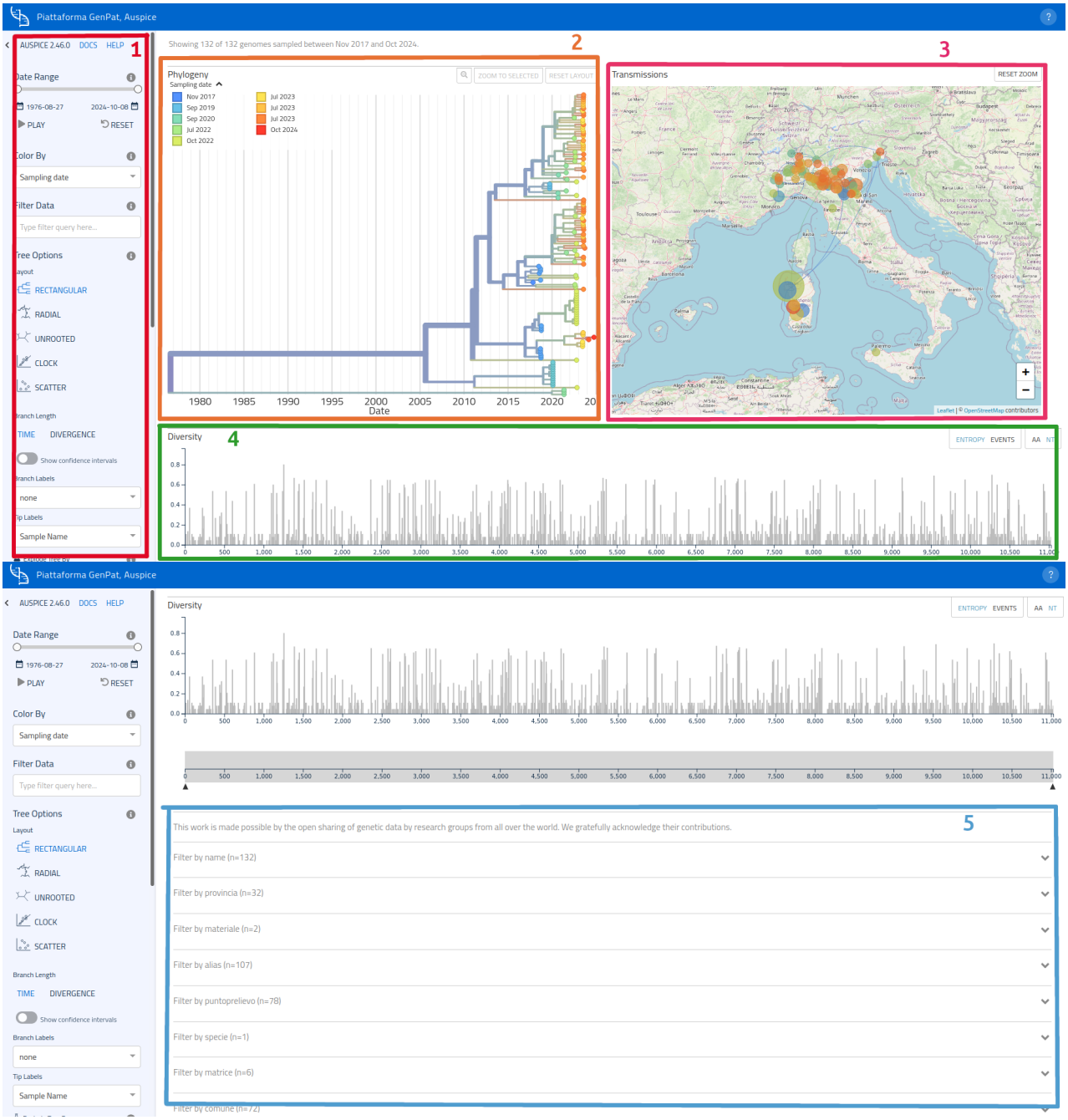
The following video provides a quick guide to Auspice's interface, with demonstration of menu options, map and graph features.

type:video-tag Auspice_demo.mp4

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 thanks to 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 (or menu) links to Auspice's official manual pages and docs and groups various options in categories, in order to manage different aspects of how data is displayed:

- Visualization, colors and data filter;
- Tree options;
- Map options;
- Animation options;
- Layout and panels.

The menu is thoroughly described in Auspice's official guide, furthermore each element of the menu has an associated **info** button ( ), which displays a description of the respective option when hovering the mouse pointer above it.

Usage of the options bar is demonstrated in the introductory video.

> In this manual page you will find 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 side panel on the left (default) or over the entire width of the page ("**FULL**" layout mode, enabled thanks to the **Panel Options** of the side bar).

The side bar options also allow to choose from 5 different layouts for the phylogenetic tree.

Tree branch options are available in the side bar too, and include toggles to:

- switch between representation of sample distances as time distance scale and divergence scale;
- show or hide confidence intervals and branch labels;
- select displayed information in the leaf labels;
- display the tree as a break-out chart based on the associated metadata.

Legend

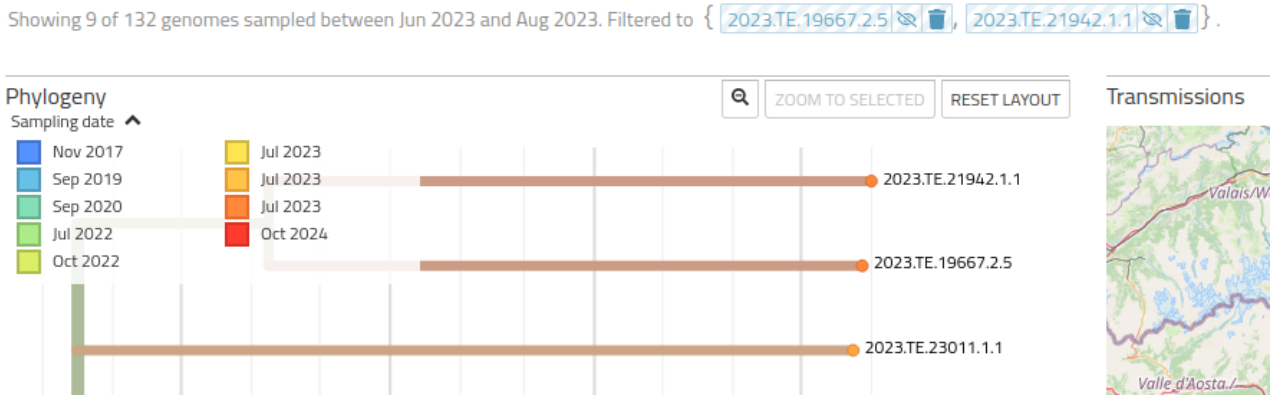
The **legend** (in the top-right corner) can be displayed or hidden by clicking on the arrow button (^); furthermore, tree leaves are highlighted when hovering the mouse over the legend elements.

Zoom and focused samples

Hovering the mouse over samples in the tree will show a summary of the sample's related information, while clicking on one of them will focus on that area and display the full information sheet of the selected sample.

Zoom options are 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 above the graph):



This interaction empowers users with the ability to hide or show samples selectively, as well as to remove single samples from the filter. Samples can also be removed in bulk from the filter using options in the **Data Filter** section of the side bar, which allow to activate/deactivate visualization of selected samples or to remove the filter altogether.

Map

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 click on the map is equivalent to pressing the **+** button;
- Hold the `Shift` key (or `Maiusc` key, depending on keyboard layout) while creating a selection area with the mouse (hold the left mouse button and drag) to zoom in the selected area;
- Use the button in the upper-left corner of the map to reset the zoom.

Map and graph are tied to each other: node selection from the graph will also highlight the position on the map.

Options for the map available in the side bar include adjustment of geographical resolution (depending on level of accuracy of geographic data) and hide/show toggle of transmission lines.

Diversity chart

A third graph, the "Diversity graph", located below the map and phylogenetic tree, displays sample diversity based on nucleotides or amino acids (the latter is not available in the platform's integration). The chart can show either the "Entropy" page or the "Events" page (the toggle is in the upper-right corner) and also provides zoom adjustment through selection of a range in the graduated line at the bottom.

Filers

Available filters (one for each metadata category) are listed further down. Filter types can be selected by clicking on the corresponding card to expand it.

> Example: to filter samples based on sampling point, open the card of "sampling point" metadata and select the desired location.

Animation

Samples can be shown on the map and graph through an animation which them displays consecutively, starting from those closer to the tree's root. Available options in the option bar include animation speed, activation of loop mode and cumulative mode (the latter keeps showing the nodes on the map after they first appear, rather than showing them just transiently).

10. FAQs

This page holds Frequently Asked Questions (FAQs) and answers about the software's usage, alongside tips for easier data management.

Before reaching out for assistance, please check out if your issue has already been solved.

This section does not contain any questions yet.

11. News

News and descriptions of fresh features added to Cohesive will be published in this section.

We don't have any news yet...

12. Resources

12.1 Keyboard shortcuts

Main functions can be launched via keyboard through particular predefined combinations. Some combinations may be valid at the global level of the application or only within certain pages. Below is the list of shortcuts defined by default.

Important

The list is constantly updated, so we invite you to visit this page often. The page can be reached from the platform with one of the following combinations: `ctrl+f1` , `f1` , `f+1` , `double ctrl` .

12.1.1 Shortcuts list

Function	Combination	Pages
Open list of shortcuts	`f1` , `ctrl+f1` , `f+1` , `doppio ctrl`	All
Enable/disable multiple selection	`ctrl+shift+m`	Classes and Views (pages with selectable grid as Samples)
Seleziona all	`ctrl+shift+a`	Classes and Views (pages with selectable grid as Samples)
Empty cart	`alt+shift+e`	All
Add to cart	`alt+shift+a`	Samples
Replace cart	`alt+shift+r`	Samples
Remove from cart	`alt+shift+d`	Samples
Disable tag	`alt+shift+t`	All

12.2 Video

This page hosts a collection of useful videos about the information system and the project that generated it.

CIS demo at Matrix WP6

Presentation of COHESIVE Information System at Matrix WP6 (24 February, 2022)

type:video-tag

cohesive_demo_at_matrix_wp6.mp4

COHESIVE Information System, Wiki