TriFusion Documentation

Release 1

Diogo N. Silva

Sep 26, 2018

Getting started

TriFusion is a modern GUI and command line application for gathering, processing and visualizing phylogenomic data. For the moment, this page is intended to host the tutorials and how-to guides for using TriFusion. The complete user guide is available as a PDF here.

CHAPTER 1

Tutorials

The structure of the tutorials is organized in a couple of sections:

- *Getting started*
- ***TriFusion GUI***
    - *Orthology (GUI)*
    - *Process (GUI)*
    - *Statistics (GUI)*
- *TriFusion CLI*

1.1 Installation

1.1.1 Binaries and Installers

Note: The installers **only provide the GUI version** of TriFusion. If you want to install both the GUI and command line versions, check the *Installation from source*.

The easiest way to install TriFusion is through binaries and installers provided for Windows, MacOS and Linux.

Windows

- TriFusion 1.0.0 64bit installer
- TriFusion 1.0.0 32bit installer

MacOS

- TriFusion-1.0.0.app

Linux

- Ubuntu and Debian based: TriFusion-1.0.0.deb
- RPM based: TriFusion-1.0.0.rpm
- General linux: TriFusion-1.0.0.tar.xz

1.1.2 Installation from source

TriFusion is available on PyPi nd can be easily installed with `pip` . This will only install the command line versions of TriFusion (`TriSeq`, `TriStats` and `orthomcl_pipeline`). Therefore, if you are interested only in the command line version of TriFusion, and assuming you have `python2.7` and `pip` on your system, installing TriFusion is simply:

```
pip install trifusion
```

The dependencies for the graphical user interface require only a few extra commands that are provided below for each operating system.

Windows

Windows does not come with a python installation by default. We recommend using a package manager, such as Anaconda, which automatically installs most of the dependencies (**Note that TriFusion requires python2.7**). After installing python, you will need to install `kivy` by executing the following commands on a command line prompt:

```
python -m pip install --upgrade pip wheel setuptools
python -m pip install docutils pygments pypiwin32 kivy.deps.sdl2 kivy.deps.glew
python -m pip install kivy.deps.gstreamer --extra-index-url https://kivy.org/
↪downloads/packages/simple/
python -m pip install kivy
```

Then, install trifusion by typing:

```
pip install trifusion
```

MacOS (using homebrew)

f you do not have homebrew yet, you'll need to install it:

```
/usr/bin/ruby -e "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/
↪master/install)"
```

Then, to install TriFusion and it's dependencies:

```
brew install sdl2 sdl2_image sdl2_ttf sdl2_mixer
pip install -I Cython==0.23
USE_OSX_FRAMEWORKS=0 pip install kivy
pip install trifusion
```

Ubuntu (and relatives)

On Ubuntu, there are PPAs available for the installation of TriFusion via `apt-get` in addition to the `pip` installation method.

Via PPA

- **Add one of the following PPAs:**

  ```
  # Stable release:
  sudo add-apt-repository ppa:o-diogosilva/trifusion
  # Daily release:
  sudo add-apt-repository ppa:o-diogosilva/trifusion-daily
  ```

- **Upgrade your package list and install TriFusion:**

  ```
  sudo apt-get update && sudo apt-get install trifusion
  ```

Via `pip`

```
sudo apt-get install python-pip build-essential python-dev libsdl2-dev
pip install cython==0.23
pip install kivy
pip install trifusion
```

Debian

As with Ubuntu, you may install TriFusion via the available PPAs or with `pip`.

Via PPA

- **Add one of the following PPAs manually to the `sources.list` file:**

  ```
  # Stable release:
  http://ppa.launchpad.net/o-diogosilva/trifusion/ubuntu trusty main
  # Daily release:
  http://ppa.launchpad.net/o-diogosilva/trifusion-daily/ubuntu trusty main
  ```

- **Add the GPG key to your apt keyring:**

  ```
  sudo apt-key adv --keyserver keyserver.ubuntu.com --recv-keys D4F1E8E6
  ```

- **Upgrade your package list and install TriFusion:**

  ```
  sudo apt-get update && sudo apt-get install trifusion
  ```

Via `pip`

```
sudo apt-get install python-pip build-essential python-dev libsdl2-dev
pip install cython==0.23
pip install kivy
pip install trifusion
```

RPM based

```
dnf install python-pip python-devel redhat-rpm-config freeglut-devel SDL* libsdl2-dev
pip install cython==0.23
pip install kivy
pip install trifusion
```

ArchLinux

There are three AUR packages for TriFusion:

- trifusion: The latest release of TriFusion, based on source code.
- trifusion-bin: The latest release of TriFusion. in binary format. Does not require dependencies to be installed, as all the necessary libs are bundled with the distributed binary
- trifusion-git: The bleeding edge version directly from git. Requires dependencies to be installed, as it is also source code based.

Just use any AUR helper to handle the packages for you, or download the *PKGBUILD* you require and use `makepkg`.

1.2 Usage

1.2.1 TriFusion GUI

If TriFusion was installed using one of the provided installers, through a the Ubuntu PPA or as a AUR package, the application should be available on the system's program list under the name `TriFusion`.

- Windows

[Windows image]

- MacOS

[MacOS image]

- Ubuntu

[Ubuntu image]

Calling from the command line

In any case, TriFusion can be executed from the command line by typing:

```
TriFusion
```

1.2.2 TriFusion CLI

If TriFusion was installed from source, The command line programs associated with each module of TriFusion are also available.

Orthology - search

The ortholog search pipeline can be executed from the command line using:

```
orthomcl_pipeline
```

Process

Most of the operations of the **Process** module can be executed in the command line using:

```
TriSeq
```

Statistics

The generation of plots from the Statistics module can be performed in the command line using:

```
TriStats
```

1.3 Load data into TriFusion

TriFusion deals with different types and formats of input files, depending on which module you want to use. The Orthology module deals with proteomes and group files while the *Process and Statistics* module deals with alignment files. Regardless, input files are loaded into the application mostly in the same way (see *How to load data into the app* below).

1.3.1 Input types and formats

Orthology - search

Proteome files can be provided as the input for the Orthology search operation. These are **Fasta** formatted files, each with the amino acid sequences of a single species. **TriFusion will interpret the name of the proteome file (minus extension) as the taxon name**, so it is recommended that these files are named accordingly (for instance, `Aspergilus_fumigatus.fasta` will appear as `Aspergilus_fumigatus` in the final ortholog files). The only requirements for the input files is that the headers must have one or more fields separated by a “|” symbol, and at least one of those fields must be different for all sequences.

A standard proteome file resulting from a genome sequencing project is usually something like this example of the fungus *Aspergilus fumigatus*:

```
>jgi|Aurde3_1|1209208|estExt_Genewise1.C_13_t10159
MTD(...)
>jgi|Aurde3_1|1326459|estExt_fgenesh1_pm.C_130053
MPP(...)
>jgi|Aurde3_1|1274305|fgenesh1_kg.13_#_18_#_isotig02263
MLY(...)
```

This is a valid input file for TriFusion with the headers containing 4 fields, the third one being the unique ID field.

Orthology - explore

Group files are one of the outputs of the Orthology search operation and the input of the Orthology explore operation. These are simple text files that contain all ortholog groups identified in the search operation by OrthoMCL:

```
Ortholog1: Afumigatus_proteins|433 Anidulans_proteins|4605 (...)
Ortholog2: Afumigatus_proteins|3278 Afumigatus_proteins|9183 (...)
Ortholog3: Anidulans_proteins|36 Anidulans_proteins|9893 (...)
(...)
```

Each line contains the name of the ortholog group and a list of sequence references separated by whitespace. Each reference (e.g., *Afumigatus_proteins|433*) corresponds to an actual protein sequence from one of the input protome files.

Process and Statistics

The **Process** and **Statistics** modules share the same input, which are **sequence alignment files**. The supported input formats are:

- **Fasta**
- **Phylip**
- **Nexus**
- **Loci (PyRAD)**
- **Stockholm**

The input format, sequence type (nucleotide or protein) and string formatting (leave or interleave) of the provided alignment files are **automatically detected by TriFusion**. The missing data symbol used in the input alignments will also be automatically detected from the three possible symbols of x, n or ?.

Note: Is there any constraint on how formats and sequence types can be loaded?

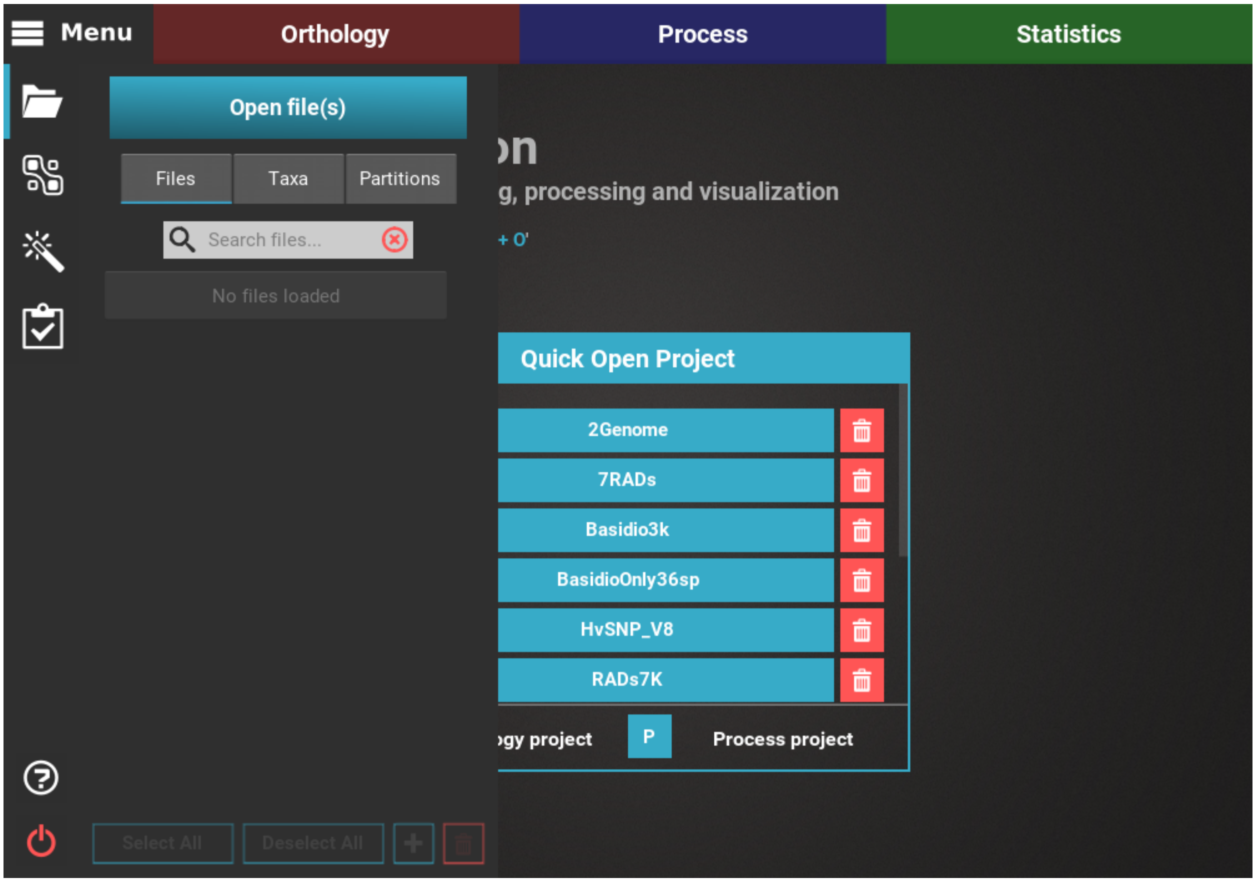
No. You can load files of multiple formats and sequence types all at once. All information will be automatically detected for each input alignments separately.

1.3.2 How to load data into the app

Note: Data availability for this tutorial: the small data set of 7 alignment files is available here.

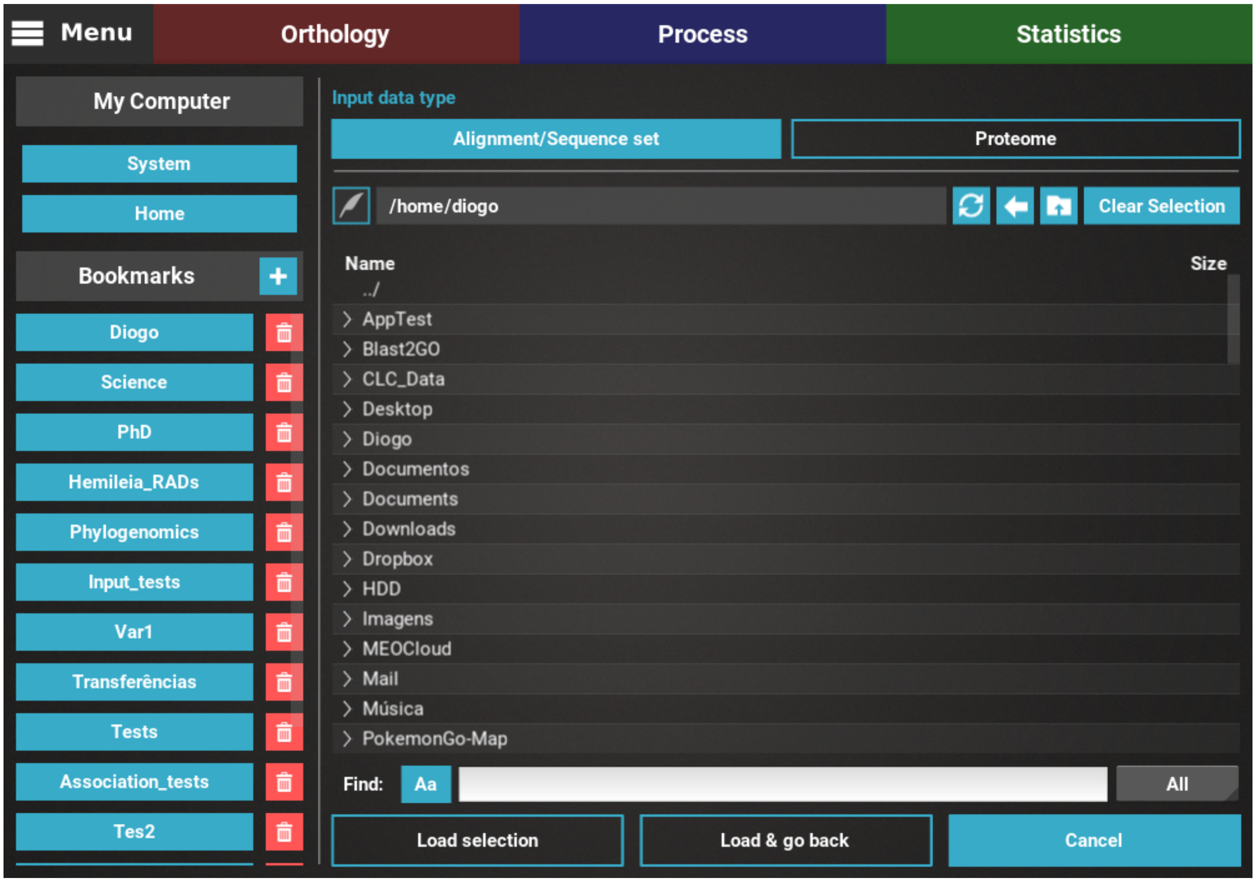
Filechooser

Proteome and **sequence alignment** files can be loaded through the application's file browser. To do so, navigate to `Menu -> Open/View Data` and click the `Open file(s)` button.

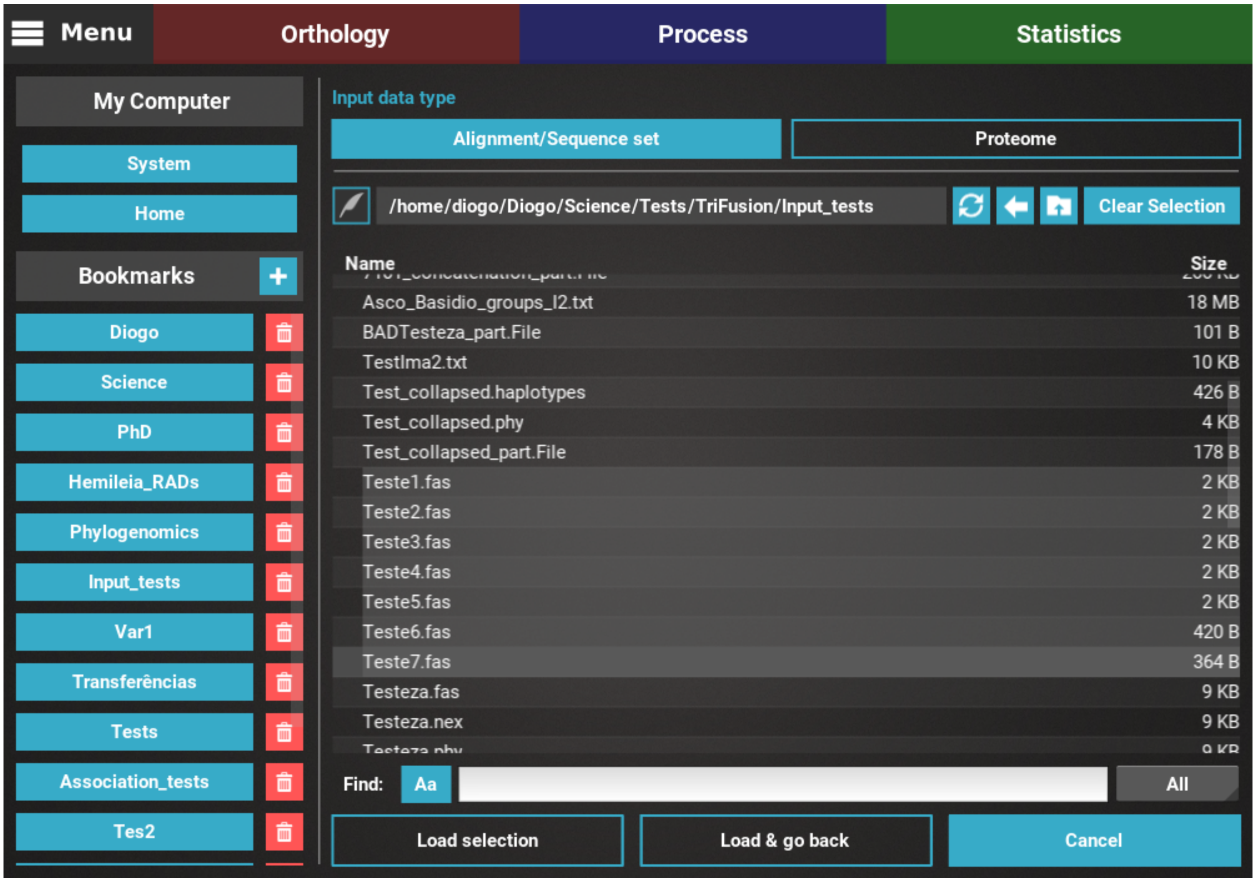


This will open the main file browser, which supports a couple of features:

- A list of **bookmarks** is displayed on the left, and any directory can be added to this list by opening it and clicking the + button or pressing `Ctrl + D`.
- On the top of the screen, you can choose the input data type (whether you are loading **proteome** or **alignment files**).
- Below you can find the path of the current directory and several utility buttons to navigate the file browser.
- At the bottom of the file browser, there is a text field that searches folders and files in the current directory. There is also a drop down menu that filters files according to their extension.



Navigate through the file browser by double clicking directories or clicking on the > symbol. Multiple files can be selected by pressing either the `Ctrl` or `Shift` keys. After completing you selection, click the `Load & go back` button to load the data and go back to the previous screen. If you wish to load additional data, click the `Load selection` button, which will load the data but remain in the file browser screen. In the example below, 7 files have been selected and are ready to be loaded.



/home/diogo/Diogo/Science/Tests/TriFusion/Input_tests

Name	Size
Asco_Basidio_groups_I2.txt	18 MB
BADTesteza_part.File	101 B
TestIma2.txt	10 KB
Test_collapsed.haplotypes	426 B
Test_collapsed.phy	4 KB
Test_collapsed_part.File	178 B
Teste1.fas	2 KB
Teste2.fas	2 KB
Teste3.fas	2 KB
Teste4.fas	2 KB
Teste5.fas	2 KB
Teste6.fas	420 B
Teste7.fas	364 B
Testeza.fas	9 KB
Testeza.nex	9 KB
Testeza.phy	9 KB

Note: TriFusion also supports the selection of one or more directories instead of files!

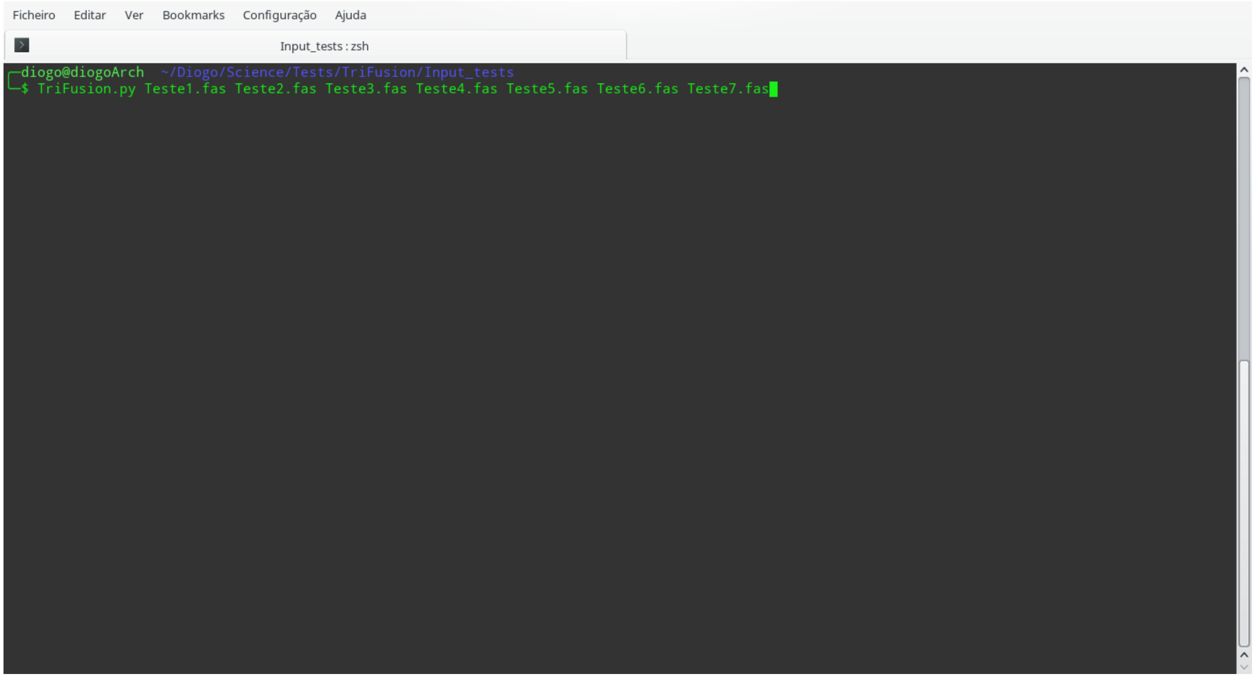
When directories are selected, all files contained in those directories will be loaded into TriFusion. If you are worried that not all files in a directory are alignments/proteomes, do not worry. TriFusion will ignore invalid input files while successfully loading valid alignment/proteome files.

Drag and Drop

Input files can be provided to TriFusion's window directly from your systems' file manager. After selecting the files, drag them into TriFusion's window, which will display a popup informing of how many files will be loaded and asking whether the files represent **alignments**, **proteomes** or **groups**. Directories can also be dragged as well. In the example below, 7 sequence alignment files are loaded using this method.

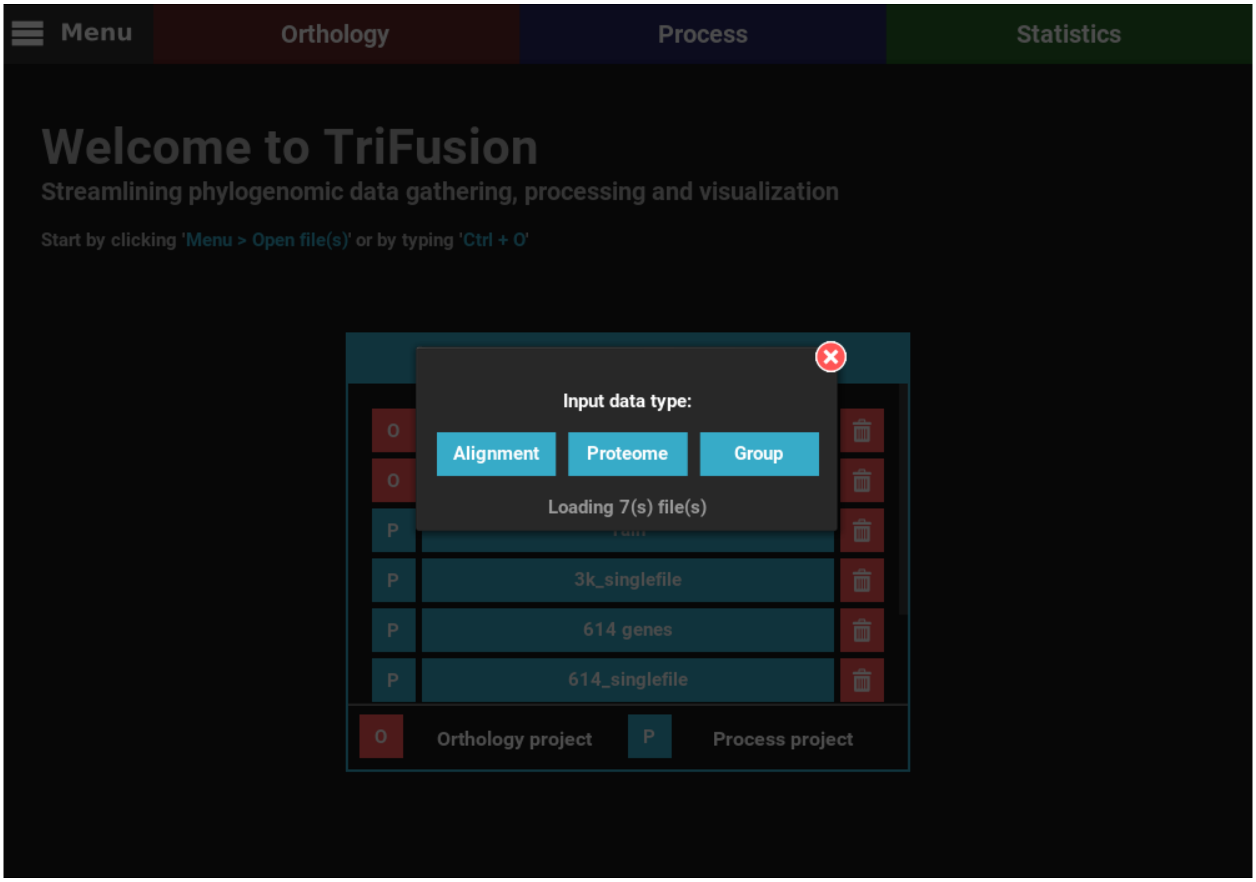
Via terminal

For terminal lovers (<3) files can be loaded automatically when executing the TriFusion application. If TriFusion's executable is already in you `$PATH` environmental variable, you can write it in the terminal and then provide any number of files.



```
┌─diogo@diogoArch  ~/Diogo/Science/Tests/TriFusion/Input_tests
└─$ TriFusion.py Teste1.fas Teste2.fas Teste3.fas Teste4.fas Teste5.fas Teste6.fas Teste7.fas
```

This will open TriFusion and automatically open a popup informing that 7 files will be loaded into TriFusion and asking whether the files represent **alignment**, **proteome** or **group** files. In this case, the data files correspond to alignments.



Welcome to TriFusion

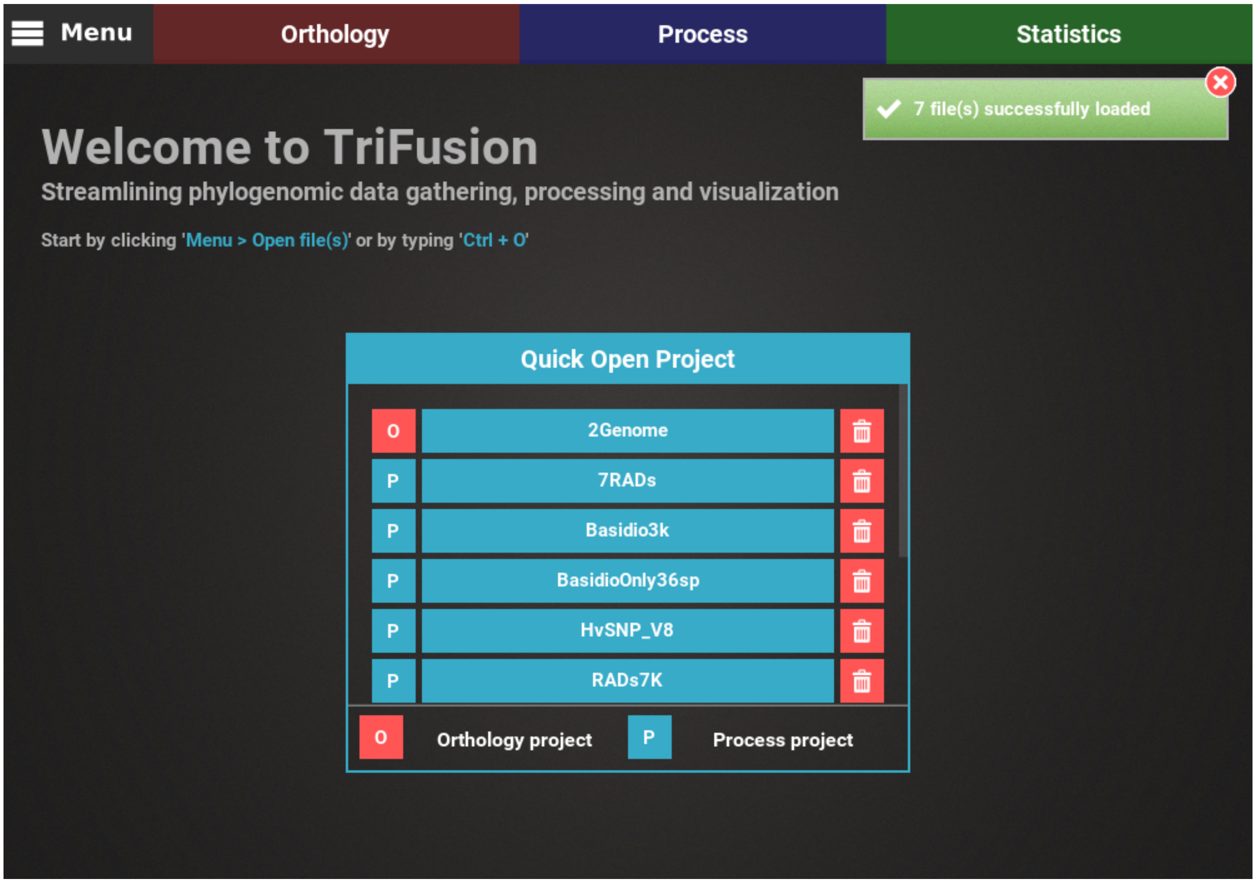
Streamlining phylogenomic data gathering, processing and visualization

Start by clicking 'Menu > Open file(s)' or by typing 'Ctrl + O'

Input data type:

Loading 7(s) file(s)

Once the sequence type is selected, the selected files will be loaded normally into TriFusion.



7 file(s) successfully loaded

Welcome to TriFusion

Streamlining phylogenomic data gathering, processing and visualization

Start by clicking 'Menu > Open file(s)' or by typing 'Ctrl + O'

1.4 Data set groups

Note: Data availability for this tutorial: the medium sized data set of 614 genes and 48 taxa that will be used can be downloaded here.

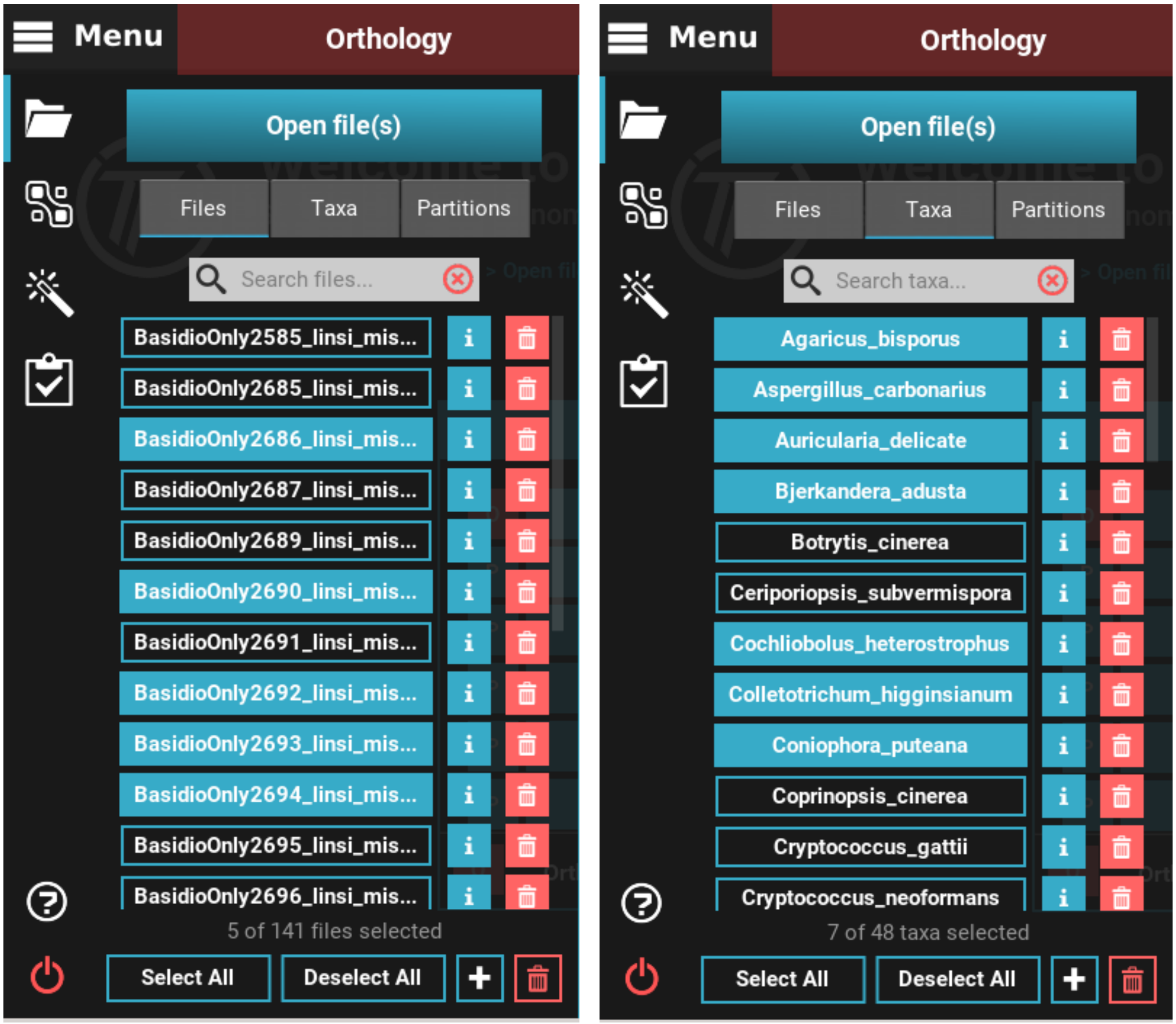
1.4.1 What are active data sets

Most operations in TriFusion can be applied to either the **total data set** (all files and taxa currently loaded) or to custom made data sets, named **active data sets**. When a custom data set is specified, operations will be applied only on the active files and/or taxa and ignore all others. These active data sets can defined in TriFusion in several ways and serve to quickly apply different operations on different sets of files/taxa.

1.4.2 How to define active data sets

Active data sets can be created/modified in two main ways:

- *Toggle file/taxa buttons in side panel*
    - *Mouse click toggling*



BasidioOnly2585_linsi_mis...
BasidioOnly2685_linsi_mis...
BasidioOnly2686_linsi_mis...
BasidioOnly2687_linsi_mis...
BasidioOnly2689_linsi_mis...
BasidioOnly2690_linsi_mis...
BasidioOnly2691_linsi_mis...
BasidioOnly2692_linsi_mis...
BasidioOnly2693_linsi_mis...
BasidioOnly2694_linsi_mis...
BasidioOnly2695_linsi_mis...
BasidioOnly2696_linsi_mis...

5 of 141 files selected

Agaricus_bisporus
Aspergillus_carbonarius
Auricularia_delicate
Bjerkandera_adusta
Botrytis_cinerea
Ceriporiopsis_subvermispora
Cochliobolus_heterostrophus
Colletotrichum_higginsianum
Coniophora_puteana
Coprinopsis_cinerea
Cryptococcus_gattii
Cryptococcus_neoformans

7 of 48 taxa selected

Fig. 1: Example of custom active file (left) and taxa (right) data sets.

    - *Import selection from file*
- ***Create data set groups***
    - *Manual creation in TriFusion*
    - *Group creation from file*

1.4.3 Toggle file/taxa buttons in side panel

Mouse click toggling

By default, when data is loaded into TriFusion **all** files/taxa are active. Therefore, the total and active data sets are the same. The quickest way to modify the active data set is by navigating to `Menu -> Open/View Data` and toggle the corresponding file/taxa buttons. `Shift + click` is also supported to select multiple contiguous files/taxa.

Active files/taxa will appear with a blue background, while inactive buttons will have no background. A label below the button list displays how many files/taxa are currently active.

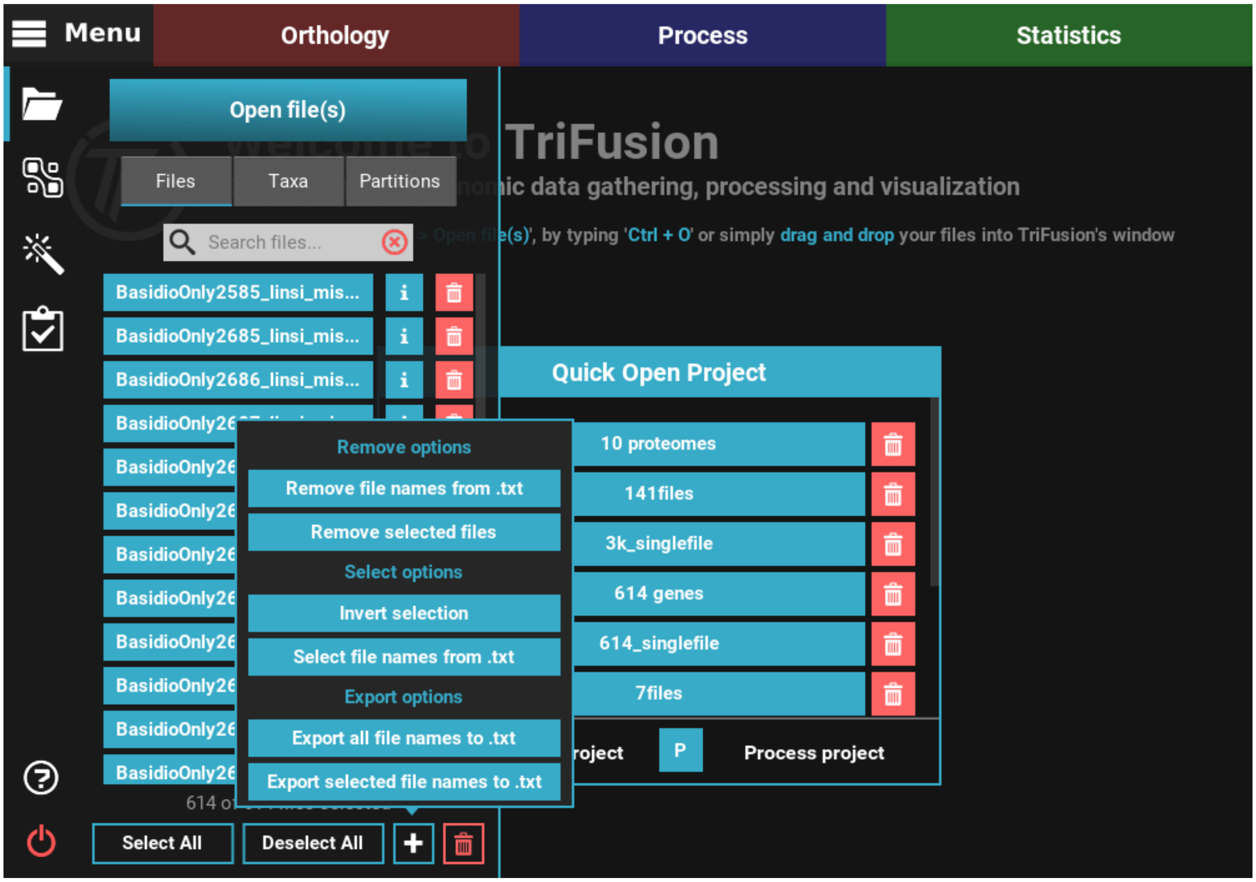
Import selection from file

When dealing with a larger number of files/taxa it may be more convenient to provide the active data set through a text file. This should be a simple text file containing the names of the desired files/taxa in each line. You can create it yourself, or download an example from here.

```
# Example of a text file for taxa selection in TriFusion
Agaricus_bisporus
Botrytis_cinerea
Coniophora_puteana
```

```
# Example of a text file for file selection in TriFusion (note the extension)
BasidioOnly2585_linsi_missingFilter_concPrep.fasta
BasidioOnly2685_linsi_missingFilter_concPrep.fasta
BasidioOnly2686_linsi_missingFilter_concPrep.fasta
```

Open the `Menu -> Open/View Data` side panel and click on the + button at the bottom of either the *Files* or *Taxa* tabs. This will open a sub-menu with several options, one of which is `Select file/taxa names from .txt`. Clicking this button will open a file browser where you can provide the file containing the file/taxa names. Once you select the text file, the the active file/taxa names will update.

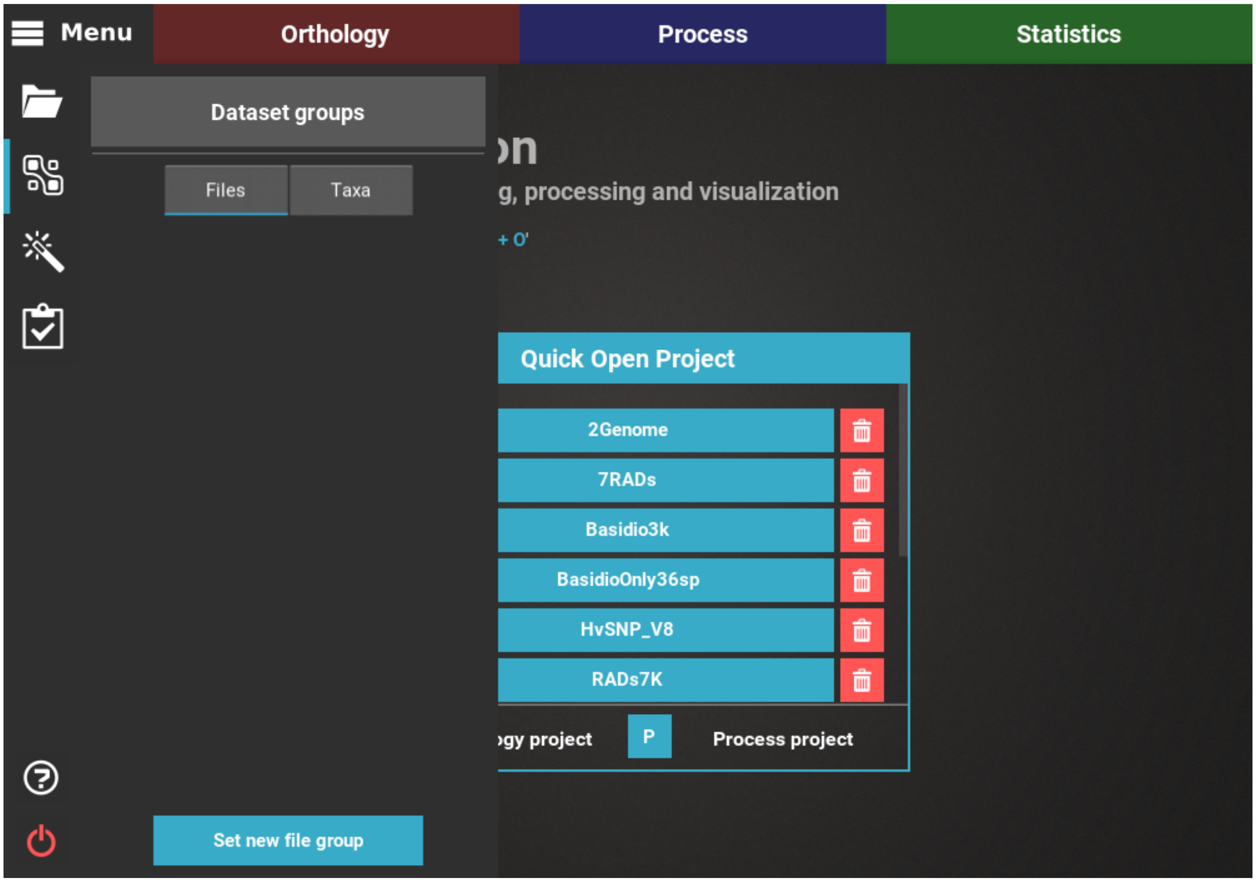

Menu
Orthology
Process
Statistics
Open file(s)
Files
Taxa
Partitions
Search files...
TriFusion
ic data gathering, processing and visualization
e(s)', by typing 'Ctrl + O' or simply drag and drop your files into TriFusion's window
BasidioOnly2585_linsi_mis...
BasidioOnly2685_linsi_mis...
BasidioOnly2686_linsi_mis...
BasidioOnly26
Quick Open Project
10 proteomes
141files
3k_singlefile
614 genes
614_singlefile
7files
roject
P
Process project
Remove options
Remove file names from .txt
Remove selected files
Select options
Invert selection
Select file names from .txt
Export options
Export all file names to .txt
Export selected file names to .txt
614 o
Select All
Deselect All
+

Warning: After loading the file, **ONLY** the specified items will become active, regardless of the previous active data set. Names that do not match any of the files/taxa present in TriFusion will be ignored.

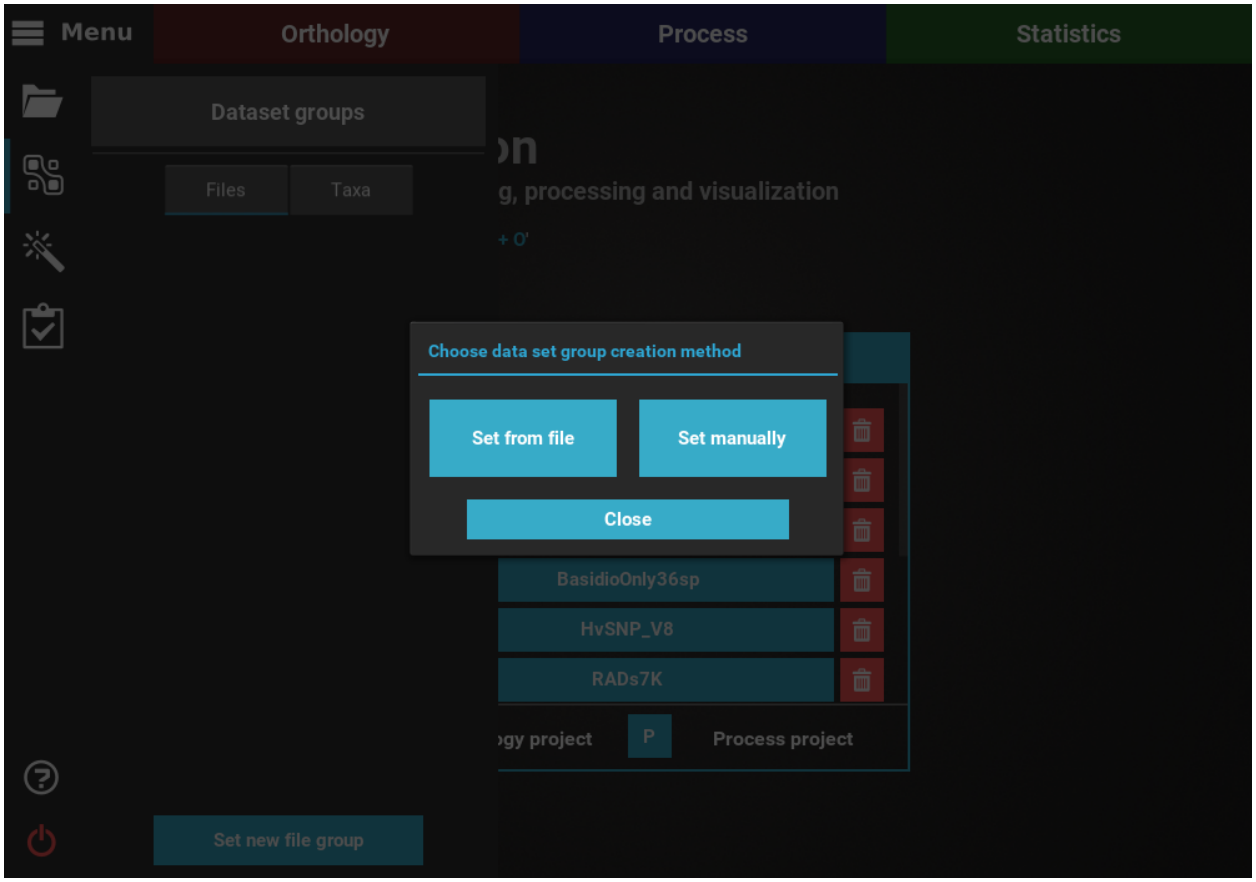
Note: You can also save any active files/taxa on the side panel to a text file by clicking the `Export selected file/taxa names to .txt`.

1.4.4 Create data set groups

When the workflow requires the execution of operations to multiple taxa/files data sets, it is more convenient to define all data set groups and then use the dropdown menus (see *How to apply data set groups* below) to select the desired active data set. Data set groups can be defined in TriFusion by navigating to `Menu > Dataset Groups`.



File and taxa groups are sorted into two tabs, like in the `Open/View Data` panel, and clicking the `Set new file/taxa group` button will start the creation of the group.



Here you can choose to create the data set group either manually in TriFusion, or by providing the names of the files/taxa in a text file.

Manual creation in TriFusion

> **Warning:** This option is discouraged for larger data sets (>500 items). In these cases, it is recommended to use the *Group creation from file* method.

The creation of groups is the same for both files and taxa. In this tutorial, we will create a taxa group by clicking in the *Taxa* tab and then the `Set new taxa group` button at the bottom of the side panel. Here, groups can be created by selecting the desired taxa from the *All taxa* column and using the arrow buttons to move them to the *Selected taxa* column. Once the group is complete, give it a unique name and the group is ready to be defined. If you wish to create multiple groups in one sitting, click the `Apply` button to create the group but remain in the dialog.

Any previously created group will be listed under the *Created groups* column. These can be selected to move their corresponding taxa to the *Selected taxa* column and continue a new group definition from there.

Group creation from file

Here, we only have to provide a text file with the names of the files/taxa we wish to select for the group. The text file is the same as the one described in the *Import selection from file* example.

```
# Example of a text file for taxa selection in TriFusion
Agaricus_bisporus
Botrytis_cinerea
Coniophora_puteana
```

After providing the file with the group names, specify a unique name of the new data set group, and that's it!

1.4.5 How to apply data set groups

Now that we know how to create active data set groups, the final step is how can they be specified.

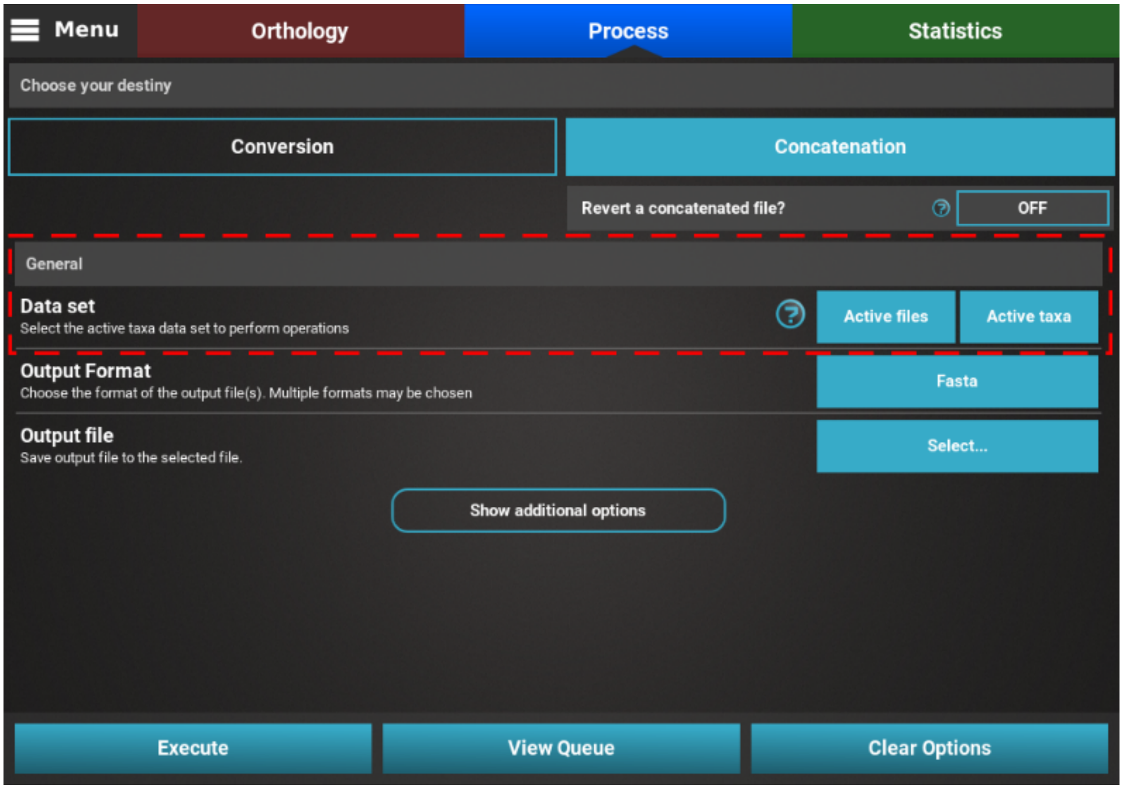
Orthology

When using the **Orthology** module, only the active proteome files are used for the Orthology search operation.

Process and Statistics

For both **Process** and **Statistics** modules, the *active* data set is selected by default (that is, the file/taxa buttons active in the side panel). You can change to the *total* data set or to any user made data set by clicking the group's name in the corresponding dropdown menu.

Dropdown menu in the **Process** screen:



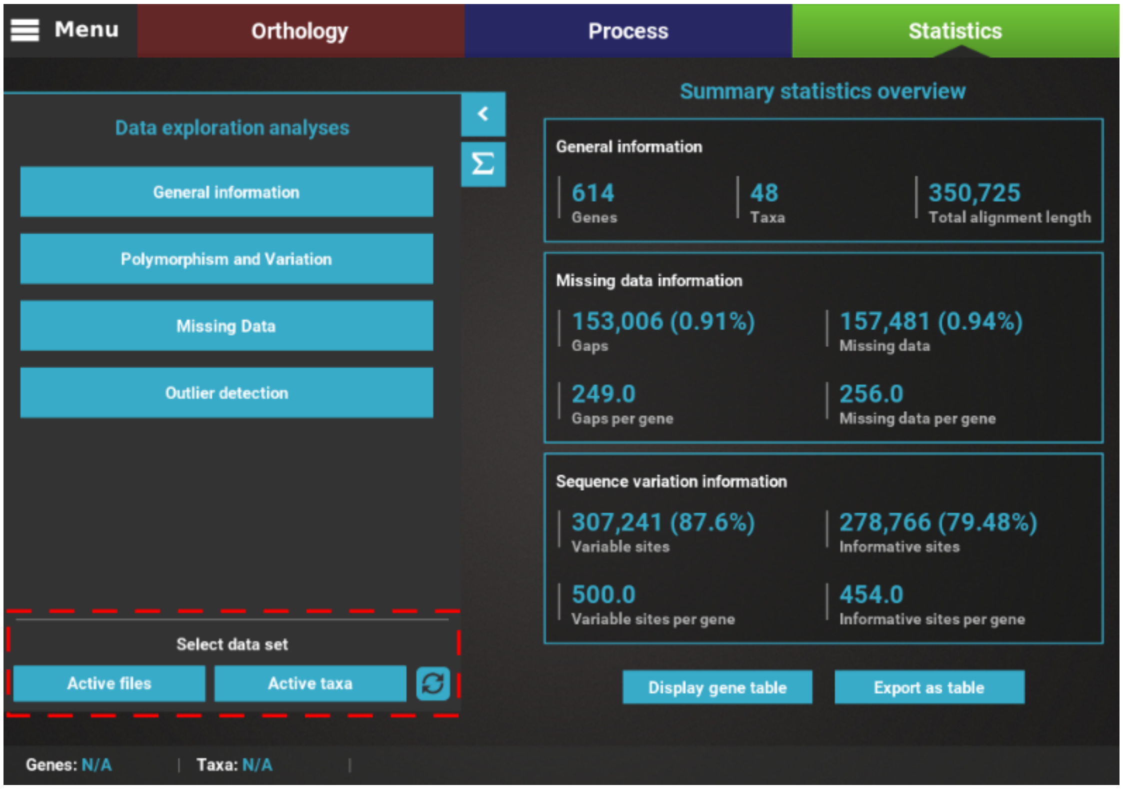
Revert a concatenated file?

Data set
Select the active taxa data set to perform operations

Output Format
Choose the format of the output file(s). Multiple formats may be chosen

Output file
Save output file to the selected file.

Dropdown menu in the **Statistics** screen:



Summary statistics overview

General information

614 Genes | 48 Taxa | 350,725 Total alignment length

Missing data information

153,006 (0.91%) Gaps | 157,481 (0.94%) Missing data

249.0 Gaps per gene | 256.0 Missing data per gene

Sequence variation information

307,241 (87.6%) Variable sites | 278,766 (79.48%) Informative sites

500.0 Variable sites per gene | 454.0 Informative sites per gene

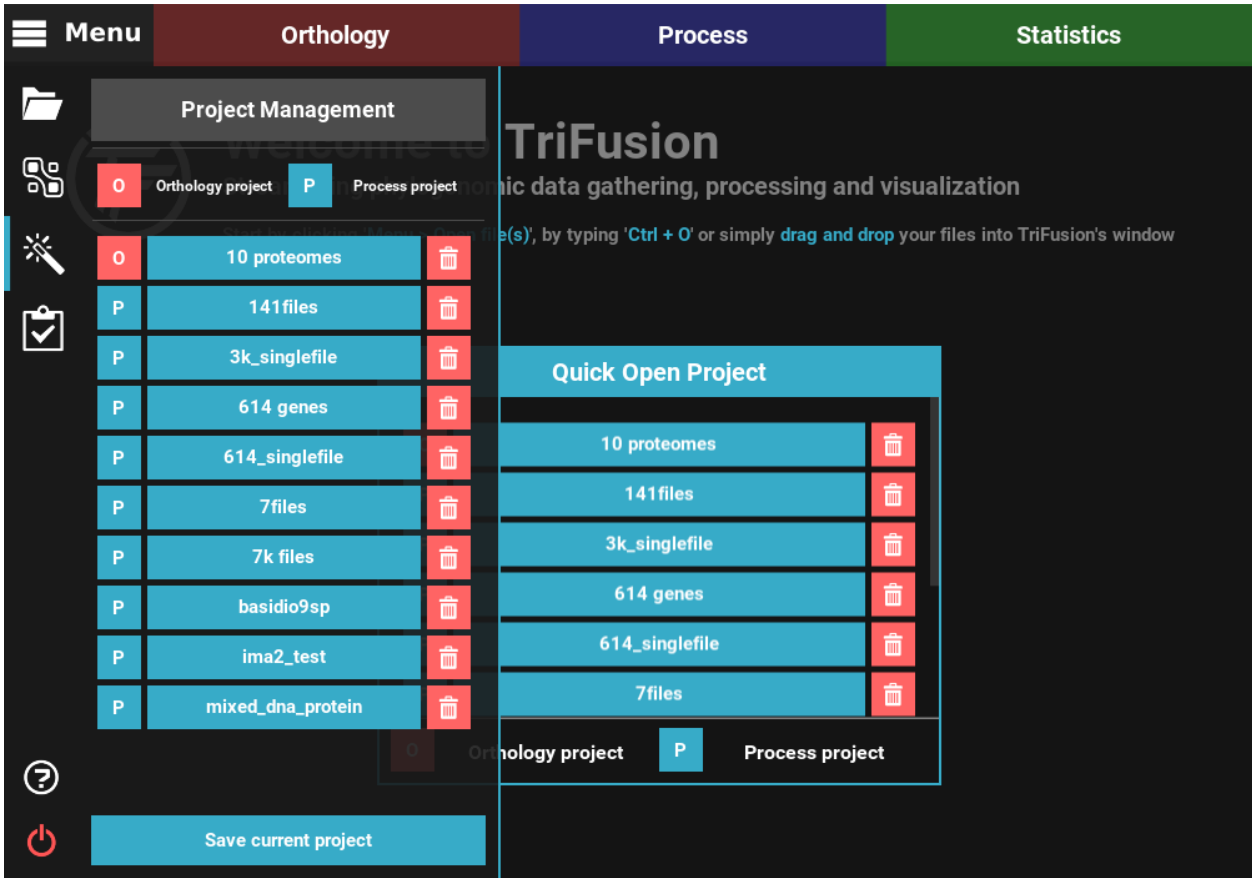
Genes: N/A | Taxa: N/A

1.5 Projects

Any data set that is loaded in TriFusion (be it proteomes or alignments) can be saved as a *Project*, which allows it to be quickly loaded in future separate sessions. As soon as TriFusion opens, it displays a list of previously save project for quick loading.

1.5.1 Save a data set as a project

Once a particular data set has been loaded into TriFusion, navigate to `Menu -> Project Management` and click the `Save current project` button. Provide a unique and descriptive name for your project and click `Ok`.

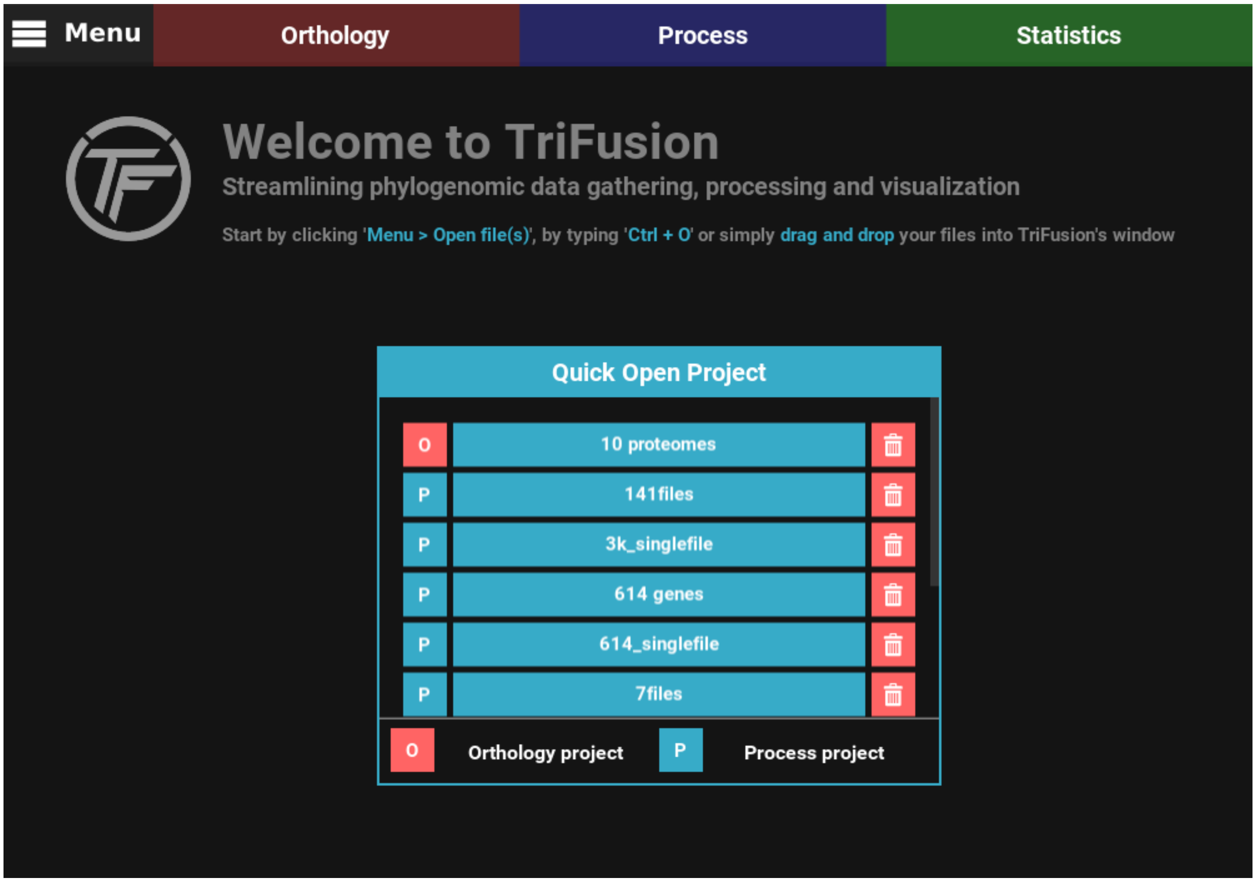


Saved projects will be stored and listed in this sub-menu of the side panel, besides the list in the Home screen of TriFusion. A small label will be associated with each project: A `O` label represents an **Orthology** project (proteomes), whereas a `P` label represents a **Process** and **Statistics** project (sequence alignments).

1.5.2 Load a project

Warning: When a new project is loaded, any previously loaded files are removed from the current session!

There are two places where saved projects can be loaded. In the home screen of TriFusion, there is a *Quick Open Project* box:

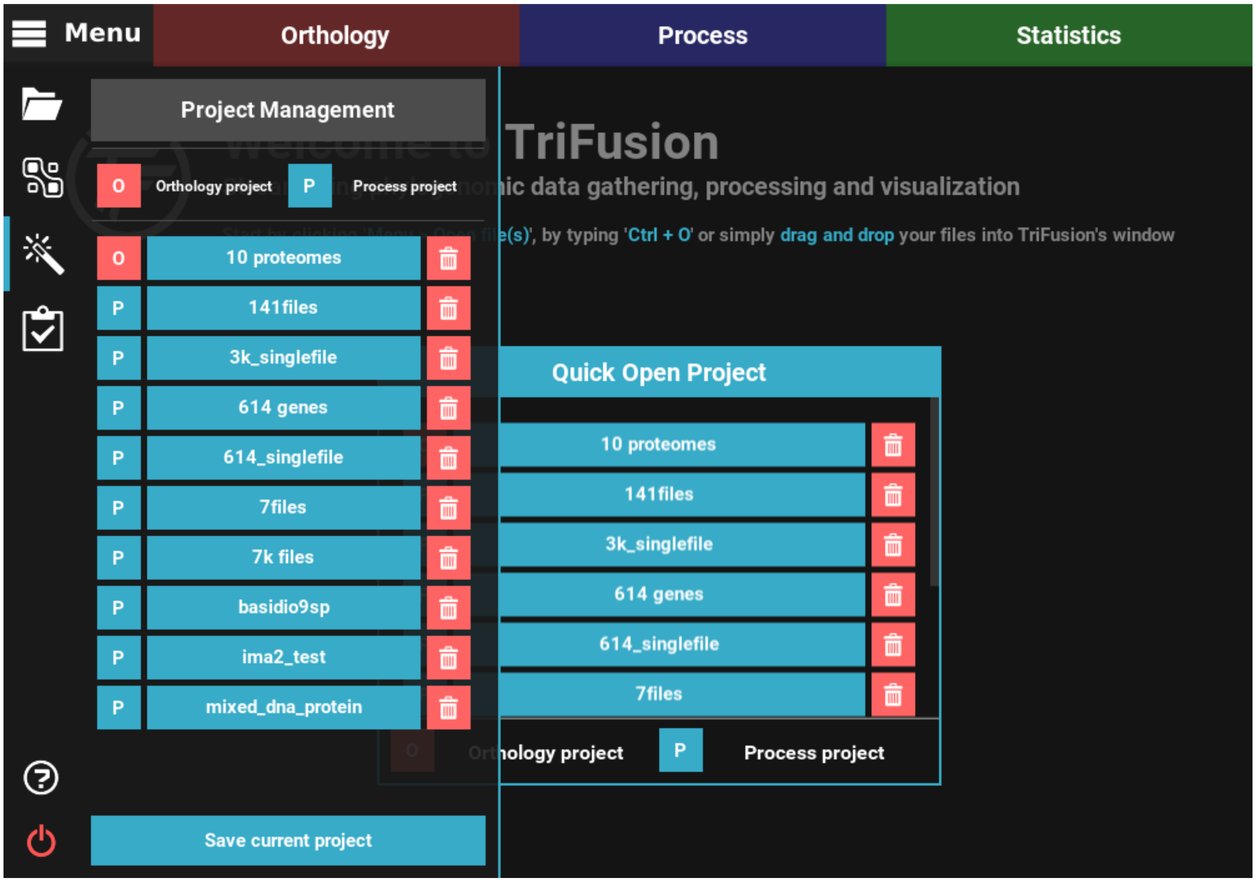


Welcome to TriFusion

Streamlining phylogenomic data gathering, processing and visualization

Start by clicking 'Menu > Open file(s)', by typing 'Ctrl + O' or simply drag and drop your files into TriFusion's window

Alternatively, navigating to `Menu -> Project Management` will also list the projects in the side panel:

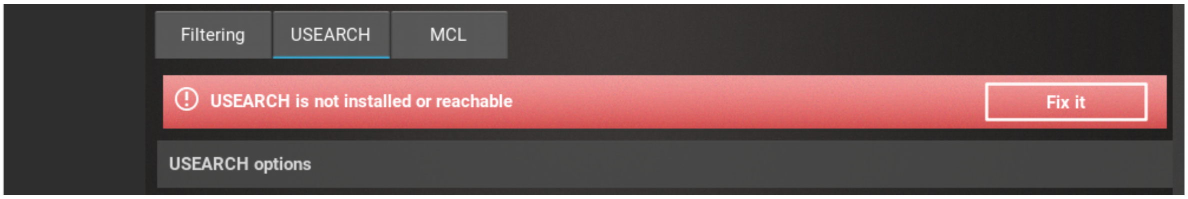


e(s)', by typing 'Ctrl + O' or simply drag and drop your files into TriFusion's window

1.6 Setup of USEARCH

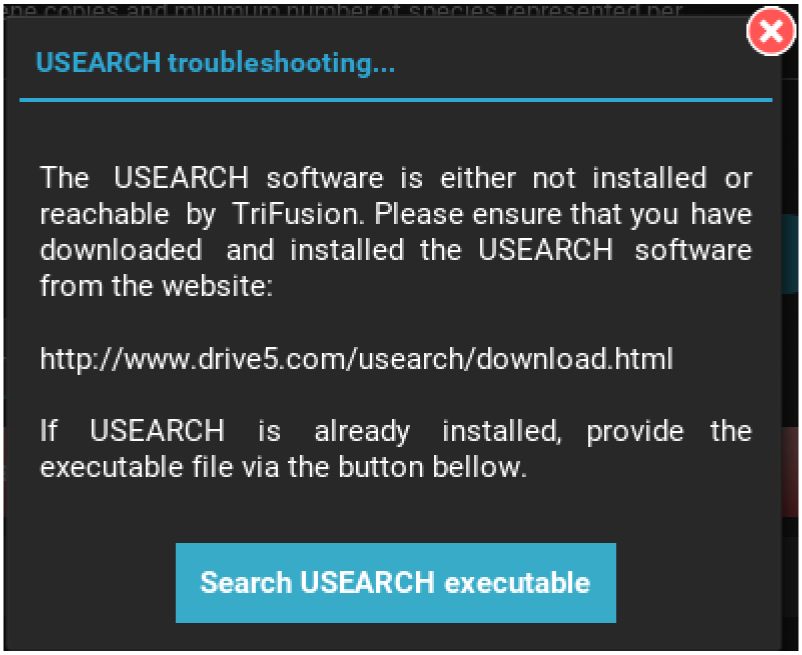
The USEARCH software is required to perform the **Orthology** search operation and to export ortholog groups into nucleotide sequences. However, due to licensing issues, USEARCH cannot be bundled with Triusion, so it requires some user intervention to setup. But don't fret! Everything can be up and running with just a few simple steps. Moreover, after the initial setup, TriFusion will store the USEARCH executable internally and use it for all subsequent sessions.

- Step 1: Download the USEARCH executable for your corresponding operating system here.
- Step 2: If USEARCH is not reachable by TriFusion, you will see a warning like this when you navigate to `Orthology -> Show additional options -> USEARCH`:

  USEARCH is not installed or reachable

  Click the `Fix it` button, and then the `Search USEARCH executable` button.



USEARCH troubleshooting...

The USEARCH software is either not installed or reachable by TriFusion. Please ensure that you have downloaded and installed the USEARCH software from the website:

http://www.drive5.com/usearch/download.html

If USEARCH is already installed, provide the executable file via the button bellow.

- Step 3: Search for the executable you have downloaded in **Step 1** and click the `Save` button.

And that's it. When a valid USEARCH executable is provided, the previous warning should be replaced with a green box saying "*USEARCH is installed and reachable*". You are good to go!

1.7 Search orthologs

Note: Data availability for this tutorial: the data set of 10 fungal proteomes that will be used can be downloaded here.

Warning: Before following this tutorial, make sure that USEARCH is correctly setup on your system and reachable by TriFusion (see *Setup of USEARCH*).

1.7.1 Load proteomes

As already covered in a separate tutorial (see *Load data into TriFusion*), proteome files can be loaded in three different ways. Here, we'll use the file browser to load 10 proteome files.

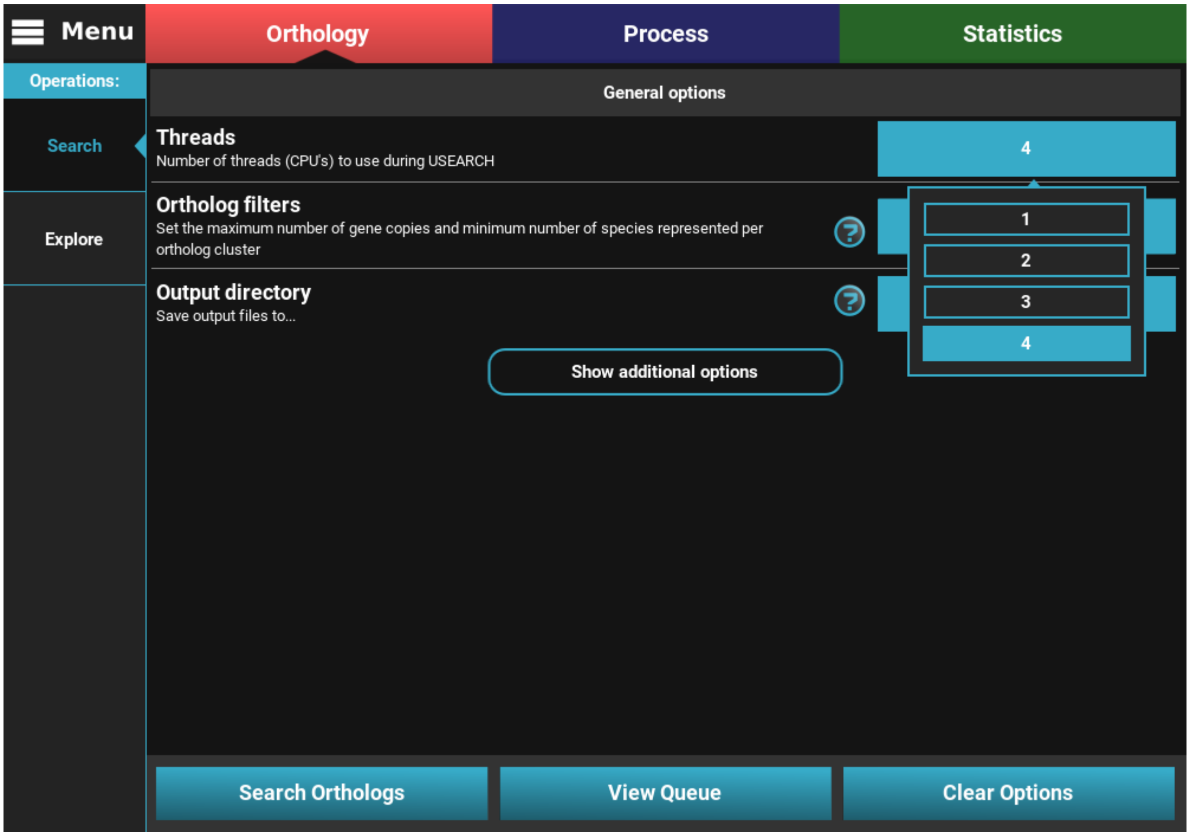
Navigate to `Menu -> Open/View Data` and click the `Open file(s)` button. This will open the main file browser. Set the *Input data type* at the top of the screen as **Proteome**. Then, go to the directory containing the protome files, select them and click `Load & go back`.

If the files are correctly formatted (see proteome format) they should be successfully loaded and appear in the `Open/View Data` sidepanel under the *Files* tab.

1.7.2 Orthology search options

Now let's set the general options for the orthology search by navigating to the `Orthology` screen. There are three general options:

- *Threads*: Sets the maximum number of CPU's that will be used by USEARCH during the most computationally intensive phase of the search. TriFusion automatically detects the number of CPU's on your system and sets it as the maximum value available. In this example, I'll choose 4 CPU's, which is the maximum of my system.



Threads
Number of threads (CPU's) to use during USEARCH
4

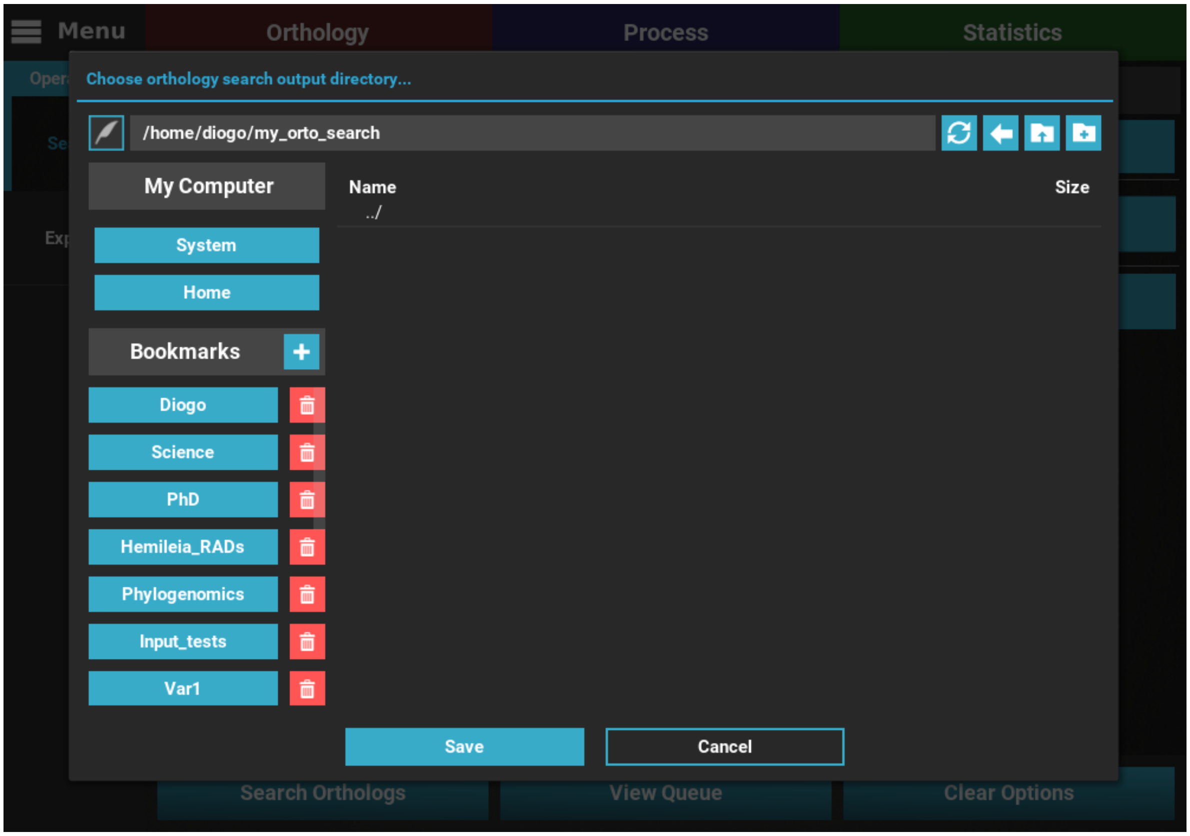
Ortholog filters
Set the maximum number of gene copies and minimum number of species represented per ortholog cluster

Output directory
Save output files to...

- *Ortholog filters*: Sets the filters that will be applied to the orthologs at the end of the search operation. Here you can set the maximum number of gene copies for each ortholog group, and the minimum number of species that must be contained in an ortholog group. Here, I'll set a maximum number of gene copies of 1 (only single copy genes) and the minimum number of taxa to 5 (50%).

  Note: These filters will not be permanent. They will be used to export the fasta sequence files at the end of the search operation, by they can can still be changed after the end in the **Explore** section. The final ortholog group files will contain **all** orthologs, regardless of the filters used here.

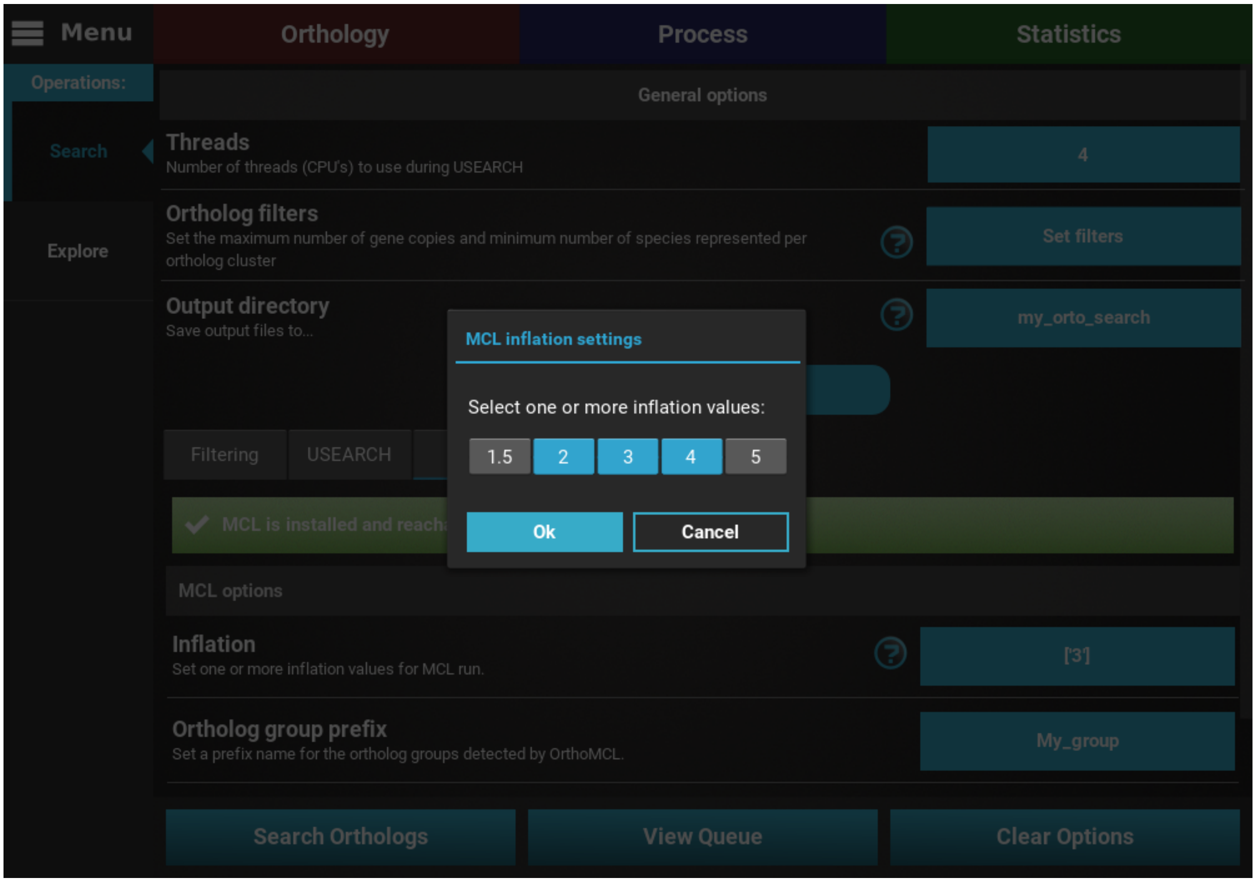
- *Output directory*: Sets the directory where all output files (intermediary and final sequence files) will be generated. I'll create a directory named my_orto_search on my home directory and set it.



Choose orthology search output directory...

/home/diogo/my_orto_search

Setting these options would be sufficient to start our search operation. However, I'm still interested in experimenting multiple inflation values to see it's impact on the final number of orthologs. To set multiple inflation values, click on the `Show additional options` button, then click on the *MCL* tab, and finally on the button of the `Inflation` option. Here you can choose multiple pre-defined inflation values. I'll select three: 2, 3 and 4.



Threads
Number of threads (CPU's) to use during USEARCH
4

Ortholog filters
Set the maximum number of gene copies and minimum number of species represented per ortholog cluster
Set filters

Output directory
Save output files to...
my_orto_search

MCL inflation settings
Select one or more inflation values:
1.5 2 3 4 5

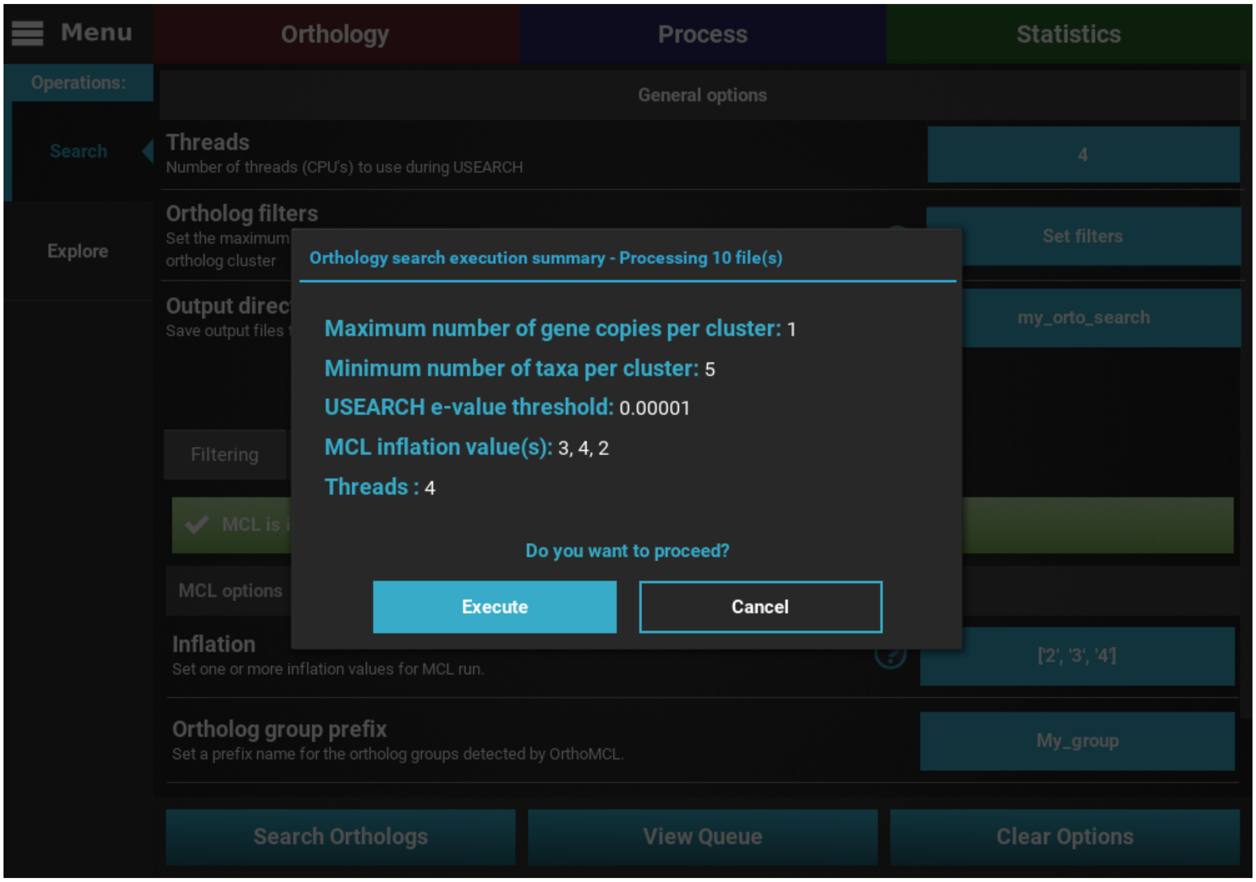
Inflation
Set one or more inflation values for MCL run.
[3]

Ortholog group prefix
Set a prefix name for the ortholog groups detected by OrthoMCL.
My_group

1.7.3 Start the search

Now we're all set to start our ortholog search. Click the `Search Orthologs` button at the bottom of the screen to display the execution summary dialog. Here you can see that 10 files will be processed, the values of the two ortholog filters, the e-value threshold for the USEARCH operation, the selected MCL inflation values and the number of CPU threads for the USEARCH execution. Click on the `Execute` button to start the search.

Note: The execution of the search operation may take a considerable amount of time, depending on the number of proteomes and their corresponding size. In my system (Intel i5-3317U @ 1.70Ghz, 4Gb RAM, HD4000) the search of the 10 proteome files took about 50 minutes. If you are only testing and wish to obtain the results sooner, try to deselect some protome files from the active data set in `Menu -> Open/View Data`.



Threads
Number of threads (CPU's) to use during USEARCH
4

Ortholog filters
Set the maximum
ortholog cluster
Set filters

Output direc
Save output files
my_orto_search

Orthology search execution summary - Processing 10 file(s)

Maximum number of gene copies per cluster: 1

Minimum number of taxa per cluster: 5

USEARCH e-value threshold: 0.00001

MCL inflation value(s): 3, 4, 2

Threads : 4

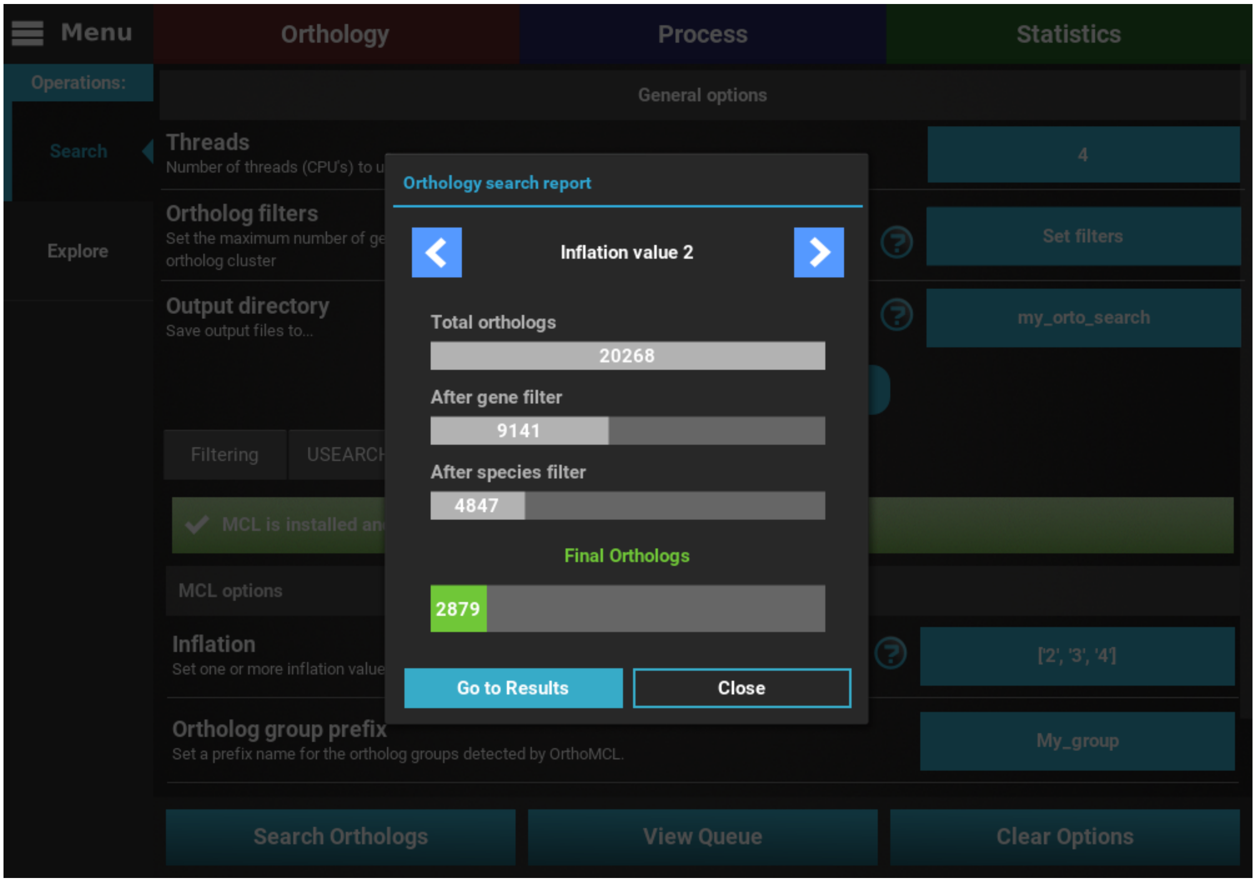
Do you want to proceed?

Inflation
Set one or more inflation values for MCL run.
['2', '3', '4']

Ortholog group prefix
Set a prefix name for the ortholog groups detected by OrthoMCL.
My_group

1.7.4 The orthology search report

At the end of the search operation, a report dialog will appear with the search results for each inflation value.



Orthology search report

Inflation value 2

Total orthologs
20268

After gene filter
9141

After species filter
4847

Final Orthologs
2879

Threads
Number of threads (CPU's) to u
4

Ortholog filters
Set the maximum number of ge
ortholog cluster

Output directory
Save output files to...
my_orto_search

Inflation
Set one or more inflation value
['2', '3', '4']

Ortholog group prefix
Set a prefix name for the ortholog groups detected by OrthoMCL.
My_group

You can use the top arrow buttons to cycle through all selected inflation values. For each inflation value the number of total and filtered orthologs appear in graphical format. The orthologs that pass the maximum gene and minimum species filters appear individually, so that you can assess the impact of each filter. At the bottom, in green, the final number of orthologs that passed both filters is shown.

From this point, you can either further explore your newly detected orthologs by clicking the `Go to Results` button, or close the dialog and proceed on your own with the new result files.

1.7.5 Output directories and files

The results of the orthology search will be stored in the directory that you specified in the *Output directory* option. Inside, you will have two directories: a *backstage_files*, where the proteome database and all intermediate files were stored, and a *Orthology_results*, where the final output files were generated. Inside the *Orthology_results* directory, a **groups** file and a directory with the ortholog group Fasta files will be created for each inflation value specified before the search.

The ortholog group Fasta files already have the sequence name headers normalized for each taxa (or proteome). This means that the Fasta headers will be something like:

```
>TaxonA
MDG(...)
>TaxonB
MGF(...)
```

Instead of the original headers in the proteome files. However, if you wish to make the correspondence of particular sequence with their original names in the proteome files, a directory named *header_correspondance* is created with a list for each ortholog group.

1.8 Explore ortholog search results

Note: Data availability for this tutorial: the three group files used in this tutorial can be downloaded here

1.8.1 Load group files

Note: There are three ways of loading data in TriFusion. Here we'll use the file browser.

The input data of the *Explore* operation of the **Orthology** module are the group files that are generated at the end of the ortholog search operation. These are simple text files that contain the definition of an ortholog group in each line. A typical group file should start with something like:

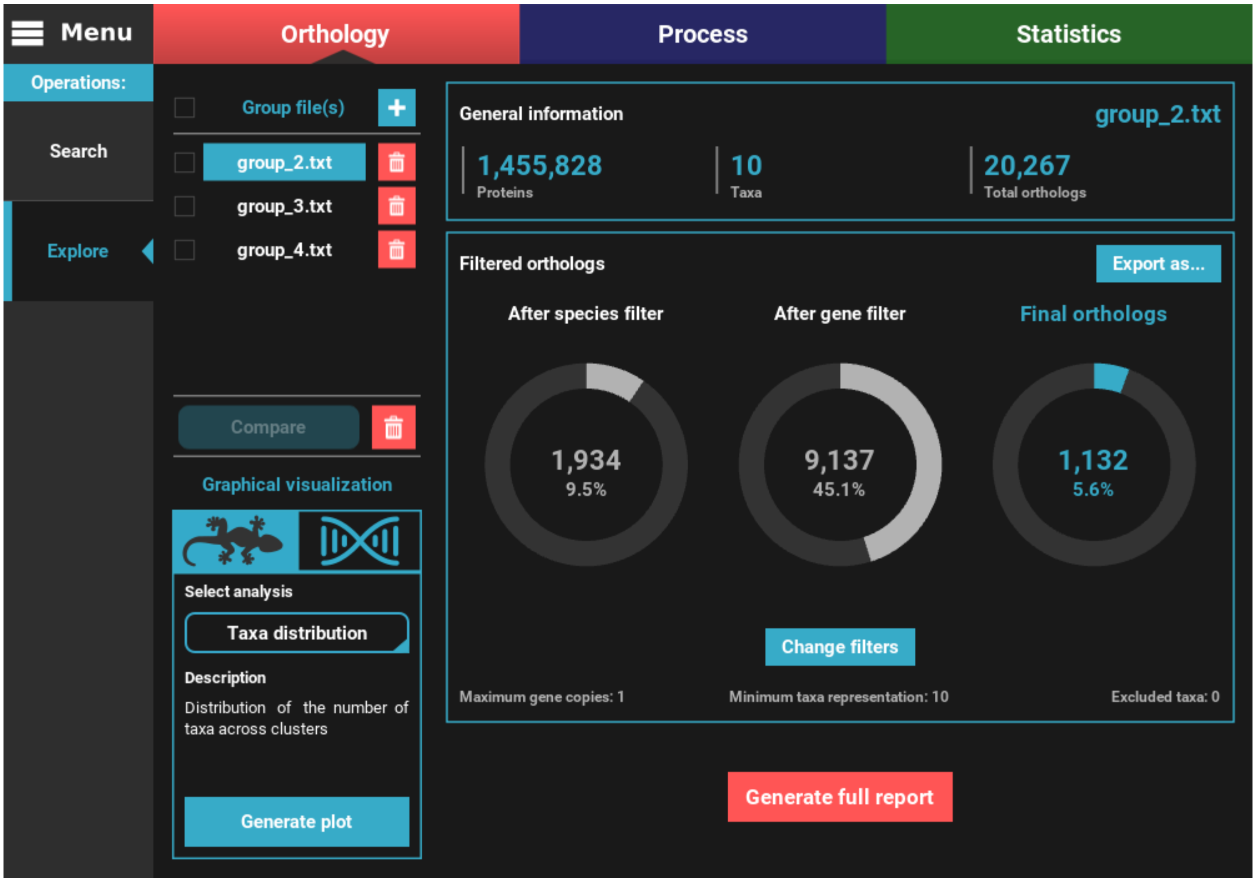
```
Ortholog1: Afumigatus_proteins|433 Anidulans_proteins|4605 (...)
Ortholog2: Afumigatus_proteins|3278 Afumigatus_proteins|9183 (...)
Ortholog3: Anidulans_proteins|36 Anidulans_proteins|9893 (...)
(...)
```

If you are loading group files from previous ortholog search runs, they will be found inside the specified output directory, in the *Orthology_results* directory.

To load the data, navigate to the `Orthology` screen and click the `Explore` operation on the left of the screen. Then click the + button on the top left of the screen to open the file browser. Navigate to the directory containing the group files and then select files. In this case, we will select the three group files generated in a previous search operation that was performed with inflation values 2, 3 and 4.

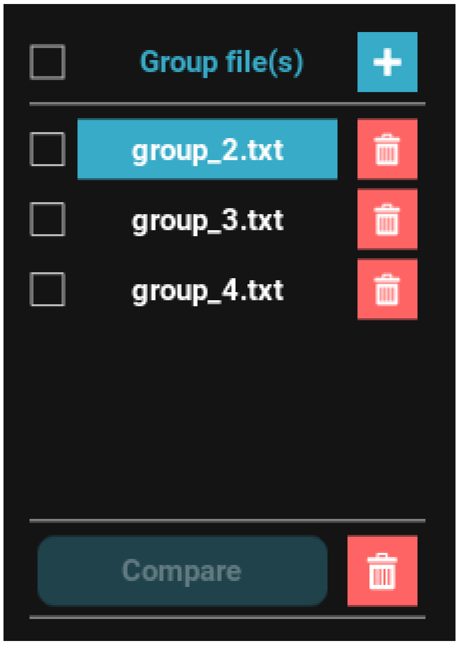
1.8.2 The orthology explore screen

Once the group files are loaded into TriFusion, several descriptive statistics will populate the screen.

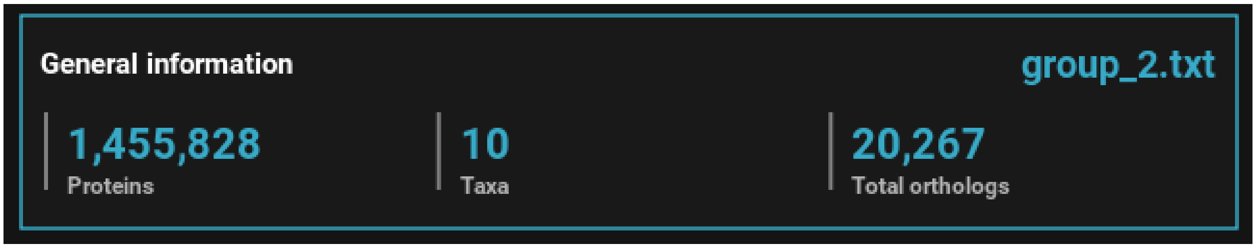


Distribution of the number of taxa across clusters

To the left, the **loaded group files** are listed under the **Group file(s)** section, where they can be selected to visualize the statistics specific to that group. They can also be removed by clicking the trash bin red button.



On the remaining of the screen, general statistics and information on the filtered orthologs are presented for the currently selected group file. The **General information** section informs the total number of proteins, taxa and ortholog groups contained in the group file.



General information

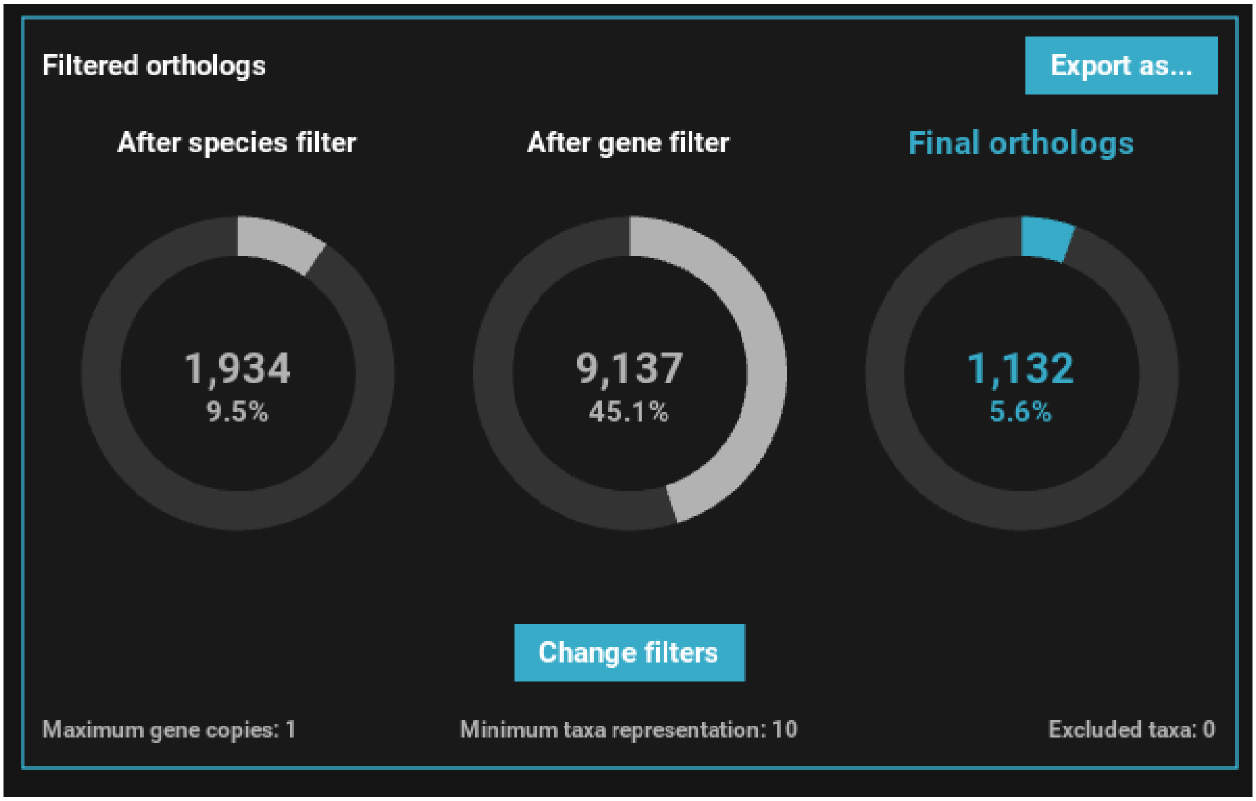
group_2.txt

1,455,828 Proteins

10 Taxa

20,267 Total orthologs

Below, in the **Filtered orthologs** section, the number of orthologs after applying the filters is displayed in gaussian plots. The values displayed are for the default ortholog filters, wich are set to single copy genes (maximum gene copies of 1) and with all taxa present (minimum number of taxa equal to the number of taxa).



Filtered orthologs

After species filter: 1,934 9.5%

After gene filter: 9,137 45.1%

Final orthologs: 1,132 5.6%

Maximum gene copies: 1

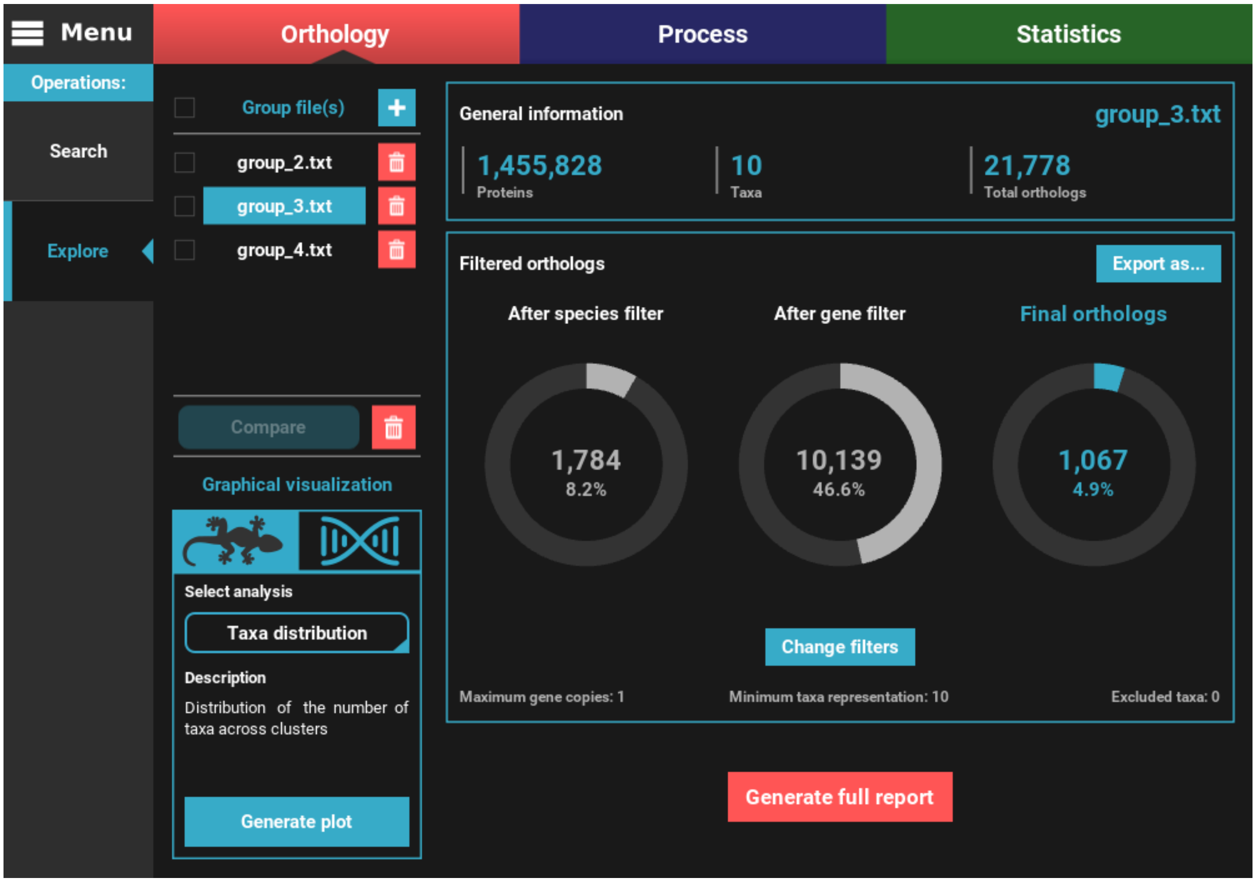
Minimum taxa representation: 10

Excluded taxa: 0

In our case, we can see that the **group_2.txt** file contains around 1.5M proteins for 10 taxa, clustered in 20k ortholog groups. From these 20k ortholog groups, 1 934 passed the species filter (minimum number of taxa), 9 137 passed the gene filter (maximum number of gene copies) and **1 132** passed both filters. This indicates that the species filter is the major limiting factor in the final number of ortholog groups.

1.8.3 Change the active group

To change the active group file, simply click the group button in the **Group file(s)** list section in the top left of the screen. Let's change the active group file to the **group_3.txt** file.



As you can see, the numbers of total and filtered ortholog groups changed slightly, which is a result from using different inflation values during the search operation.

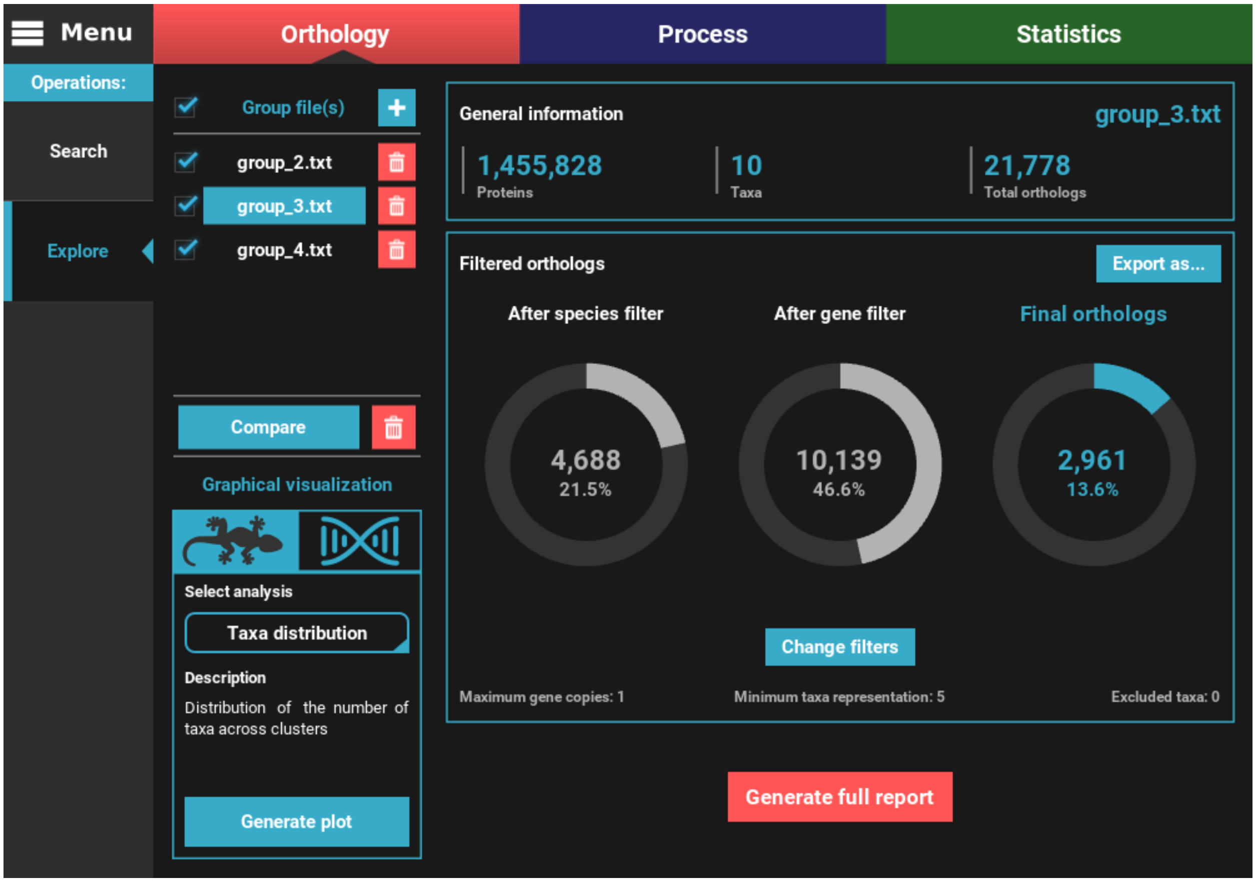
1.8.4 Change the orthology filters

A common procedure during the exploration of the orthology search results is the modification of the *ortholog filters*. To change the filters, click the `Change filters` button in the bottom of the **Filtered orthologs** section. This will open the ortholog filters dialog where you can change the maximum number of gene copies, minimum number of taxa for the ortholog groups and exclude/include taxa from the ortholog groups. Let's maintain the gene copy filter and only allow for single copy genes, but relax the minimum number of taxa to half of the data set (5). The `Apply filter to all group files` check box will also remain active to update all group files with the new filter. When all filters are set, click the `Ok` button to update.

After the application of the new filters, you can see that the number of filtered orthologs changes. The number of final orthologs for the **group_3.txt** file almost tripled when we relaxed the number of minimum taxa per ortholog group. You can also see that the filter values were updated at the bottom of the **Filtered orthologs** section.

1.8.5 Compare group files

To easily compare the number of total and filtered ortholog groups among different group files, you can check the boxes to the left of the group files in the **Group file(s)** list section. To select/deselect all group files, you can also check the top checkbox. Here, let's compare all group files by selecting all and then clicking on the `Compare` button.



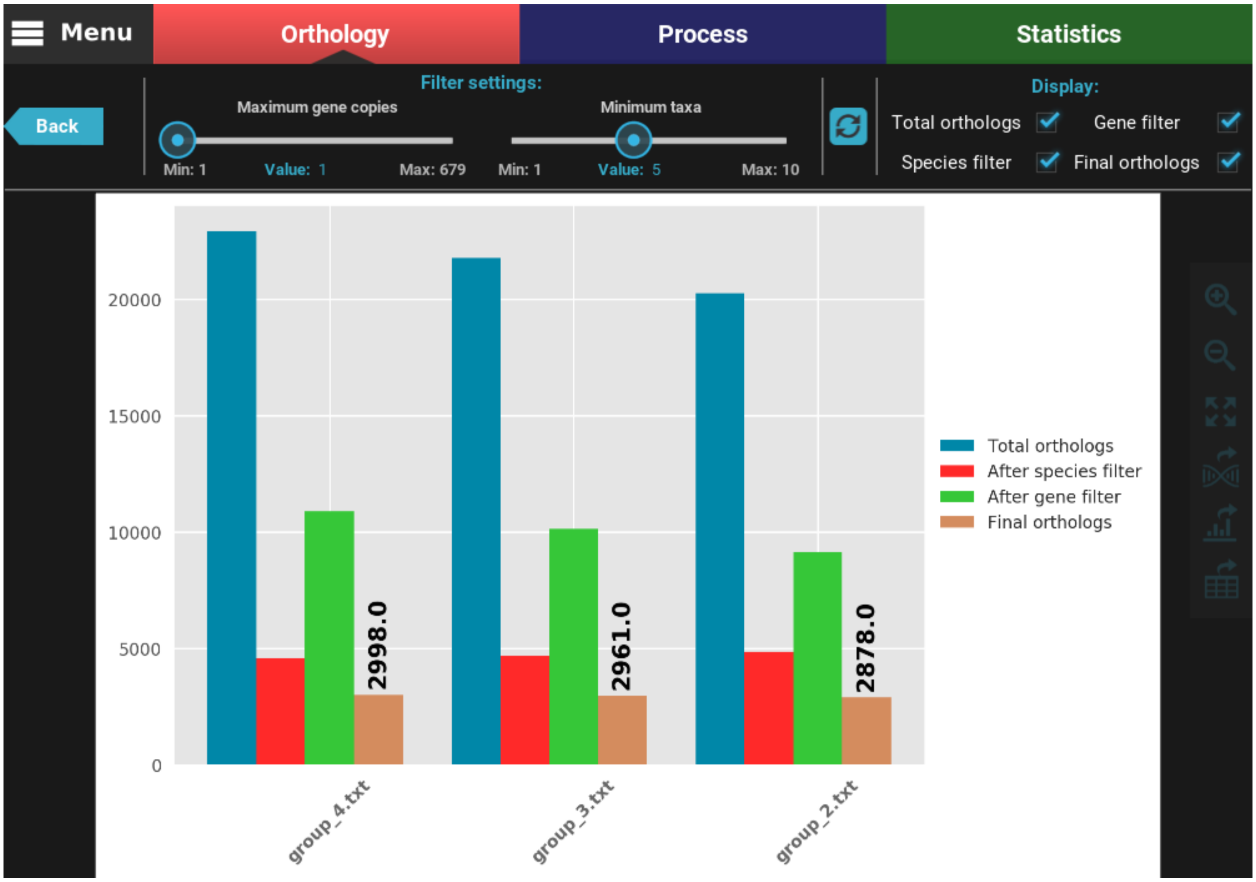
Distribution of the number of taxa across clusters

Maximum gene copies: 1

Minimum taxa representation: 5

Excluded taxa: 0

This will bring you to a plotting screen, where a bar plot will be displayed with the number of total and filtered ortholog groups for each group file. You can interact with the plot by pressing the left mouse button and dragging the plot. You can also zoom in and out using Ctrl + mouse wheel or by clicking the corresponding buttons on the right side panel.



Filter settings:

Maximum gene copies — Min: 1 Value: 1 Max: 679

Minimum taxa — Min: 1 Value: 5 Max: 10

Display:

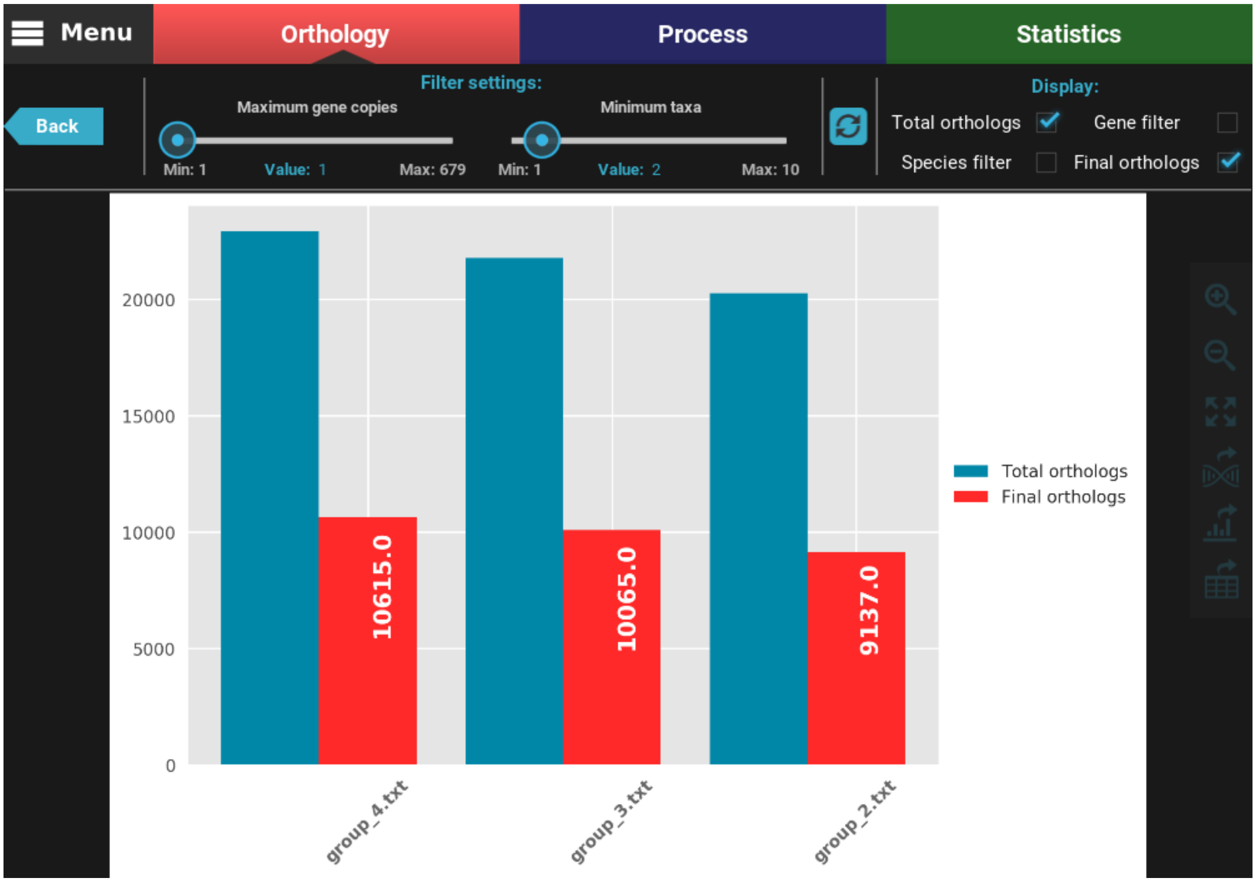
Total orthologs ☑ Gene filter ☑

Species filter ☑ Final orthologs ☑

At the top of the screen, you can see the currently active filters, which are the same we set in the previous section. **Note if taxa were excluded previously for the active group file, those taxa will also be excluded here.** You can change the filter values using the sliders. Let's try to relax even further the minimum number of taxa to 2. After changing the slider value (or changing the "Value" number), you can see that the *refresh* button turned red, which means that you have set different filter. To update the plot, click the refresh button.

After clicking the refresh button, the plot values will be updated. You can see now that the total number of orthologs is almost 10k for all group files and that there is almost no different between the gene filtered and final ortholog groups. Indeed, we can see that the final number of orthologs does not deviate much between group files (range between 9 137 and 10 615).

You can also change which type of ortholog groups are displayed by ticking the check boxes in the **Display** section on the top right of the screen. Let's visualize only the total and final number of orthologs. To accomplish this, uncheck the **Gene filter** and **Species filter** boxes.

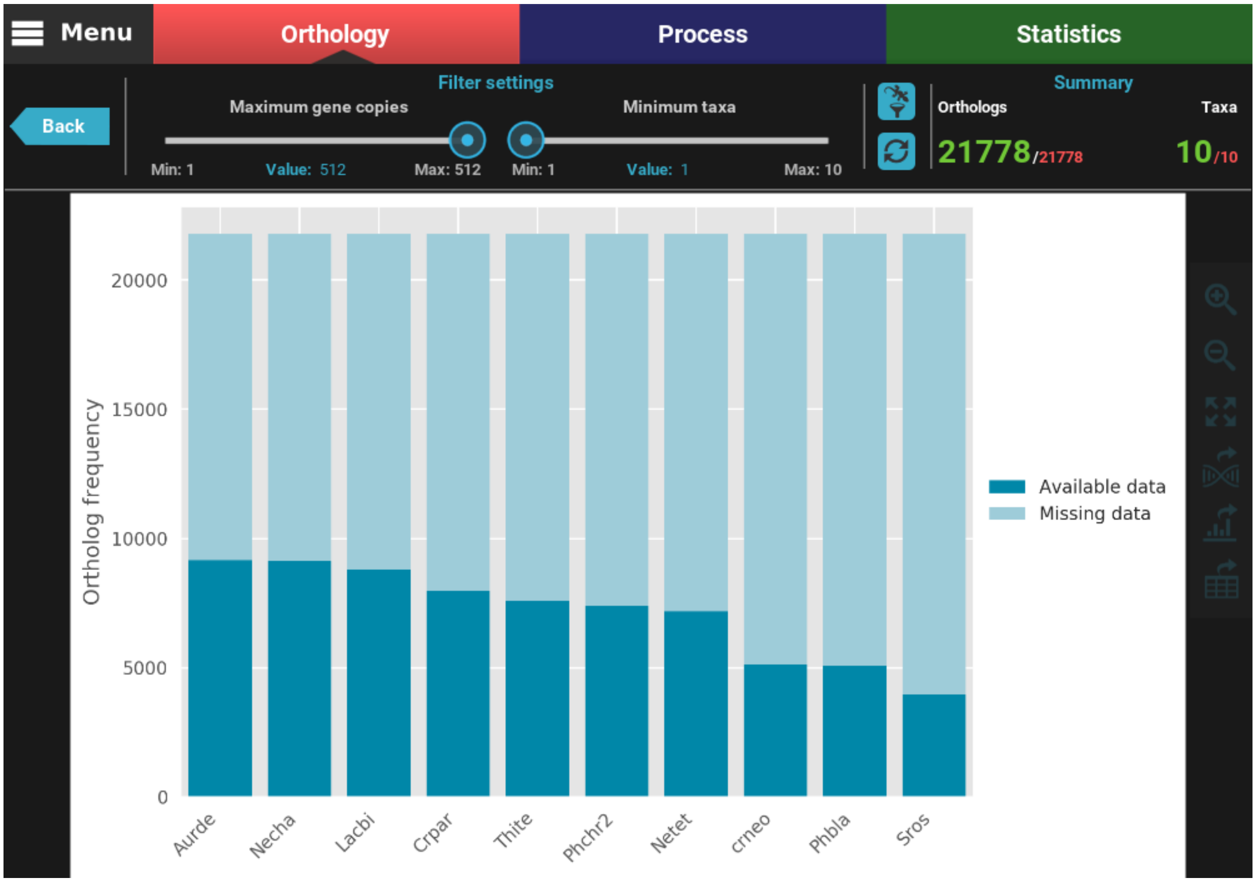


At any time, you can export the current plot in figure or table format by clicking the `Export as graphics` or `Export as table` buttons, respectively, in the right side panel.

1.8.6 Graphical visualization of group files

Individual group files can also be further visually explored using the plotting tools under the **Graphical visualization** section in the bottom left of the screen. Graphical visualization options are sorted into **Species focused** exploration and **Ortholog focused** exploration. Clicking on either option will present a drop down menu where specific plotting options are available. When one of these options is selected, a short description is shown below. Let's investigate the taxa coverage of the currently active group file, by selecting the **Species focused** exploration and the *Taxa coverage* plot option. Then, click on the `Generate plot` button.

This will open a plot screen akin to the one displayed when comparing different group files. In this specific plot you can see, for each taxa, the proportion of ortholog groups where they are present (dark blue) or missing (light blue). In the top right of the screen, under the **Summary** section, you can see the total (red) and filtered (green) number of ortholog groups and taxa that are being used to generated the plot. In this case, a total of 21 777 ortholog groups across 10 taxa are being used. As you can see, by default, all plotting options will set the filters to their most relaxed values (allowing for all gene copy numbers and any taxa representation).



The plot can be interacted with by clicking and dragging and by zooming in and out. In the header of the screen, the ortholog filters can be changed. Let's change the filter setting so that only single copy genes with at least 5 taxa represented are considered. When the filters are modified, the refresh button should turn red and must be clicked in order to update the plot.

After the plot is updated, you can see that the values in the **Summary** section of the header have also updated. This plot is now being generated with 2 691 ortholog groups across 10 taxa. We can also see that, using these filter values, all taxa have a pretty decent proportion of available data. However, you have also the option to remove specific taxa from this analyses, by clicking the *filter taxa* button in the header above the refresh button. Clicking this button will display all taxa listed. These can be toggled in or out by clicking the respective buttons. For exampled, let's remove the last two taxa, `Thite` and `crneo`, by clicking them once.

As you can see, the bars of the removed taxa are no longer in the plot and the numbers in the **Summary** section of the header were updated to 8 active taxa.

As in the compare groups plot screen, all plots in the **Graphical visualization** section can be exported into figures or table formats by clicking the `Export as graphics` or `Export as table` buttons, respectively. The filtered ortholog groups can also be exported to a new groups file, to protein or nucleotide sequences, by clicking the `Export group` button (see *Export ortholog groups as protein or nucleotide sequences*).

1.8.7 Generation of full report for single groups

All plotting options in the **Graphic visualization** section can be automatically generated into a HTML file by clicking the `Generate full report` button at the bottom of the **Explore** screen. Then select the directory where the report will be generated. In that directory, an HTML file will be created where all plots will be visualized for the currently set ortholog filters.

1.9 Export ortholog groups as protein or nucleotide sequences

Note: Data availability for this tutorial: The data required to complete this tutorial include:

- ortholog group files
- protein database
- CDS files

This tutorial demonstrates how to export ortholog groups from a previous **Orthology** search operation as protein and nucleotide sequences.

1.9.1 Load group files

Let's import the results from the previous search of orthologs across 10 genomes (see tutorial Basic search of orthologs among 10 proteome files). Navigate to the **Orthology** screen, **Explore** section, and click the + button at the top left of the screen. Go to the directory containing the group files from the corresponding ortholog search operation and select one or more files. Here we'll select only one. Once loaded, the basic information of the group file will be displayed for the default orthology filters (only single copy genes present in all species).

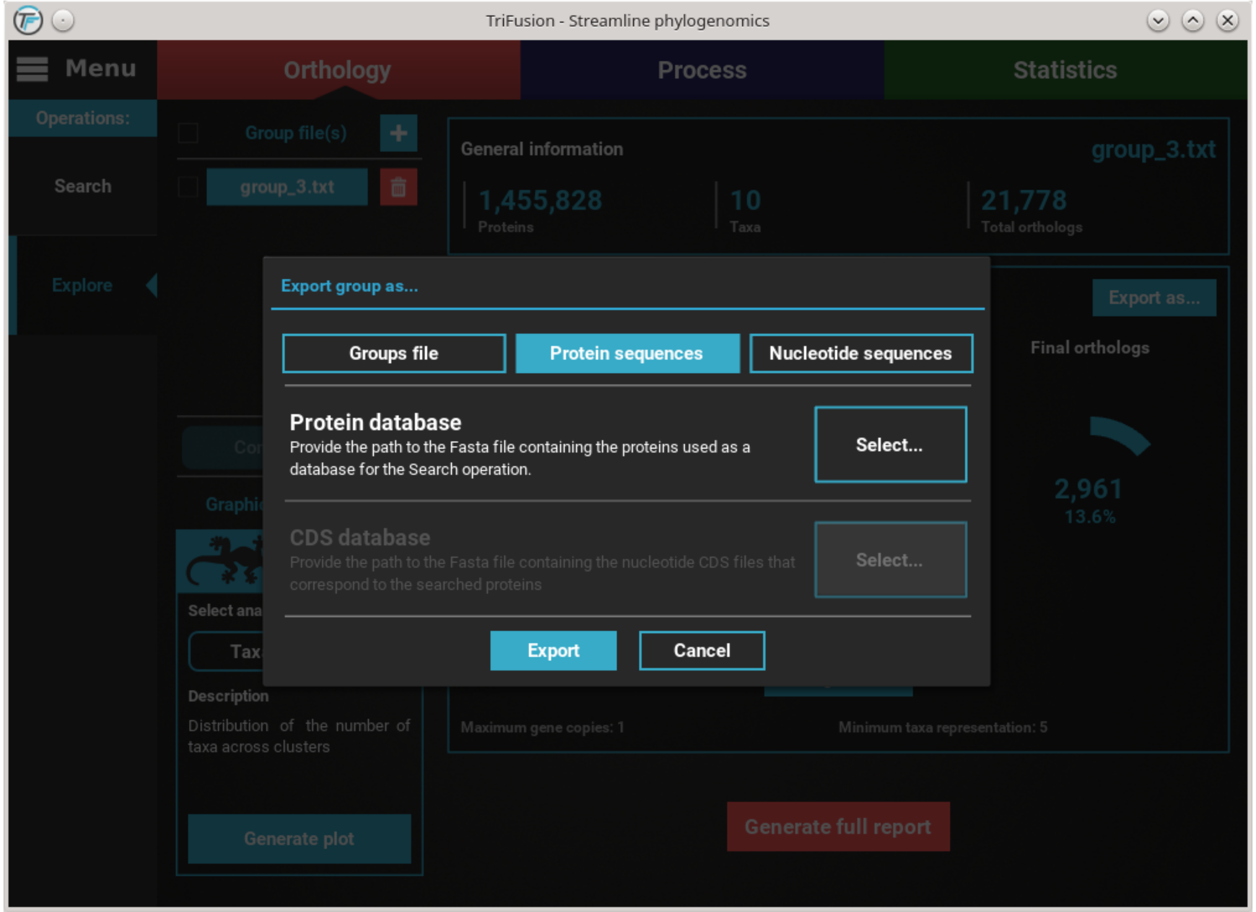
However, let's change the filters for something more permissive in terms of minimum taxa representation. Click the `Change filters` button, and change the minimum number of taxa value to 5 (50% of taxa representation). Click `Ok` and the information on the screen should be updated to something like this.

1.9.2 Export into protein sequences

First click the `Export as...` button in the **Explore** section screen. This will open the export group dialog. To export the ortholog groups into protein sequence files (in Fasta format), a **protein database** of all input genomes must be provided. This file is automatically generated during the **Orthology** search operation and is stored in the backstage_files directory, with the default name of *goodProteins_db* (this name can be change by the *Database name* option). **If you have just finished an Orthology search operation in the current session of TriFusion, this database file is already set**. However, if you are executing a different session of TriFusion, you'll need to provide this file.

A **protein** database file is simply a Fasta file that contains all sequences used during the ortholog search procedure, with simplified headers. TriFusion will look for the sequence headers in the *groups* file and fetch the corresponding sequence from this database file.

Click the `Protein sequences` button. This will make the Protein database base option available. To search and select the database file, click the `Select...` button.



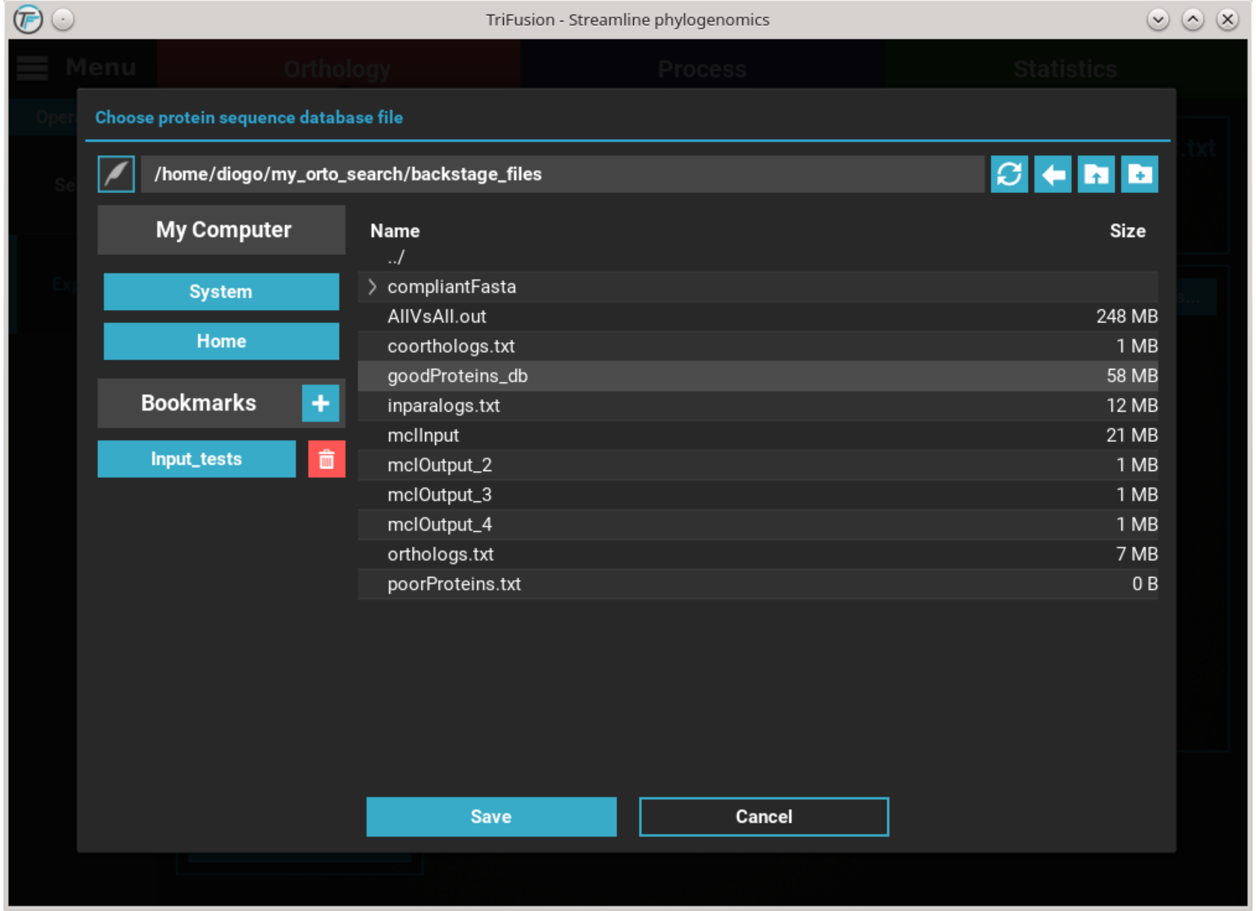
Protein database

Provide the path to the Fasta file containing the proteins used as a database for the Search operation.

CDS database

Provide the path to the Fasta file containing the nucleotide CDS files that correspond to the searched proteins

Notice that I navigated to the results directory of my previous ortholog search and then to the *backstage_files* directory. Since I did not change the *Database name* option value in TriFusion, I have a *goodProteins_db* file in this directory. If you are using the downloaded tutorials data, select the protein database file. Then click `Save`.

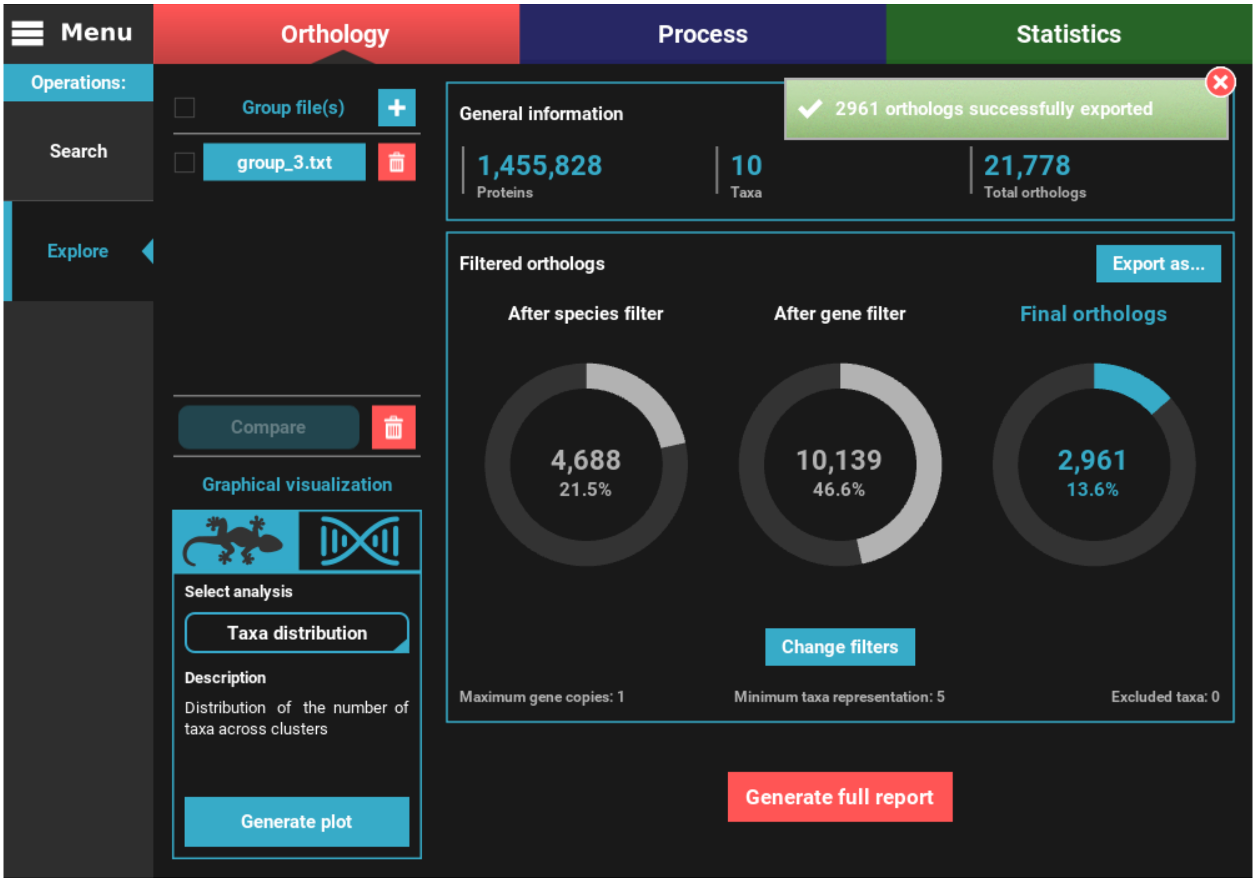


Choose protein sequence database file

/home/diogo/my_orto_search/backstage_files

Name	Size
../	
compliantFasta	
AllVsAll.out	248 MB
coorthologs.txt	1 MB
goodProteins_db	58 MB
inparalogs.txt	12 MB
mclInput	21 MB
mclOutput_2	1 MB
mclOutput_3	1 MB
mclOutput_4	1 MB
orthologs.txt	7 MB
poorProteins.txt	0 B

You'll notice that the `Protein database` button changed in accordance to the name of the protein database file. Finally, to export the ortholog groups click the `Export` button. Select or create a directory where the new files will be generated and then click `Ok`. At the end of the export operation, a success popup should appear informing the number of ortholog groups exported.



Select analysis

Taxa distribution

Description

Distribution of the number of taxa across clusters

General information

2961 orthologs successfully exported

1,455,828 Proteins

10 Taxa

21,778 Total orthologs

Filtered orthologs

After species filter: 4,688 21.5%

After gene filter: 10,139 46.6%

Final orthologs: 2,961 13.6%

Maximum gene copies: 1

Minimum taxa representation: 5

Excluded taxa: 0

Your protein sequence files are ready to be used in the specified directory. Notice that TriFusion will set the same name for each taxon/species across the protein sequence files. For instance, sequence references from a given species in multiple ortholog groups of *Necoc|153* and *Necoc|646* will be appear as *Necoc* in all sequence files. The correspondence between each taxon sequence and the original header in the groups file will be written in the *header_correspondance* directory, for each protein sequence file.

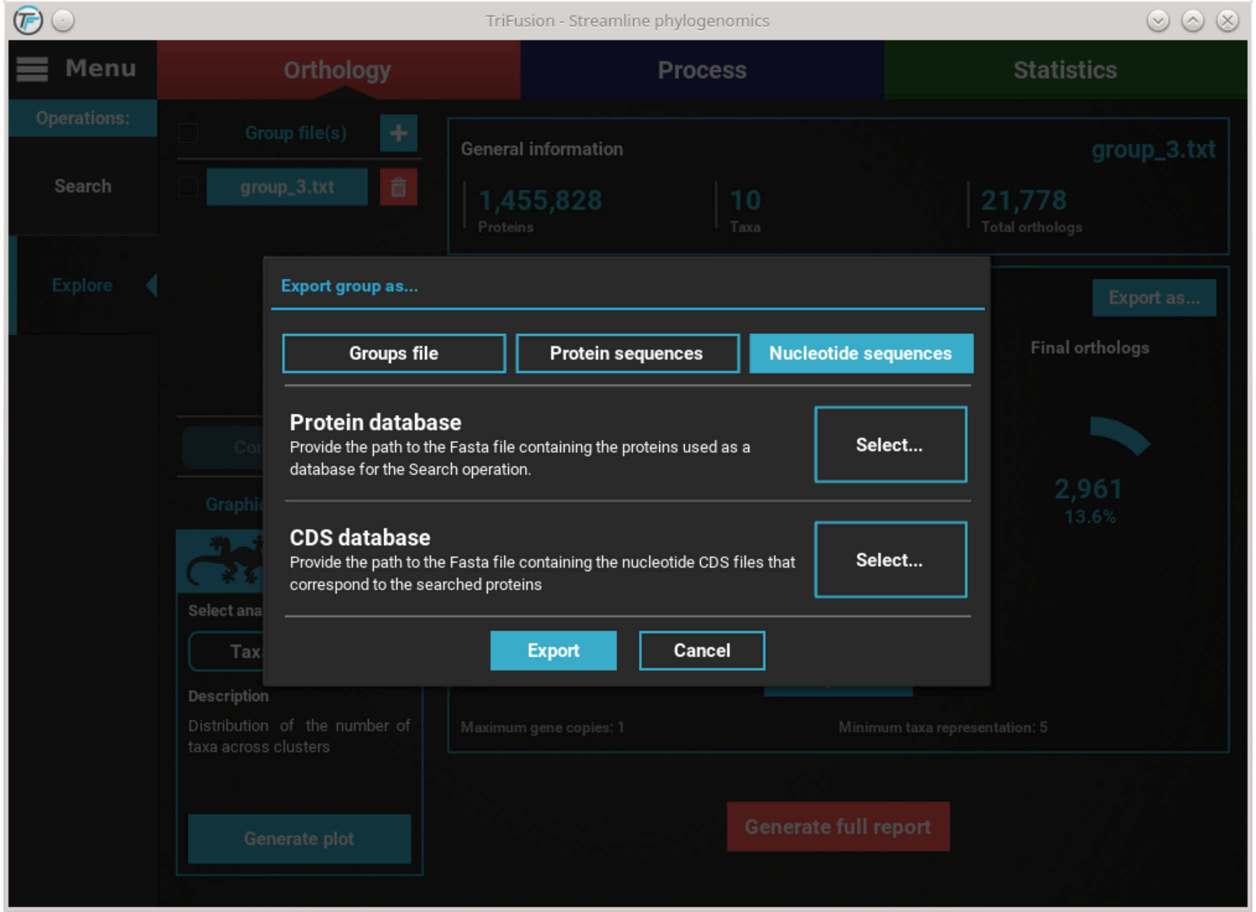
1.9.3 Export into nucleotide sequences

Note: To export ortholog groups, a working executable of USEARCH is required. See the *Setup of USEARCH* tutorial.

First click the `Export as...` button in the **Explore** screen. This will open the export group dialog. To export the ortholog groups into nucleotide sequence files (in Fasta format), a protein database **AND** cds/transcript files must be provided.

The CDS/transcript files are usually associated with the proteome files in genome sequencing projects.

Click the `Nucleotide sequences` button. This will make available the *Protein database* and *CDS database* options.



Protein database

Provide the path to the Fasta file containing the proteins used as a database for the Search operation.

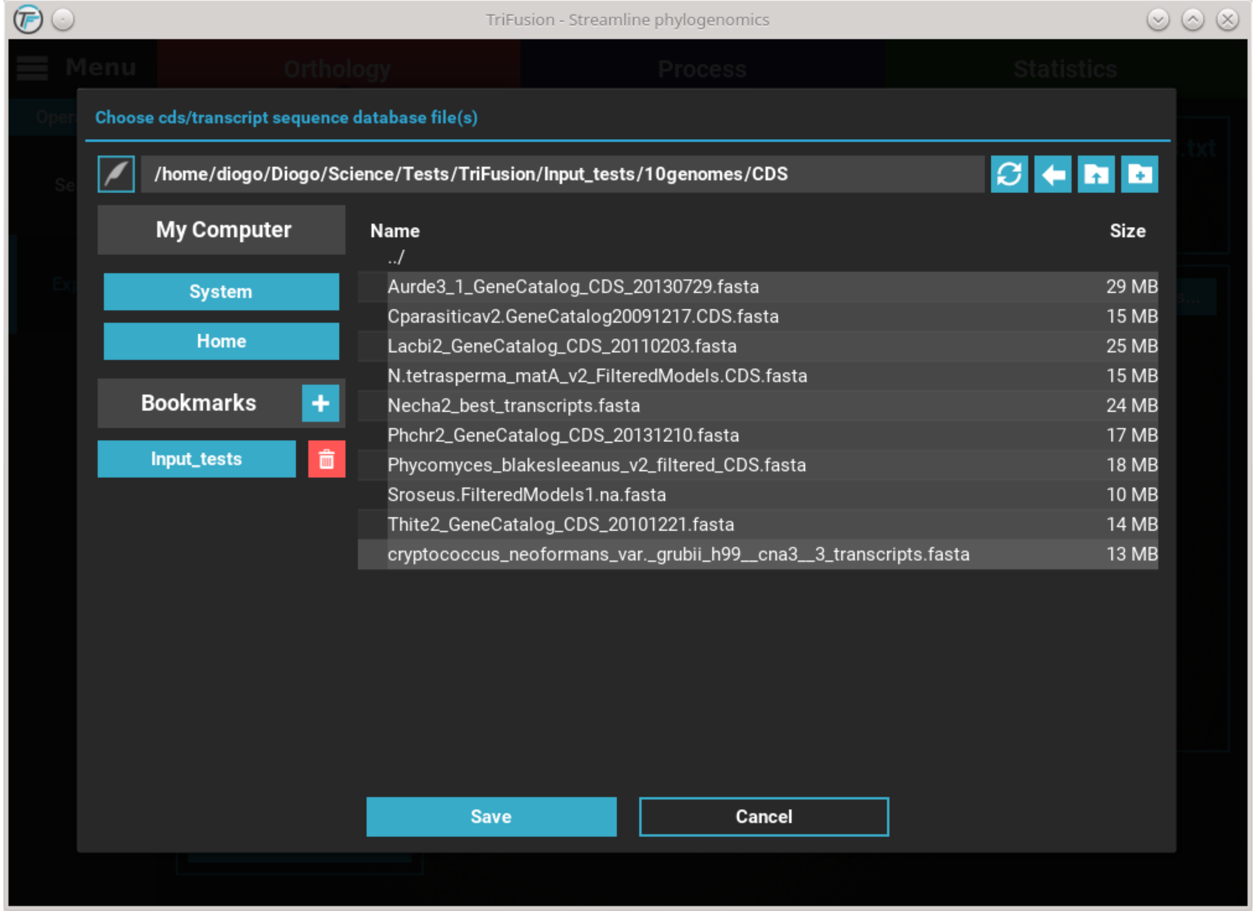
CDS database

Provide the path to the Fasta file containing the nucleotide CDS files that correspond to the searched proteins

Description

Distribution of the number of taxa across clusters

Refer to the previous *Export into protein sequences* section on how to set the protein database file. After setting this file, the cds/transcripts that correspond to the proteomes used during the **Orthology** search operation, must be also provided. You can have an individual cds/transcript file for each species, or concatenate all files into a single master file. Click the `Select...` button of the CDS database option and search for the cds/transcript files. If you are using the tutorial's material, provide the CDS files.

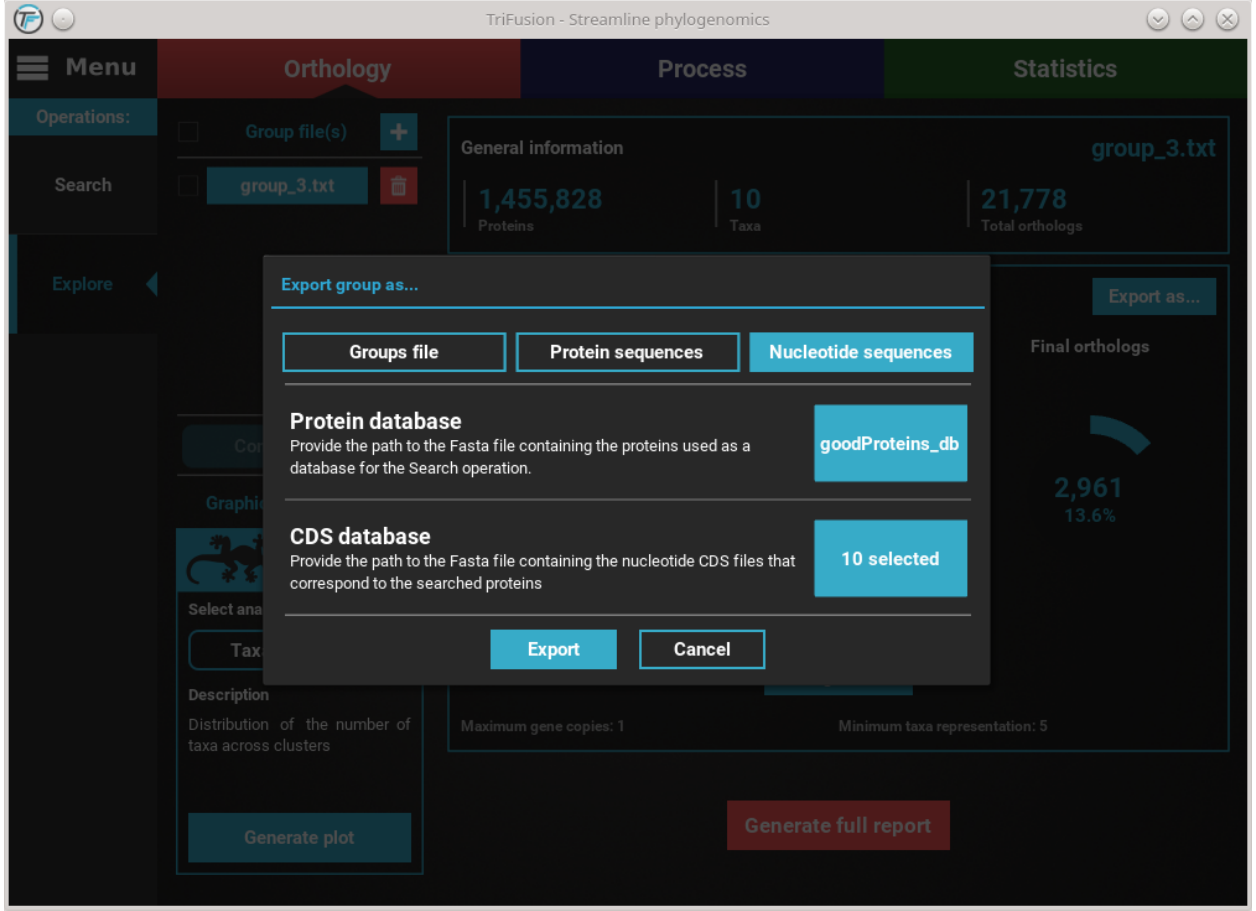


Choose cds/transcript sequence database file(s)

/home/diogo/Diogo/Science/Tests/TriFusion/Input_tests/10genomes/CDS

Name	Size
../	
Aurde3_1_GeneCatalog_CDS_20130729.fasta	29 MB
Cparasiticav2.GeneCatalog20091217.CDS.fasta	15 MB
Lacbi2_GeneCatalog_CDS_20110203.fasta	25 MB
N.tetrasperma_matA_v2_FilteredModels.CDS.fasta	15 MB
Necha2_best_transcripts.fasta	24 MB
Phchr2_GeneCatalog_CDS_20131210.fasta	17 MB
Phycomyces_blakesleeanus_v2_filtered_CDS.fasta	18 MB
Sroseus.FilteredModels1.na.fasta	10 MB
Thite2_GeneCatalog_CDS_20101221.fasta	14 MB
cryptococcus_neoformans_var._grubii_h99__cna3__3_transcripts.fasta	13 MB

Here, I have the CDS and transcript data for each of the 10 species in their respective individual files. Select them all with `shift + click` and click `Save`. You should notice that the CDS database button changed in accordance to the number of files select, which is 10 in this case.



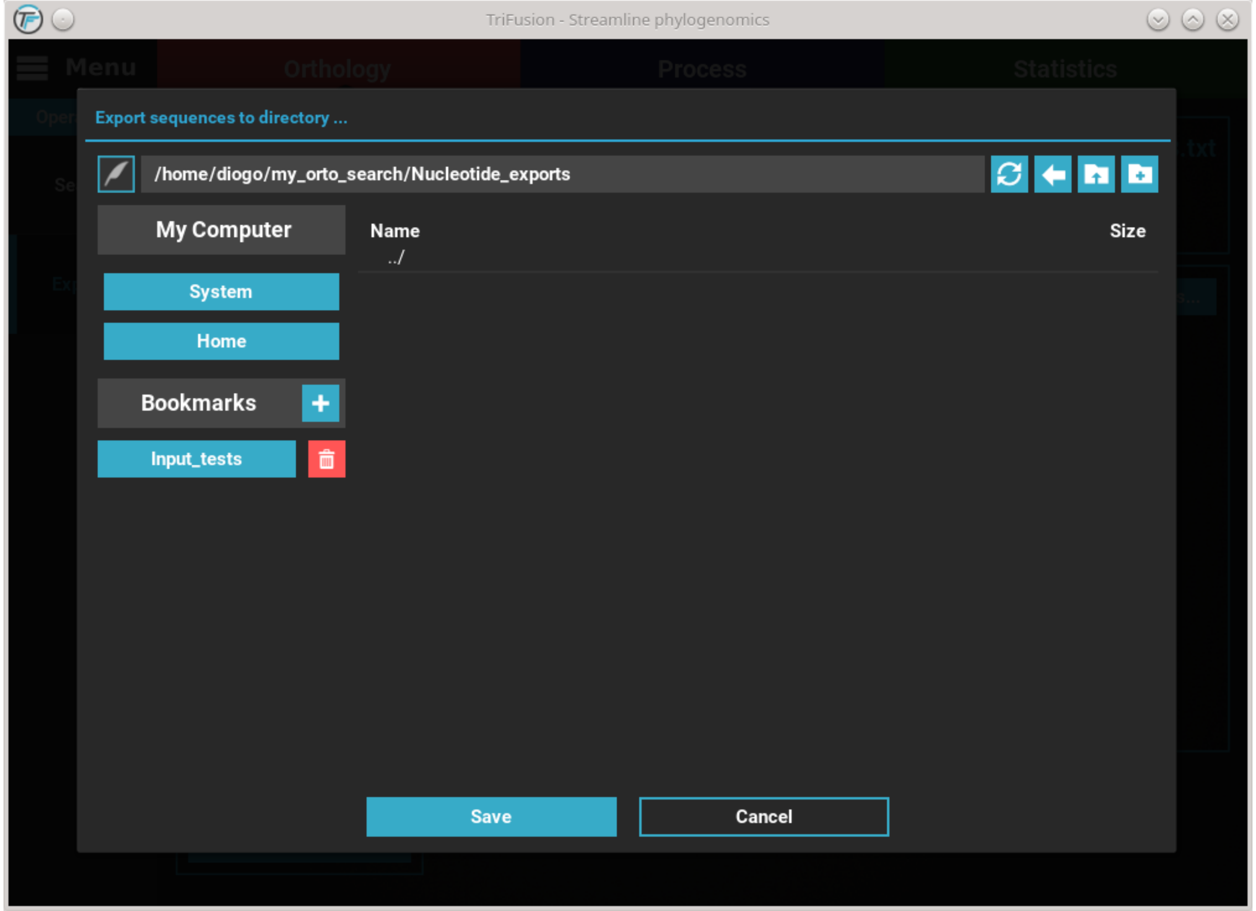
Protein database

Provide the path to the Fasta file containing the proteins used as a database for the Search operation.

CDS database

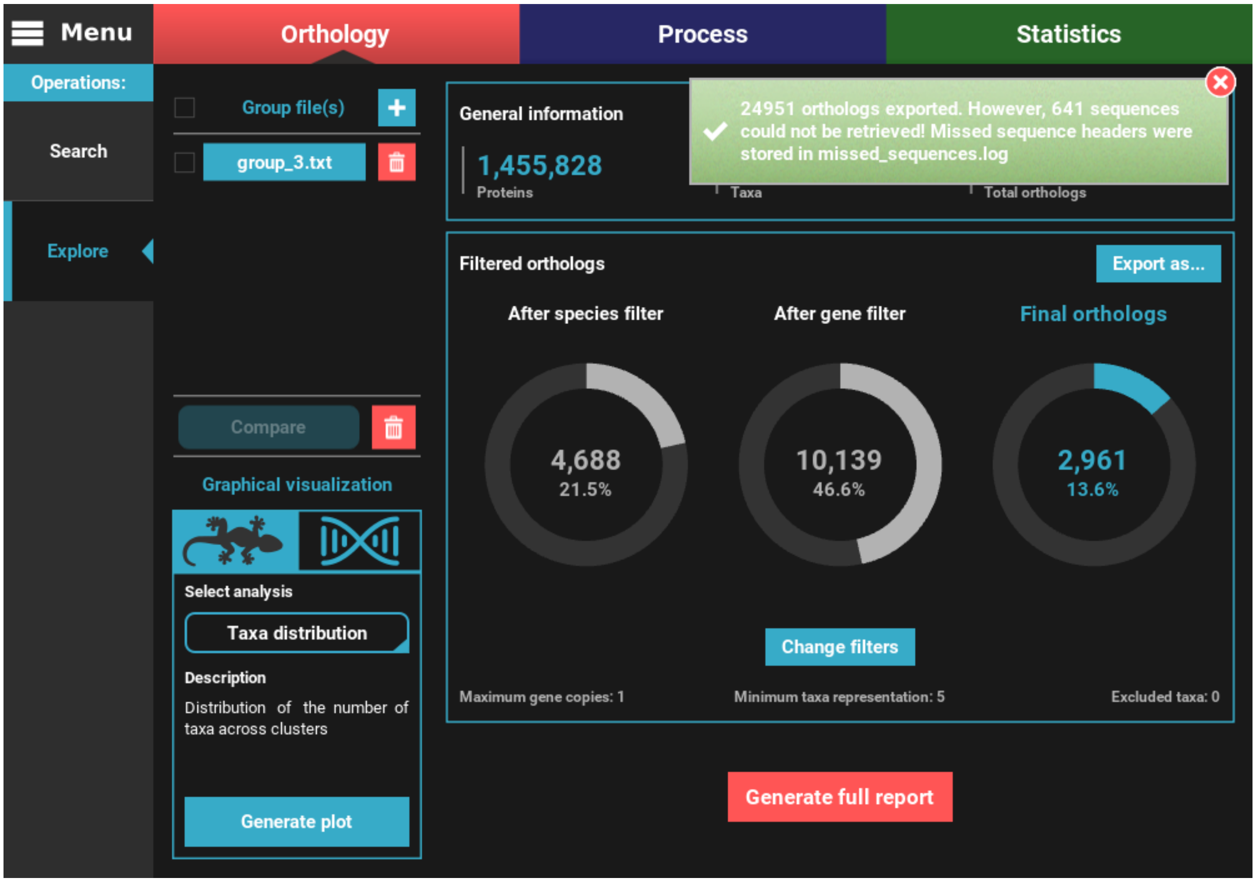
Provide the path to the Fasta file containing the nucleotide CDS files that correspond to the searched proteins

With both the protein database and cds/transcript files selected, we are ready to begin the ortholog export. Click the `Export` button and select or create the directory where you want to generate the nucleotide sequence files.



/home/diogo/my_orto_search/Nucleotide_exports

At the end of the export operation, a success popup should appear informing the number of sequences that were successfully exported.



24951 orthologs exported. However, 641 sequences could not be retrieved! Missed sequence headers were stored in missed_sequences.log

1,455,828 Proteins

4,688 21.5%

10,139 46.6%

2,961 13.6%

Maximum gene copies: 1

Minimum taxa representation: 5

Excluded taxa: 0

Description

Distribution of the number of taxa across clusters

Note: Note on the sequences that could not be retrieved:

TriFusion converts groups into nucleotide sequences by searching the proteins from the main output of the Search operation in CDS/transcript databases provided by the user. The reason why this search is done instead of simply looking for sequence headers that are the same in the protein and nucleotide databases is because sometimes there is no such cross reference. Therefore, TriFusion creates two different databases and then uses *USEARCH* to search for perfect hits between the protein and nucleotide sequences. This ensures that the nucleotide sequences correspond exactly to the proteins referenced during the **Orthology** search operation. However, even with this method, some nucleotide sequences may be absent from the databases. Fortunately, this represents only a minority of the cases. In this example, 641 protein sequence had no match in the nucleotide databases provided by the user, which represents only 2.8% of the total dataset. In most cases, this occurs only on a limited number of species but in any case, make sure that the proteome and CDS/transcript files correspond to the same version of the genome sequencing project.

1.10 Limitations for input files

The **Process** module deals with several input formats and sequence types, which begs the question of whether there are limitations on the type of files that can be loaded simultaneously into TriFusion.

The answer is almost none.

TriFusion was designed to capture all the details about your files automatically and to handle any combination you can throw at it. In the example below, 8 alignment files of nucleotide and protein sequences in Fasta, Nexus, Phylip and Stockholm formats are loaded simultaneously. Then, these files are easily concatenated into a single file with just a few clicks.

Moreover, defining partitions when there are multiple files and sequence types can be extremely time consuming and error prone to perform manually. That is why TriFusion handles all of that automatically. Even though we did not dealt with partitions in the above example, when you open a Nexus alignment file, you can see that the header and partitions block are correctly defined without any user intervention:

```
#NEXUS
Begin data;
    dimensions ntax=101 nchar=7030 ;
    format datatype=mixed(dna:1-3934,protein:3935-7030) interleave=no gap=-;

(... DATA ...)

begin mrbayes;
    charset DNAfas = 1-668;
    charset DNAnex = 669-1140;
    charset DNAphy = 1141-1808;
    charset DNAstockholm = 1809-2476;
    charset PROTEINphy = 2477-3934;
    charset PROTEINfasta = 3935-4966;
    charset PROTEINnex = 4967-5998;
    charset PROTEINstockholm = 5999-7030;
    partition part = 8: DNAfas, DNAnex, DNAphy, DNAstockholm, PROTEINphy,␣
↪PROTEINfasta, PROTEINnex, PROTEINstockholm;
    set partition=part;
end;
```

1.11 Basic conversion/concatenation

Note: Data availability for this tutorial: the medium sized data set of 614 genes and 48 taxa that will be used can be downloaded here.

1.11.1 Which input alignments can be used?

TriFusion was designed to impose as little limitations when loading alignment data as possible. All of the supported input formats and sequence types can be provided simultaneously to TriFusion. If you have nucleotide and protein sequence alignments in multiple formats, such as fasta, nexus, phylip, etc, you can load them simultaneously and all of the relevant information will be automatically detected.

When using the **Concatenation** operation, and if you are interested in generating the partitions definition in Nexus or Phylip formats, TriFusion will handle the partition ranges for you. If a mixture of nucleotide and protein alignments is loaded, the nucleotide and amino acid residue ranges will be sorted by sequence type, updating the partition ranges and generating the correct Nexus header.

The bottom line is that regardless of the type and format in which you have your data, it should be fine to load it into the application and TriFusion will deal with all the details automatically.

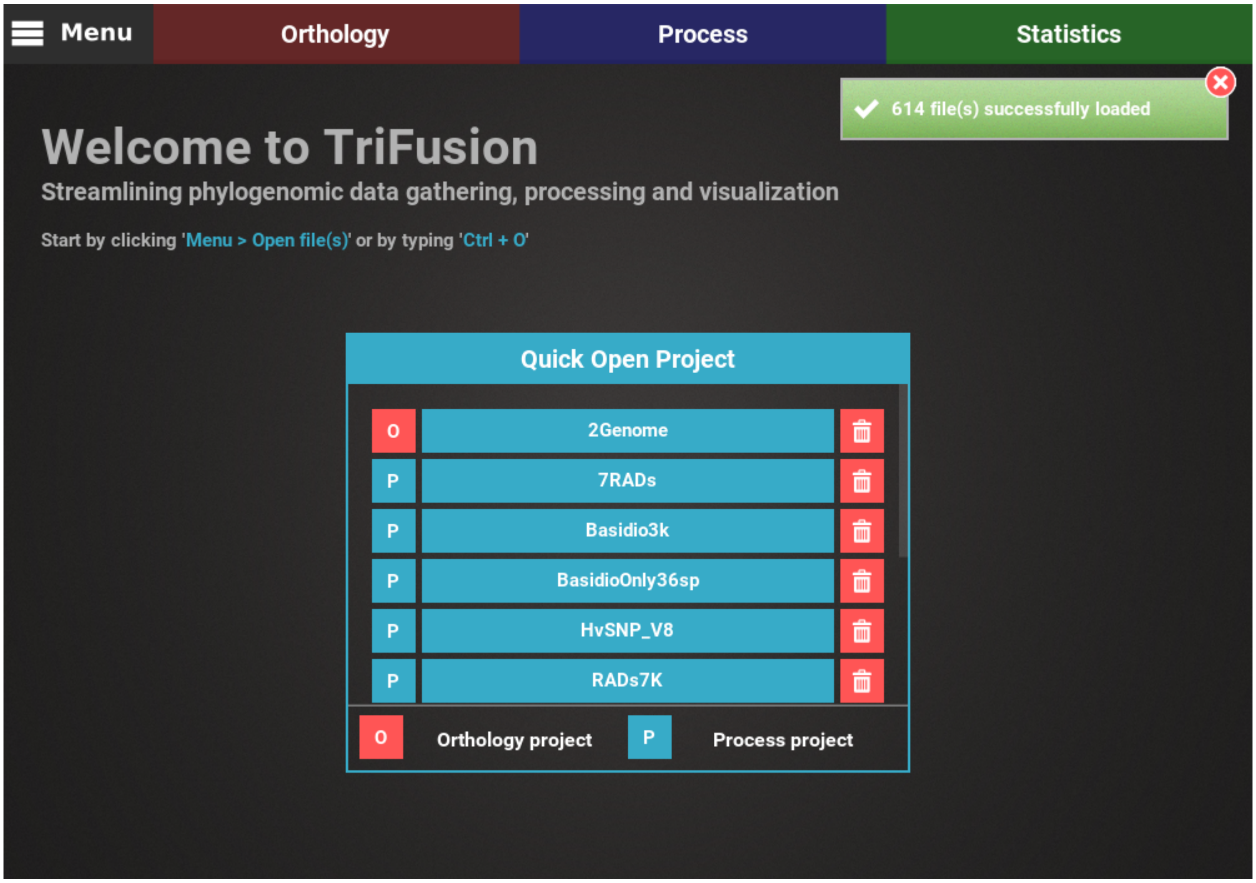
1.11.2 Load alignments

As already covered in a separate tutorial (see *Load data into TriFusion*), alignment data can be loaded into TriFusion in three different ways. Here we will use the file browser to load an entire directory where 614 alignments files are stored.

Navigate to `Menu > Open/View Data` and click the `Open file(s)` button. This will open the main file browser.

The input data type is already correctly set to **Alignment/Sequence set**, so we'll leave that as it is. Then, navigate the file browser until you find the directory containing the alignment files. In this case, all alignments are stored in a directory named *Version2*. Since TriFusion supports the selection of directories (in which case all files inside the specified directory will be loaded), I will only select the *Version2* directory and click `Load & go back` button. At the end of the data loading, a popup informs how many files were loaded.

Note: If you know that not all files in the selected directory are alignments, you could still load that particular directory. All invalid alignment files will be ignored when the data is loaded.



✓ 614 file(s) successfully loaded

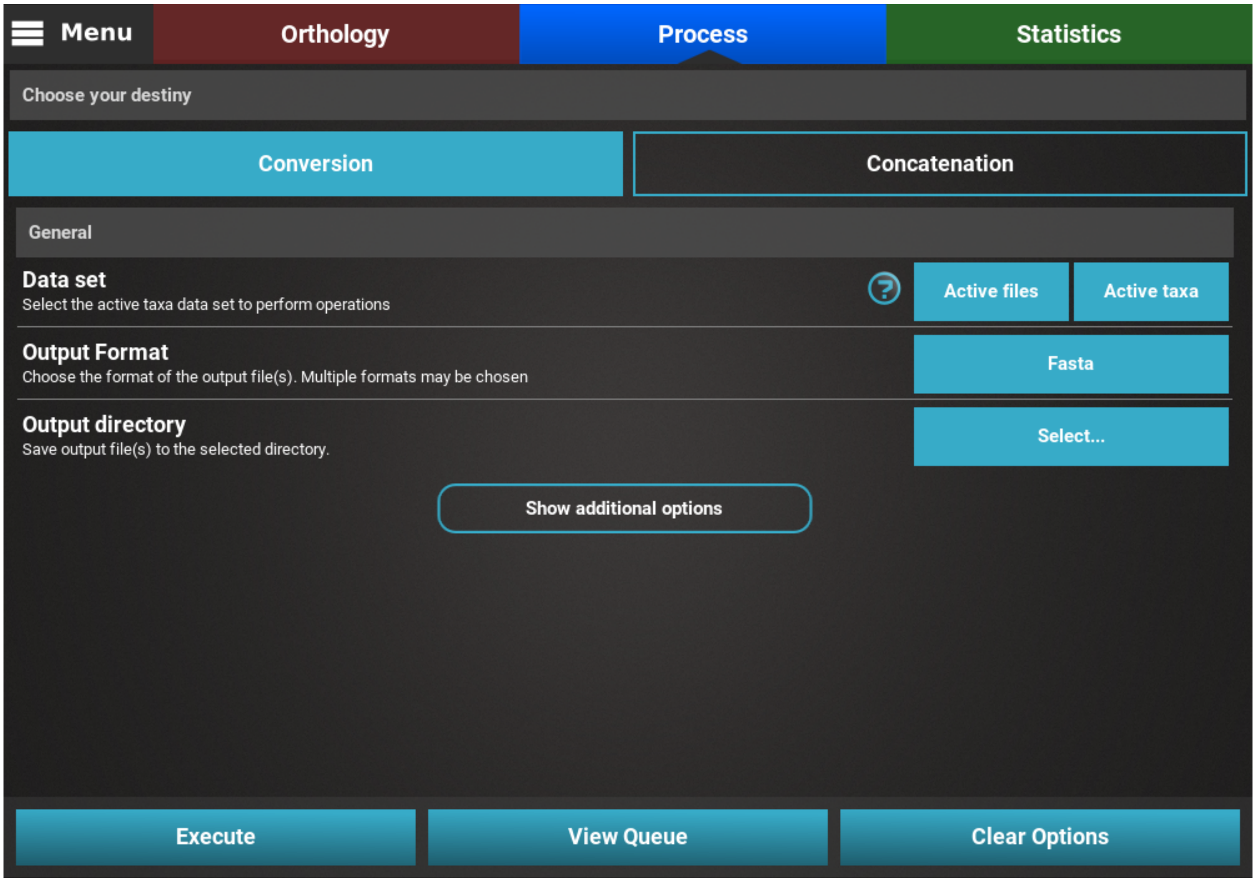
Welcome to TriFusion

Streamlining phylogenomic data gathering, processing and visualization

Start by clicking 'Menu > Open file(s)' or by typing 'Ctrl + O'

1.11.3 Conversion/concatenation

The **Conversion** and **Concatenation** options are found in the **Proces** screen. In this screen, select either `Conversion` or `Concatenation` to reveal the **General options**, which are mostly the same for both operations.



Data set
Select the active taxa data set to perform operations

Output Format
Choose the format of the output file(s). Multiple formats may be chosen

Output directory
Save output file(s) to the selected directory.

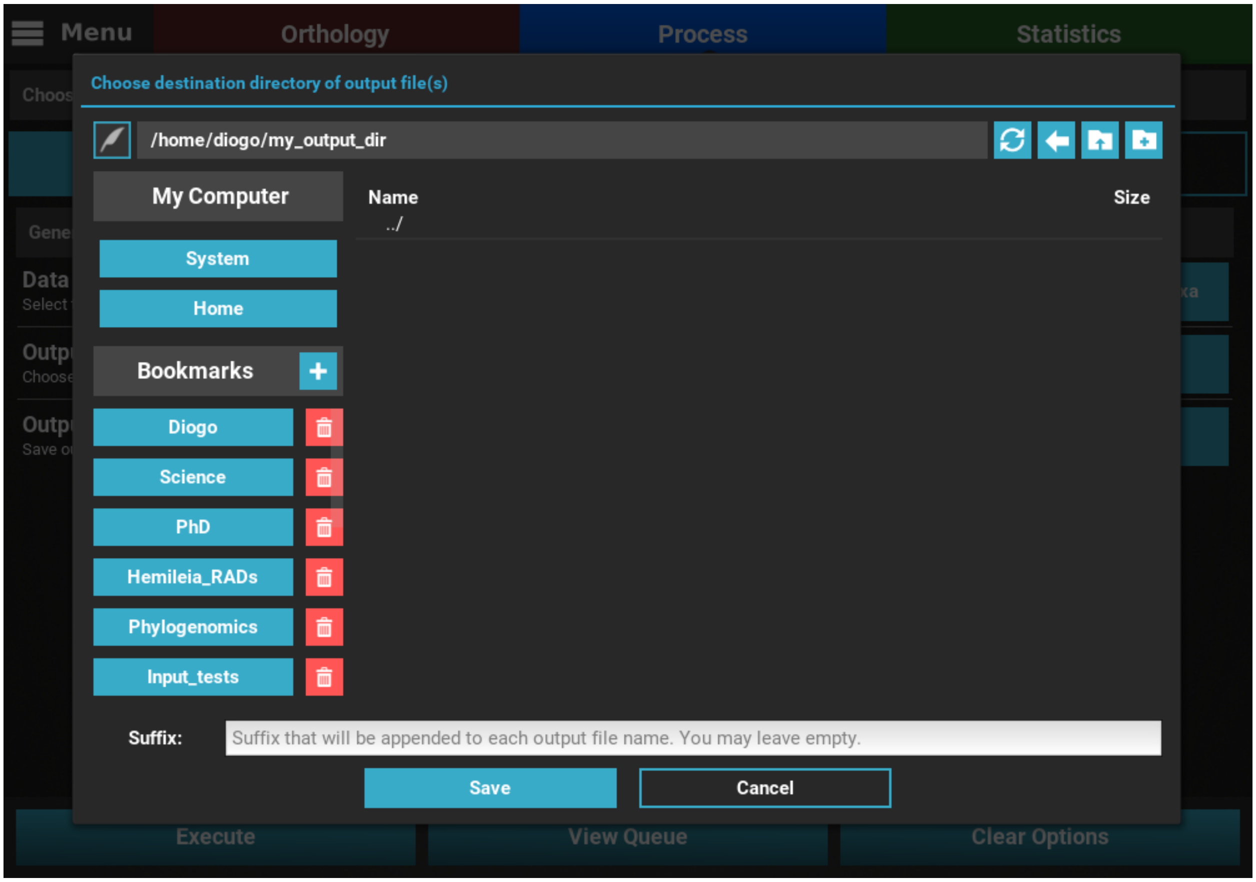
General options

The first option, **Data set**, specifies which active data set will be used for the conversion operation (see *Data set groups* tutorial or *Concatenation with custom active data sets* below). For now we'll leave it in the default value.

In the second option, **Output format**, you can choose one or more output formats to convert the input data. In this case we will choose 4 output formats (fasta, phylip, nexus and stockholm). Some output formats also contains specific additional options that can be viewed by clicking the corresponding *settings* button. Also note that some formats can only be used with the concatenation operation.

THe final general option is used to specify where you want to generate the output alignment(s).

In the case of **Conversion**, the **Output directory** option is used to select the directory where the output files will be generated. Here, the name of each output file will be based on the corresponding input file (for instance, the input *alignment.fas* will be converted into *alignment.nex* when the Nexus output format is specified). However, you can specify a suffix that will be appended to the end of every output file in the **Suffix** text box. For example, specifying *"_variant1"* as the suffix will create output files like *alignment1_variant1.nex*.

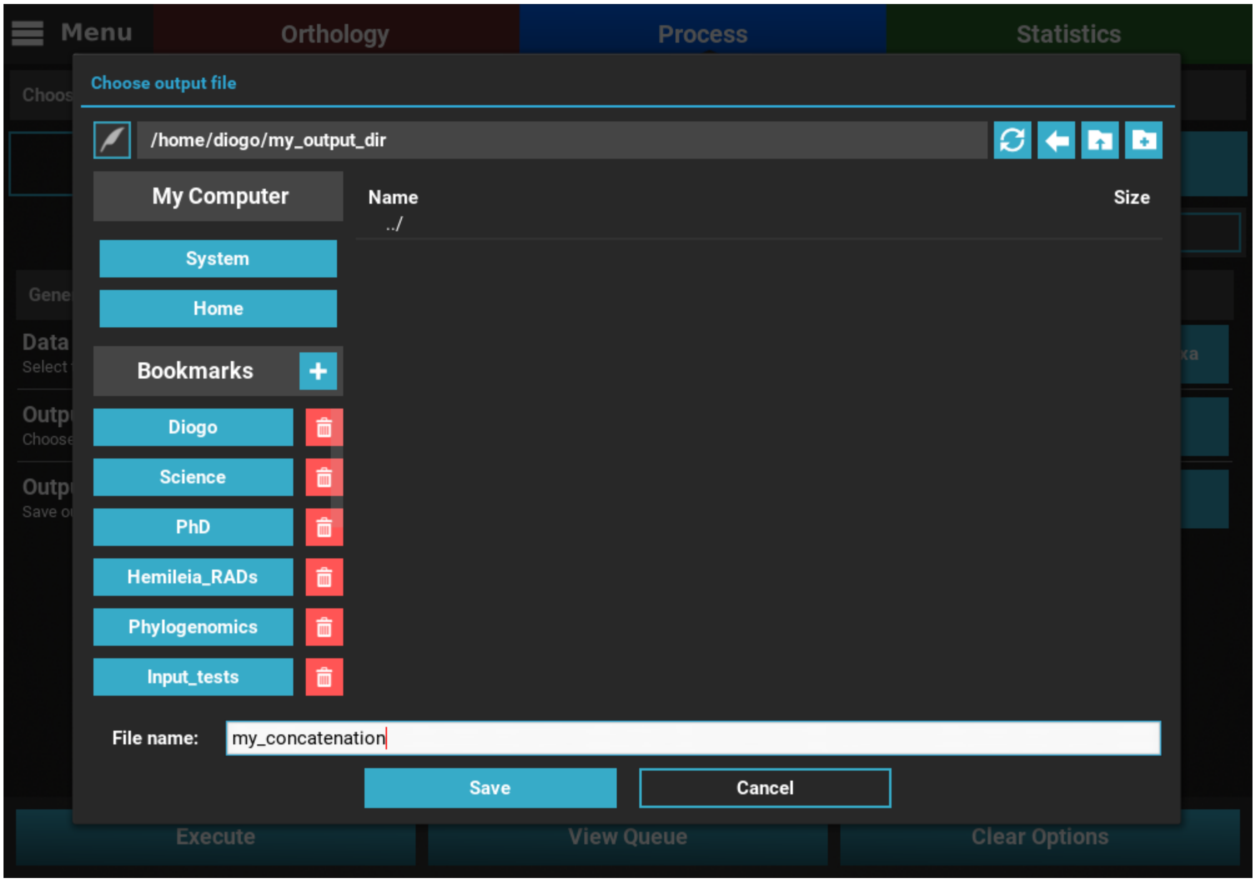


Choose destination directory of output file(s)

/home/diogo/my_output_dir

Suffix: Suffix that will be appended to each output file name. You may leave empty.

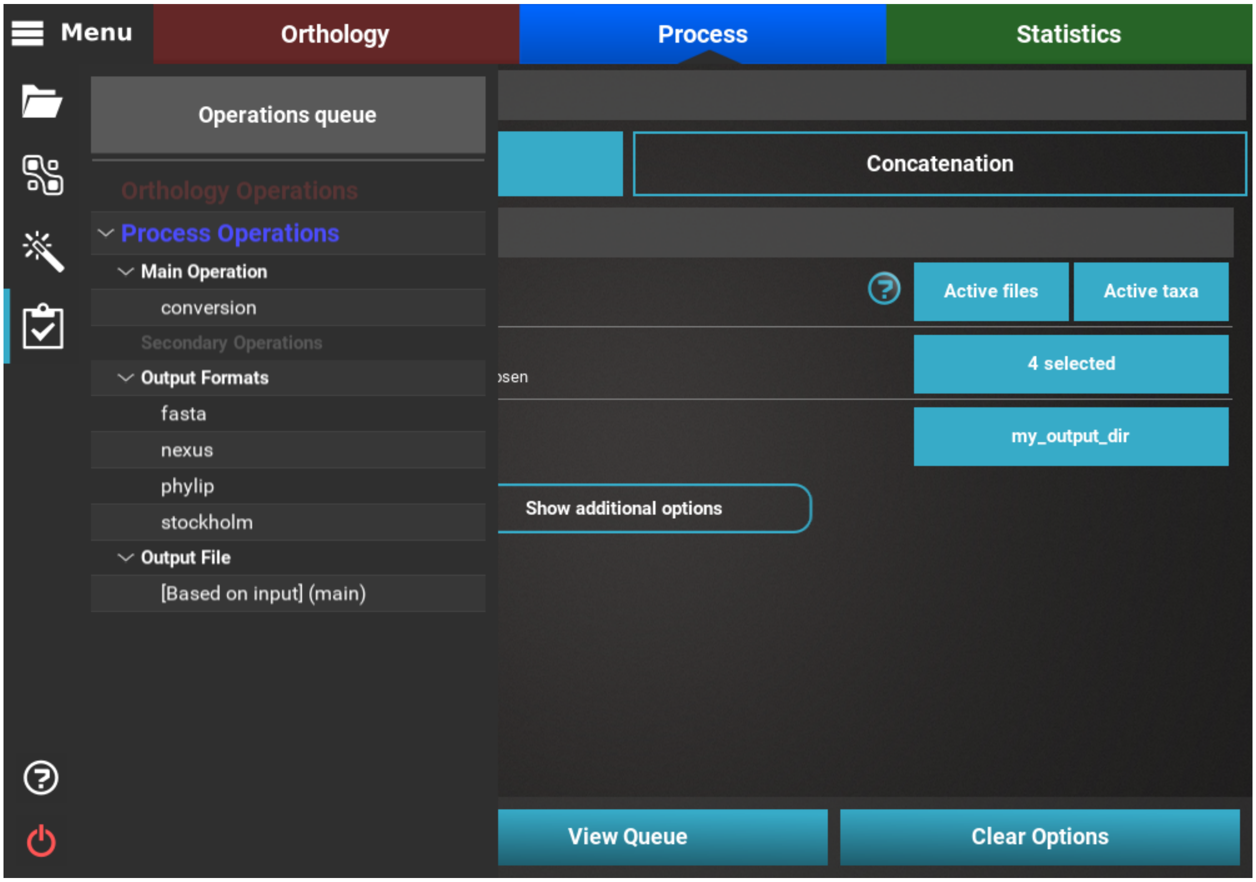
In the case of **Concatenation**, the **Output file** option is used to specify the directory **AND** name of the output file. For example, we could name our concatenated output file "my_concatenation". The extension is automatically added.



/home/diogo/my_output_dir

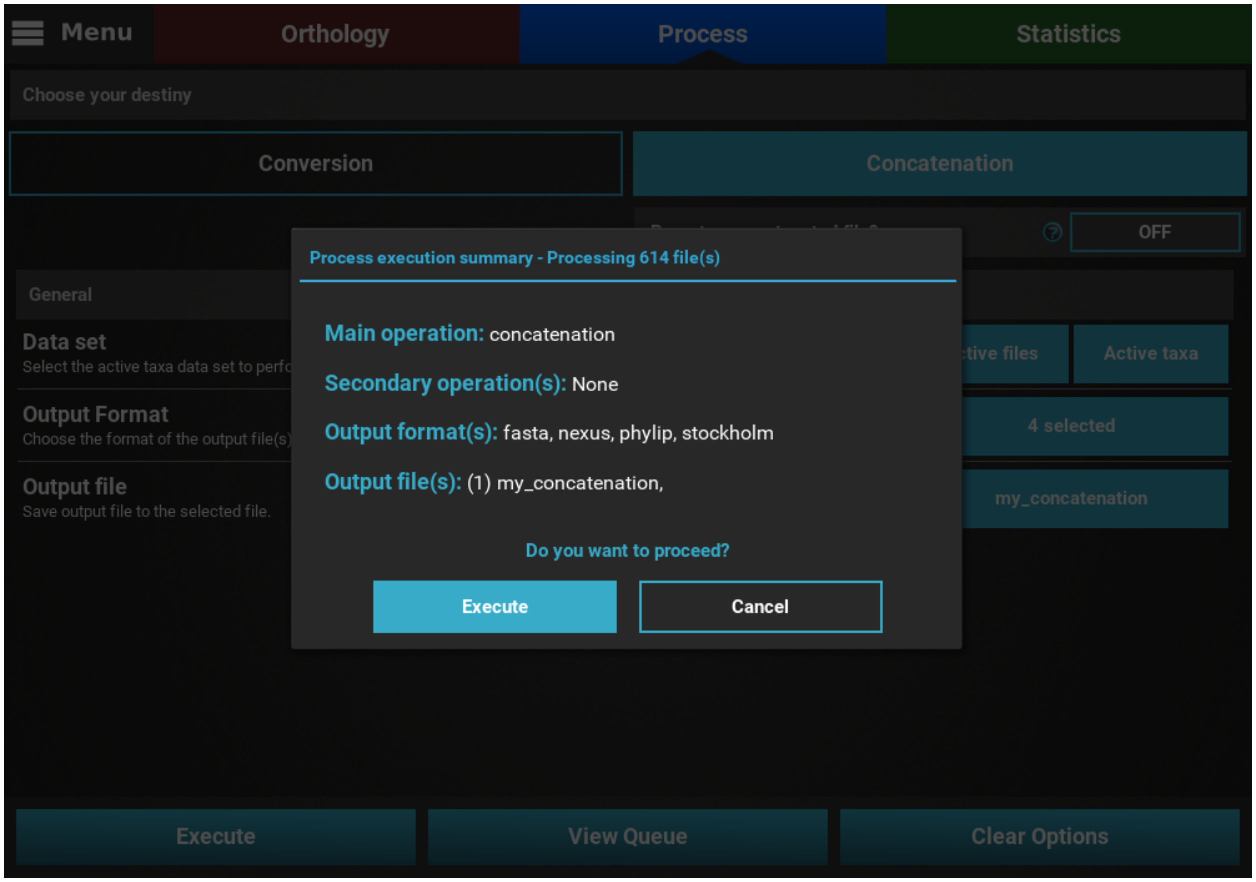
File name: my_concatenation

After setting up these general options, you can click the `View Queue` button at the bottom of the screen to get an overview of the selected options. There you'll see that the 614 files are set to be converted into 4 output formats in a number of output formats whose name will be based on the input.



Execution

The execution of either **Conversion** or **Concatenation** operations is started by clicking the `Execute` button at the bottom of the screen. This will open an execution summary with information on the selected main operation, the selected secondary operations (if any), the selected output formats and the expected number of output files. In the case of **Concatenation** the actual output file name should appear.



Process execution summary - Processing 614 file(s)

Main operation: concatenation

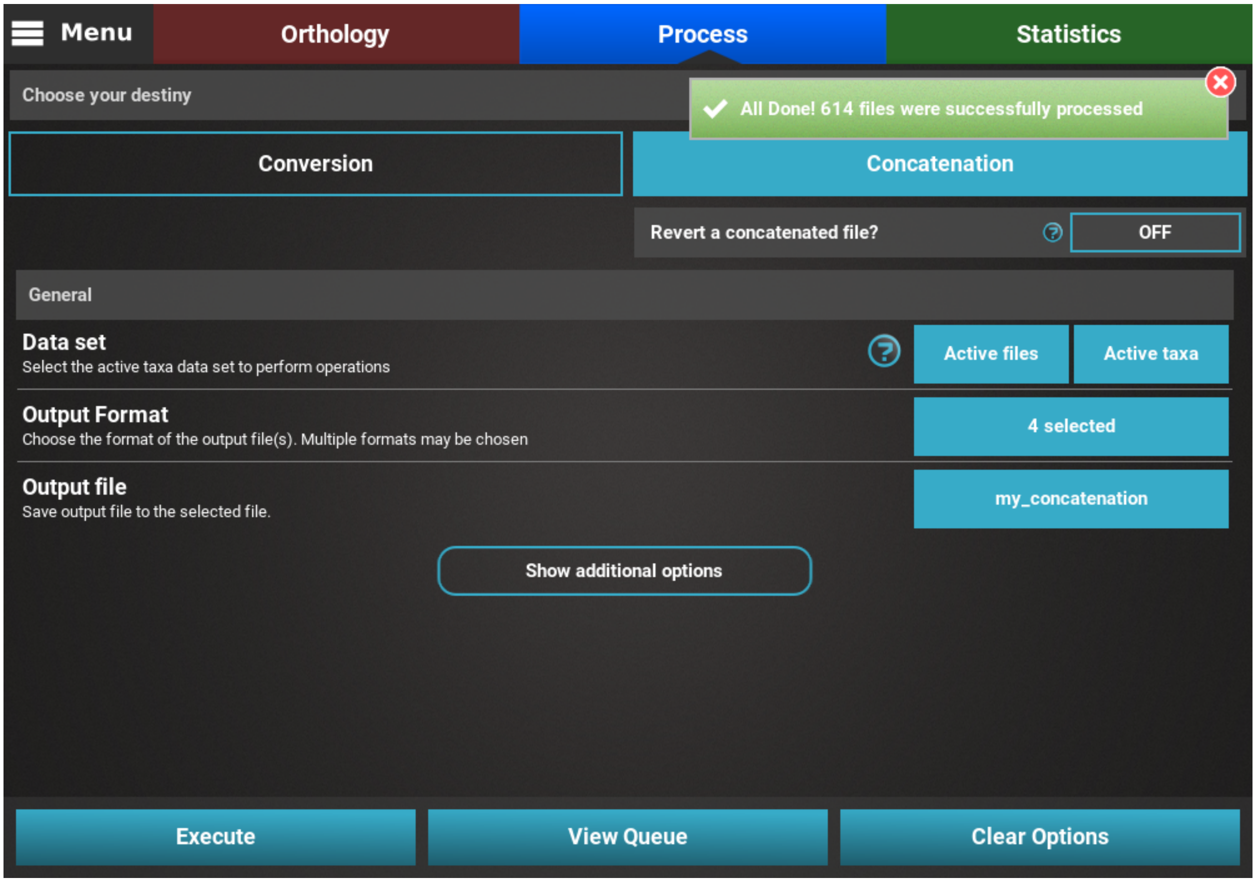
Secondary operation(s): None

Output format(s): fasta, nexus, phylip, stockholm

Output file(s): (1) my_concatenation,

Do you want to proceed?

If you're happy with these settings, click the `Execute` button, and the **Conversion/Concatenation** operation will be carried out. At the end of the execution, an informative popup should appear with a notification that all files were successfully processed.


Menu
Orthology
Process
Statistics
Choose your destiny
All Done! 614 files were successfully processed
Conversion
Concatenation
Revert a concatenated file?
OFF
General
Data set
Select the active taxa data set to perform operations
Active files
Active taxa
Output Format
Choose the format of the output file(s). Multiple formats may be chosen
4 selected
Output file
Save output file to the selected file.
my_concatenation
Show additional options
Execute
View Queue
Clear Options

1.11.4 Concatenation with custom active data sets

Note: Operations on custom data sets can also be applied with the **Conversion** operation. In this case, however, it just means that the alignments and taxa that are not converted.

In many cases, additional operations may be desired on specific subsets of the total loaded dataset. Here we'll see one way of performing an additional concatenation operation on a custom made data set. More information is available in the *Data set groups* tutorial.

Creating and changing the active data set

Suppose we were interested in concatenating the same 614 files, but only for taxa whose names start with the letter "A". And after that for taxa whose names start with the letter "C". Since I need to create two taxa groups (say, *A_taxa* and *C_taxa*), we will also explore two methods of creating these data sets.

Using the side panel toggling method

To create an active data set that contains, for example, only taxa whose names start with an "A", go to `Menu > Open/View Data` and selected the *Taxa* tab. There are three taxa whose name starts with an "A". The quickest way to selected only these taxa would be to click the `Deselect All` button and then toggling ON the desired three taxa.

Using the data set creation dialog

To create the *C_taxa* via the data set creation dialog, go to `Menu > Dataset Groups`, click the *Taxa* tab, and then the `Set new taxa group` button. Since we're dealing with a small number of taxa, we will set the taxa group manually in TriFusion. In the taxa group creation dialog, select the taxa with names starting with a "C" (here using `Shift + clicking` to selected the seven taxa is convenient), specify the group name and click `OK`.

Execution with custom active data sets

We'll start with the execution of the **Concatenation** of the 614 files for the *A_taxa* taxa group. We need to make sure that the value of the **Data set** general option is set to *Active taxa*, so that TriFusion will use the three active taxa previously defined. Then, click `Execute` and complete the concatenation operation as before.

Now for the *C_taxa* group, select the name of this group in the drop down menu of the **Data set** general option.

Once the C_taxa group is selected, click the `Execute` button and complete the concatenation as before.

1.11.5 Concatenation with custom partitions

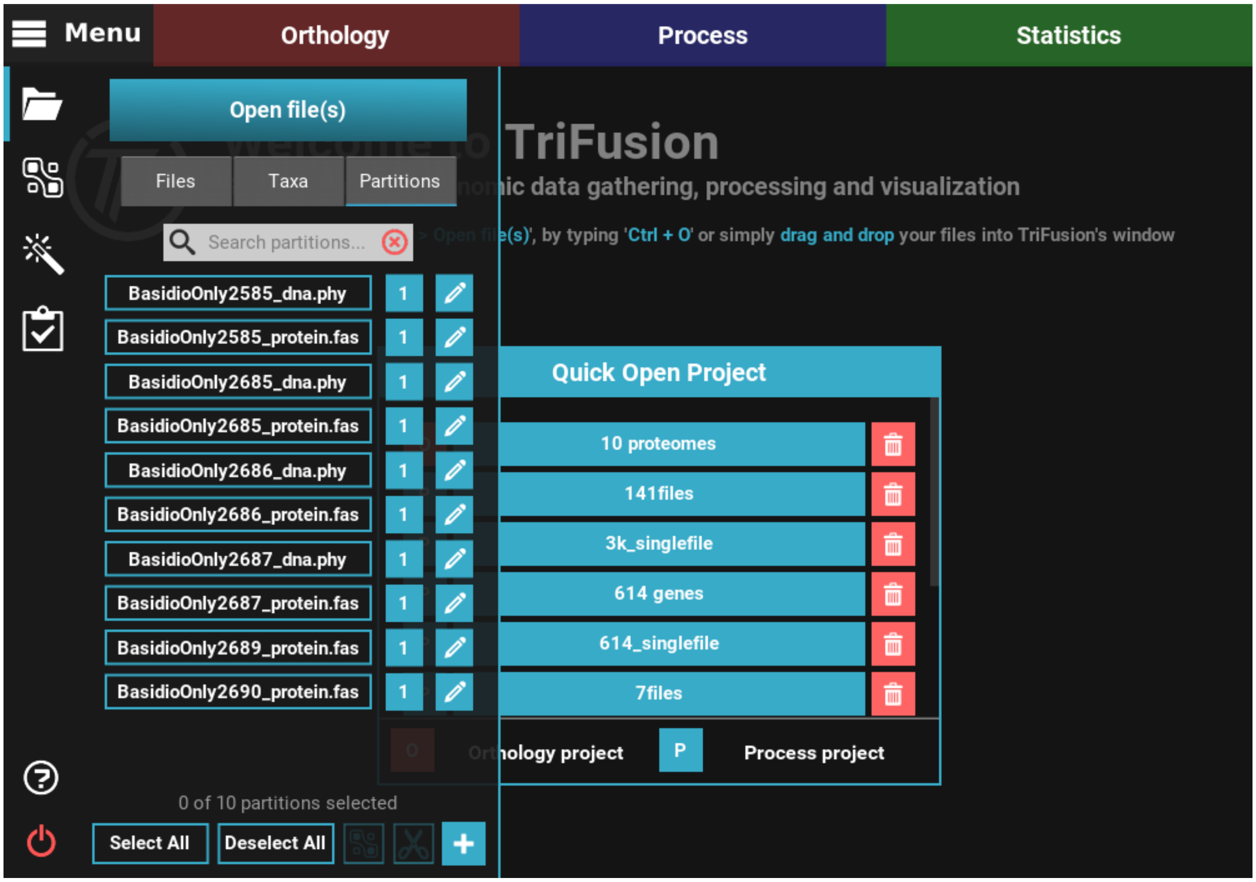
One of the convenient features of TriFusion is that it allows you to easily edit or import from a text file the partitions of your current data set. You don't really have to worry about the range, order, size of the partitions, as long as you don't mix partitions of different sequences types (e.g. protein and nucleotide). You can also specify some substitution models for your partitions for output formats that support that kind of information (Nexus and Phylip). You can check the more detailed *Partitions and substitution models* tutorial.

Load data

Here we'll see how the concatenation operation can seamlessly deal with any partition scheme you provide, with or without information on the substitution model. For this part of the tutorial we'll use a smaller data set of 10 alignments so that it is easier to follow the changes. Nevertheless, TriFusion is able to deal with thousands of partitions as easily.

This is a mixed data set containing Fasta and Phylip alignments of protein and nucleotide sequences. Let's import the data using the drag and drop method.

If you navigate to `Menu -> Open/View Data` and click on the *Partitions* tab you can see that TriFusion attributes a partition to each individual input file by default (unless partition schemes are provided when loading Nexus files).


Menu
Orthology
Process
Statistics
Open file(s)
Welcome to TriFusion
Genomic data gathering, processing and visualization
Files
Taxa
Partitions
Search partitions...
> Open file(s)', by typing 'Ctrl + O' or simply drag and drop your files into TriFusion's window
BasidioOnly2585_dna.phy
BasidioOnly2585_protein.fas
BasidioOnly2685_dna.phy
BasidioOnly2685_protein.fas
BasidioOnly2686_dna.phy
BasidioOnly2686_protein.fas
BasidioOnly2687_dna.phy
BasidioOnly2687_protein.fas
BasidioOnly2689_protein.fas
BasidioOnly2690_protein.fas
1
Quick Open Project
10 proteomes
141files
3k_singlefile
614 genes
614_singlefile
7files
O
Orthology project
P
Process project
0 of 10 partitions selected
Select All
Deselect All

Basic concatenation

Loading a mixed data set (nucleotide and protein sequences) raises the immediate issue that, in formats such as Nexus, the ranges of the nucleotide and protein sequences has to be defined in the header, in addition to the partitions definition. TriFusion does this for you and simplifies the issue by grouping nucleotide and protein files/partitions together, regardless of their input order.

First, let's perform a **Concatenation** operation without further modification of the default partitions. Specify the *Nexus* as the output format, provide an output file and click `Execute`.

If you inspect the output Nexus file, you can see that the header now has the information on the mixed data set:

```
#NEXUS
Begin data;
    dimensions ntax=49 nchar=6134 ;
    format datatype=mixed(dna:1-2790,protein:2791-6134) interleave=no gap=-;
```

With the concatenated alignment having the first 2790 characters as nucleotides and the remaining as amino acid residues. At the end of the file, the partitions are also correctly defined and ready for downstream software like MrBayes:

```
begin mrbayes;
    charset BasidioOnly2585dnaphy = 1-1458;
    charset BasidioOnly2685dnaphy = 1459-1722;
    charset BasidioOnly2686dnaphy = 1723-2259;
    charset BasidioOnly2687dnaphy = 2260-2790;
```

(continues on next page)

(continued from previous page)

```
    charset BasidioOnly2585proteinfas = 2791-3837;
    charset BasidioOnly2685proteinfas = 3838-3959;
    charset BasidioOnly2686proteinfas = 3960-4153;
    charset BasidioOnly2687proteinfas = 4154-4373;
    charset BasidioOnly2689proteinfas = 4374-5178;
    charset BasidioOnly2690proteinfas = 5179-6134;
    partition part = 10: BasidioOnly2585dnaphy, BasidioOnly2685dnaphy,␣
↪BasidioOnly2686dnaphy, BasidioOnly2687dnaphy, BasidioOnly2585proteinfas,␣
↪BasidioOnly2685proteinfas, BasidioOnly2686proteinfas, BasidioOnly2687proteinfas,␣
↪BasidioOnly2689proteinfas, BasidioOnly2690proteinfas;
    set partition=part;
end;
```

Merge partitions

Partitions can be merged in any number and order, provided that they share the same sequence type (nucleotide partitions can only be merged with nucleotide). We can, for instance, merge all protein partitions together and the first and last nucleotide partitions. To accomplish this, select all partitions you wish to merge and click the *merge partitions* button at the bottom of the panel.

When we repeat the **Concatenation** operation, we can see that the Nexus header remains the same, but the partitions have been updated. **Notice that even though we merged non-contiguous partitions, they appear with the same range**. This is because TriFusion will first sort the partition sequences so that they become contiguous and only then it will write the output file:

```
begin mrbayes;
    charset nuc1 = 1-1989;
    charset BasidioOnly2685dnaphy = 1990-2253;
    charset BasidioOnly2686dnaphy = 2254-2790;
    charset proteinparts = 2791-6134;
    partition part = 4: nuc1, BasidioOnly2685dnaphy, BasidioOnly2686dnaphy,␣
↪proteinparts;
    set partition=part;
end;
```

1.12 Reverse concatenation

Note: Data availability for this tutorial: a small concatenated alignment with the corresponding partition files is available here.

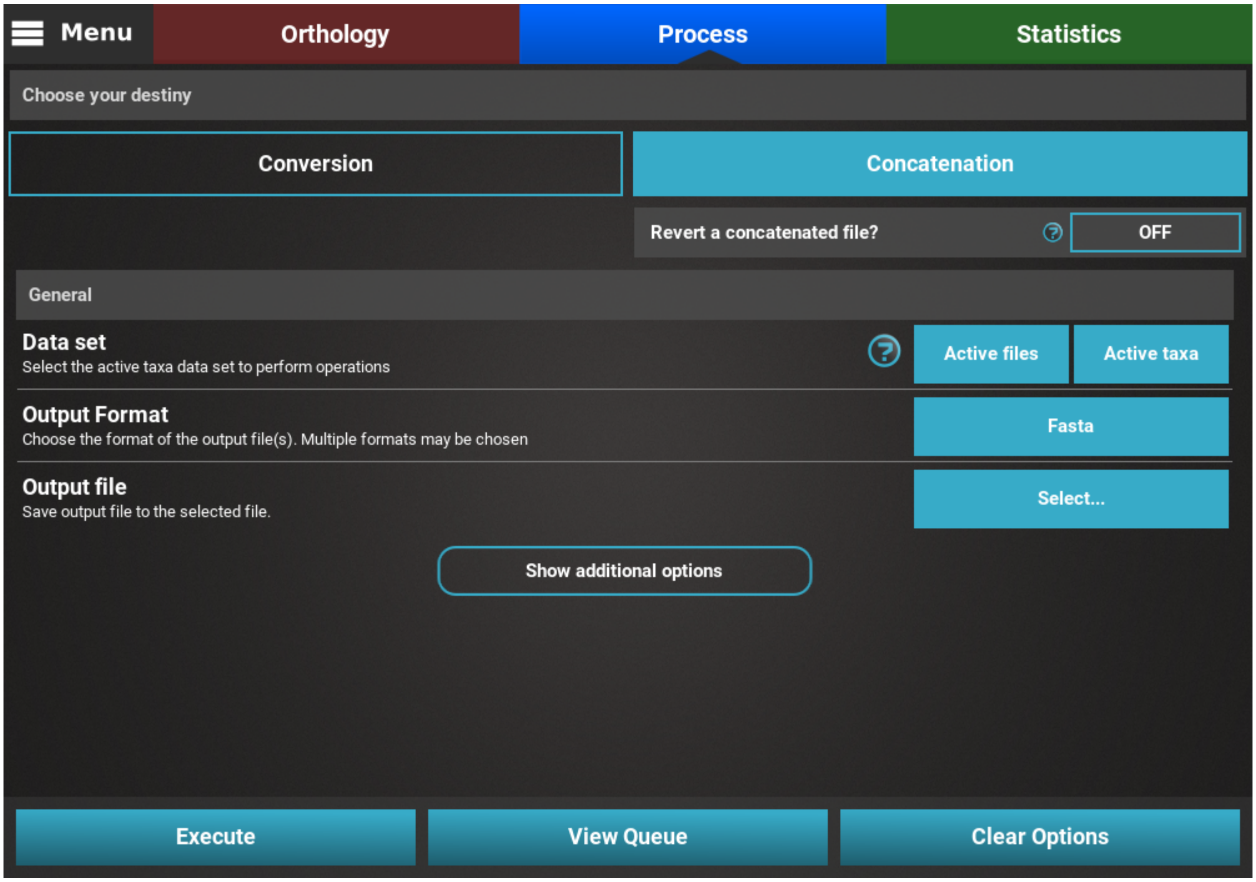
Here we'll reverse a concatenated file into its original alignment files. TriFusion offers two main ways of doing this, but both require an **input alignment file with partitions defined**. As you will see, reverse concatenation is essentially the **split of a single alignment file into multiple output alignments** based on a given partitions file. These partitions can be anything you want, provided that they have the same sequence type (nucleotide or protein).

Note: At the end of this tutorial we'll also see how secondary operations work when reversing a concatenated file.

1.12.1 Manual selection from a partition file

Note: With this method, more input files may be loaded in TriFusion besides the file that you wish to reverse the concatenation.

To open the reverse concatenation settings, click in the `Concatenation` button in the **Process** screen, which will reveal the option to *Revert a concatenated file*.

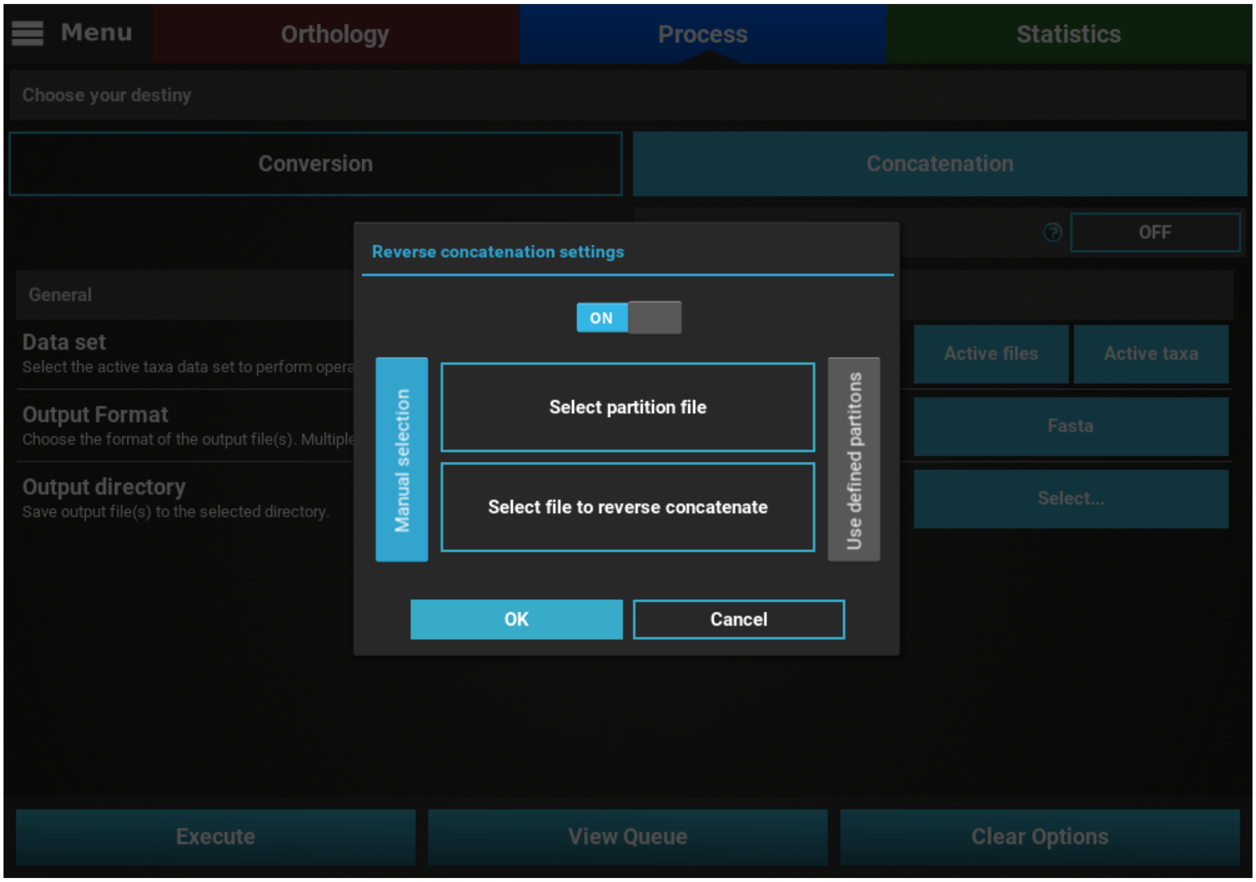


Data set
Select the active taxa data set to perform operations

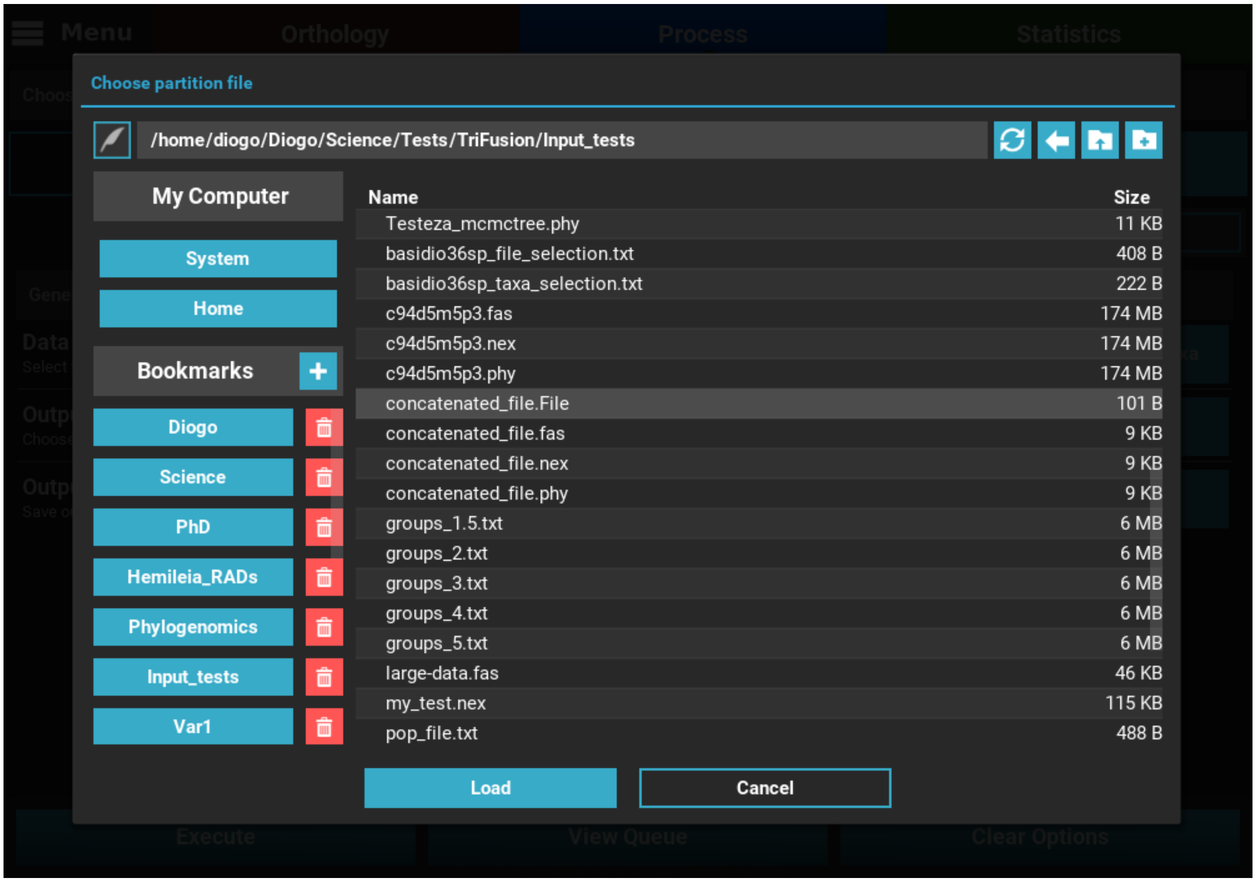
Output Format
Choose the format of the output file(s). Multiple formats may be chosen

Output file
Save output file to the selected file.

In the **reverse concatenation settings** dialog, turn the switch `ON` to active the **reverse concatenation** operation. The *Manual selection* method is already expanded by default and asks the user for the partition file and the input file that will be reversed.



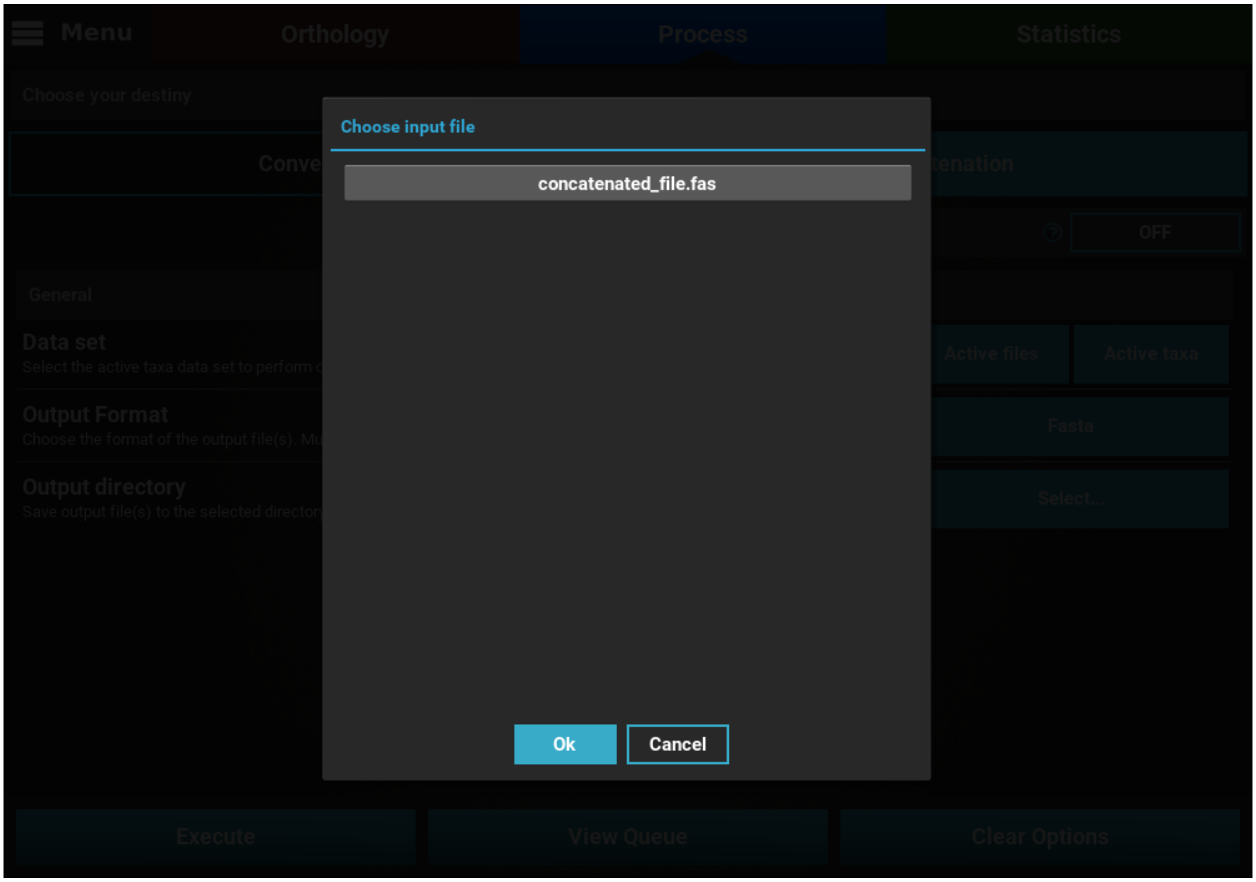
First, click in the `Select partition file` button, and navigate the file browser until you find the partition file that corresponds to the input alignment. In our case, it is the file *concatenated_file.File*.



/home/diogo/Diogo/Science/Tests/TriFusion/Input_tests

Name	Size
Testeza_mcmctree.phy	11 KB
basidio36sp_file_selection.txt	408 B
basidio36sp_taxa_selection.txt	222 B
c94d5m5p3.fas	174 MB
c94d5m5p3.nex	174 MB
c94d5m5p3.phy	174 MB
concatenated_file.File	101 B
concatenated_file.fas	9 KB
concatenated_file.nex	9 KB
concatenated_file.phy	9 KB
groups_1.5.txt	6 MB
groups_2.txt	6 MB
groups_3.txt	6 MB
groups_4.txt	6 MB
groups_5.txt	6 MB
large-data.fas	46 KB
my_test.nex	115 KB
pop_file.txt	488 B

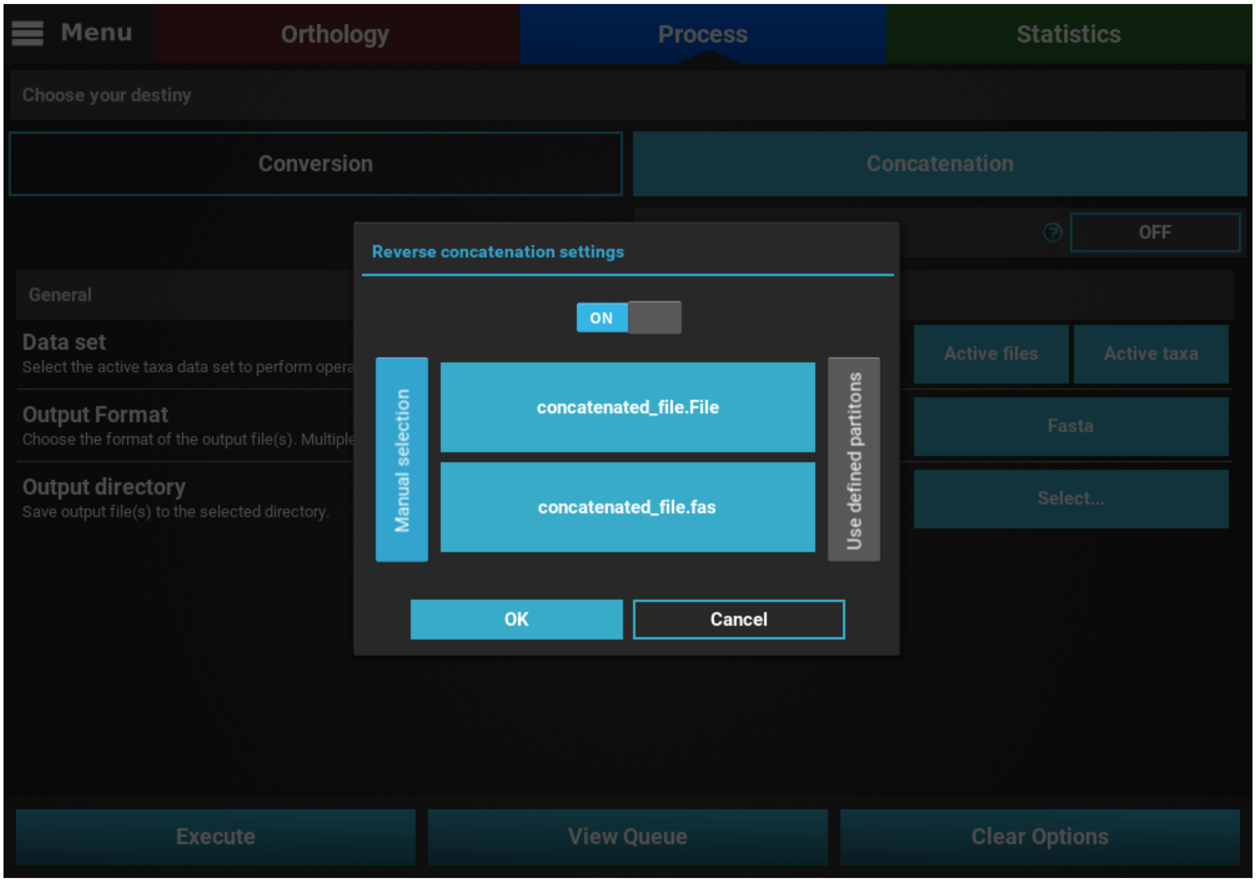
Then, click the `Select file to reverse concatenate` button to choose the concatenated file that will be reversed. This will open a popup listing all input files currently loaded into TriFusion. In our case, the list contains only the single concatenated file. Click on the alignment button to select it.



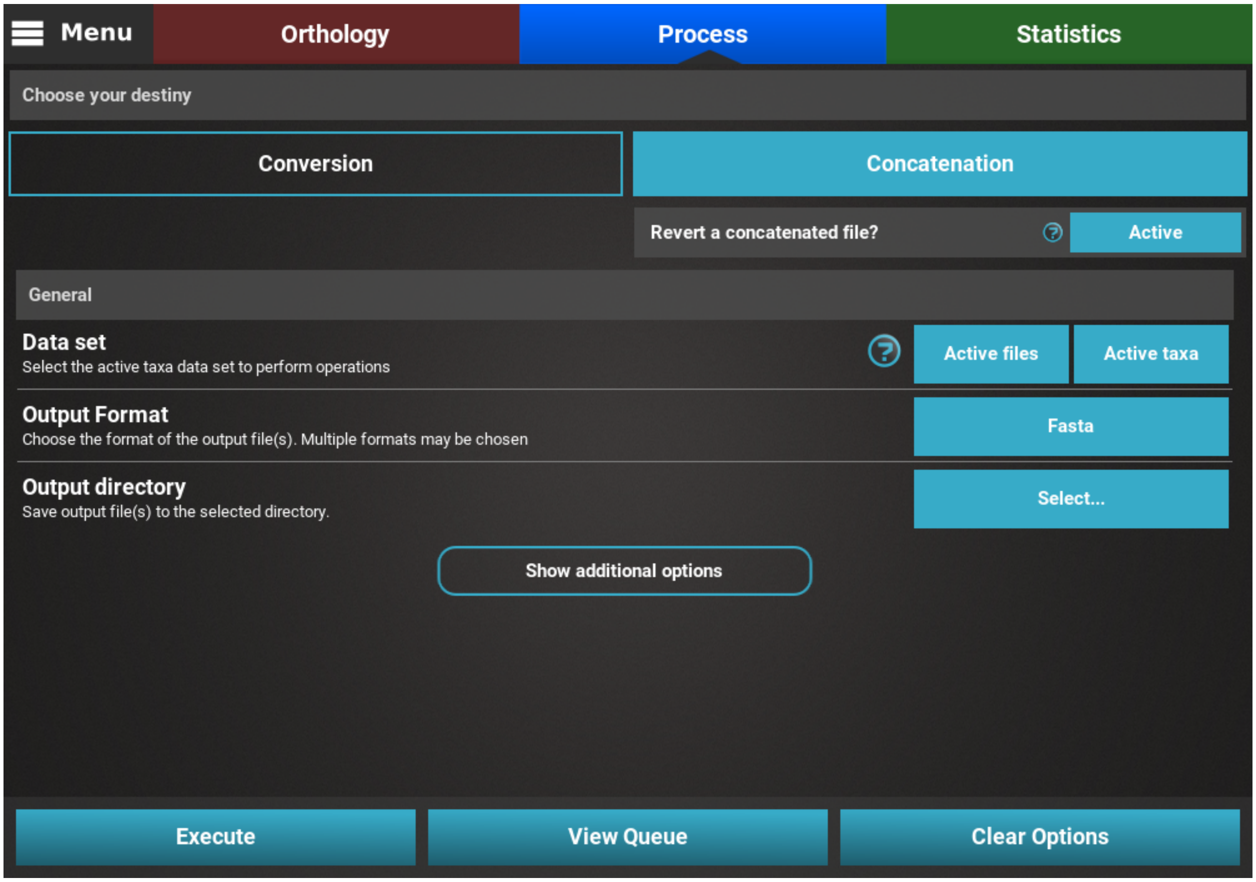
Choose input file

concatenated_file.fas

After setting these two requirements, the reverse concatenation settings dialog should look something like this.



When you click *'OK'* TriFusion will check if the partition file is compliant with the concatenated file. If it detects issues, such as missing partitions or the defined partitions being out of range from the alignment file, an informative error will popup. However, if all checks out the *Revert a concatenated file?* button will now say **Active**.

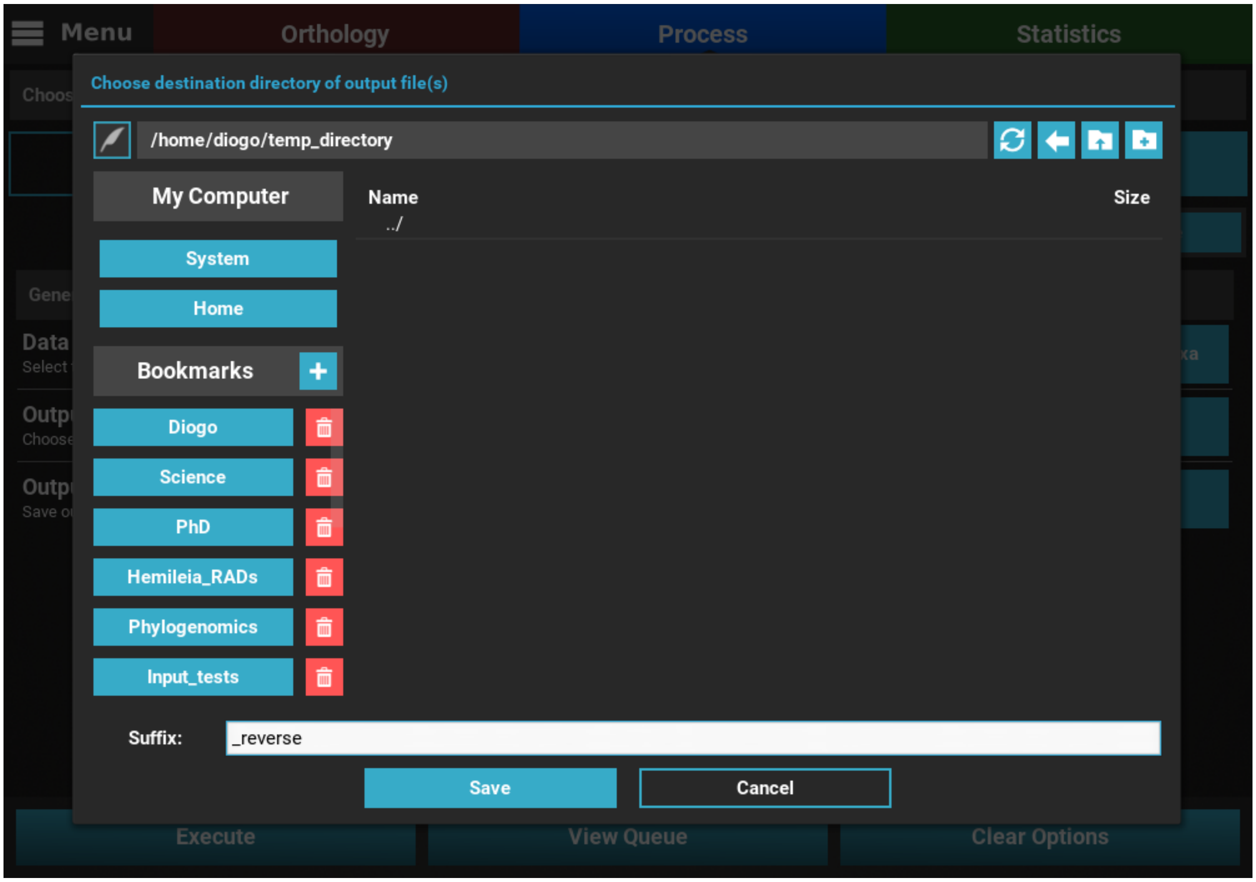


Data set
Select the active taxa data set to perform operations

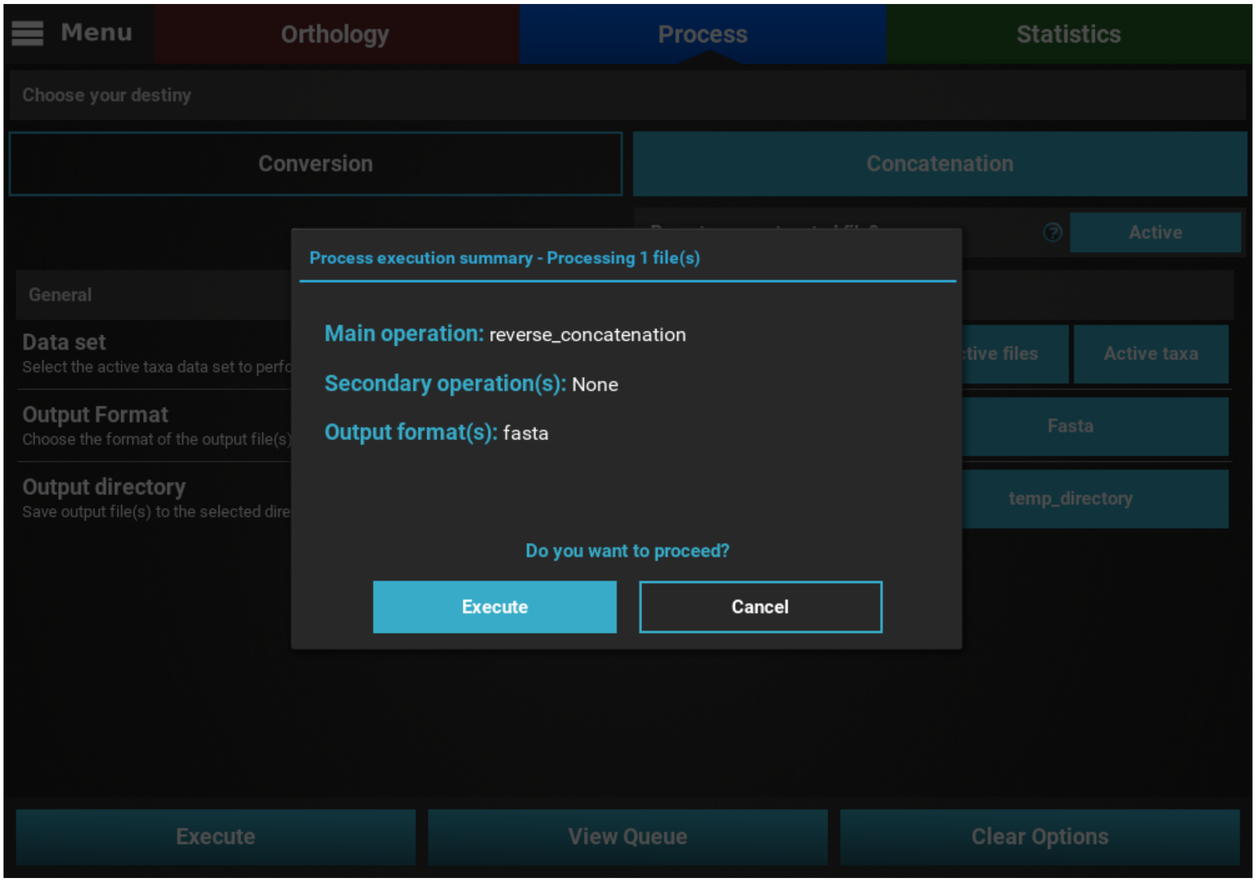
Output Format
Choose the format of the output file(s). Multiple formats may be chosen

Output directory
Save output file(s) to the selected directory.

Now select an output directory where the individual alignments will be generated by clicking the *'Select'* button of the *Output directory* option. Note that the output files will be named according to the names of the defined partitions. Optionally, you may specify a suffix that will be appended to the end of every output file, but before the output format extension. Here we will specify the *"reverse"* suffix.



Finally, click the `Execute` button to display the execution summary dialog, which will inform that a reverse concatenation operation will be performed, with no additional secondary operations and the output files will be in Fasta format. To begin the reverse concatenation, click the `Execute` button.



Process execution summary - Processing 1 file(s)

Main operation: reverse_concatenation

Secondary operation(s): None

Output format(s): fasta

Do you want to proceed?

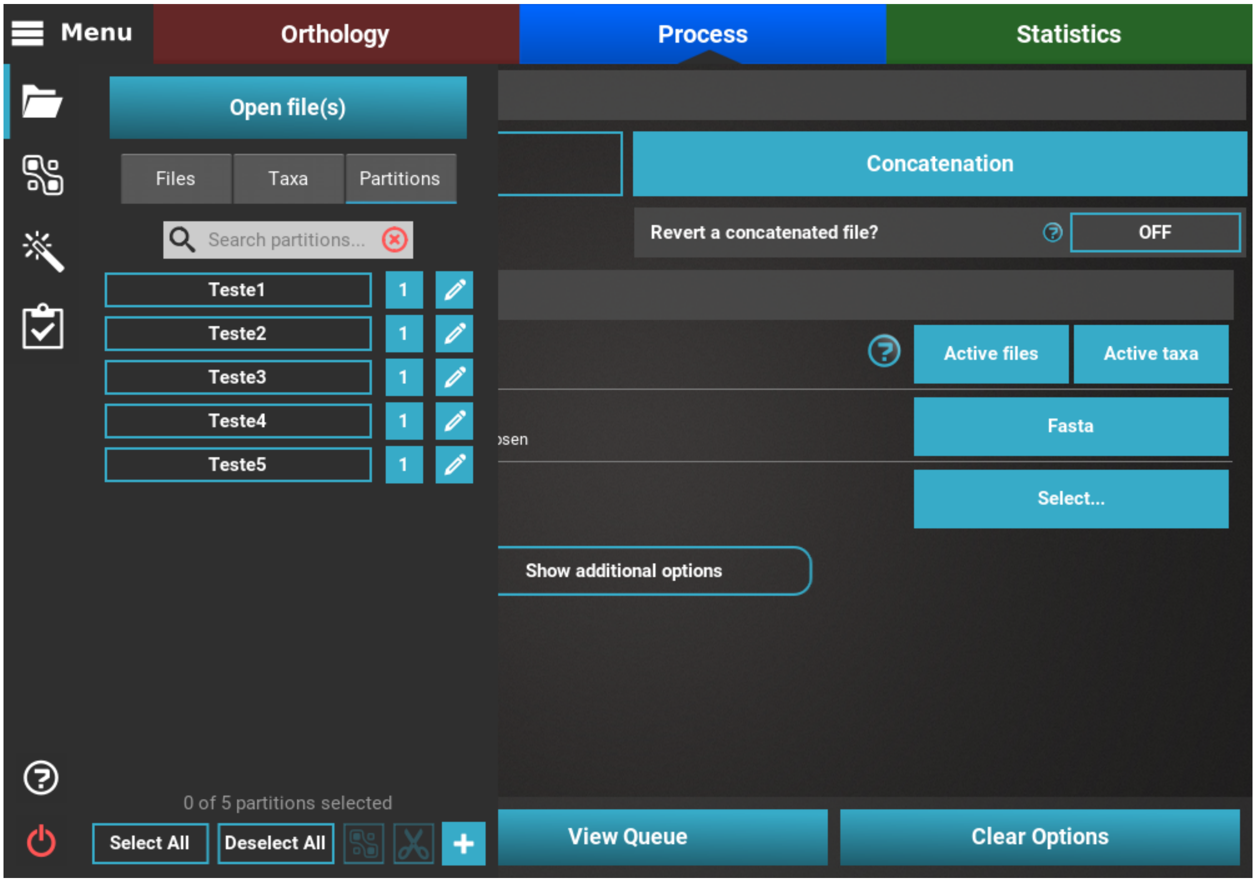
1.12.2 Reverse using partitions defined in TriFusion

This method uses the partitions defined within TriFusion to reverse concatenate a single input alignment file. Therefore, if you use this method, make sure only one alignment is loaded. There are several ways to import or create partitions in TriFusion (check the User Guide, Section 4.3.2 Partitions tab). For instance, partitions may already be defined in a Nexus input file, in which case TriFusion will automatically detect them and set them in the Partitions tab of the side panel.

In our case the input alignment is in Fasta format, so we'll have to set the partitions first in a different way. Navigate to `Menu > Open/View Data` and click in the *Partitions* tab. You can see that one partition is already present, since TriFusion automatically attributes a partition for every input alignment file.

Since we already have a partition file for this concatenated alignment, we do not need to create all partitions by hand. To import a partition scheme from a file, click the + button at the bottom of the panel. In the file browser, navigate to the directory containing the partition file and load it. In our case, the partition file is named *concatenated_file.File*.

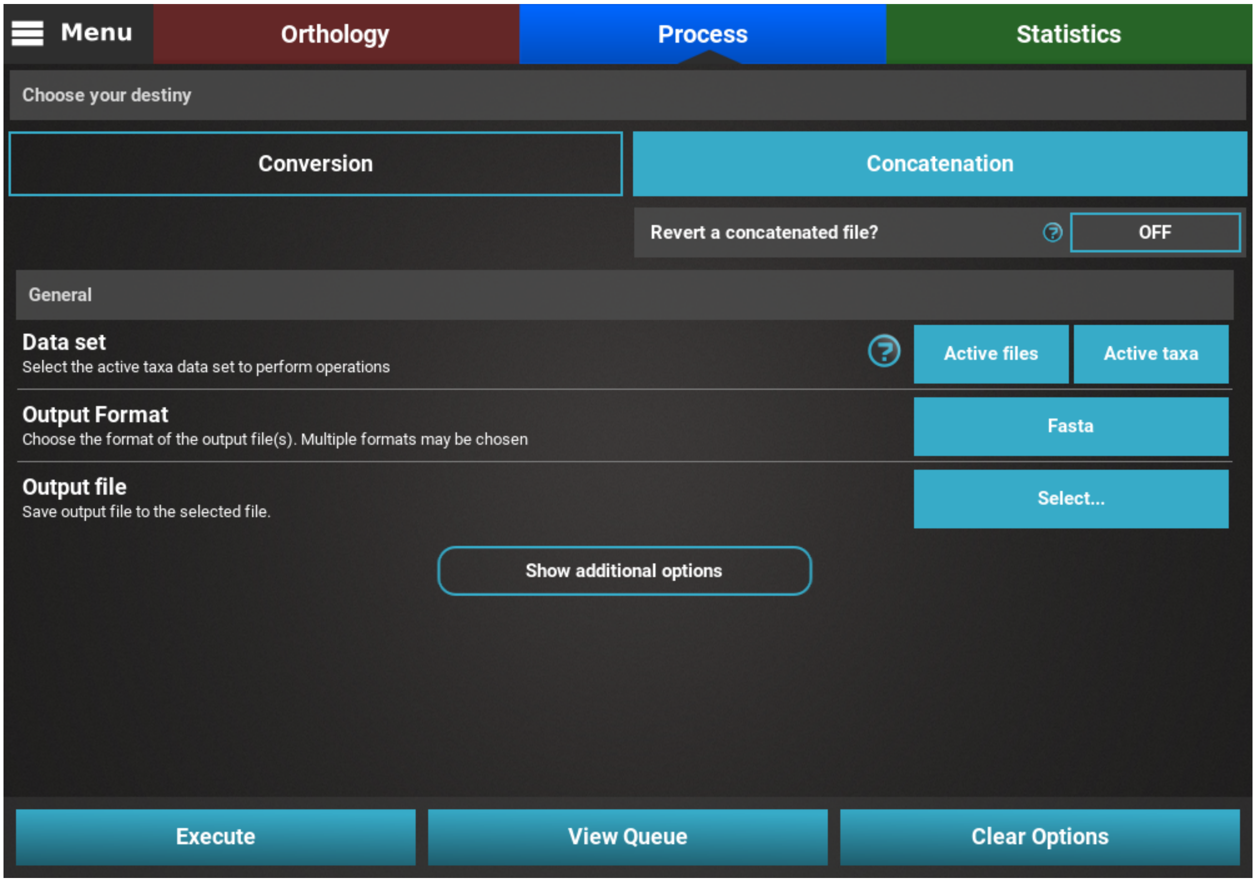
After loading the appropriate partition file, the list in the *Partitions* tab will update with the new partitions.



Revert a concatenated file? OFF

0 of 5 partitions selected

At this point you can still edit the partitions any way you want (change ranges, merge partitions, change names, etc.). When you are ready to select the reverse concatenation settings, click in the `Concatenation` button in the Process screen to reveal the option to *Revert a concatenated file*.



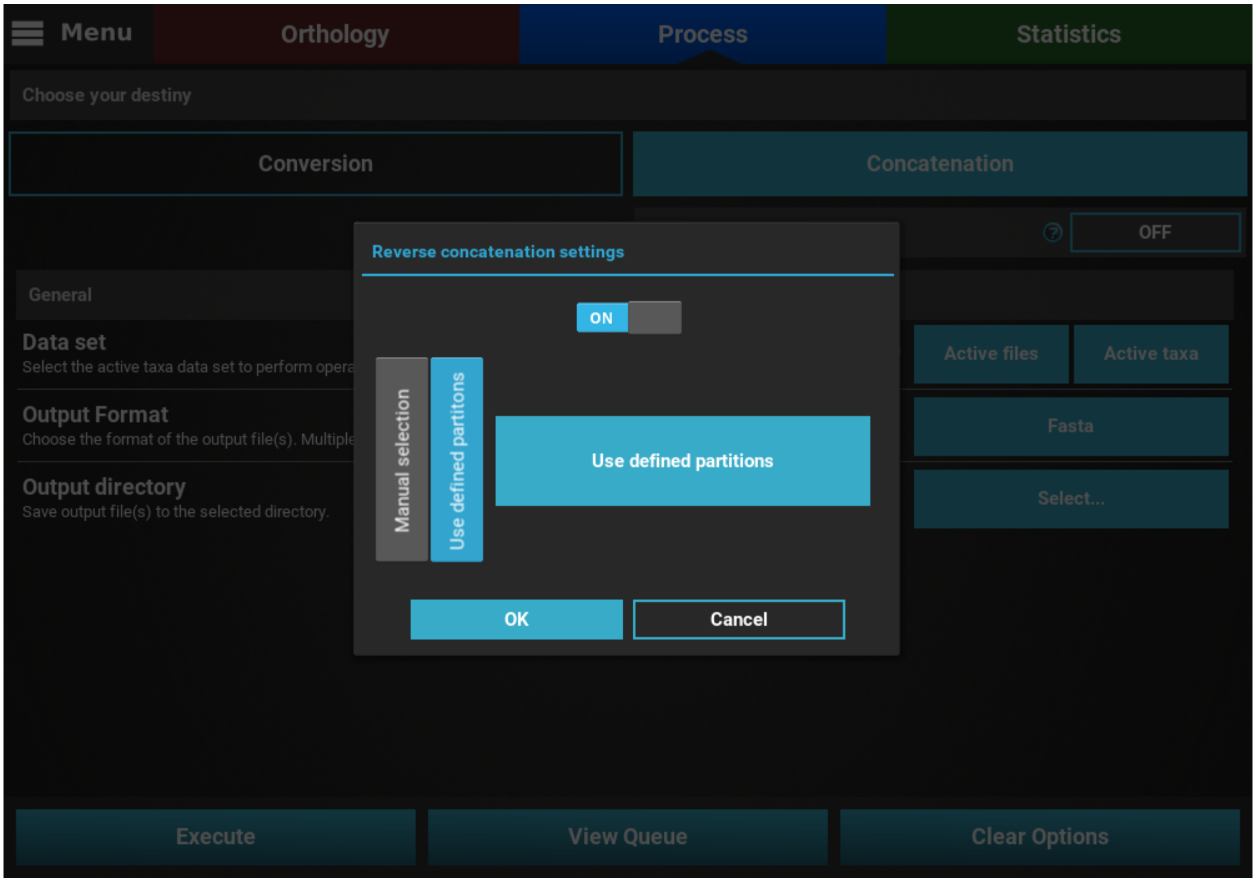
Revert a concatenated file? OFF

Data set
Select the active taxa data set to perform operations

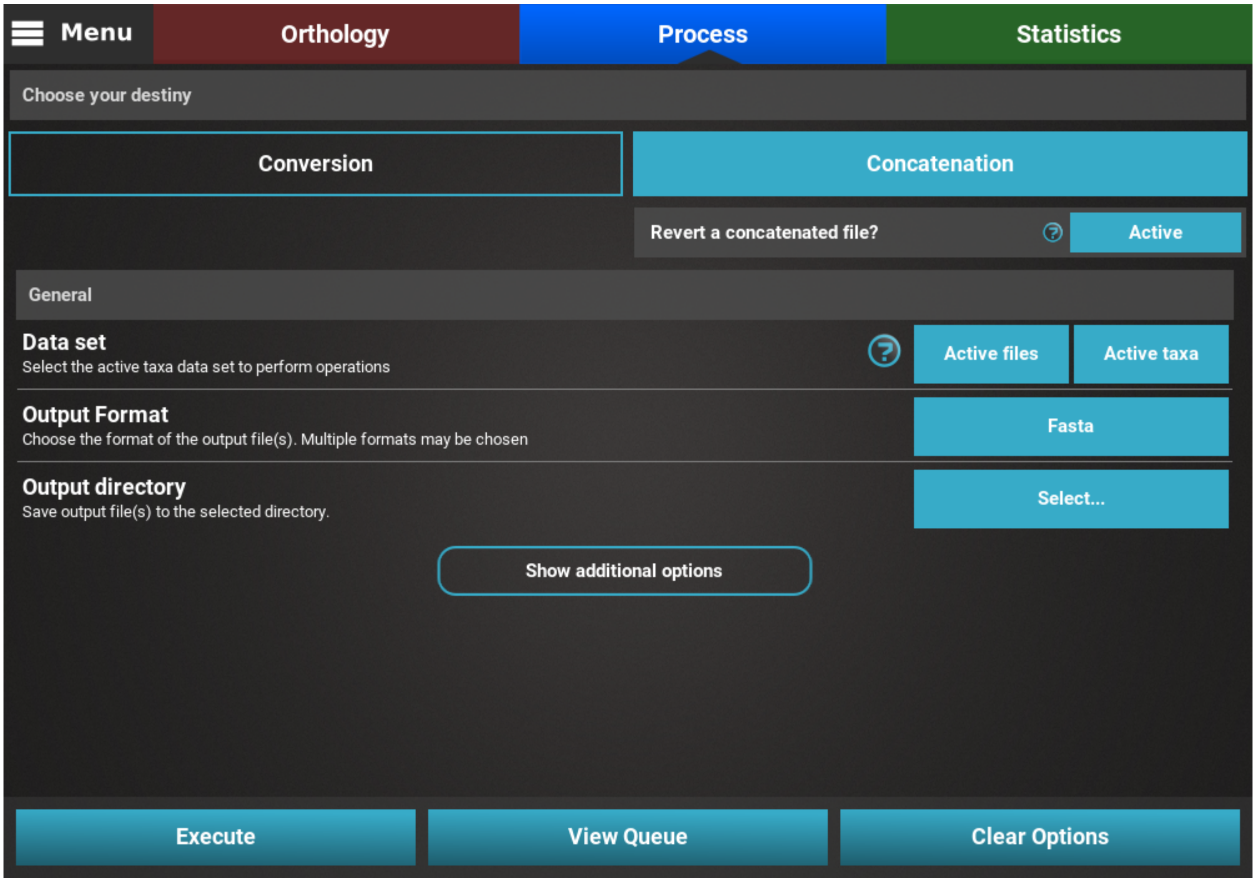
Output Format
Choose the format of the output file(s). Multiple formats may be chosen
Fasta

Output file
Save output file to the selected file.

In the reverse concatenation settings dialog, click the *Use defined partitions* tab, and then activate the operation by clicking the `Use defined partitions` button.



After clicking OK, make sure the Revert a concatenated file button has changed to Active.

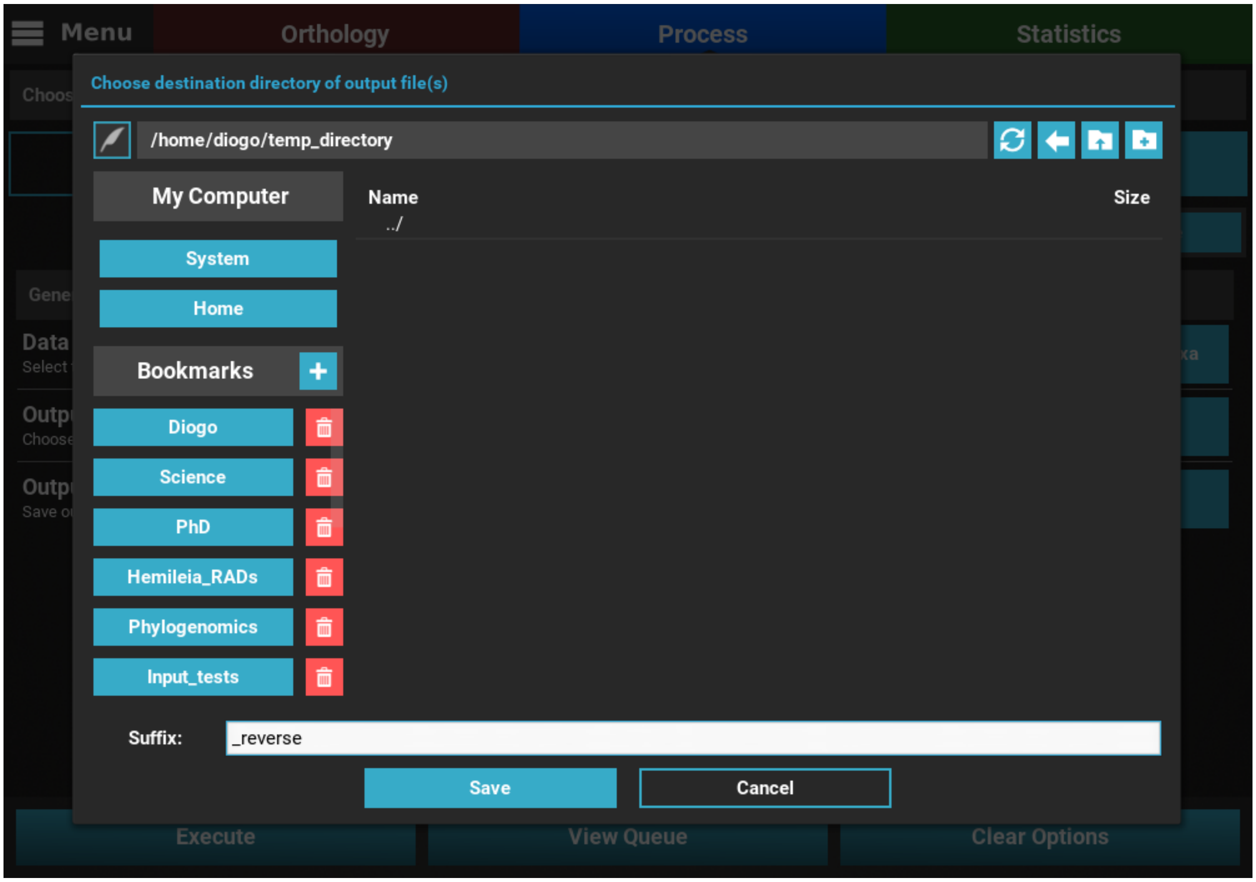


Data set
Select the active taxa data set to perform operations

Output Format
Choose the format of the output file(s). Multiple formats may be chosen

Output directory
Save output file(s) to the selected directory.

Now select an output directory where the individual alignments will be generated by clicking the `Select` button of the *Output directory* option. Note that the output files will be named according to the names of the defined partitions. Optionally, you may specify a suffix that will be appended to the end of every output file, but before the output format extension. Here we will specify the *"reverse"* suffix.

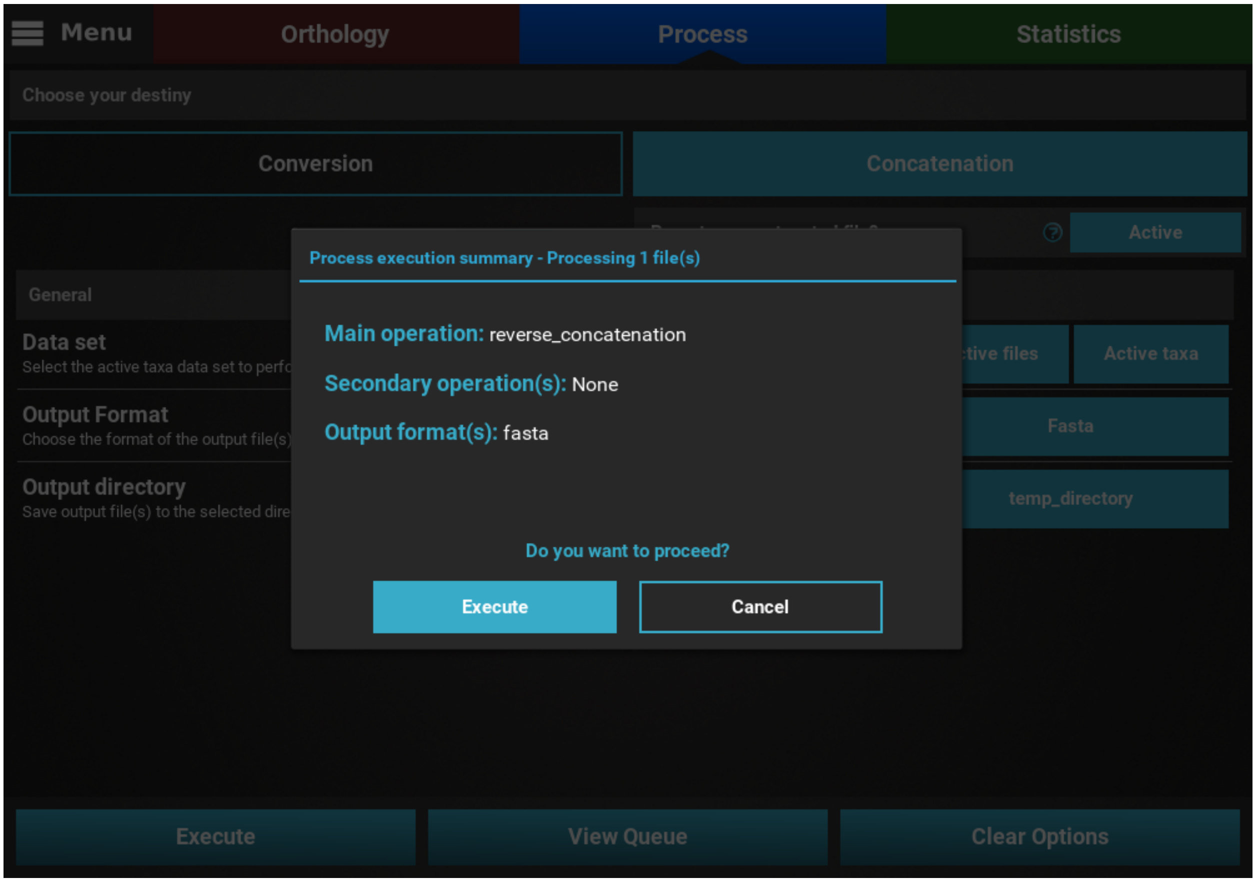


Choose destination directory of output file(s)

/home/diogo/temp_directory

Suffix: _reverse

Finally, click the `Execute` button to display the execution summary dialog, which will inform that a reverse concatenation operation will be performed, with no additional secondary operations and the output files will be in Fasta format. To begin the reverse concatenation, click the `Execute` button.



Process execution summary - Processing 1 file(s)

Main operation: reverse_concatenation

Secondary operation(s): None

Output format(s): fasta

Do you want to proceed?

1.12.3 Secondary operations after reversing a concatenated file

Secondary operations can also be performed in the same run when reversing a concatenated file. However, note that ALL secondary operations are performed after the reverse concatenation. This means that they will be applied to a set of individual alignments files and not to the initial concatenated file (see How main and secondary operations interact).

1.13 Secondary operations

Note: Data availability for this tutorial: the medium sized data set of 614 genes and 48 taxa that will be used can be downloaded here.

In addition to one of the main operations of TriFusion (**Conversion**, **Concatenation** and **Reverse concatenation**), one or more secondary operations can be applied during the processing of alignment files.

1.13.1 How main and secondary operations interact

Before starting with the secondary operations that are available on TriFusion, it is worth clarifying how the main and secondary operations interact:

- **Conversion**: Each secondary operation is applied independently on each active input alignment that will be converted.

- **Concatenation**: With the exception of the **Filter** secondary operation, all remaining secondary operation are performed on the single concatenated alignment file.
- **Reverse concatenation**: Secondary operations will be applied after the reverse concatenation, which means that they will be applied to each partition (output file) that will be generated. It's similar to the **Conversion** operation.

1.13.2 The order of operations

For performance reasons, operations in TriFusion are executed in a specific order:

1. **Reverse concatenation** [**main**]
2. **Filters** [*secondary*]
3. **Concatenation** [**main**]
4. **Collapse** [*secondary*]
5. **Gap coding** [*secondary*]
6. **Consensus** [*secondary*]
7. **Write to file**

Ok, so let's start with the tutorial.

1.13.3 Load data

As already covered in a separate tutorial (see *Load data into TriFusion*), alignment data can be loaded into TriFusion in three different ways. Here we will use the file browser to load an entire directory where 614 alignments files are stored.

Navigate to `Menu -> Open/View Data` and click the `Open file(s)` button. This will open the main file browser.

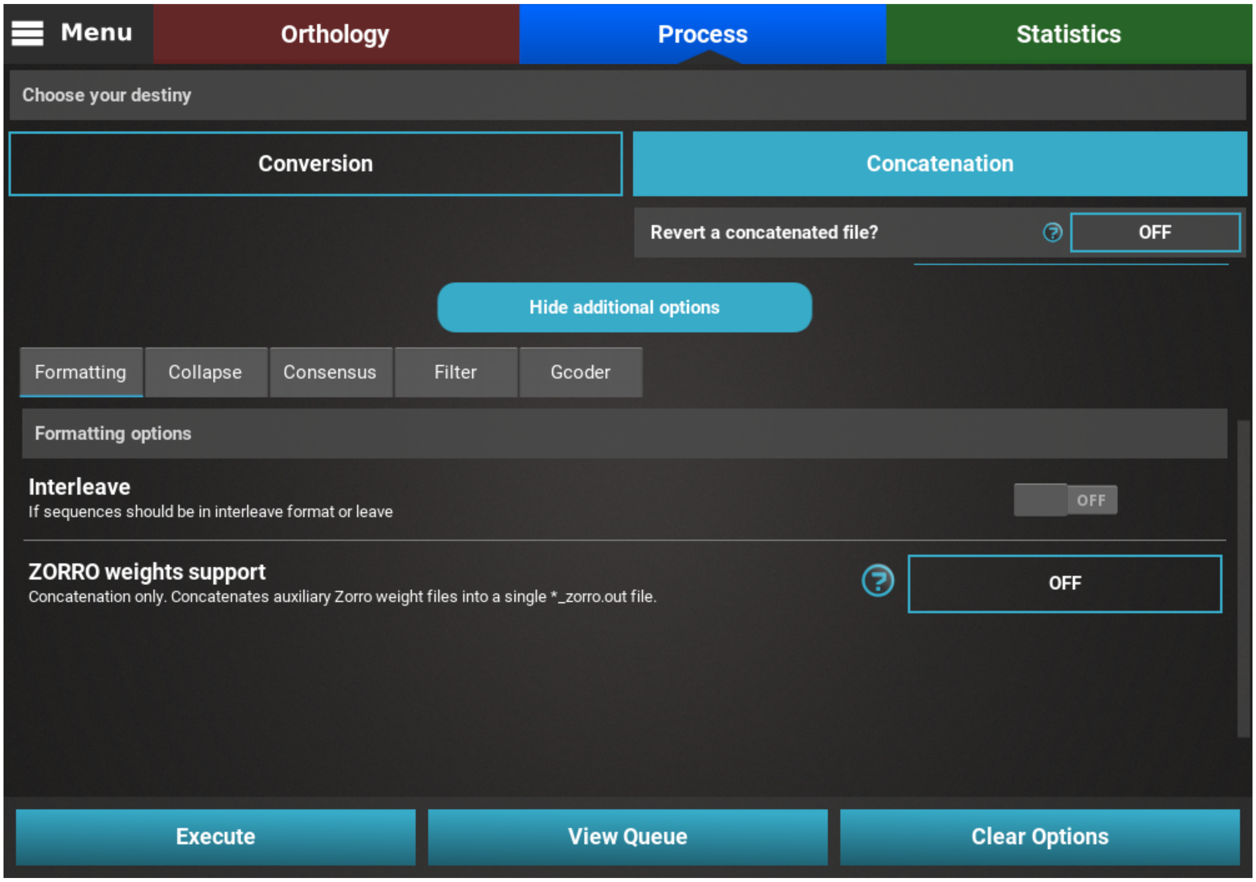
The input data type is already correctly set to **Alignment/Sequence set**, so we'll leave that as it is. Then, navigate the file browser until you find the directory containing the alignment files. In this case, all alignments are stored in a directory named *Version2*. Since TriFusion supports the selection of directories (in which case all files inside the specified directory will be loaded), I will only select the *Version2* directory and click `Load & go back` button. At the end of the data loading, a popup informs how many files were loaded.

Note: If you know that not all files in the selected directory are alignments, you could still load that particular directory. All invalid alignment files will be ignored when the data is loaded.

You'll also need to check the general options that are common to all operations.

1.13.4 Displaying secondary operations

To display all secondary operations, click the `Show additional options` button. This will reveal a tabbed panel, where the secondary operations are sorted into categories (with the exception of the *Formatting* tab, which is not a secondary operation per se).



Revert a concatenated file? OFF

Interleave
If sequences should be in interleave format or leave
OFF

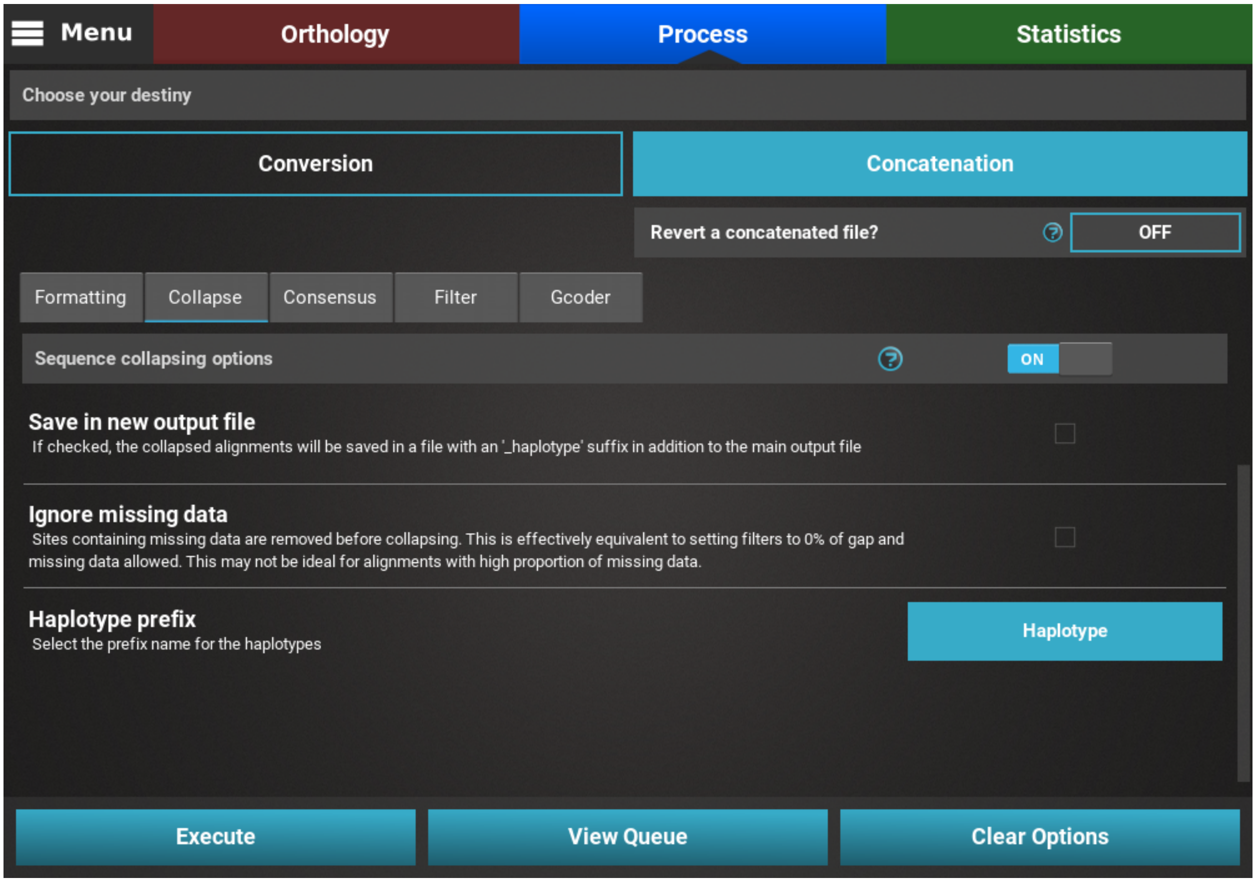
ZORRO weights support
Concatenation only. Concatenates auxiliary Zorro weight files into a single *_zorro.out file.
OFF

1.13.5 Collapse

Turn the collapse switch `ON` to activate the operation.

The **Collapse** secondary operation contains three options:

- *Save in new output file*: This will save the collapse alignment in another output file, separated from the main concatenated/converted output file. Checking this option will effectively produce two output files - a main output that is only concatenated/converted and another output file with the suffix "_collapsed" that will be concatenaded AND collapses. For now, we will not check this option.
- *Ignore missing data*: If this option is checked, sequences will be collapsed based on alignment columns that do not contain missing data and the output alignment will also contain 0% of missing data. The currently loaded data set has a fair amount of missing data, and is most likely not appropriate for collapsing using this option, so we will also leave this unchecked.
- *Haplotype prefix*: Sets the prefix for the haplotypes that will appear as the taxa names in the output file. An auxiliary file with the suffix "_haplotypes" will also be generated when performing this operation matching the new haplotype prefix to the original taxon names. Here we can change the default value to anything, like Haplotype



Revert a concatenated file? OFF

Sequence collapsing options ON

Save in new output file
If checked, the collapsed alignments will be saved in a file with an '_haplotype' suffix in addition to the main output file □

Ignore missing data
Sites containing missing data are removed before collapsing. This is effectively equivalent to setting filters to 0% of gap and missing data allowed. This may not be ideal for alignments with high proportion of missing data. □

Haplotype prefix
Select the prefix name for the haplotypes: Haplotype

Note: You can click the `Execute` button to execute the **Collapse** operation alone, or combine other secondary operations before.

1.13.6 Consensus

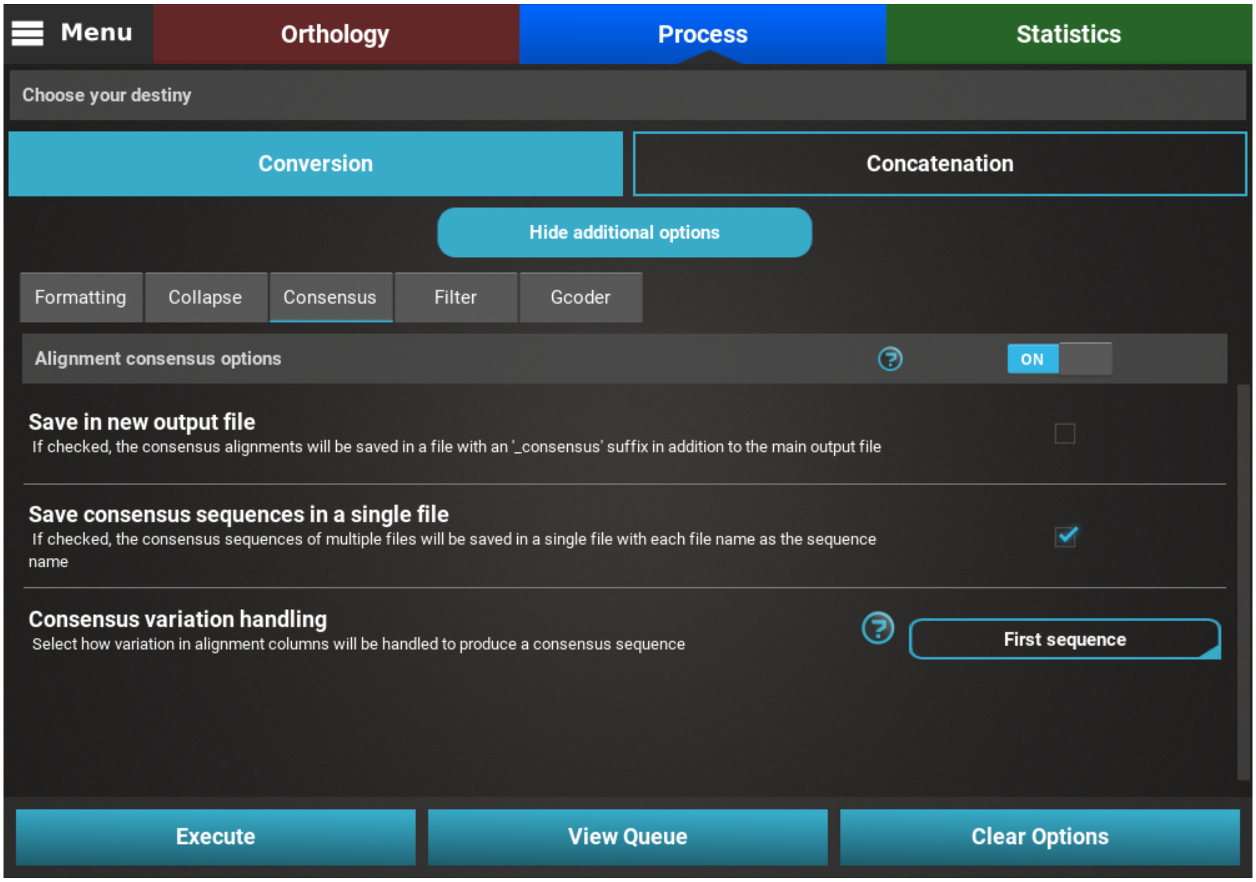
Turn the consensus switch `ON` to activate the operation.

The consensus operation is mainly used to compress multiple sequences in an alignment into one representative sequence. While it can be done on top of the **Concatenation** main operation, what this will do is concatenate all 614 alignments into a single concatenated one and then create a consensus of that large alignment. However, in the majority of the cases, users are more interested in creating a consensus sequence for each input alignment. With this in mind, this secondary operation should be done with the **Conversion** main operation.

The **Consensus** secondary operation contains three options:

- *Save in new output file*: This will save the consensus alignment in another output file, separated from the main concatenated/converted output file. Checking this option will effectively produce two output files - a main output that is only concatenated/converted and another output file with the suffix "_consensus" that have the **consensus** performed. For now, we will not check this option.
- *Save consensus in a single file*: This option can be checked to merge all consensus from each input alignment in a single file. In this case, if this option is left unchecked, 614 output files will be created using this option, each with a single representative consensus sequence of the corresponding alignment. However, here we are more interested in merging all consensus sequences in a single file that will be later provided for functional annotation analyses. So we'll check this option.

- *Consensus variation handing*: Select how you would like to handle variation within each alignment. The appropriate choice is highly dependent on subsequent analyses. In our case, since we want to create a dataset for Blast2GO and our alignment data is fairly variable, we'll select the First sequence value, where the first sequence of each alignment is selected as a representative.



Alignment consensus options ON

Save in new output file □
If checked, the consensus alignments will be saved in a file with an '_consensus' suffix in addition to the main output file

Save consensus sequences in a single file ☑
If checked, the consensus sequences of multiple files will be saved in a single file with each file name as the sequence name

Consensus variation handling: First sequence
Select how variation in alignment columns will be handled to produce a consensus sequence

Note: You can click the `Execute` button to execute the **Consensus** operation alone, or combine other secondary operations before.

1.13.7 Filters

There are several **Filter** operations that can be applied to the alignments. Turn the filter main switch `ON` to activate the operation. Now you can specify one or more filters to execute in the same run. Whenever a particular filter is active, the button of the corresponding operation will display **Filters set**.

Note: The **Codon** filter operation can only be executed on nucleotide alignments, so it will be disabled when protein alignments are loaded. You can use the small 7 alignment data set for this tutorial.

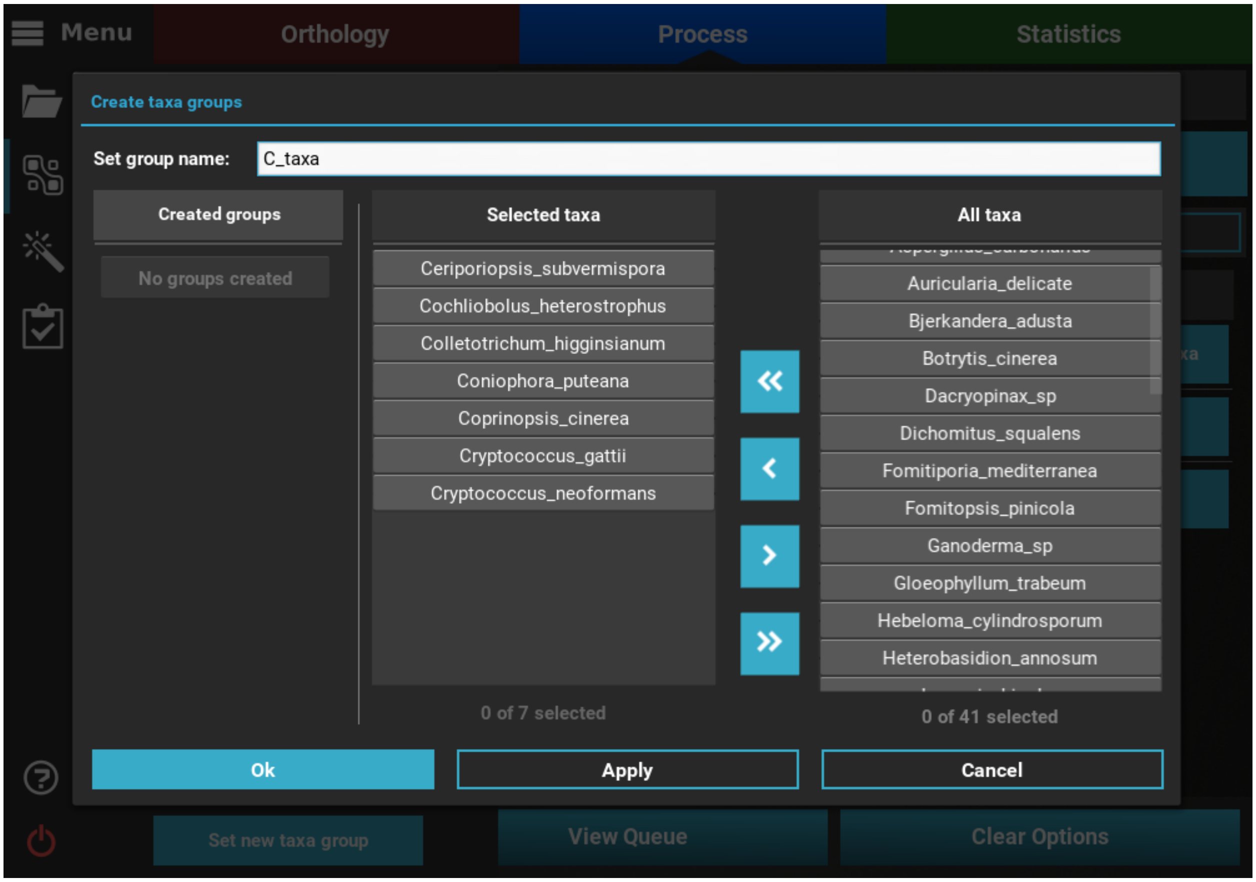
Taxa filter

Click on the button of the *Taxa filter* option and turn the switch on the popup of this operation `ON`.

The **Taxa filter** operation allows users to filter entire alignments if they contain or exclude a given set of taxa. Here, we will create a fictional case where we are interested in concatenating only alignments that contain at least all taxa with names beginning on a "C".

The *filter mode* sets whether the alignments should be filtered if they **contain** or **exclude** the taxa group. By default, it is set to *Containing*, so we'll leave that unchanged.

As you can see, there are no taxa groups yet defined so we'll need to create a new one. Click the `Set taxa group` button to start the data set group creation process and then click the `Set manually` button. Here, select the desired taxa with names starting with the letter "C" and save the taxa group by clicking `Ok`.



Set group name: C_taxa

Selected taxa:

Ceriporiopsis_subvermispora
Cochliobolus_heterostrophus
Colletotrichum_higginsianum
Coniophora_puteana
Coprinopsis_cinerea
Cryptococcus_gattii
Cryptococcus_neoformans

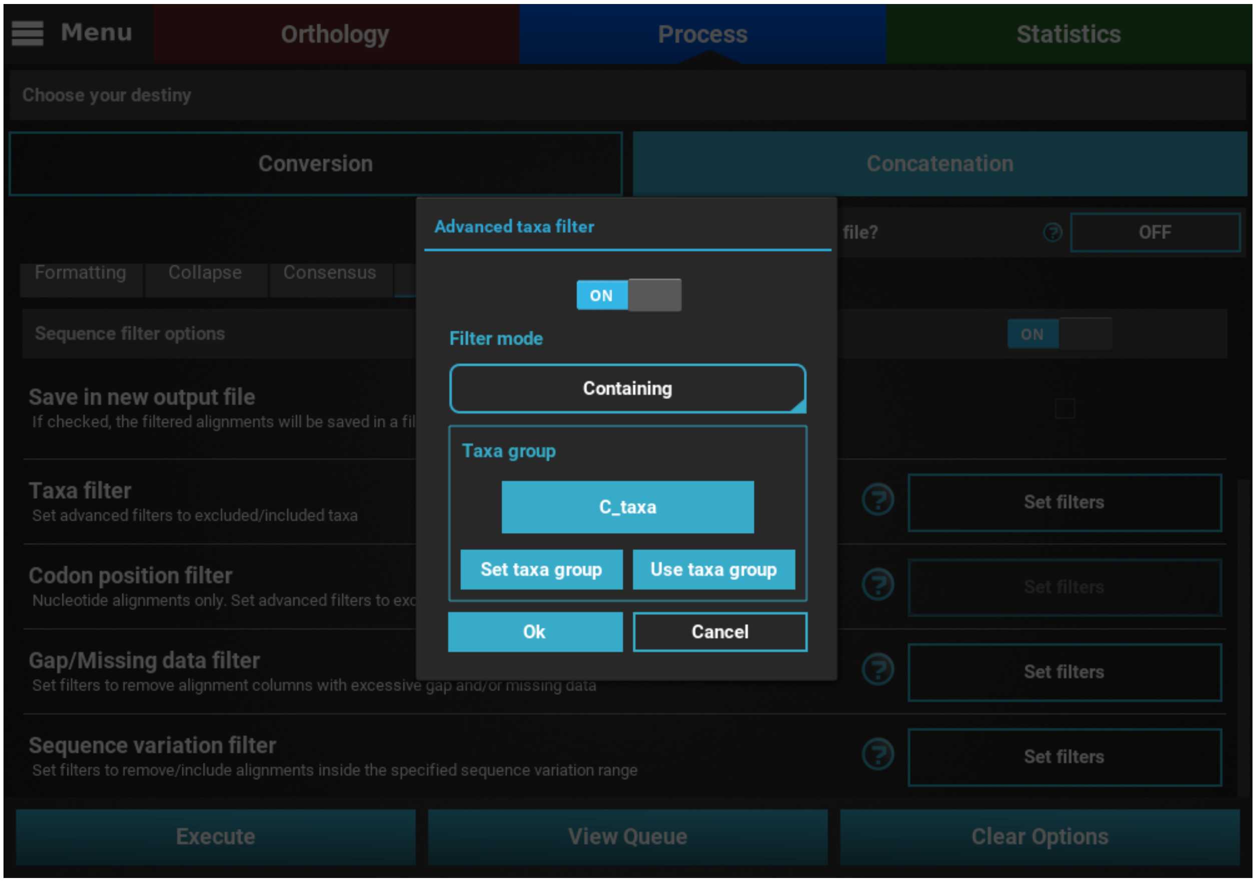
0 of 7 selected

All taxa:

Auricularia_delicate
Bjerkandera_adusta
Botrytis_cinerea
Dacryopinax_sp
Dichomitus_squalens
Fomitiporia_mediterranea
Fomitopsis_pinicola
Ganoderma_sp
Gloeophyllum_trabeum
Hebeloma_cylindrosporum
Heterobasidion_annosum

0 of 41 selected

Once the group has been created, it will be automatically selected in the **Taxa filter** dialog. Additional groups can be created in the same way. When multiple groups have been defined, they can be selected by clicking the `Use taxa group` button, and then selecting the desired group.

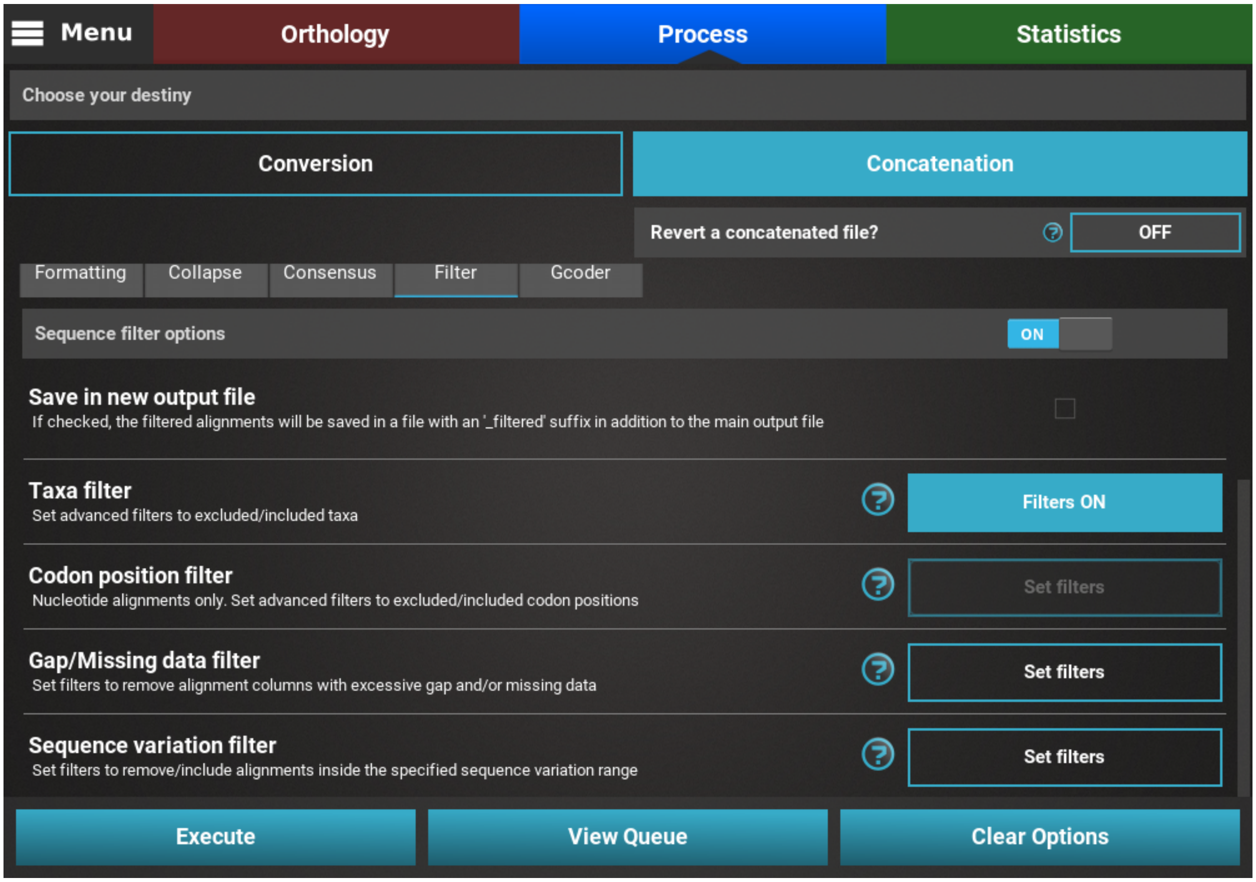

Menu
Orthology
Process
Statistics
Choose your destiny
Conversion
Concatenation
file?
OFF
Formatting
Collapse
Consensus
Sequence filter options
ON
Save in new output file
If checked, the filtered alignments will be saved in a fil
Taxa filter
Set advanced filters to excluded/included taxa
Set filters
Codon position filter
Nucleotide alignments only. Set advanced filters to exc
Set filters
Gap/Missing data filter
Set filters to remove alignment columns with excessive gap and/or missing data
Set filters
Sequence variation filter
Set filters to remove/include alignments inside the specified sequence variation range
Set filters
Execute
View Queue
Clear Options
Advanced taxa filter
Set taxa group
Use taxa group
Ok
Cancel

Advanced taxa filter: ON

Filter mode: Containing

Taxa group: C_taxa

When you are happy with the **Taxa filter** settings, click the `Ok` button. If the Taxa filter switch was turned `ON`, the button of the *Taxa filter* option should change to *Filters ON*.



Menu | Orthology | Process | Statistics

Choose your destiny

Conversion | Concatenation

Revert a concatenated file? OFF

Formatting | Collapse | Consensus | Filter | Gcoder

Sequence filter options ON

Save in new output file
If checked, the filtered alignments will be saved in a file with an '_filtered' suffix in addition to the main output file □

Taxa filter
Set advanced filters to excluded/included taxa — Filters ON

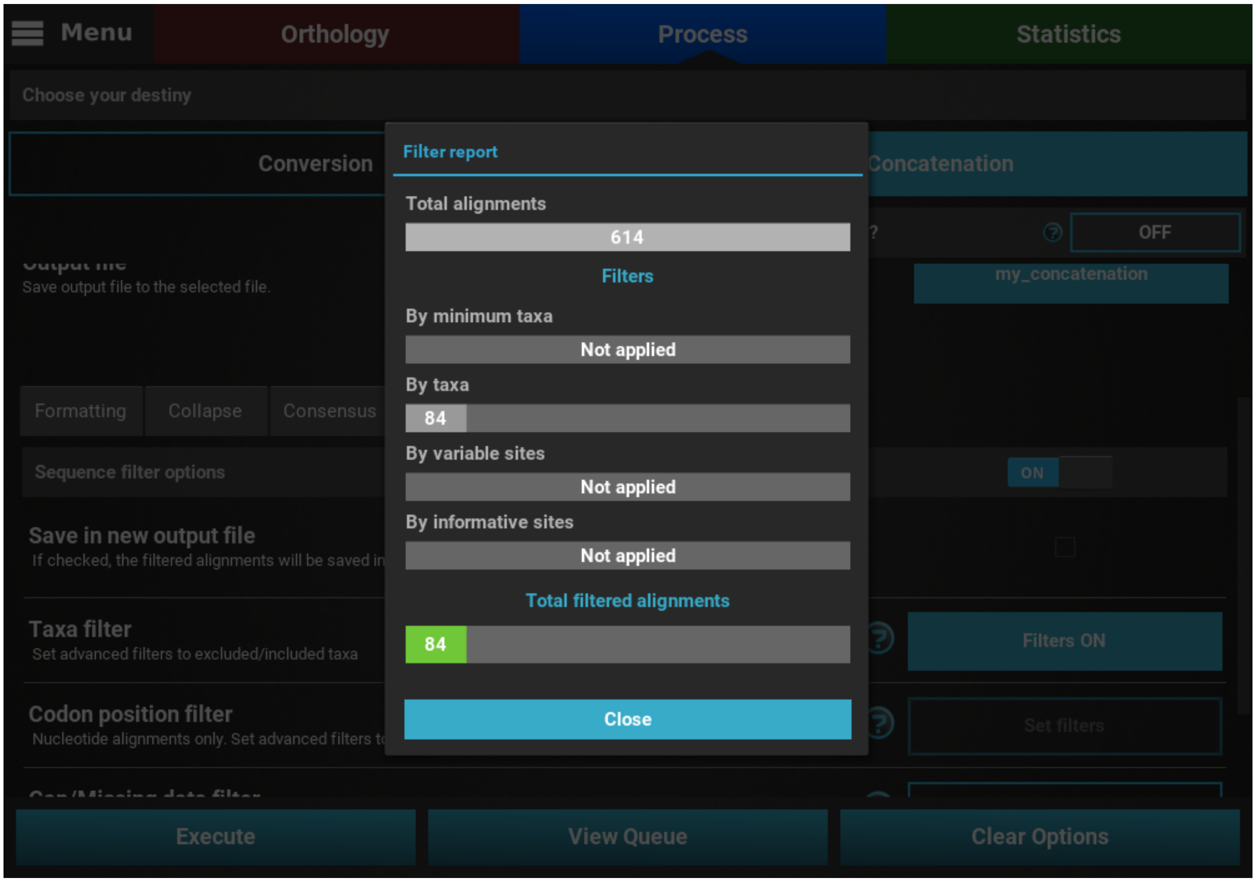
Codon position filter
Nucleotide alignments only. Set advanced filters to excluded/included codon positions — Set filters

Gap/Missing data filter
Set filters to remove alignment columns with excessive gap and/or missing data — Set filters

Sequence variation filter
Set filters to remove/include alignments inside the specified sequence variation range — Set filters

Execute | View Queue | Clear Options

Finally, press the `Execute` button at the bottom of the **Process** screen to execute the filter operation. At the end of **Filter** operations that may remove alignment files from the final output, a *Filter report* will popup informing how many alignments were filtered. In our case, 84 alignments were filtered (By taxa filter) from the final output.



Filter report

Total alignments

614

Filters

By minimum taxa

Not applied

By taxa

84

By variable sites

Not applied

By informative sites

Not applied

Total filtered alignments

84

Close

Codon position filter

Note: This filter is only available for nucleotide alignments.

Turn `ON` the filter switch to activate the operation. Then click the `Set filters` button for the *Codon position* filter option and turn `ON` the switch on the popup as well.

The **Codon position filter** operation allows you to remove certain codon positions from the output alignment. Consequently, this option is only available for nucleotide sequences. In many nucleotide alignments it is common to remove the third codon position, as it is generally much more variable and could introduce a substantial amount of phylogenetic noise. However, this option removes the same codon positions in all input alignments. For example, if you load 10 alignments in TriFusion and exclude the 3rd codon position, you must make sure that all 10 alignments start in the 1st codon position. However, if all alignments start in the 2nd codon position, for instance, removing the 3rd codon position is still possible in TriFusion, by excluding the 2nd positions (which will actually correspond to the 3rd positions in the alignment).

To exclude a given codon position, simply toggle the corresponding button off. Included position button always have a blue background.

Gap/Missing data filter

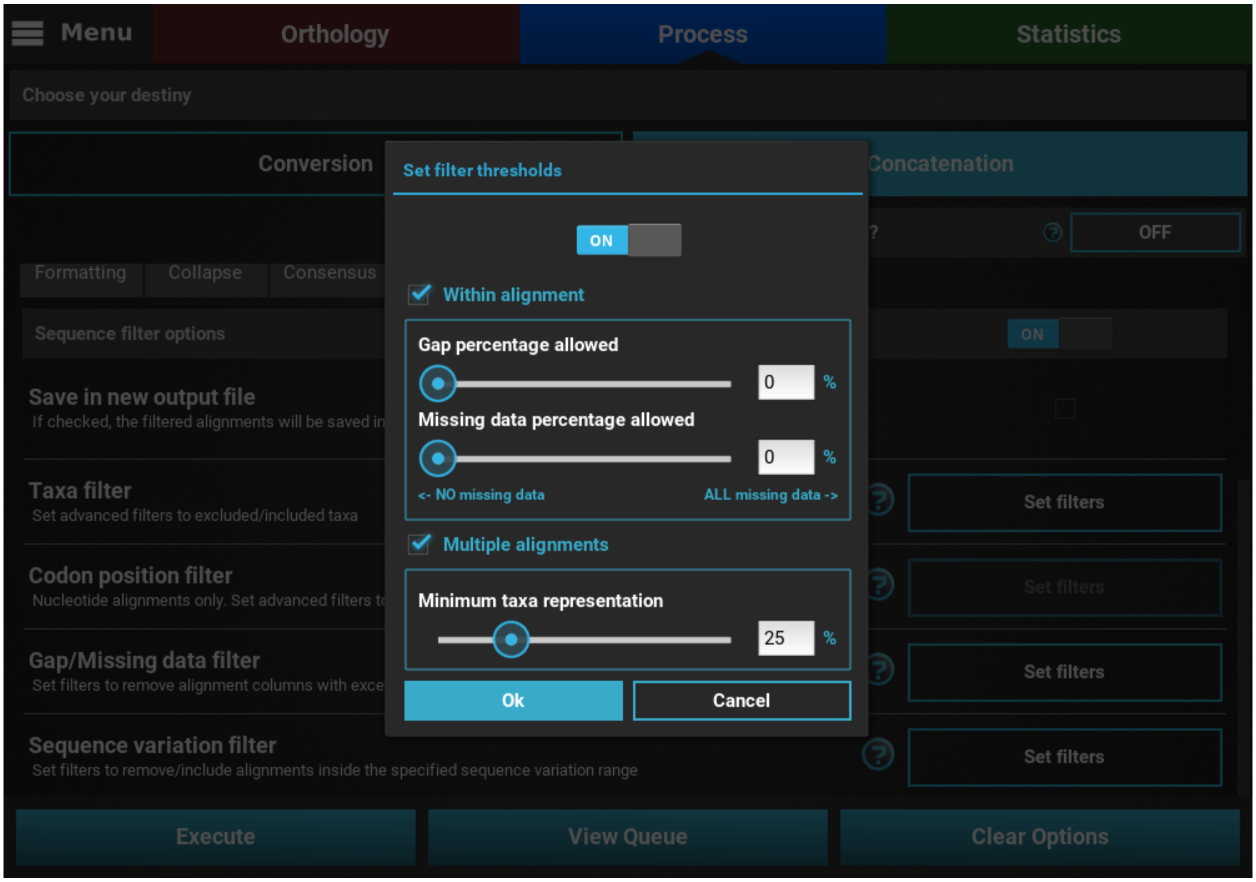
Turn ON the filter switch to activate the operation. Then click the `Set filters` button for the **Gap/Missing data filter** option and turn ON the switch on the popup as well.

The **Gap/Missing data filter** allows user to filter alignment columns (within alignment) and/or alignments (multiple alignments) based on their missing data content. Both filters can be used in combination, if both within alignment and multiple alignments checkboxes are active, or only one of them.

In this example, we will filter both alignment columns and alignment files, so both checkboxes will remain active. **Within an alignment**, columns can be filtered depending on the amount of gaps or missing data. **Gaps** refer to the usual gap symbol ("-") while **missing data** refers to the sum of gap symbols **AND** true missing data ("N" for nucleotides or "X" for proteins). These filters provide maximum threshold values in percentages, above which alignment columns are filtered. For example, if the **gap percentage** allowed option is set to 25% and the **missing data percentage** allowed option is set to 50%, then alignment columns with more than 25% of gaps OR more than 50% of gaps + true missing data are filtered.

In our case, we are interested in producing an output matrix that contains no missing data, so we will set both sliders to 0%.

Concerning the multiple alignments option, we will be more relaxed. We'll set the slider to 25%, which means that only alignments with more than 25% of the total data set taxa (12 out of 48 in this case) will be further processed.



Set filter thresholds

ON

☑ Within alignment

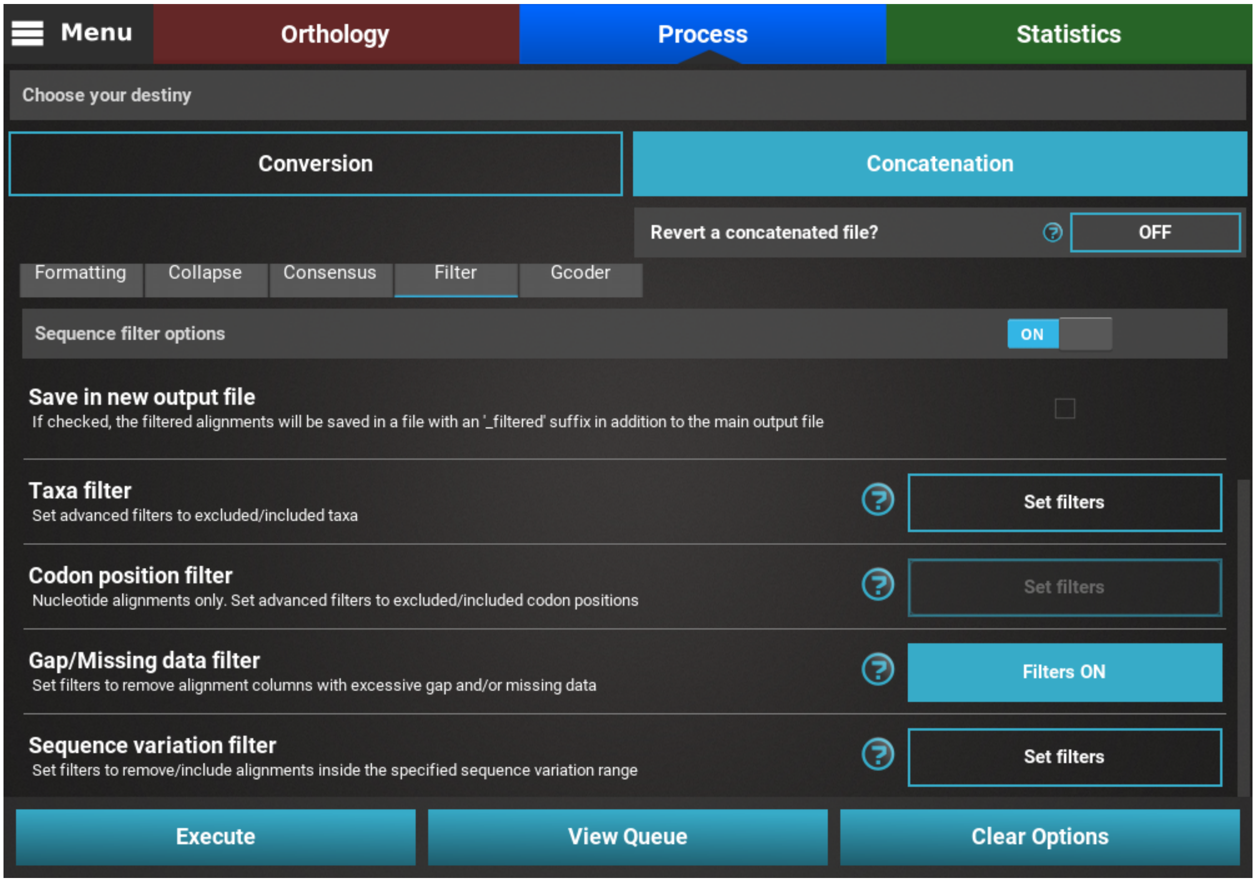
Gap percentage allowed 0 %

Missing data percentage allowed 0 %

☑ Multiple alignments

Minimum taxa representation 25 %

When you are happy with the gap/missing data filter settings, click the `Ok` button. If the **Gap/Missing data filter** switch was turned ON, the button of the **Gap/Missing data filter** option should change to *Filters ON*.



Revert a concatenated file? OFF

Sequence filter options ON

Save in new output file □
If checked, the filtered alignments will be saved in a file with an '_filtered' suffix in addition to the main output file

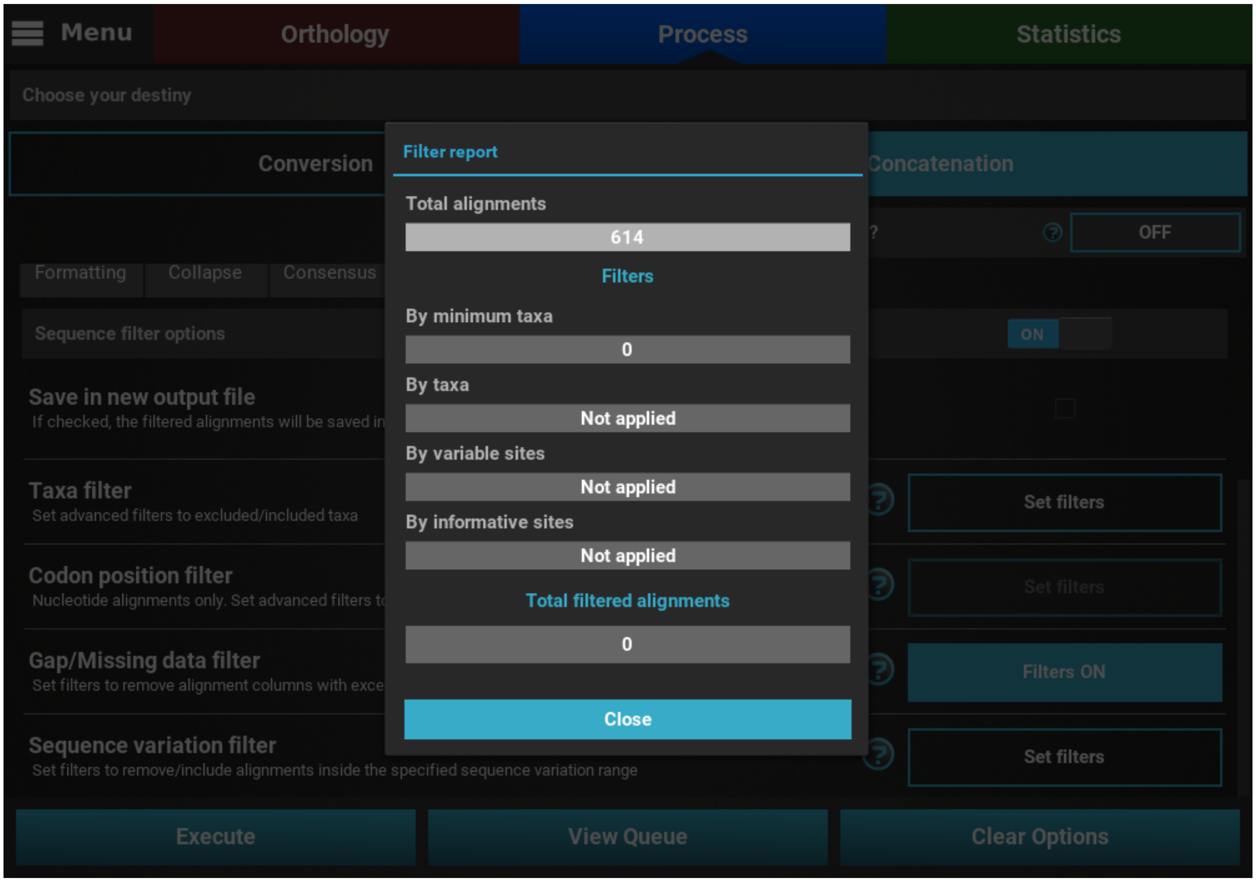
Taxa filter — Set filters
Set advanced filters to excluded/included taxa

Codon position filter — Set filters
Nucleotide alignments only. Set advanced filters to excluded/included codon positions

Gap/Missing data filter — Filters ON
Set filters to remove alignment columns with excessive gap and/or missing data

Sequence variation filter — Set filters
Set filters to remove/include alignments inside the specified sequence variation range

Finally, press the `Execute` button at the bottom of the **Process** screen to execute the filter operation. At the end of Filter operations that may remove alignment files from the final output, a Filter report will popup informing how many alignments were filtered. In our case, there were actually no filtered alignments, which means that all input alignments already contained more than 25% of the total taxa.


Menu
Orthology
Process
Statistics
Choose your destiny
Conversion
Concatenation
OFF
Formatting
Collapse
Consensus
Sequence filter options
ON
Save in new output file
If checked, the filtered alignments will be saved in
Taxa filter
Set advanced filters to excluded/included taxa
Set filters
Codon position filter
Nucleotide alignments only. Set advanced filters to
Set filters
Gap/Missing data filter
Set filters to remove alignment columns with exce
Filters ON
Sequence variation filter
Set filters to remove/include alignments inside the specified sequence variation range
Set filters
Execute
View Queue
Clear Options

Filter report

Total alignments

614

Filters

By minimum taxa

0

By taxa

Not applied

By variable sites

Not applied

By informative sites

Not applied

Total filtered alignments

0

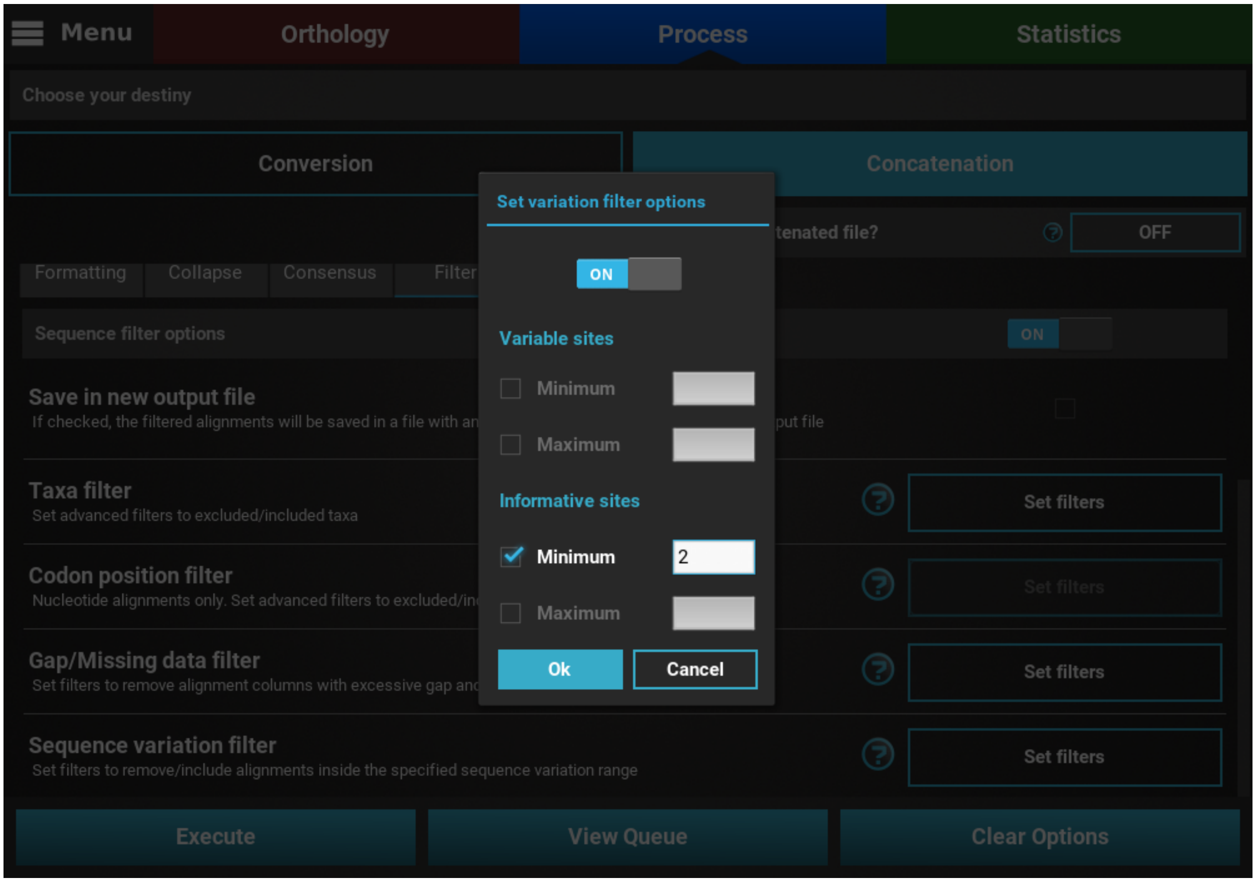
Close

Sequence variation filter

Turn `ON` the filter switch to activate the operation. Then click the `Set filters` button for the *Sequence variation filter* option and turn `ON` the switch on the popup as well.

The **sequence variation filter** allows users to filter alignment files based on the amount of sequence variation. The two supported types of sequence variation are **variable sites** and **informative sites**. The different between these types is that variable sites includes all columns with at least one variant, while informative sites only includes variable columns where at least one alternative allele has two or more copies.

Here, you can specify multiple combination of maximum and minimum values for each variation type. When a checkbox is left inactive, it is assumed that there is no boundary for that specific value. For instance, let's filter our alignments so that only alignments with at least 2 informative sites are processed. To achieve this, check the *Minimum* box of the **informative sites** option and set it to 2, but leave the *Maximum* box unchecked.


Menu
Orthology
Process
Statistics
Choose your destiny
Conversion
Concatenation
tenated file?
OFF
Formatting
Collapse
Consensus
Filter
Sequence filter options
ON
Save in new output file
If checked, the filtered alignments will be saved in a file with an
put file
Taxa filter
Set advanced filters to excluded/included taxa
Set filters
Codon position filter
Nucleotide alignments only. Set advanced filters to excluded/in
Set filters
Gap/Missing data filter
Set filters to remove alignment columns with excessive gap and
Set filters
Sequence variation filter
Set filters to remove/include alignments inside the specified sequence variation range
Set filters
Execute
View Queue
Clear Options

Set variation filter options

ON

Variable sites

☐ Minimum

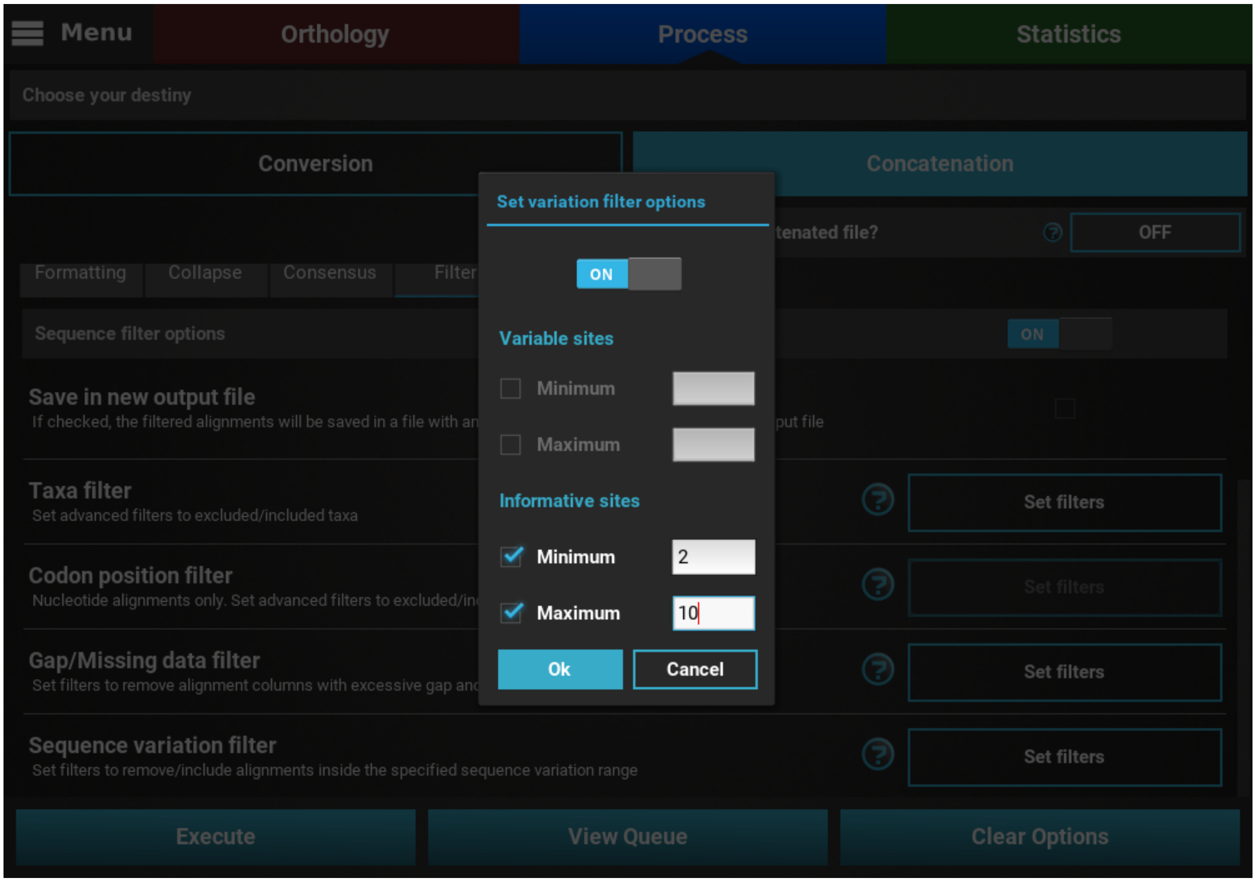
☐ Maximum

Informative sites

☑ Minimum 2

☐ Maximum

If you would like to set an upper limit to the number of **informative sites**, just check the *Maximum* box and set a number higher than 2. In this case, let's put an upper limit of 10 informative sites.


Menu
Orthology
Process
Statistics
Choose your destiny
Conversion
Concatenation
tenated file?
OFF
Formatting
Collapse
Consensus
Filter
Sequence filter options
ON
Save in new output file
If checked, the filtered alignments will be saved in a file with an
put file
Taxa filter
Set advanced filters to excluded/included taxa
Set filters
Codon position filter
Nucleotide alignments only. Set advanced filters to excluded/in
Set filters
Gap/Missing data filter
Set filters to remove alignment columns with excessive gap anc
Set filters
Sequence variation filter
Set filters to remove/include alignments inside the specified sequence variation range
Set filters
Execute
View Queue
Clear Options
Set variation filter options
ON
Variable sites
Informative sites
Ok
Cancel

Variable sites

□ Minimum

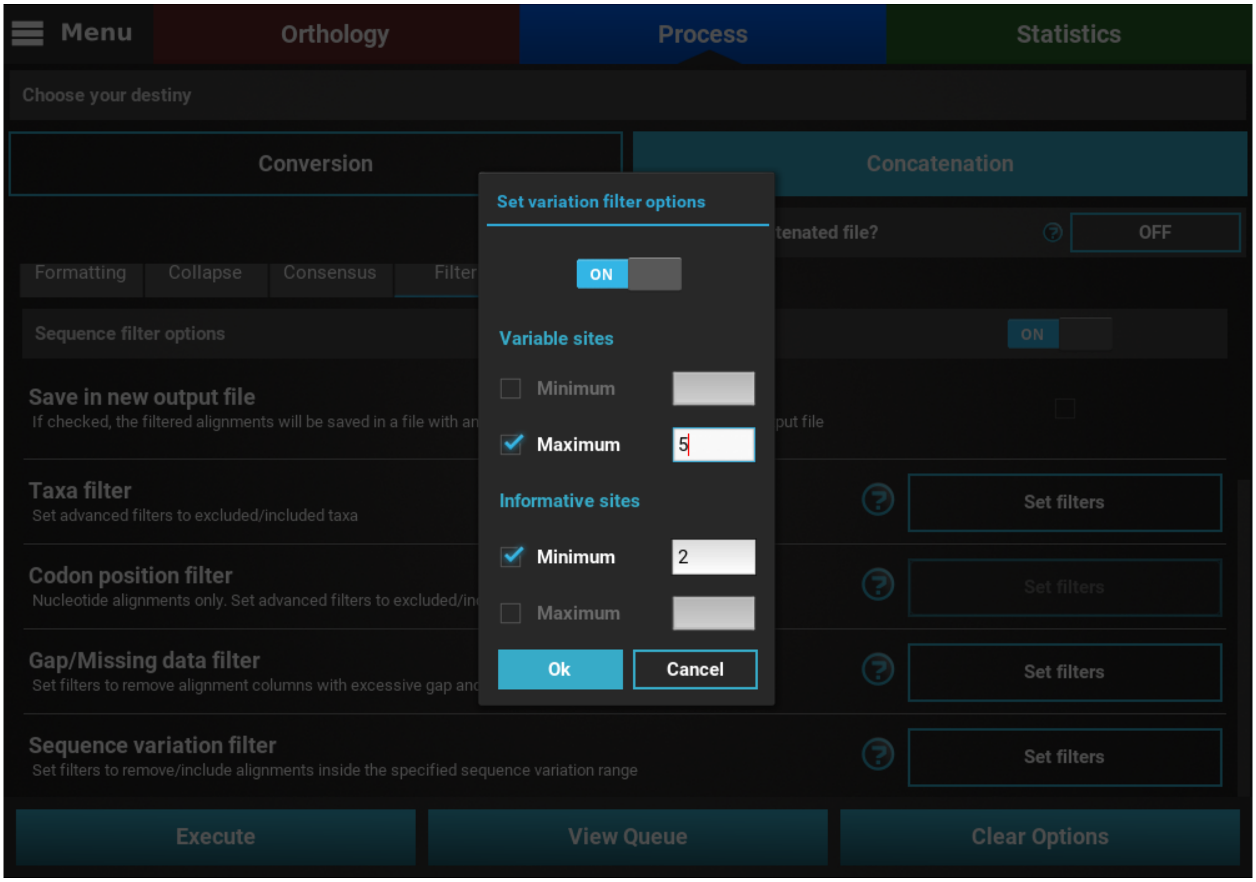
□ Maximum

Informative sites

☑ Minimum 2

☑ Maximum 10

It is also possible to mix both types of sequence variation. For instance, we may want to filter alignments with more than 2 informative sites and less than 200 variable sites.


Menu
Orthology
Process
Statistics
Choose your destiny
Conversion
Concatenation
tenated file?
OFF
Formatting
Collapse
Consensus
Filter
Sequence filter options
ON
Save in new output file
If checked, the filtered alignments will be saved in a file with an
put file
Taxa filter
Set advanced filters to excluded/included taxa
Set filters
Codon position filter
Nucleotide alignments only. Set advanced filters to excluded/in
Set filters
Gap/Missing data filter
Set filters to remove alignment columns with excessive gap and
Set filters
Sequence variation filter
Set filters to remove/include alignments inside the specified sequence variation range
Set filters
Execute
View Queue
Clear Options
Ok
Cancel

Set variation filter options

ON

Variable sites

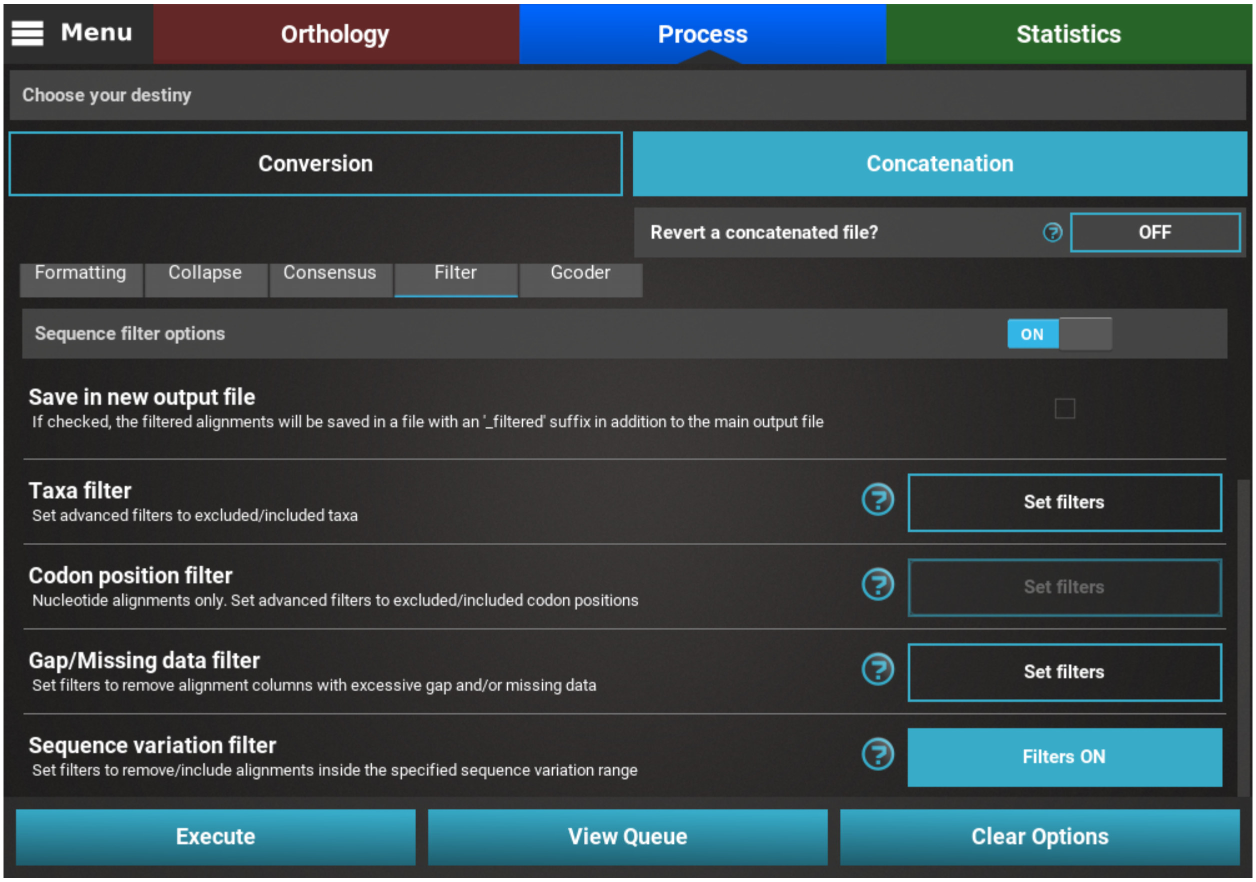
- □ Minimum
- ☑ Maximum 5

Informative sites

- ☑ Minimum 2
- □ Maximum

However, note that certain combination are redundant. For instance, if you set a minimum of informative sites to 2, setting a minimum of variable sites to 1 will have no effect on the final output.

When you are happy with the sequence variation filter settings, click the `Ok` button. If the **Sequence variation filter** switch was turned `ON`, the button of the *Sequence variation filter* option should change to *Filters ON*.



Revert a concatenated file? OFF

Sequence filter options ON

Save in new output file
If checked, the filtered alignments will be saved in a file with an '_filtered' suffix in addition to the main output file □

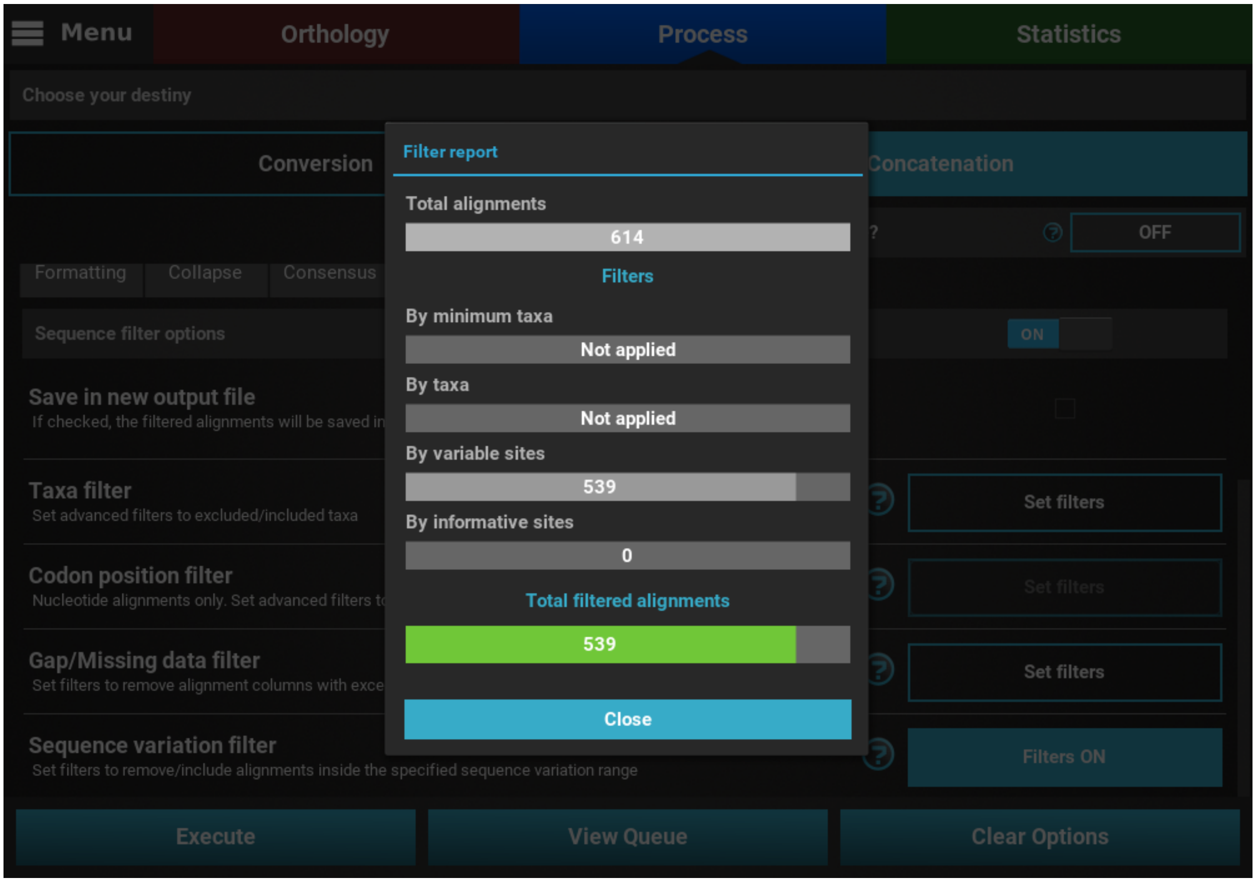
Taxa filter
Set advanced filters to excluded/included taxa

Codon position filter
Nucleotide alignments only. Set advanced filters to excluded/included codon positions

Gap/Missing data filter
Set filters to remove alignment columns with excessive gap and/or missing data

Sequence variation filter
Set filters to remove/include alignments inside the specified sequence variation range

Finally, press the `Execute` button at the bottom of the **Process** screen to execute the filter operation. At the end of **Filter** operations that may remove alignment files from the final output, a Filter report will popup informing how many alignments were filtered. In our case, if we execute filter options of a least 2 informative sites and less than 200 variable sites, a total of 539 alignments will be filtered.


Menu
Orthology
Process
Statistics
Choose your destiny
Conversion
Concatenation
OFF
Formatting
Collapse
Consensus
Sequence filter options
ON
Save in new output file
If checked, the filtered alignments will be saved in
Taxa filter
Set advanced filters to excluded/included taxa
Set filters
Codon position filter
Nucleotide alignments only. Set advanced filters t
Set filters
Gap/Missing data filter
Set filters to remove alignment columns with exce
Set filters
Sequence variation filter
Set filters to remove/include alignments inside the specified sequence variation range
Filters ON
Execute
View Queue
Clear Options

Filter report

Total alignments

614

Filters

By minimum taxa

Not applied

By taxa

Not applied

By variable sites

539

By informative sites

0

Total filtered alignments

539

Close

1.13.8 Gap coding

Turn `ON` the gap coding switch to activate the operation.

The **Gap coding** operation enables the codification of gaps as a binary matrix that is appended to the final of the alignment matrix. This option is available only when the **Nexus** format is the only output format selected. Currently, it contains a single available option:

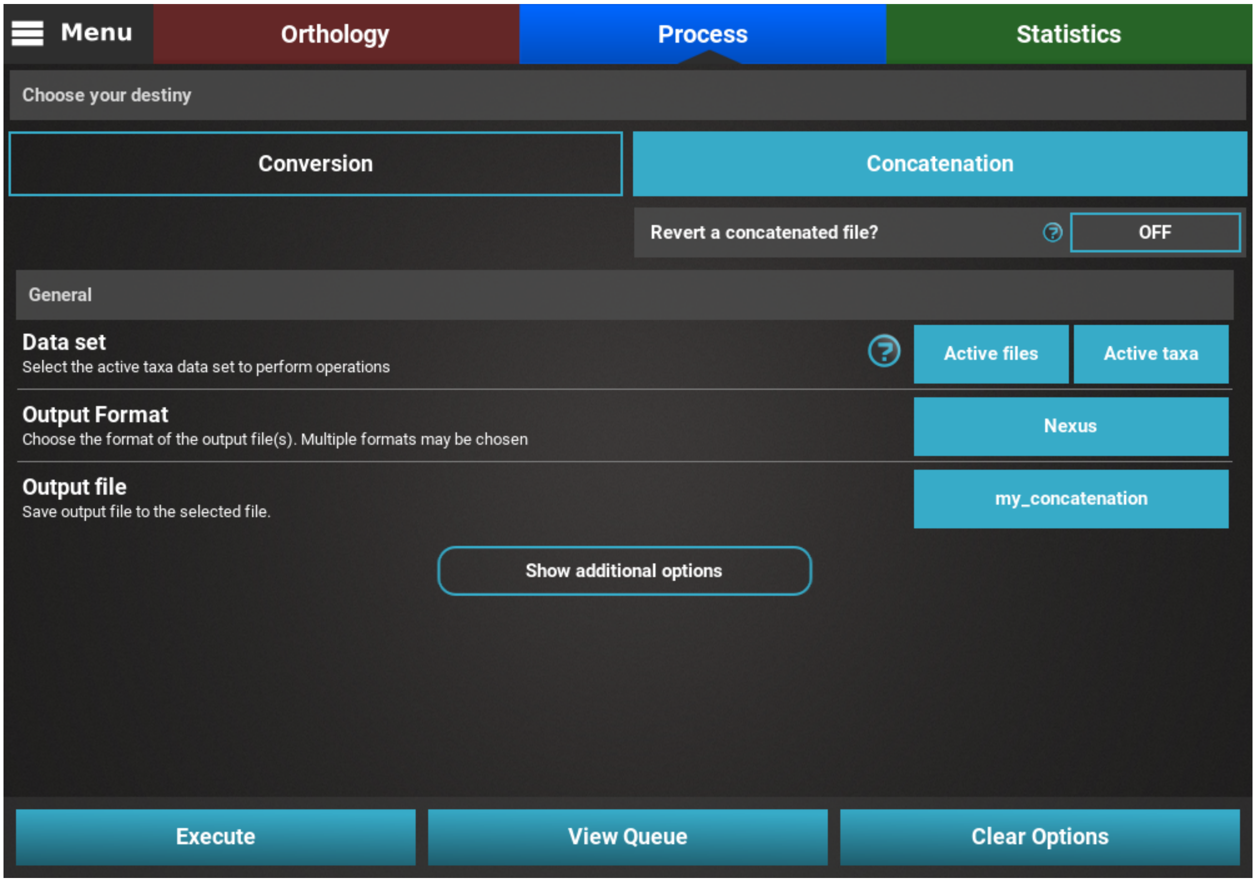
- *Save in new output file*: This will save the alignment with coded gaps in another output file, separated from the main concatenated/ converted output file. Checking this option will effectively produce two output files - a main output that is only concatenated/converted and another output file with the suffix "_gcoded" that will have the coded gaps. For now, we will not check this option.

The Gap coding method is currently restricted to the one described in Simmons and Ochotenera 2000, however additional methods are expected to be added in future releases.

1.13.9 Combination of three secondary operations

Until now, we only dealt with the activation and usage of individual secondary operations. However, many of these operations can fit rather naturally in combination. Here I'll demonstrated how a data set of 614 alignments with 48 taxa can be concatenated, collapsed and filtered in a single run, with the condition that the collapsed alignment has to be generated in an independent alignment file.

After loading the data, select the **Concatenation** main operation in the Process screen. To keep things simple, let's leave the Data set options in the default values, select only the Nexus output format and provide an output file name (here it will be *my_concatenation*).



Data set
Select the active taxa data set to perform operations

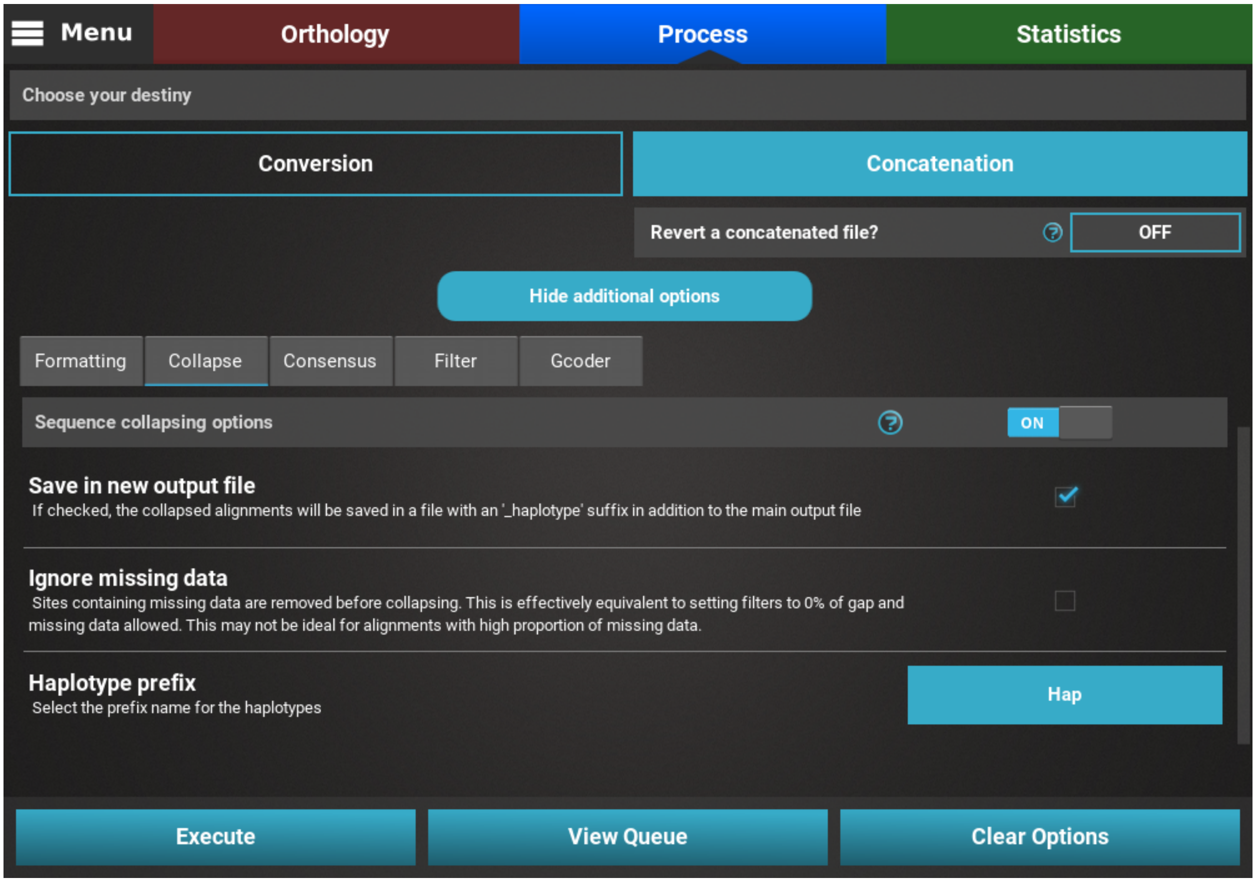
Output Format
Choose the format of the output file(s). Multiple formats may be chosen

Output file
Save output file to the selected file.

Setting up collapse operation

Open the **secondary operations** tabbed menu by clicking the `Show additional options` button, click on the *Collapse* tab and turn the switch `ON`.

Since we want to save the collapsed alignment in a separated output that is independent of the remaining operations, we'll check the *Save in new output file* box. Our data set contains a fair amount of missing data, so we'll leave the *Ignore missing data* box unchecked. Finally, we can leave the haplotype prefix in its default *Hap* value.



Revert a concatenated file? OFF

Sequence collapsing options ON

Save in new output file ☑
If checked, the collapsed alignments will be saved in a file with an '_haplotype' suffix in addition to the main output file

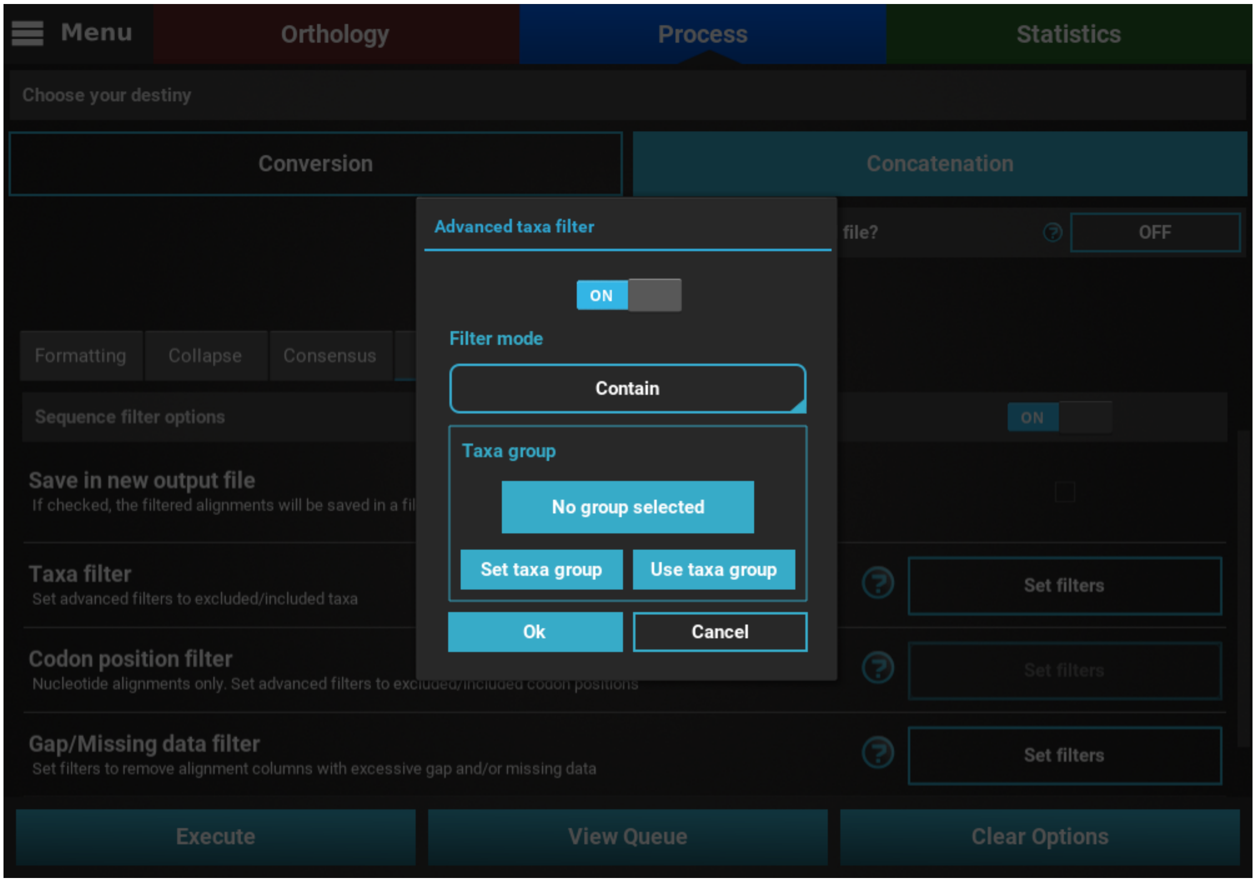
Ignore missing data ☐
Sites containing missing data are removed before collapsing. This is effectively equivalent to setting filters to 0% of gap and missing data allowed. This may not be ideal for alignments with high proportion of missing data.

Haplotype prefix Hap
Select the prefix name for the haplotypes

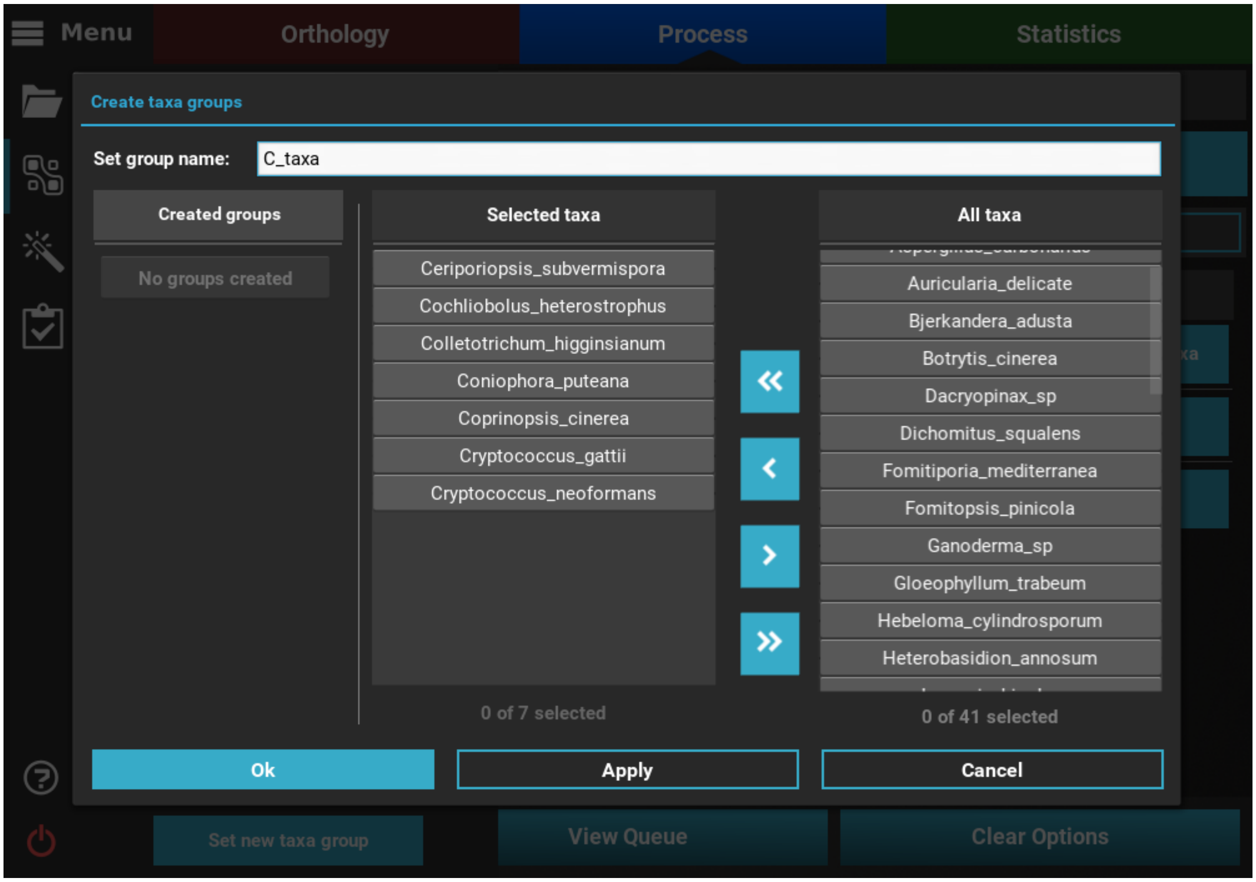
Setting up taxa filter

Here we are interested creating an output data set with alignments that contain any taxon whose name starts with the letter "C".

Click on the *Filter* tab and turn the switch `ON`. Then, click on the `Set filters button` for the *Taxa filter* option and activate the switch in the popup. Change the Filter mode to **Contain** and then click on the `Set taxa group` button to define the new taxa group.



Let's manually create a taxa group with all taxa names that start with the letter "C" by clicking the `Set manually` button.



Set group name: C_taxa

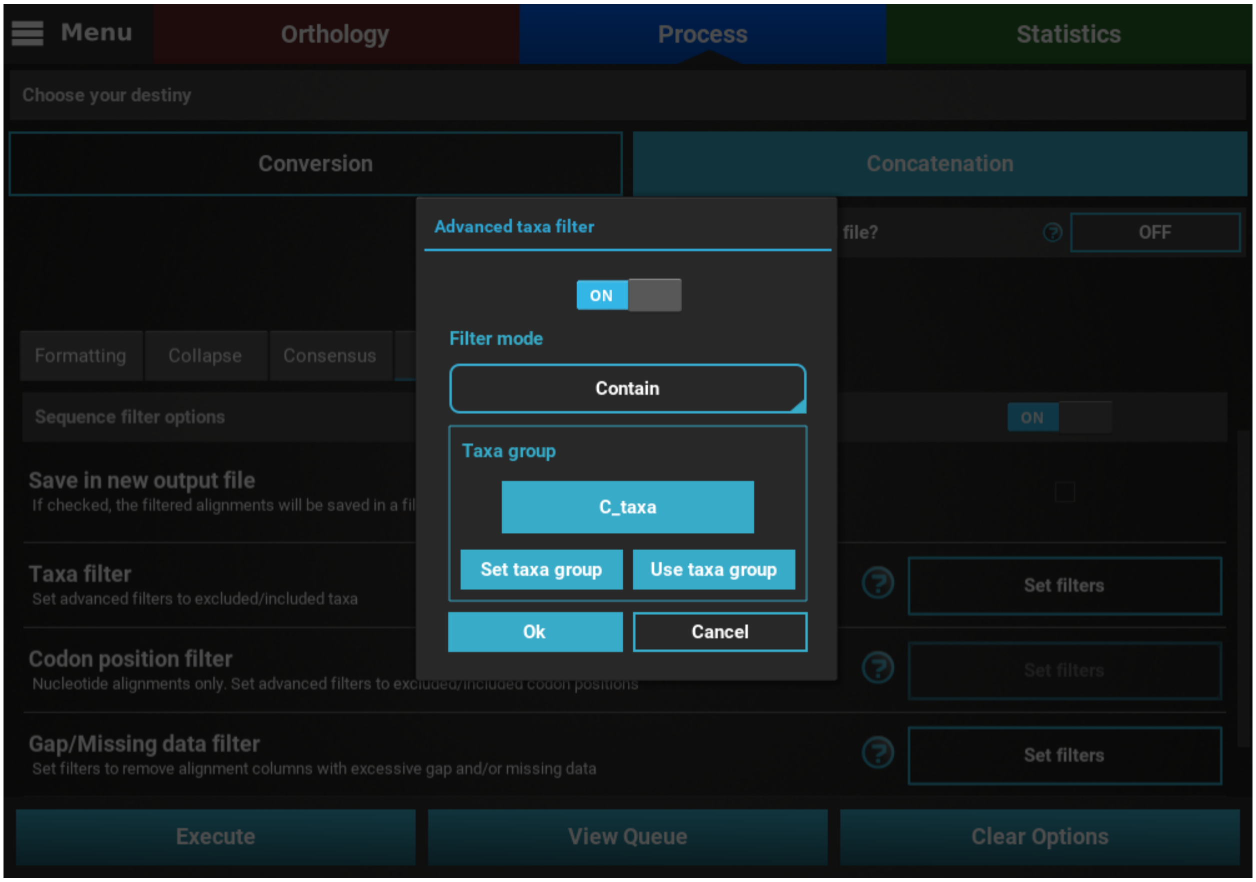
Selected taxa:
- Ceriporiopsis_subvermispora
- Cochliobolus_heterostrophus
- Colletotrichum_higginsianum
- Coniophora_puteana
- Coprinopsis_cinerea
- Cryptococcus_gattii
- Cryptococcus_neoformans

0 of 7 selected

All taxa:
- Auricularia_delicate
- Bjerkandera_adusta
- Botrytis_cinerea
- Dacryopinax_sp
- Dichomitus_squalens
- Fomitiporia_mediterranea
- Fomitopsis_pinicola
- Ganoderma_sp
- Gloeophyllum_trabeum
- Hebeloma_cylindrosporum
- Heterobasidion_annosum

0 of 41 selected

Once the group has been created, check that the this group is correctly selected in the *Taxa filter* dialog.



Advanced taxa filter

ON

Filter mode: Contain

Taxa group: C_taxa

Set taxa group | Use taxa group

Ok | Cancel

Save in new output file
If checked, the filtered alignments will be saved in a fil

Taxa filter
Set advanced filters to excluded/included taxa

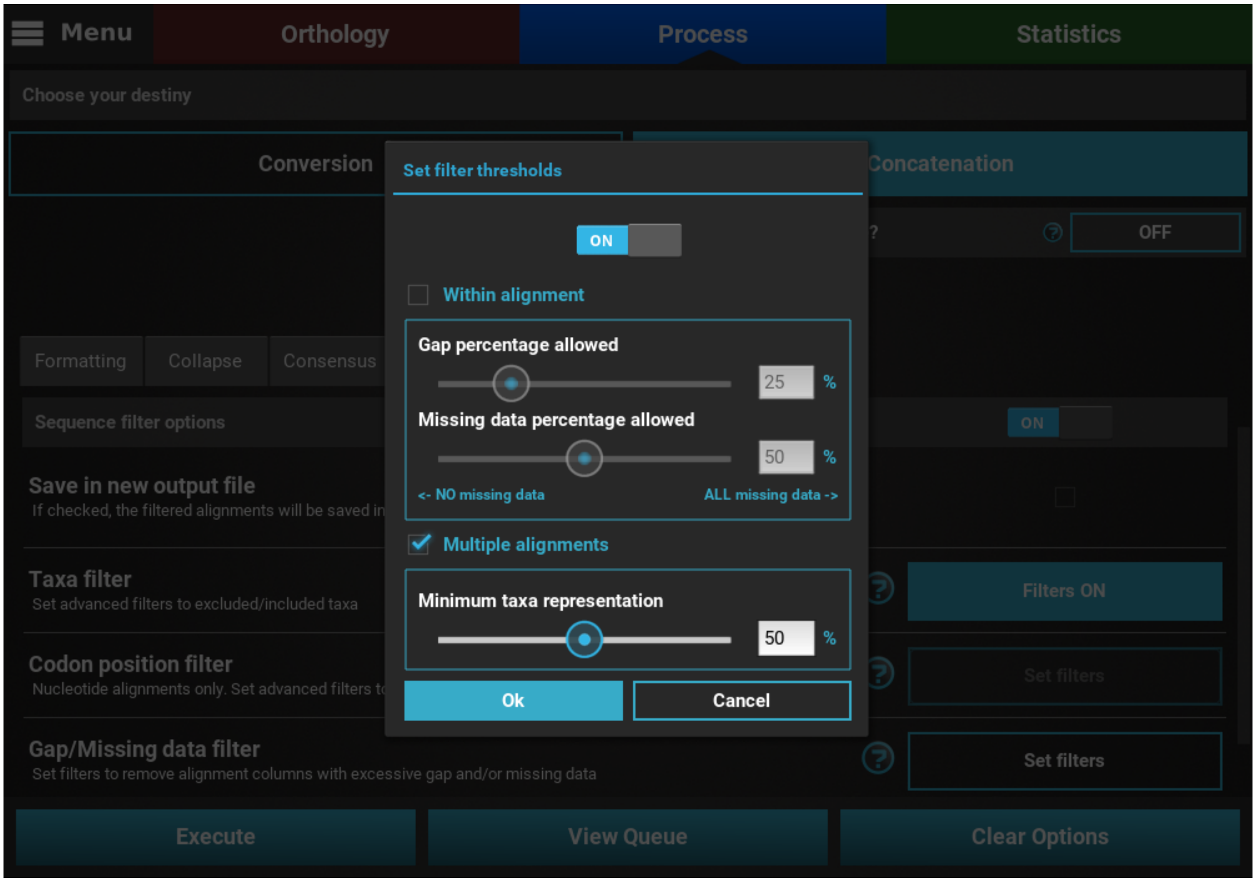
Codon position filter
Nucleotide alignments only. Set advanced filters to excluded/included codon positions

Gap/Missing data filter
Set filters to remove alignment columns with excessive gap and/or missing data

If all checks out, click `Ok` and the button of the *Taxa filter* option should now display *Filters ON*.

Setting up missing data filter

Here we are interested in filtering **ONLY** alignments that contain less than 50% of the total taxa in the data set. Since we are not interested in the within alignment filtering, let’s uncheck this box and set the *Multiple alignments* slider to 50%.



Set filter thresholds

ON

☐ Within alignment

Gap percentage allowed 25 %

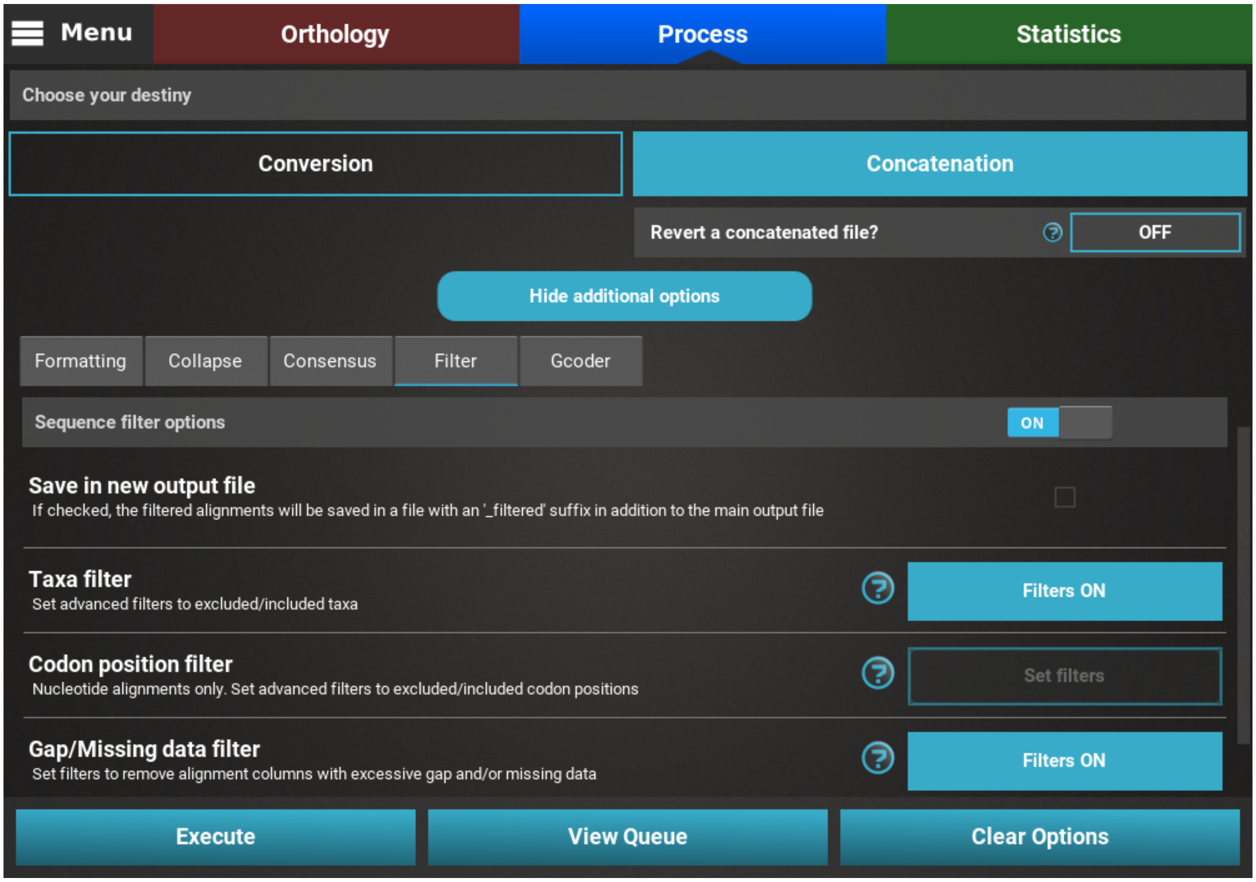
Missing data percentage allowed 50 %

<- NO missing data ALL missing data ->

☑ Multiple alignments

Minimum taxa representation 50 %

Then select the `Ok` button, and both the *Taxa filter* and *Gap/Missing data filter* buttons should now display *Filters ON*.



Revert a concatenated file? OFF

Sequence filter options ON

Save in new output file ☐
If checked, the filtered alignments will be saved in a file with an '_filtered' suffix in addition to the main output file

Taxa filter — Filters ON
Set advanced filters to excluded/included taxa

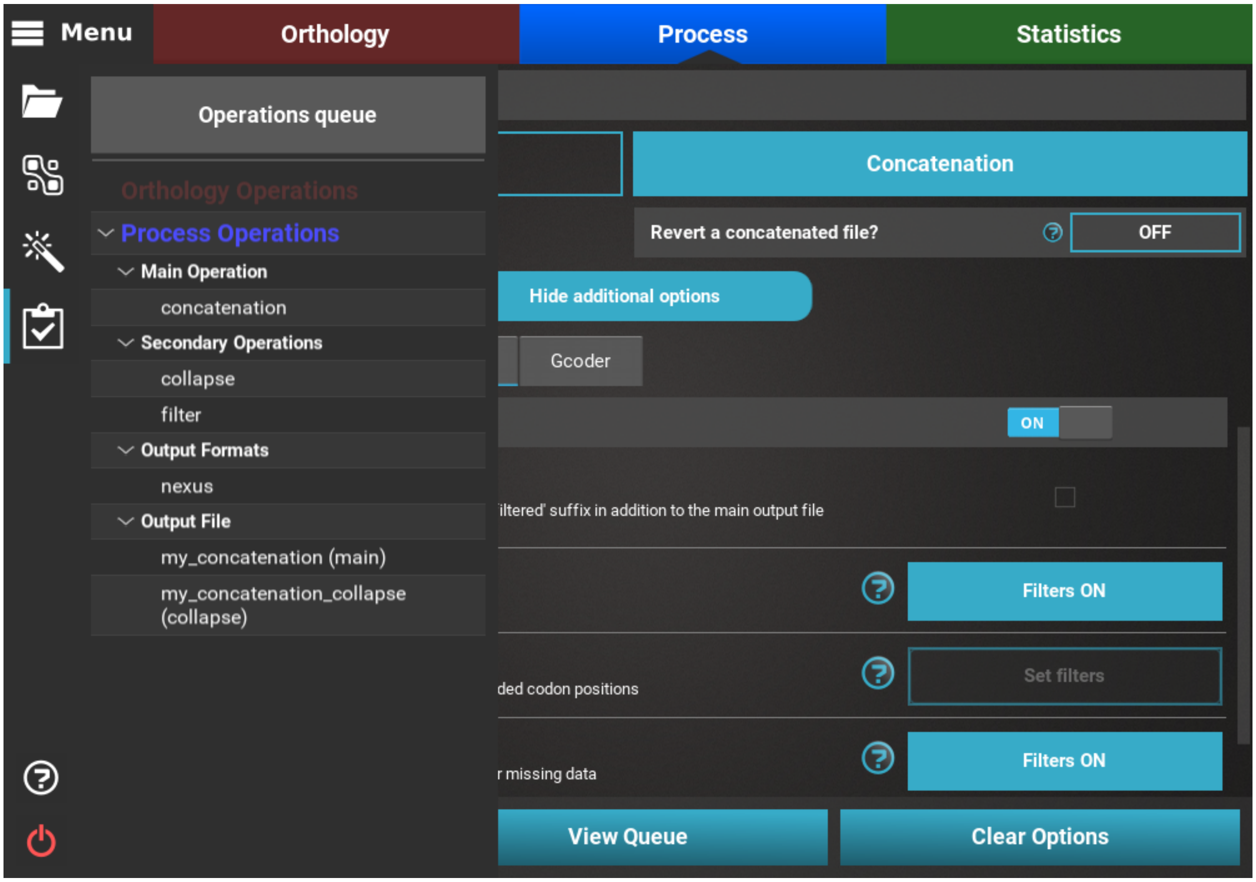
Codon position filter — Set filters
Nucleotide alignments only. Set advanced filters to excluded/included codon positions

Gap/Missing data filter — Filters ON
Set filters to remove alignment columns with excessive gap and/or missing data

Checking selected options

All currently active options can be viewed by clicking the `View Queue` button at the bottom of the **Process** screen. This will open the Menu side panel and show that:

- The main operation is **Concatenation**;
- There are two active secondary operations: **Collapse** and **Filter**;
- The **Nexus** output format is the only selected;
- There are two expected output files: The **main output**, *my_concatenation*, and the **separate output** file that will only contain the result of the concatenation and collapse operations, *my_concatenation_collapse*.



Orthology Operations

Process Operations

- Main Operation
  - concatenation
- Secondary Operations
  - collapse
  - filter
- Output Formats
  - nexus
- Output File
  - my_concatenation (main)
  - my_concatenation_collapse (collapse)

Revert a concatenated file? OFF

ON

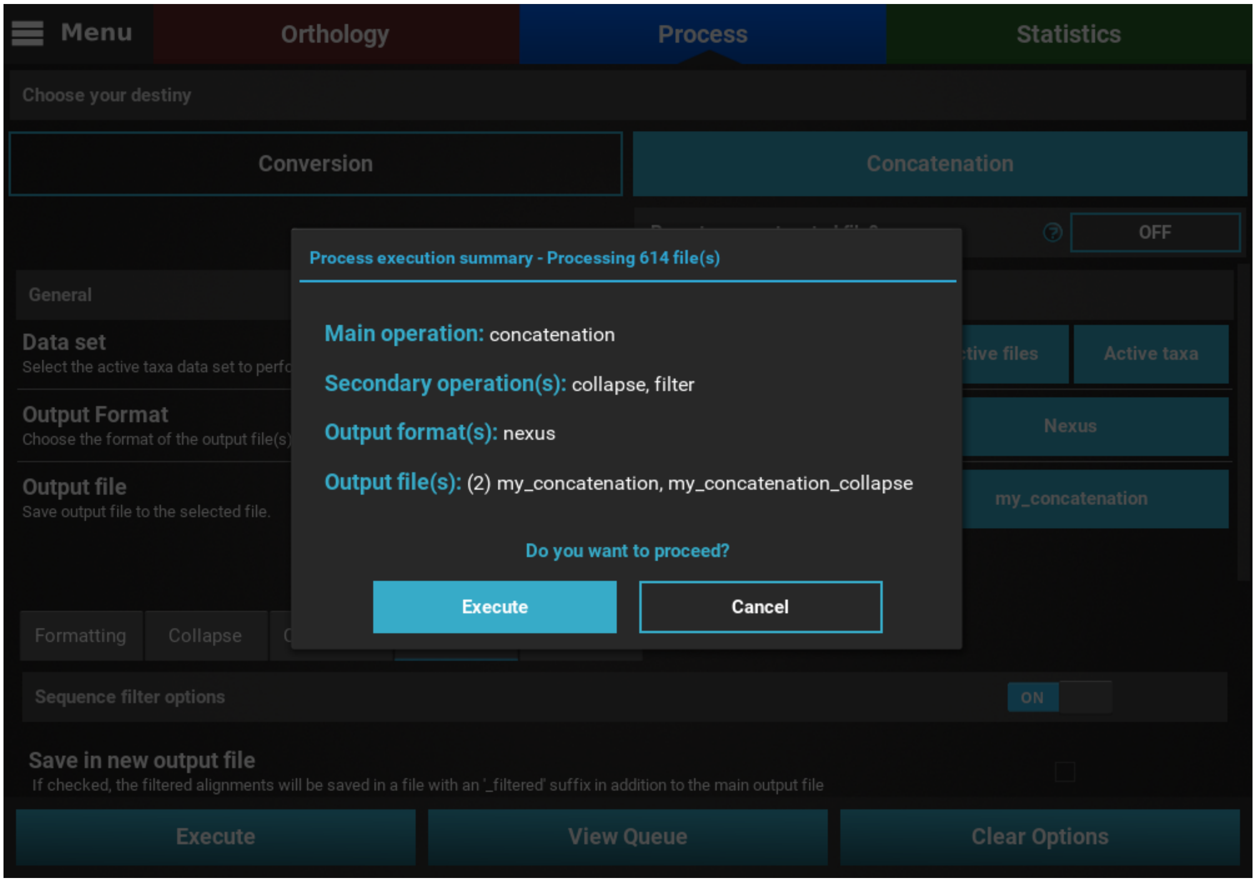
iltered' suffix in addition to the main output file □

ded codon positions

r missing data

Execution

If everything checks out, click the `Execution` button at the bottom of the **Process** screen to show the small popup that displays a summary of the process execution and then click the `Execute` button to begin the execution.



Process execution summary - Processing 614 file(s)

Main operation: concatenation

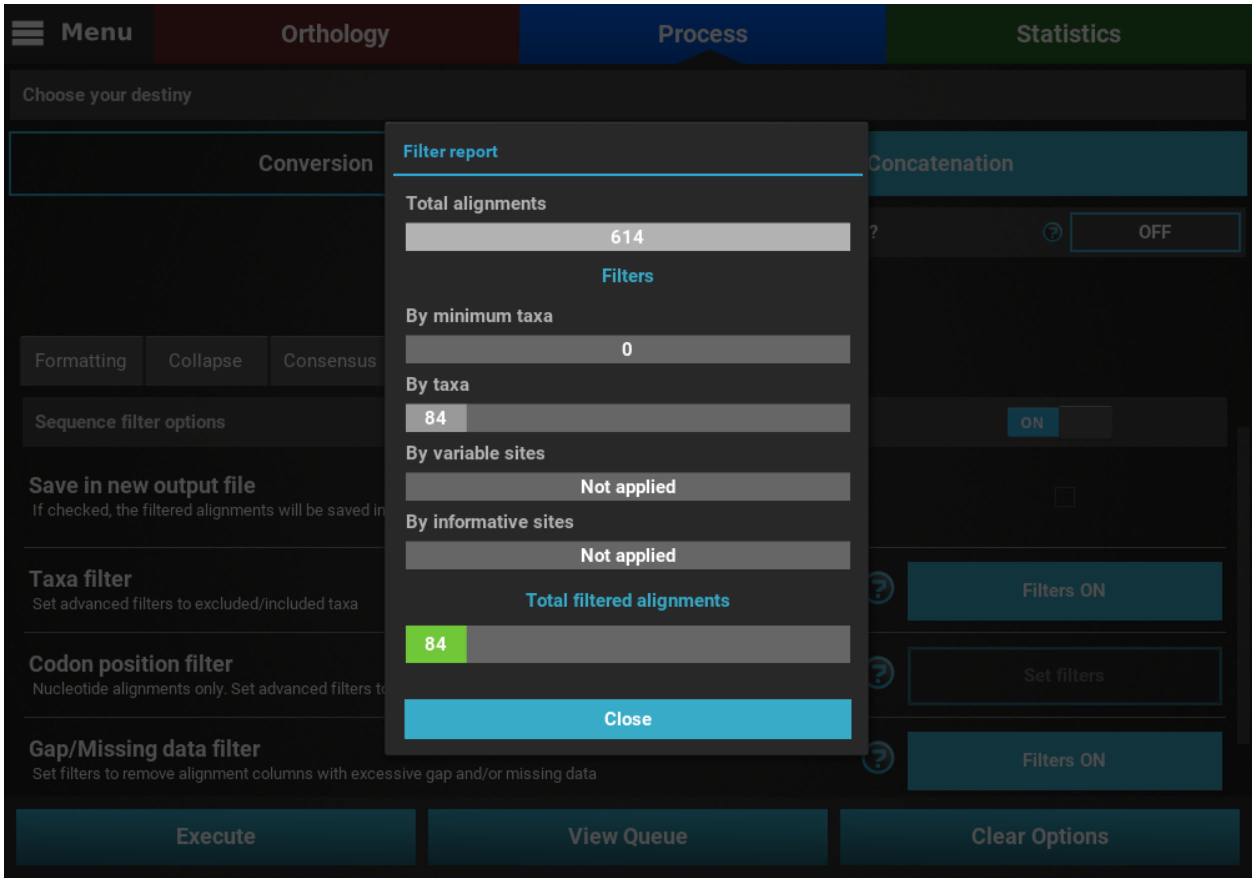
Secondary operation(s): collapse, filter

Output format(s): nexus

Output file(s): (2) my_concatenation, my_concatenation_collapse

Do you want to proceed?

At the end of the execution, a filter report will appear showing the number of alignments that were filtered by the active filters. Since we only activated two of the four filters that can remove alignments from the final output, the values for the other two filters display a *Not applied message*. For the active filters, the number of alignments removed due to that filter is displayed. In this case, no alignment was removed from the Gap/Missing data filter (it seems all alignments already contained more than 50% of the total taxa) and 84 alignments were removed by the Taxa filter.



Filter report

Total alignments
614

Filters

By minimum taxa
0

By taxa
84

By variable sites
Not applied

By informative sites
Not applied

Total filtered alignments
84

1.14 Partitions and substitution models

Note: Data availability for this tutorial: a small concatenated alignment with the corresponding partition files is available here.

TriFusions offers several features to import and handle partitions and substitution models for alignment files. Here I'll describe some of the most common operations.

1.14.1 How to import partitions

From the alignment file

Note: This is only supported for Nexus input files.

Nexus alignment files often have a **charset block** after the alignment matrix where its partitions are described:

```
# NEXUS
Begin data; dimensions ntax=20 nchar=425 ;
format datatype=DNA interleave=no gap=- missing=n ;
```

(continues on next page)

(continued from previous page)

```
matrix

(... alignment matrix...)

;
end;
begin mrbayes;
charset Teste1 = 1-85;
charset Teste2 = 86-170;
charset Teste3 = 171-255;
charset Teste4 = 256-340;
charset Teste5 = 341-425;
partition part = 5: Teste1, Teste2, Teste3, Teste4, Teste5;
set partition=part;
end;
```

In this case, 5 partitions were defined using the charset keywords. When this file is loaded into TriFusion, this block is used to define the partitions in the *Partitions* tab of TriFusion's side panel.

From a partitions file

TriFusion can import partitions schemes formatted in one of two popular formats. Here I'll exemplify how partitions can be imported in either case after loading a concatenated file of 5 alignments into TriFusion, named *concatenated_file.fas*.

Nexus charset block

A Nexus partitions file is a simple text file containing the charset block defining the partitions for an alignment file. In our case, the partition file (named *concatenated_file.nxpart*) would look something like this:

```
# charset [name of partitions] = [partition-range];
charset Teste1.fas = 1-85;
charset Teste2.fas = 86-170;
charset Teste3.fas = 171-255;
charset Teste4.fas = 256-340;
charset Teste5.fas = 341-425;
```

RAxML partition file

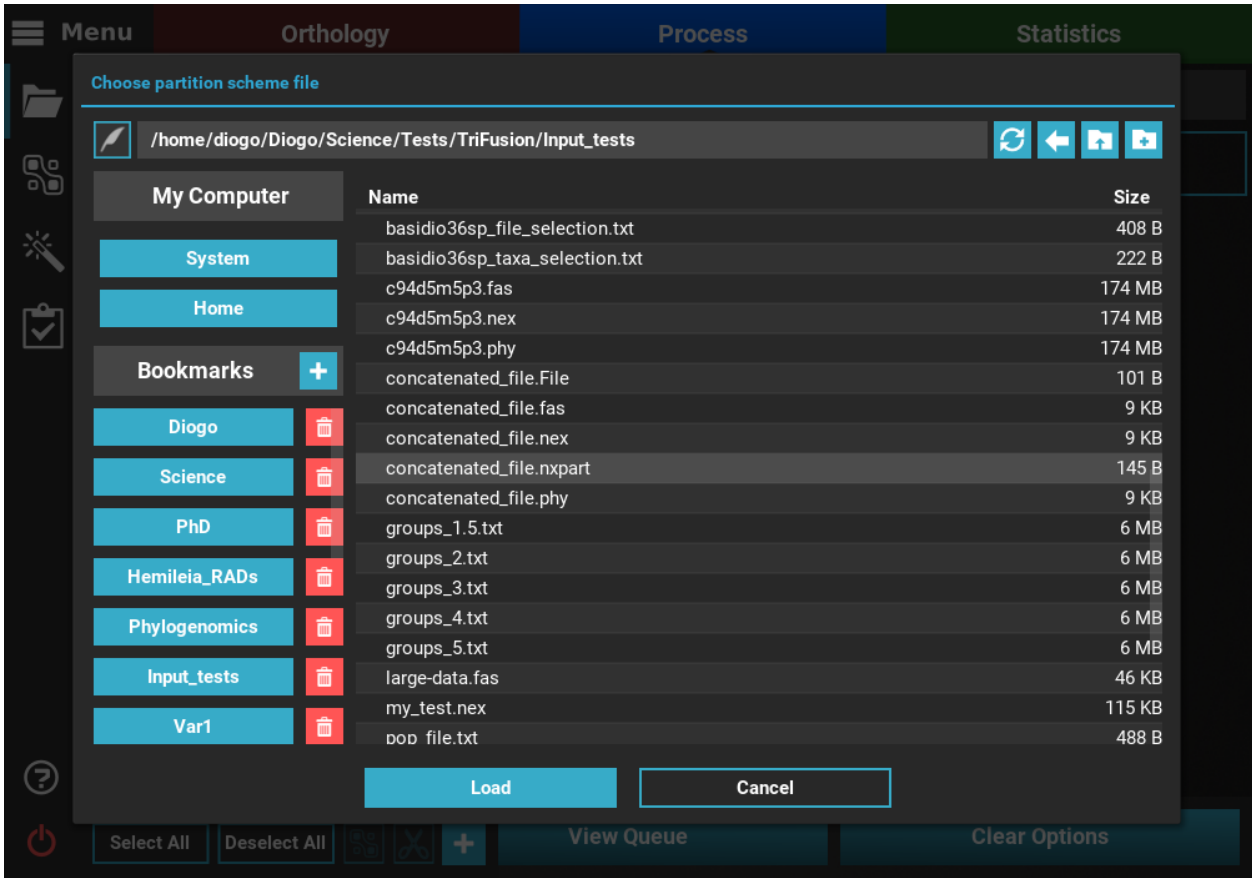
This is the partition file usually required by RAxML for partitioned alignments. Here, partitions are simply defined in each line by providing the substitution model (optional), the name of the partition and then its range. We'll name this file *concatenated_file.partFile*:

```
GTR, BaseConc1.fas = 1-85
GTR, BaseConc2.fas = 86-170
GTR, BaseConc3.fas = 171-255
GTR, BaseConc4.fas = 256-340
GTR, BaseConc5.fas = 341-425
```

Importing the partition file

To import this partition scheme, and assuming that our *concatenated_file.fas* is already loaded into TriFusion, navigate to `Menu > Open/View Data` and click the *Partitions* tab.

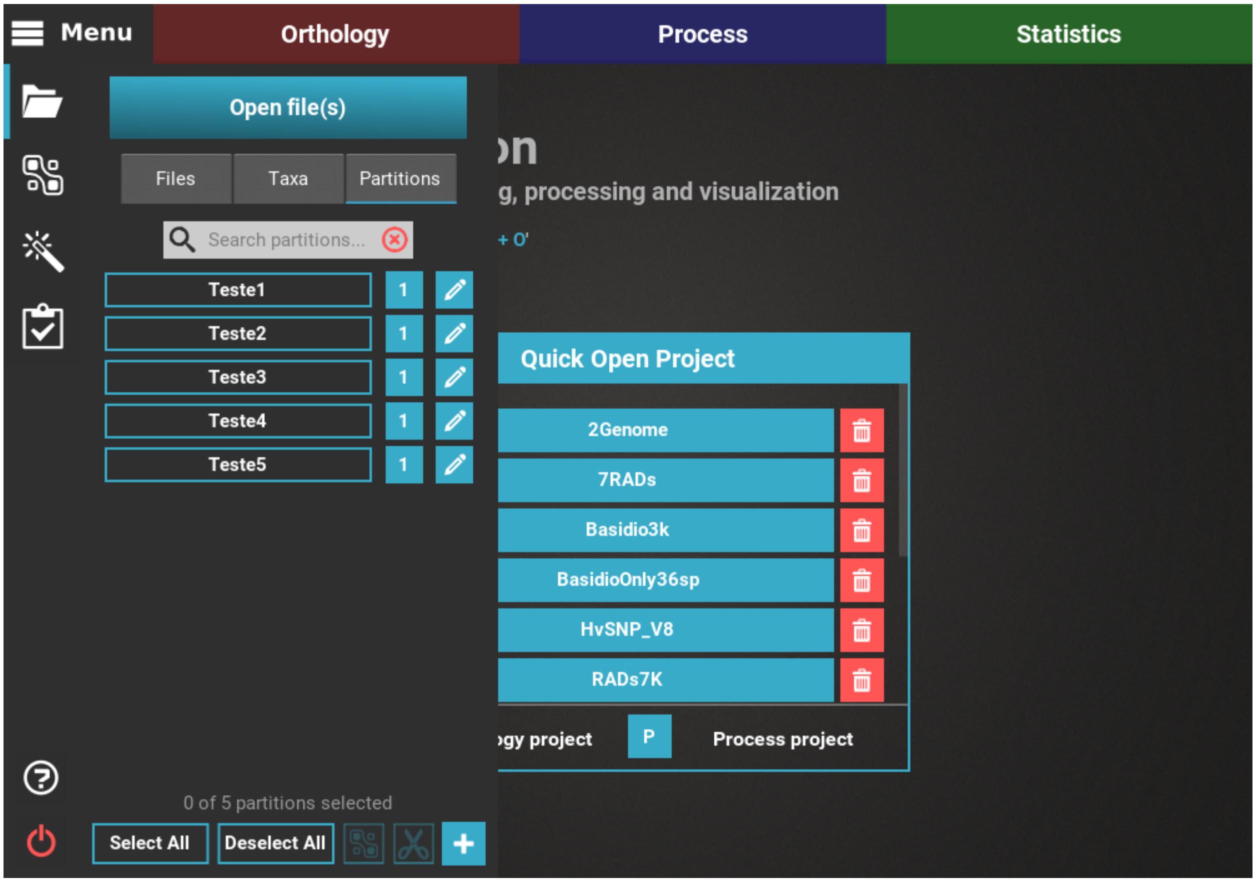
There is already a single partition defined because TriFusion always attributes one partition for each input alignment by default. However, by providing a partition scheme, any previously defined partitions will be discarded. The partition scheme can be provided by clicking the + button at the bottom of the panel and selecting the partition file in the file browser. You can try to import either the Nexus or RAxML partitions file, since the result will be the same.



/home/diogo/Diogo/Science/Tests/TriFusion/Input_tests

Name	Size
basidio36sp_file_selection.txt	408 B
basidio36sp_taxa_selection.txt	222 B
c94d5m5p3.fas	174 MB
c94d5m5p3.nex	174 MB
c94d5m5p3.phy	174 MB
concatenated_file.File	101 B
concatenated_file.fas	9 KB
concatenated_file.nex	9 KB
concatenated_file.nxpart	145 B
concatenated_file.phy	9 KB
groups_1.5.txt	6 MB
groups_2.txt	6 MB
groups_3.txt	6 MB
groups_4.txt	6 MB
groups_5.txt	6 MB
large-data.fas	46 KB
my_test.nex	115 KB
pop_file.txt	488 B

After selecting the partition file, TriFusion will perform several checks to ensure the consistency of the partitions according to the alignment file. If all checks out, the 5 defined partitions will appear in the *Partitions* tab.

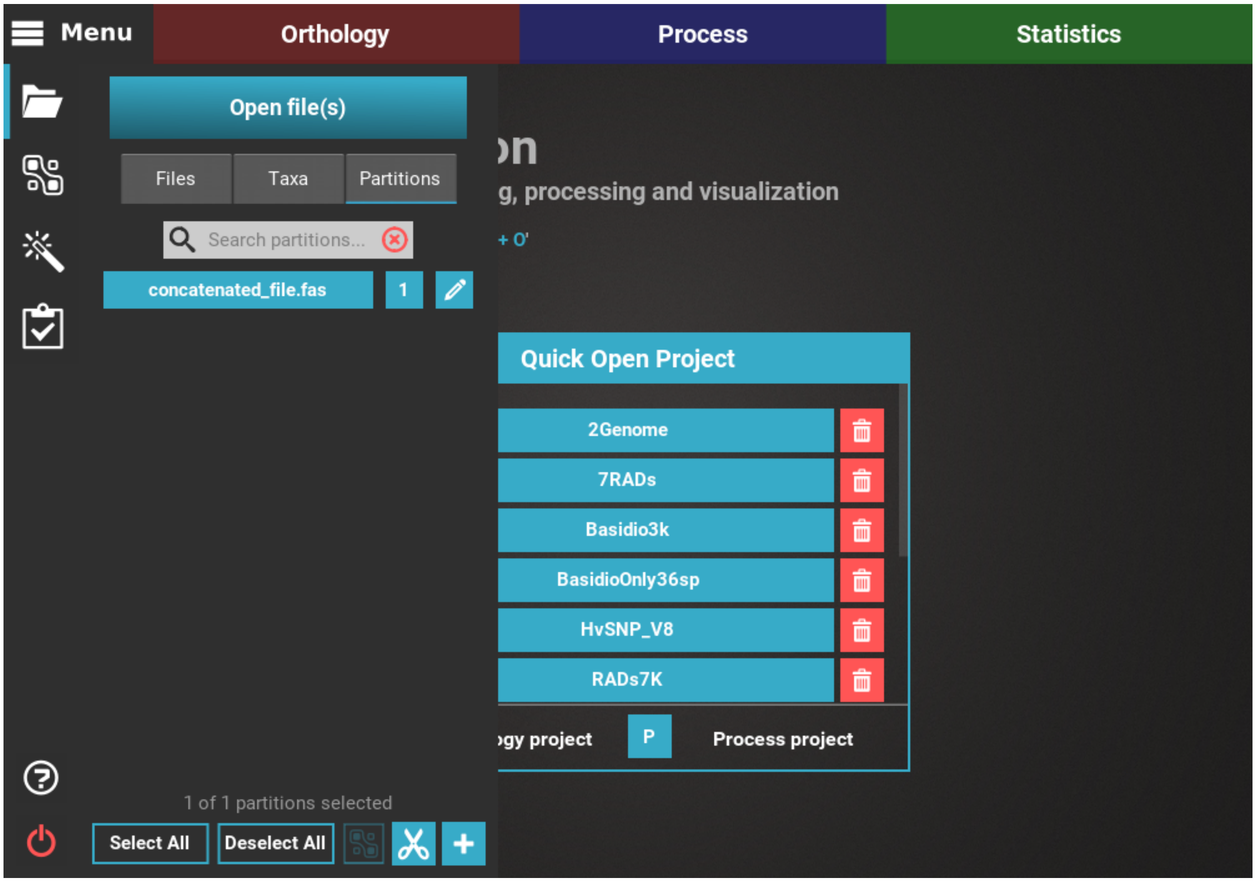


1.14.2 How to create/split partitions

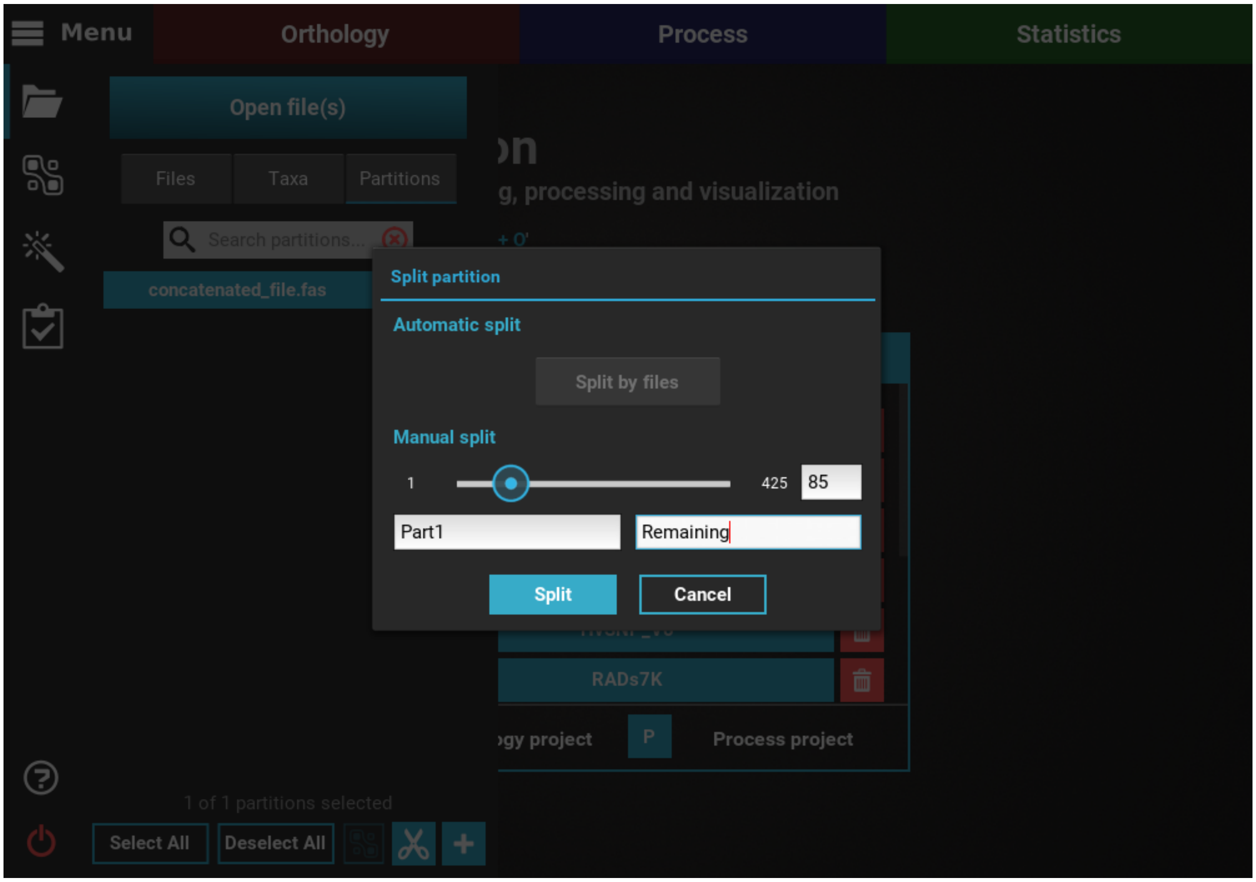
Let's assume we still have the *concatenated_file.fas* without defined partitions loaded into TriFusion. To **create/split** partitions, navigate to `Menu > Open/View Data` and click on the *Partitions* tab.

By default, TriFusion creates a single partition for each input alignment file. This means that when a new partition is created, it is actually split from an existing partition. In this way, we can re-create the 5 partitions that were defined in the sections above. However, as you will see, this taks is more suitable for small punctual modification to the partition scheme than to define partitions from scratch. For larger partitions schemes, using partition files is always easier and more convenient.

To create the first partition, which should have the range from position 1 to 85, select the *concatenated_file.fas* partition button. When you do, the **Scissor** button at the bottom of the panel should become available.



When you click it, a dialog will allow you to split the selected partition into two. You can use the slide or the text input to define the range of the first partition. Let's name this partitions *Part1* and provide a temporary *Remaining* name for the remaining range. Then, click `Split`.



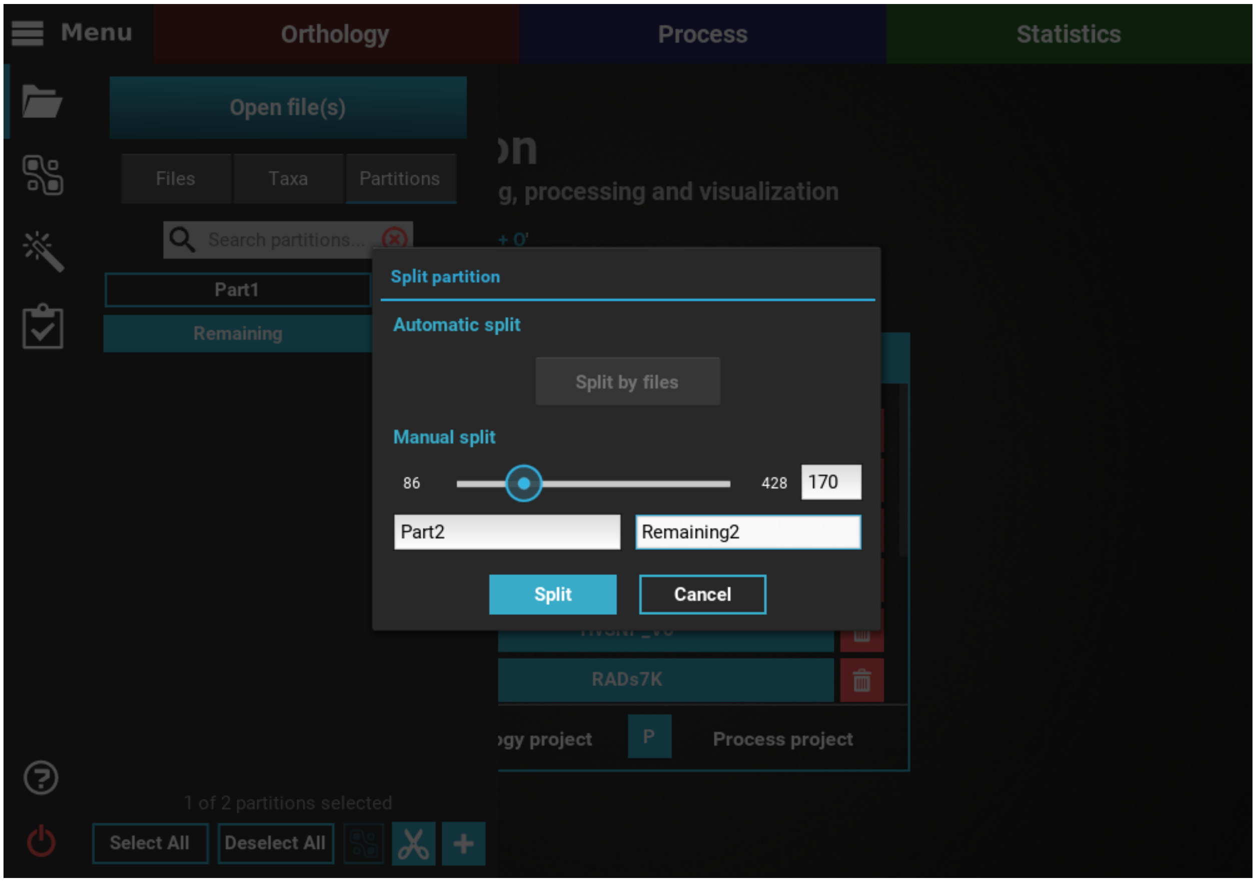
85

Part1

Remaining

As you can see, the new partition *Part1* was created. We can continue this process of creating 85bp partitions, by clicking the *Remaining* partition button, and then the **Scissors** icon to define a new partition.

Now, the *Remaining* partition will start at the 86th bp, so we'll need to add the length of the second partition.

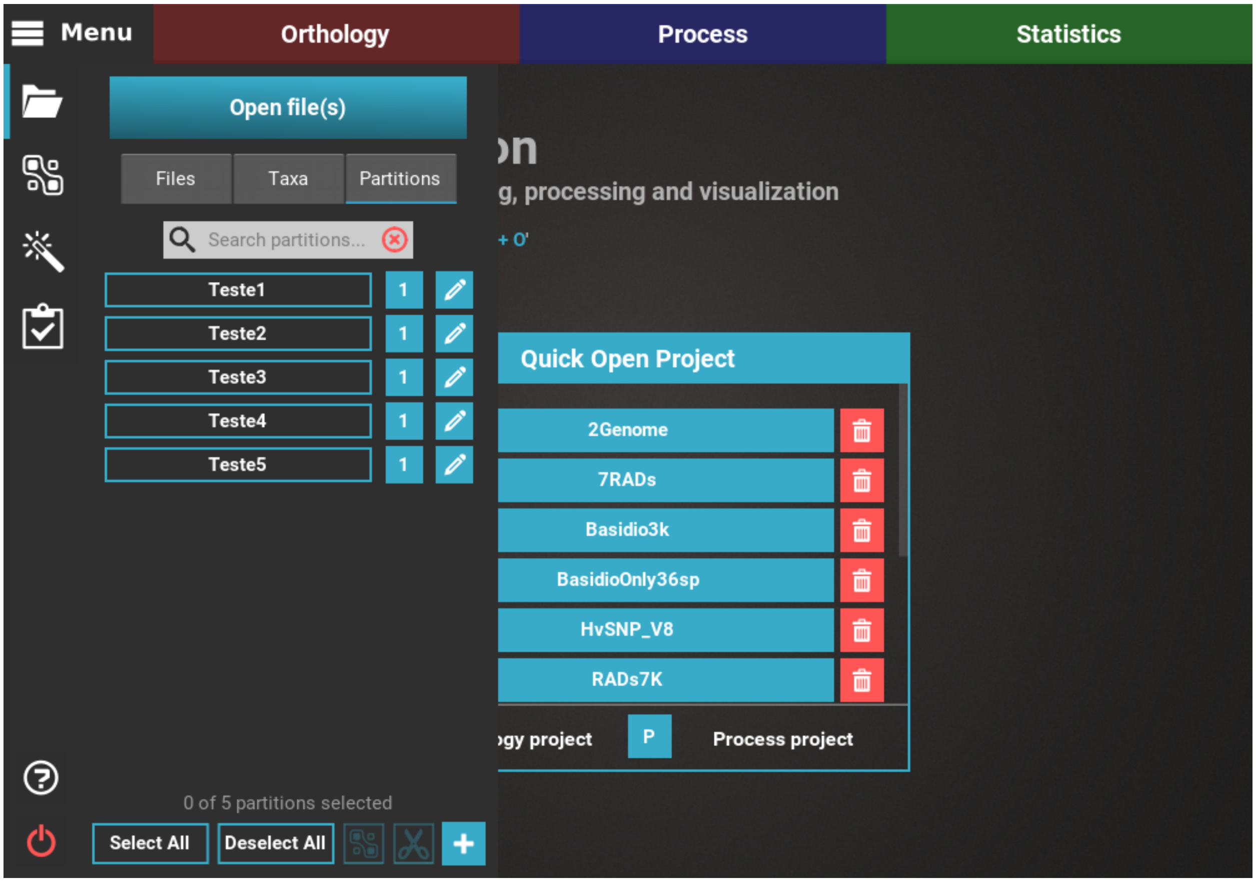


86 428 170

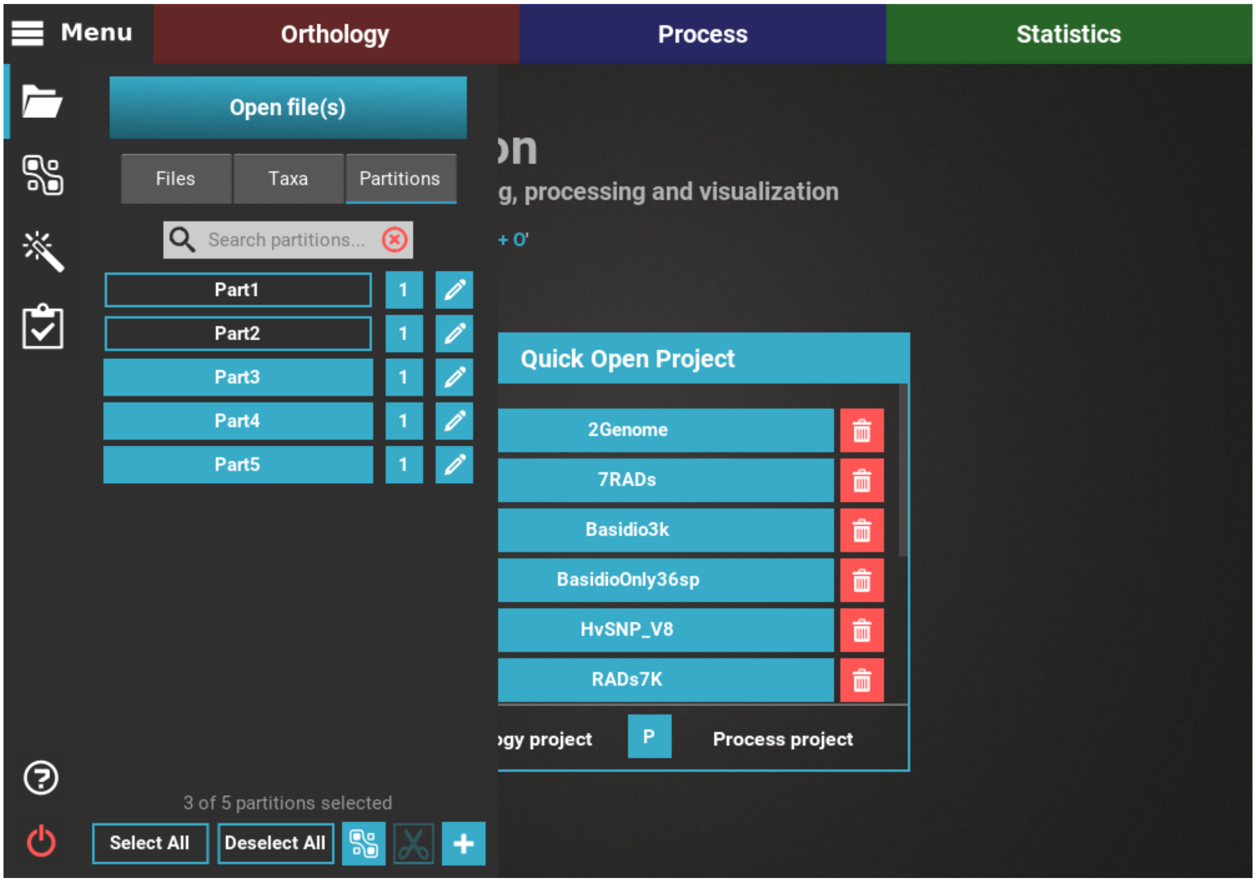
Part2 Remaining2

1.14.3 How to merge pre-existing partitions

Partitions in TriFusion cannot be actually removed, since any part of the alignment must be covered by one partition. However, partitions can be merged to produce a similar effect. For instance, if we load the *concatenated_file.nex* file into TriFusion, it will automatically set 5 partitions for this alignment.



If you want to remove, say, the last two partitions, you can merge them with the last standing partition. Click on the partition buttons *Part3*, *Part4* and *Part5* and the `Merge` button at the end of the panel should become available.



Clicking the Merge button will ask you for the name of the new partition. We'll name it *end_partition*.

This will effectively remove the last two partitions, and append their range to the previouus *Part3* partition. The merge procedure can be combined with the split procedure to fine tune partition ranges.

Ultimately, you can *"remove"* all partitions by merging all partitions in a single one. For this, simply select all partitions and click the Merge button.

Non-contiguous partitions

There is no requirement for partitions to be contiguous before merging. **The only limitation when merging partition is that they must be of the same sequence type (nucleotide or protein)**.

If we want, we could merge the first and last partitions in a new partition named *extremes*.

By merging non-contiguous partitions together, TriFusion will automatically merge the sequence data into continuous segments and the remaining partition ranges. Therefore, if you perform a **Concatenation** into a Nexus output format, you'll see that the sequence data from the last alignment will now appear merged with the sequence from the first alignment. Indeed, the order of the new merged partition is based on the starting position of the first selected partition.

As an example, the result of the concatenated nexus file of this merger will be:

```
begin mrbayes;
    charset extremes = 1-170;
```

(continues on next page)

(continued from previous page)

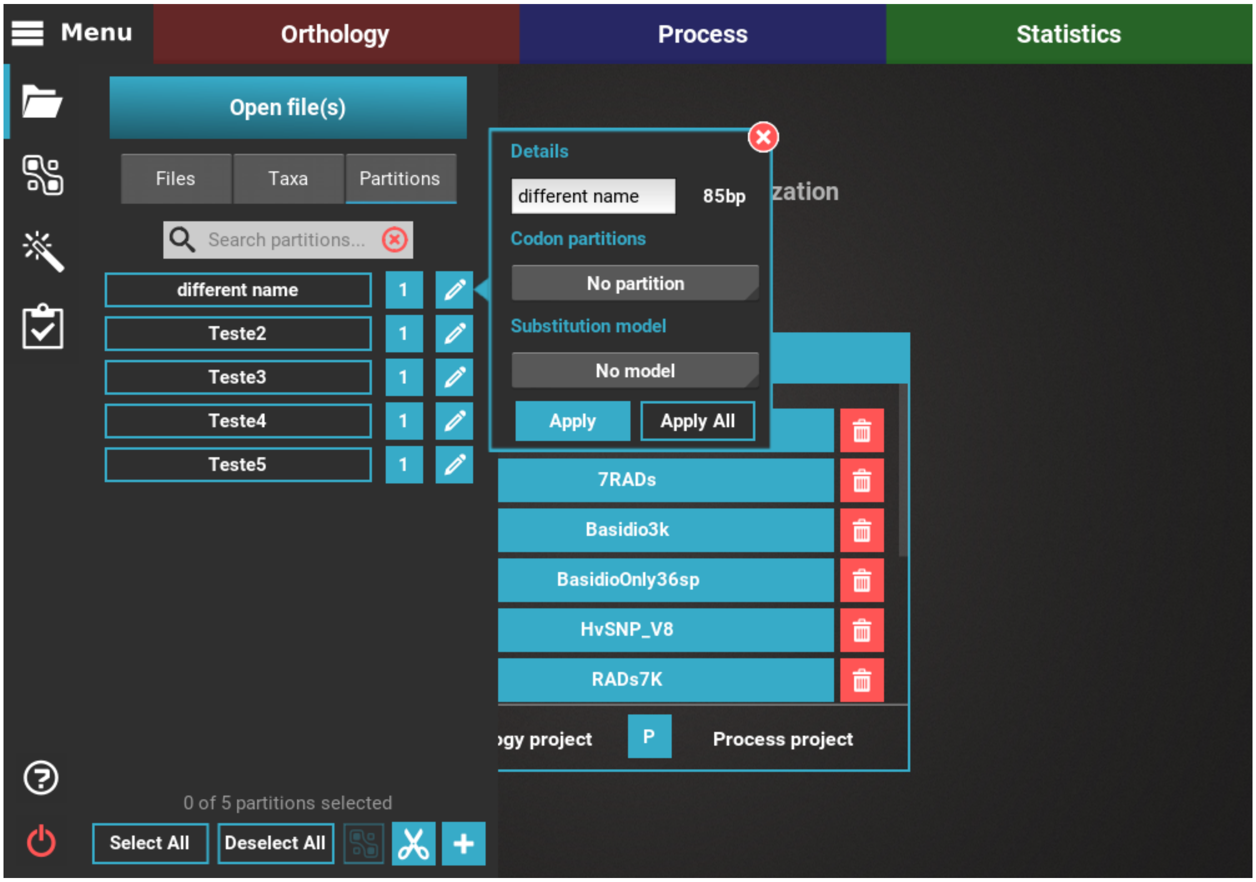
```
    charset Teste2 = 171-255;
    charset Teste3 = 256-340;
    charset Teste4 = 341-425;
    partition part = 4: extremes, Teste2, Teste3, Teste4;
    set partition=part;
end;
```

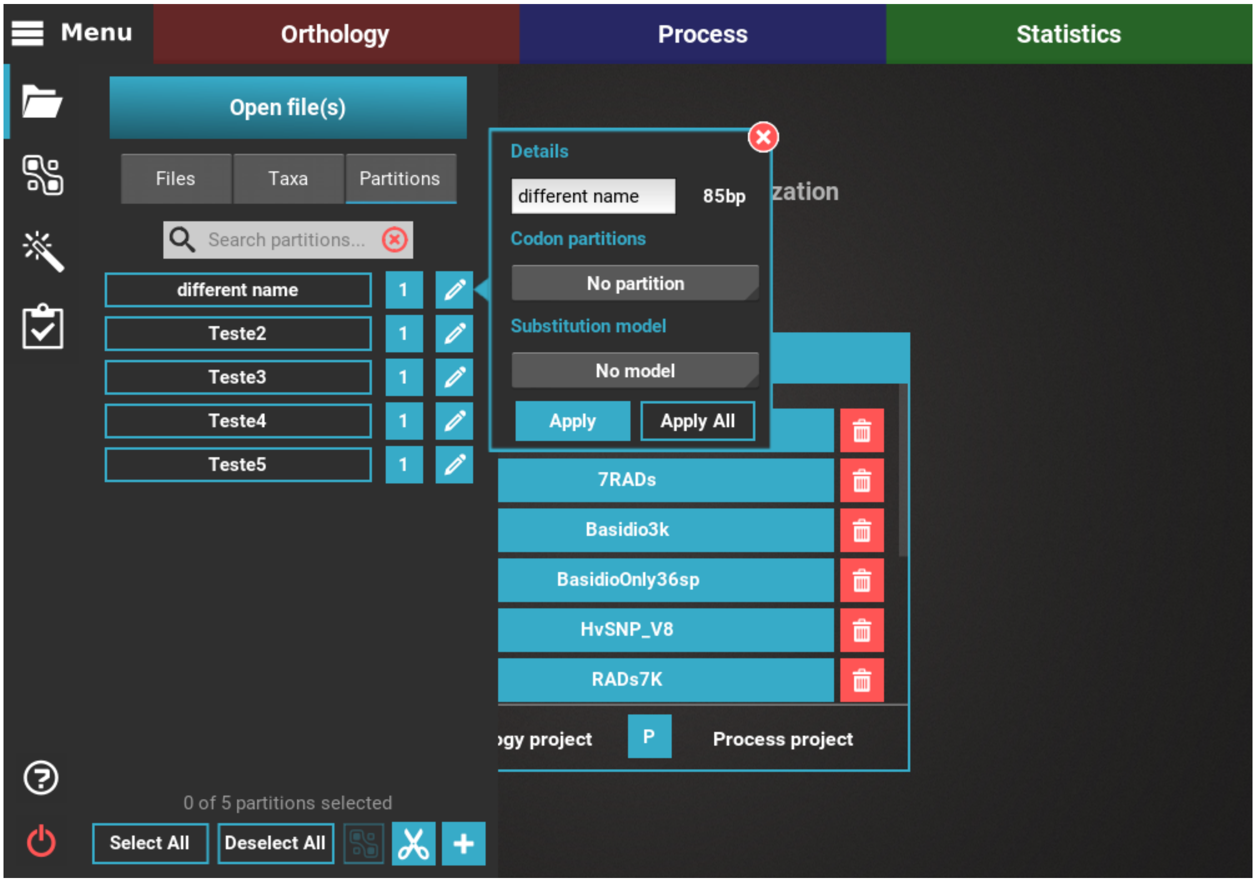
1.14.4 Change the partition's name

Partition names can be easily changed in TriFusion. Navigate to `Menu > Open/View Data` and click on the *Partitions* tab.

To change the name of one partition, say *Test1*, click on the corresponding **Pencil** button. The current name should appear in a text field under the **Details** section.

Then, modify the name no your liking and press `Enter` to change it.





Details

different name 85bp

Codon partitions

No partition

Substitution model

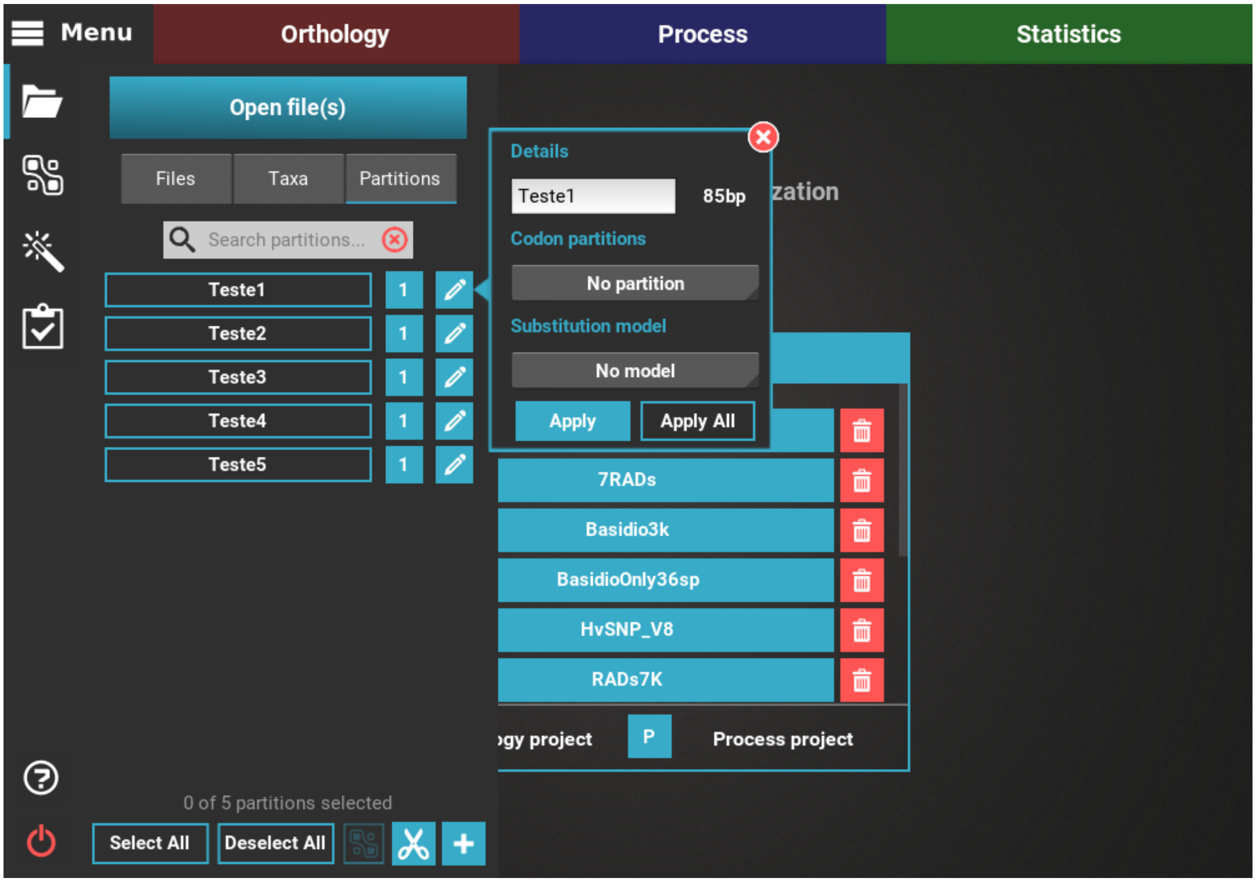
No model

1.14.5 Edit the substitution model

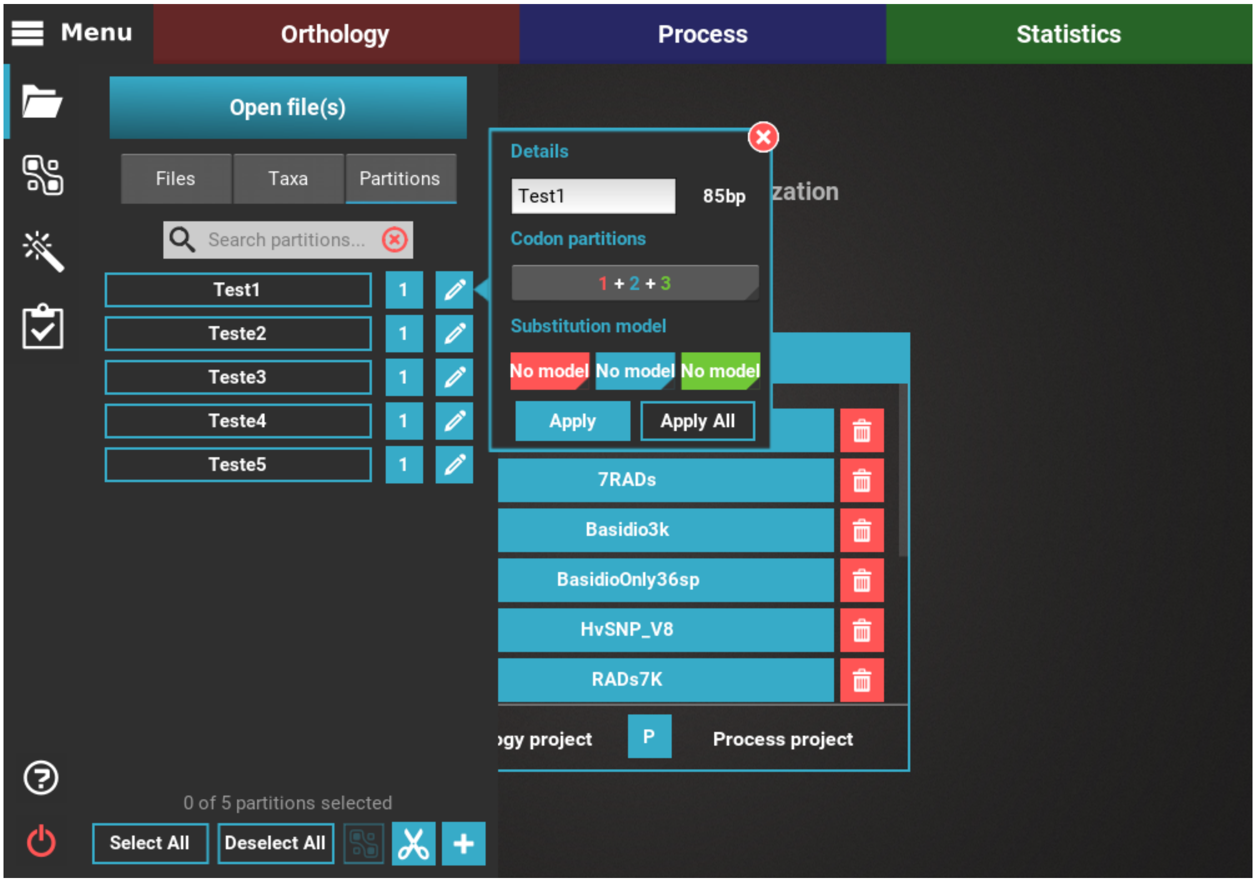
TriFusion supports the specification of substitution models and codon partitions. However, note that this information is can only be included in Nexus output formats or in the RAxML partition file that is generated for the Phylip output format.

To set/change the substitution model and/or codon partitions of a partition, navigate to `Menu > Open/View Data` and click on the *Partitions* tab.

Then, click on the **Pencil** button of any partition to open the edition dialog.



You can choose a codon partition scheme using the drop down menu under the **Codon partitions** section. All possible codon partition schemes are listed, included the option to have no sub-partitions. In this example, lets create separate partitions for each codon position by selecting the **1 + 2 + 3** value.



Details

Test1 85bp

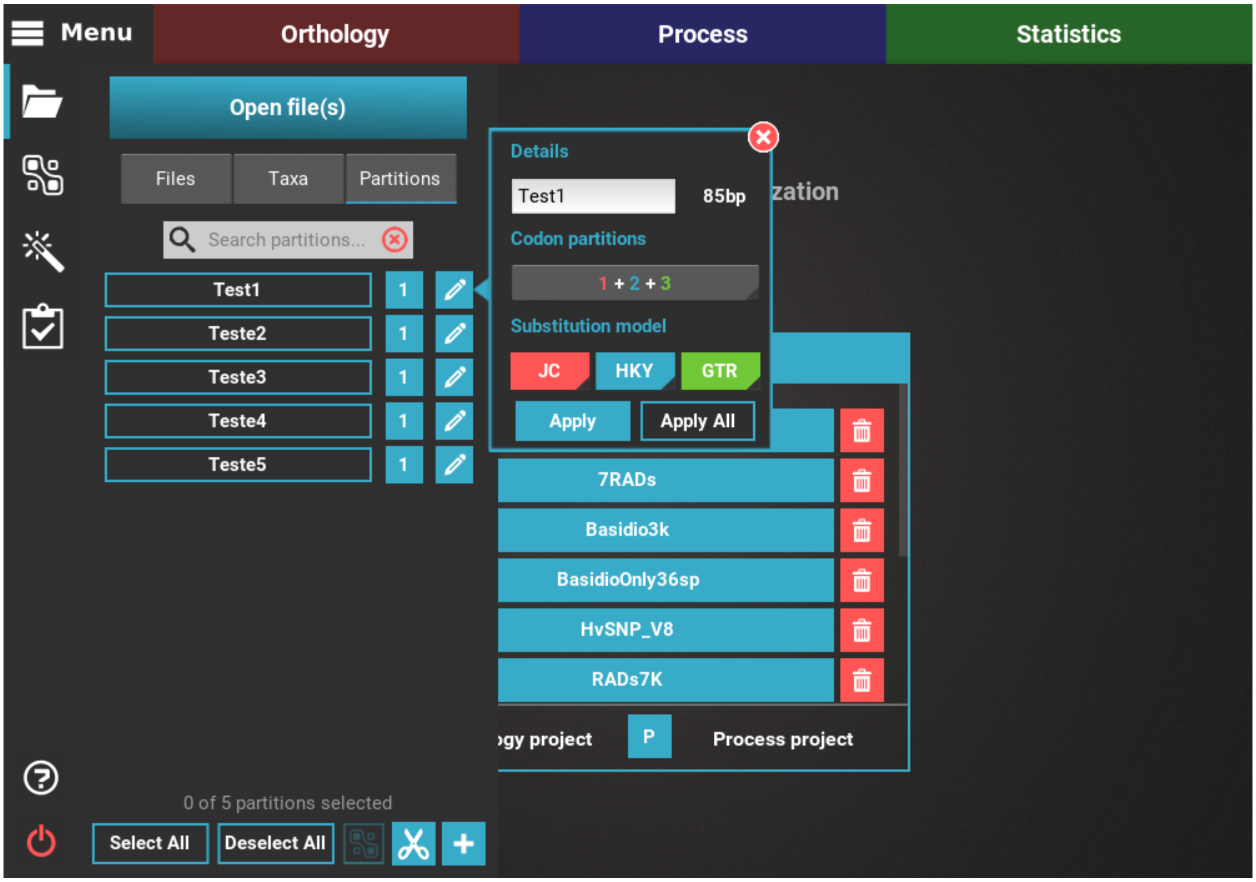
Codon partitions

1 + 2 + 3

Substitution model

No model No model No model

Then, you can choose the appropriate model for each partition, following the color code. For example, we want to set **JC** for the first codon (red), **HKY** for the second codon (blue) and **GTR** for the third codon (green).



If you want to make the change only for the current partition, click the `Apply` button. If you want to make this change for **all partitions**, click the `Apply All` button.

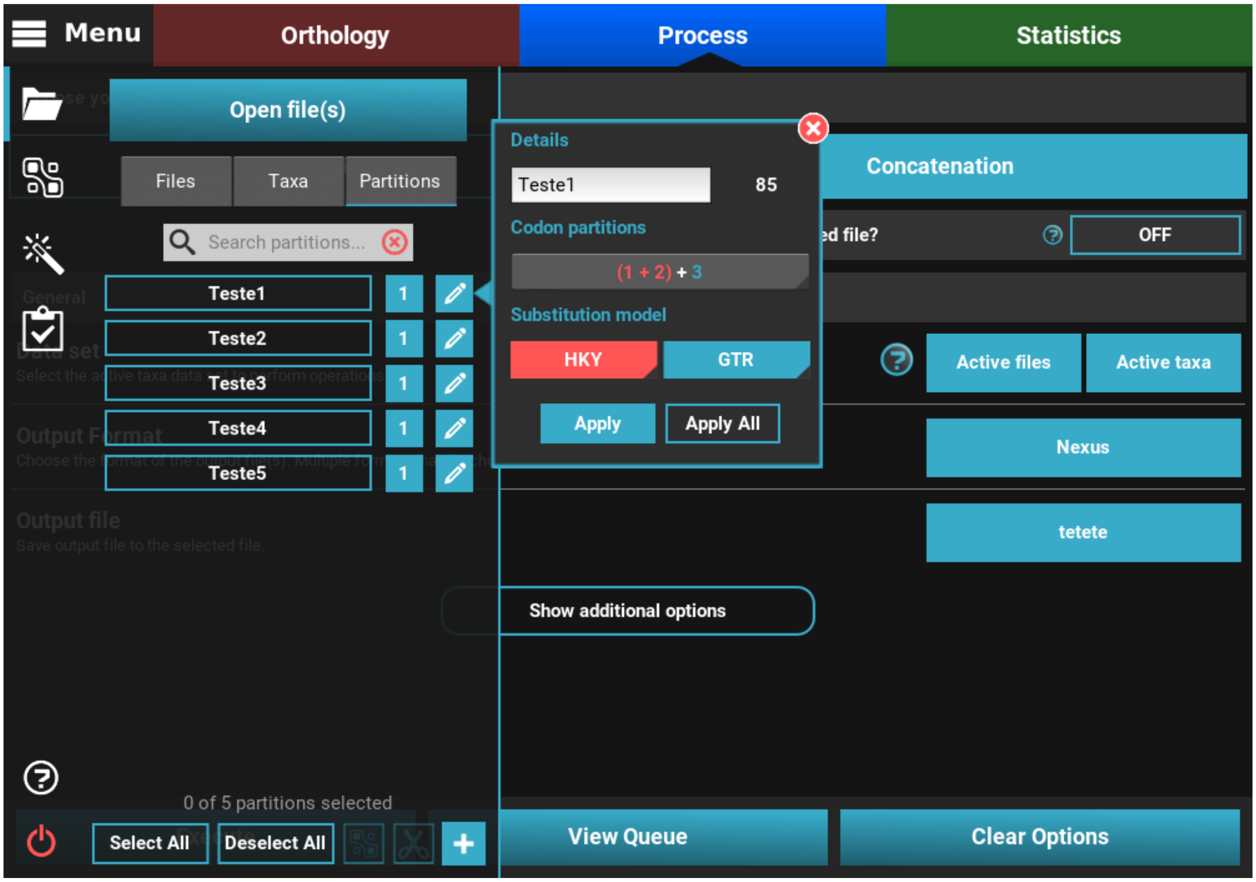
If we apply this codon partition and substitution models to all partitions, the final result in a concatenated Nexus file will have the partitions defined using the notation for codon partitions:

```
begin mrbayes;
    charset Teste1_1 = 1-85\3;
    charset Teste1_2 = 2-85\3;
    charset Teste1_3 = 3-85\3;
    charset Teste2_86 = 86-170\3;
    charset Teste2_87 = 87-170\3;
    charset Teste2_88 = 88-170\3;
    charset Teste3_171 = 171-255\3;
    charset Teste3_172 = 172-255\3;
    charset Teste3_173 = 173-255\3;
    charset Teste4_256 = 256-340\3;
    charset Teste4_257 = 257-340\3;
    charset Teste4_258 = 258-340\3;
    charset Teste5_341 = 341-425\3;
    charset Teste5_342 = 342-425\3;
    charset Teste5_343 = 343-425\3;
    partition part = 15: Teste1_1, Teste1_2, Teste1_3, Teste2_86, Teste2_87, Teste2_
↪88, Teste3_171, Teste3_172, Teste3_173, Teste4_256, Teste4_257, Teste4_258, Teste5_
↪341, Teste5_342, Teste5_343;
    set partition=part;
end;
```

Below the partitions block, the substitution models were also specified for each partition:

```
begin mrbayes;
lset applyto=(1) nst=1;
prset applyto=(1) statefreqpr=fixed(equal);
lset applyto=(2) nst=2;
prset applyto=(2) statefreqpr=dirichlet(1,1,1,1);
lset applyto=(3) nst=6;
prset applyto=(3) statefreqpr=dirichlet(1,1,1,1);
lset applyto=(4) nst=1;
prset applyto=(4) statefreqpr=fixed(equal);
lset applyto=(5) nst=2;
prset applyto=(5) statefreqpr=dirichlet(1,1,1,1);
lset applyto=(6) nst=6;
prset applyto=(6) statefreqpr=dirichlet(1,1,1,1);
lset applyto=(7) nst=1;
prset applyto=(7) statefreqpr=fixed(equal);
lset applyto=(8) nst=2;
prset applyto=(8) statefreqpr=dirichlet(1,1,1,1);
lset applyto=(9) nst=6;
prset applyto=(9) statefreqpr=dirichlet(1,1,1,1);
lset applyto=(10) nst=1;
prset applyto=(10) statefreqpr=fixed(equal);
lset applyto=(11) nst=2;
prset applyto=(11) statefreqpr=dirichlet(1,1,1,1);
lset applyto=(12) nst=6;
prset applyto=(12) statefreqpr=dirichlet(1,1,1,1);
lset applyto=(13) nst=1;
prset applyto=(13) statefreqpr=fixed(equal);
lset applyto=(14) nst=2;
prset applyto=(14) statefreqpr=dirichlet(1,1,1,1);
lset applyto=(15) nst=6;
prset applyto=(15) statefreqpr=dirichlet(1,1,1,1);
unlink statefreq=(all) revmat=(all) shape=(all) pinvar=(all) tratio=(all);
end;
```

Note that all codon partitions have unlinked models. However, you can also link codon models in TriFusion. For instance, we could choose the codon partition option of **(1 + 2) + 3** to link the same substitution model of the first two codons and keep a different one for the third codon. Let's set the **HKY** model for the first two codons and the GTR for the third.



If we repeat the concatenation to a Nexus output file, you can see that the while the partition block is the same, the definition of the substitution models has changed:

```
begin mrbayes;
lset applyto=(1) nst=2;
prset applyto=(1) statefreqpr=dirichlet(1,1,1,1);
lset applyto=(2) nst=2;
prset applyto=(2) statefreqpr=dirichlet(1,1,1,1);
lset applyto=(3) nst=6;
prset applyto=(3) statefreqpr=dirichlet(1,1,1,1);
lset applyto=(4) nst=2;
prset applyto=(4) statefreqpr=dirichlet(1,1,1,1);
lset applyto=(5) nst=2;
prset applyto=(5) statefreqpr=dirichlet(1,1,1,1);
lset applyto=(6) nst=6;
prset applyto=(6) statefreqpr=dirichlet(1,1,1,1);
lset applyto=(7) nst=2;
prset applyto=(7) statefreqpr=dirichlet(1,1,1,1);
lset applyto=(8) nst=2;
prset applyto=(8) statefreqpr=dirichlet(1,1,1,1);
lset applyto=(9) nst=6;
prset applyto=(9) statefreqpr=dirichlet(1,1,1,1);
lset applyto=(10) nst=2;
prset applyto=(10) statefreqpr=dirichlet(1,1,1,1);
lset applyto=(11) nst=2;
prset applyto=(11) statefreqpr=dirichlet(1,1,1,1);
lset applyto=(12) nst=6;
```

(continues on next page)

(continued from previous page)

```
prset applyto=(12) statefreqpr=dirichlet(1,1,1,1);
lset applyto=(13) nst=2;
prset applyto=(13) statefreqpr=dirichlet(1,1,1,1);
lset applyto=(14) nst=2;
prset applyto=(14) statefreqpr=dirichlet(1,1,1,1);
lset applyto=(15) nst=6;
prset applyto=(15) statefreqpr=dirichlet(1,1,1,1);
unlink statefreq=(all) revmat=(all) shape=(all) pinvar=(all) tratio=(all);
link statefreq=(1,2) revmat=(1,2) shape=(1,2) pinvar=(1,2) tratio=(1,2);
link statefreq=(4,5) revmat=(4,5) shape=(4,5) pinvar=(4,5) tratio=(4,5);
link statefreq=(7,8) revmat=(7,8) shape=(7,8) pinvar=(7,8) tratio=(7,8);
link statefreq=(10,11) revmat=(10,11) shape=(10,11) pinvar=(10,11) tratio=(10,11);
link statefreq=(13,14) revmat=(13,14) shape=(13,14) pinvar=(13,14) tratio=(13,14);
end;
```

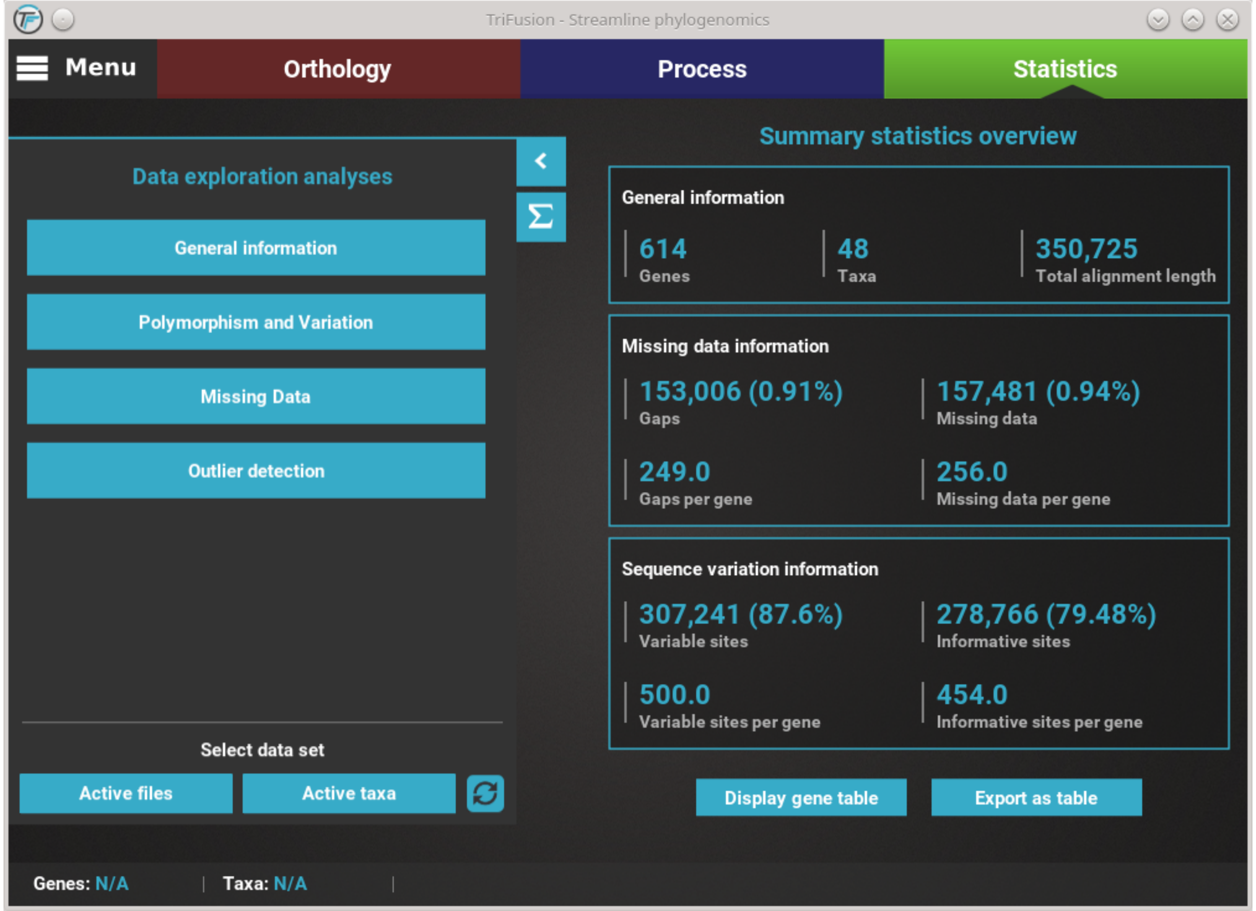
At the end of this block, the substitution parameters for all first and second codons were linked.

1.15 Summary statistics

Note: Data availability for this tutorial: the medium sized data set of 614 genes and 48 taxa that will be used can be downloaded here.

1.15.1 Summary statistics overview

As soon as you load your data into TriFusion and navigate to the **Statistics** module, the computation of general and gene specific summary statistics will start. This computation is being done in the background, and unless you start to generate a plot or load more data into TriFusion, it will continue to do so. When finished, a **summary statistic overview** for the currently active data set will be displayed in the **Statistics** screen.



Summary statistics overview

General information

614 Genes | 48 Taxa | 350,725 Total alignment length

Missing data information

153,006 (0.91%) Gaps | 157,481 (0.94%) Missing data

249.0 Gaps per gene | 256.0 Missing data per gene

Sequence variation information

307,241 (87.6%) Variable sites | 278,766 (79.48%) Informative sites

500.0 Variable sites per gene | 454.0 Informative sites per gene

Genes: N/A | Taxa: N/A |

Information is sorted in three main cateagories: **General**, **Missing data** and **Sequence variation**.

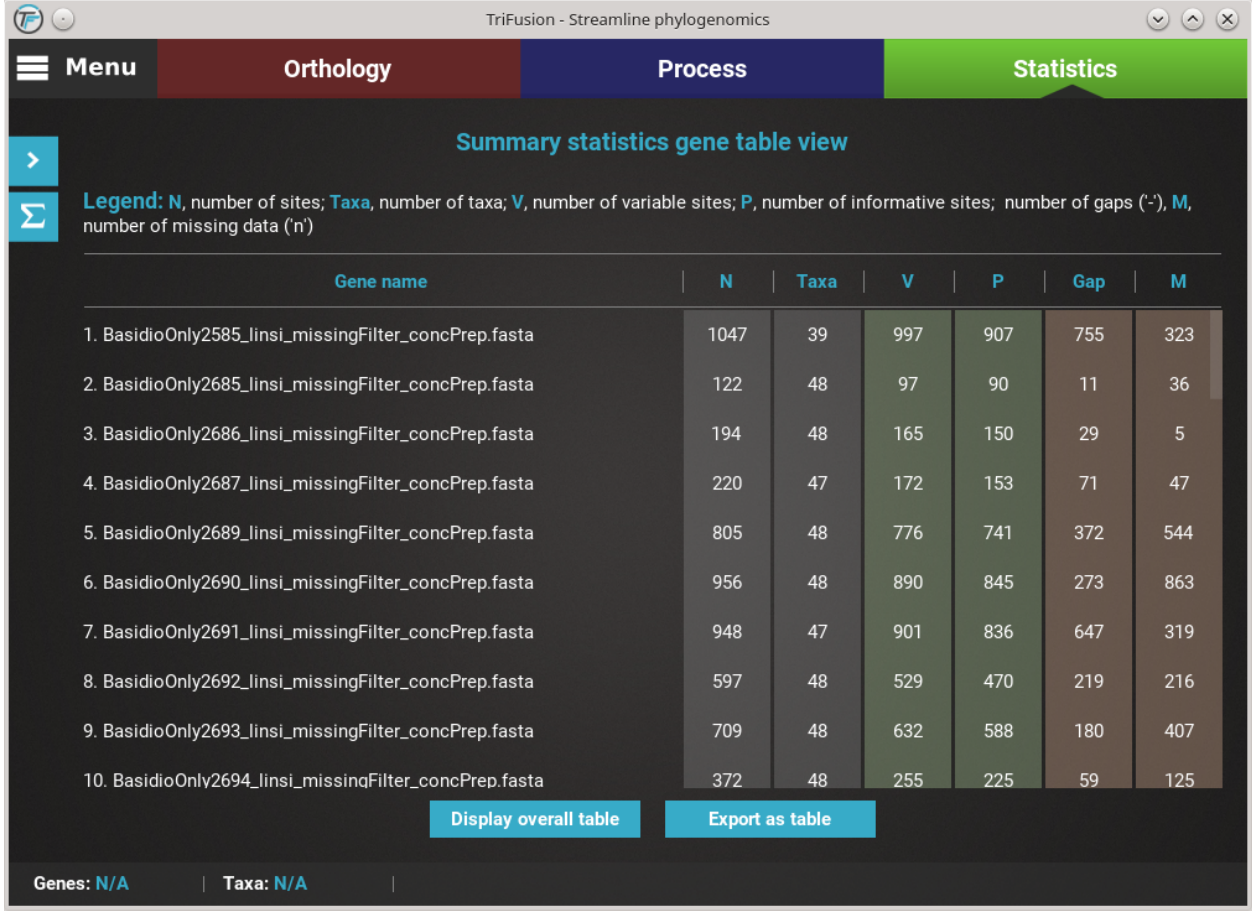
The values in the **General** section are mostly self-explanatory. We only note that the *Total alignment length* refers to the length of the alignment as a whole, not the sum of each sequence in the alignment.

The **Missing data** section separates the role of gaps (usually denoted by "-" in the alignment file) and true missing data (usually "N" in nucleotide sequences and "X" in protein sequences). The *Gaps* and *Missing* data values refer to the total number of gaps or missing data across all sequences, not alignment columns. Therefore,the associated percentages provide the relationship between these values and the sum of total characters in the alignment (in this case, 48 * 350 725).

The **Sequence variation** section provides the number of *variable* (at least one variant) and *informative* (one of the variants must be represented at least in two taxa) sites across the data set. In this case, these values correspond to the number of alignment columns, so percentages are relative to the *Total alignment length*.

1.15.2 Gene specific summary statistics

To visualize the same statistics as in the previous section discriminated for each alignment file, click the `Display gene table` at the bottom of the screen. This will change the display to show a list with individual alignment files as rows and summary statistics in the different columns.



Summary statistics gene table view

Legend: N, number of sites; Taxa, number of taxa; V, number of variable sites; P, number of informative sites; number of gaps ('-'), M, number of missing data ('n')

Gene name	N	Taxa	V	P	Gap	M
1. BasidioOnly2585_linsi_missingFilter_concPrep.fasta	1047	39	997	907	755	323
2. BasidioOnly2685_linsi_missingFilter_concPrep.fasta	122	48	97	90	11	36
3. BasidioOnly2686_linsi_missingFilter_concPrep.fasta	194	48	165	150	29	5
4. BasidioOnly2687_linsi_missingFilter_concPrep.fasta	220	47	172	153	71	47
5. BasidioOnly2689_linsi_missingFilter_concPrep.fasta	805	48	776	741	372	544
6. BasidioOnly2690_linsi_missingFilter_concPrep.fasta	956	48	890	845	273	863
7. BasidioOnly2691_linsi_missingFilter_concPrep.fasta	948	47	901	836	647	319
8. BasidioOnly2692_linsi_missingFilter_concPrep.fasta	597	48	529	470	219	216
9. BasidioOnly2693_linsi_missingFilter_concPrep.fasta	709	48	632	588	180	407
10. BasidioOnly2694_linsi_missingFilter_concPrep.fasta	372	48	255	225	59	125

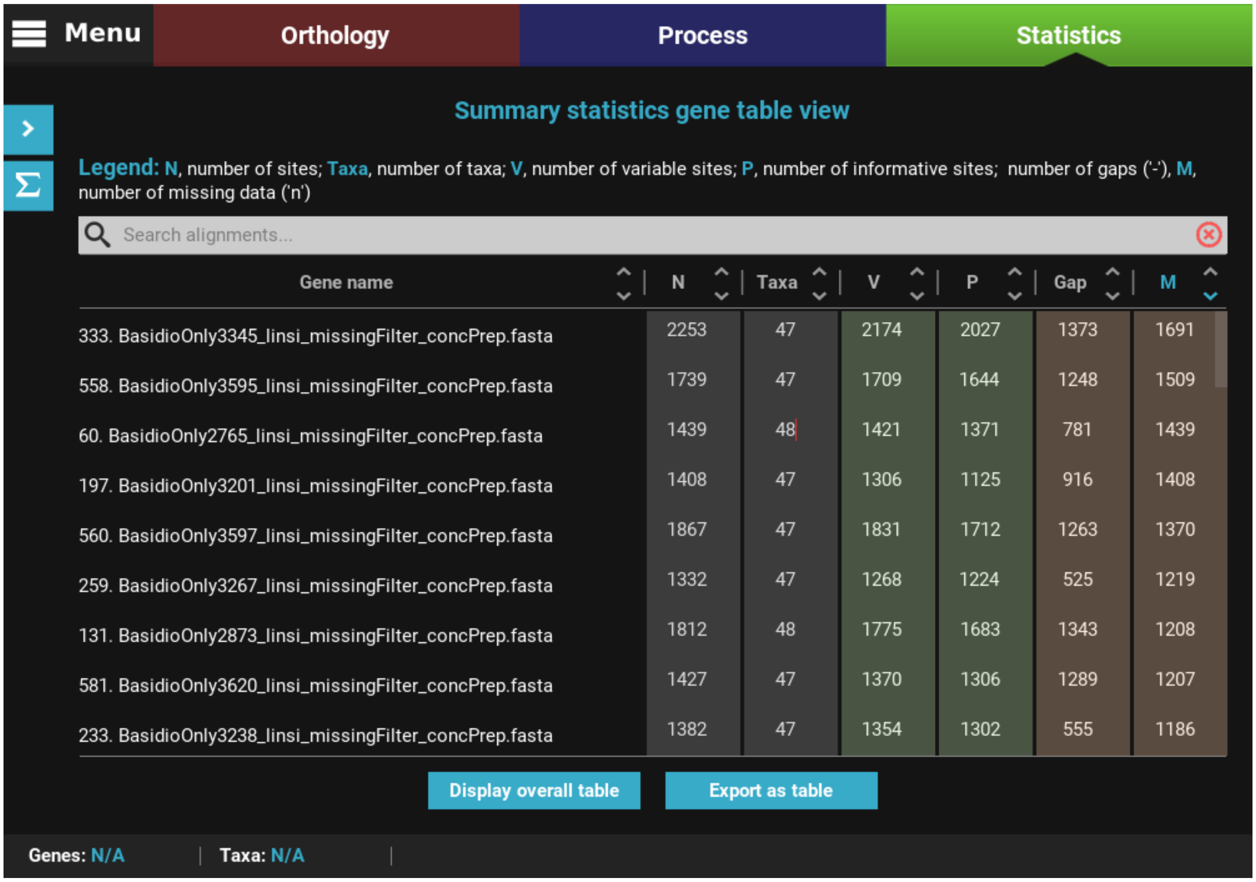
Genes: N/A | Taxa: N/A

Note that, due to performance issues, only the first 50 alignments are shown by default. You can increment the number of shown alignments by scrolling to the bottom and clicking the `Show more 25` button. Alternatively, you can export this data into a .csv file that can be read by **LibreOffice** or **MS Excel** by clicking the `Export as table` button.

As in the previous section, there are three main summary statistic categories , which are color coded along the table for convenience. A legend of each summary statistic is provided at the top of the table.

Sorting and filtering

Each column in this table can be sorted in ascending or descending order, which makes it easier to identify alignments with higher missing data or higher variation, for example. Let's try to sort our table in descending order by the missing data (**M**) column.



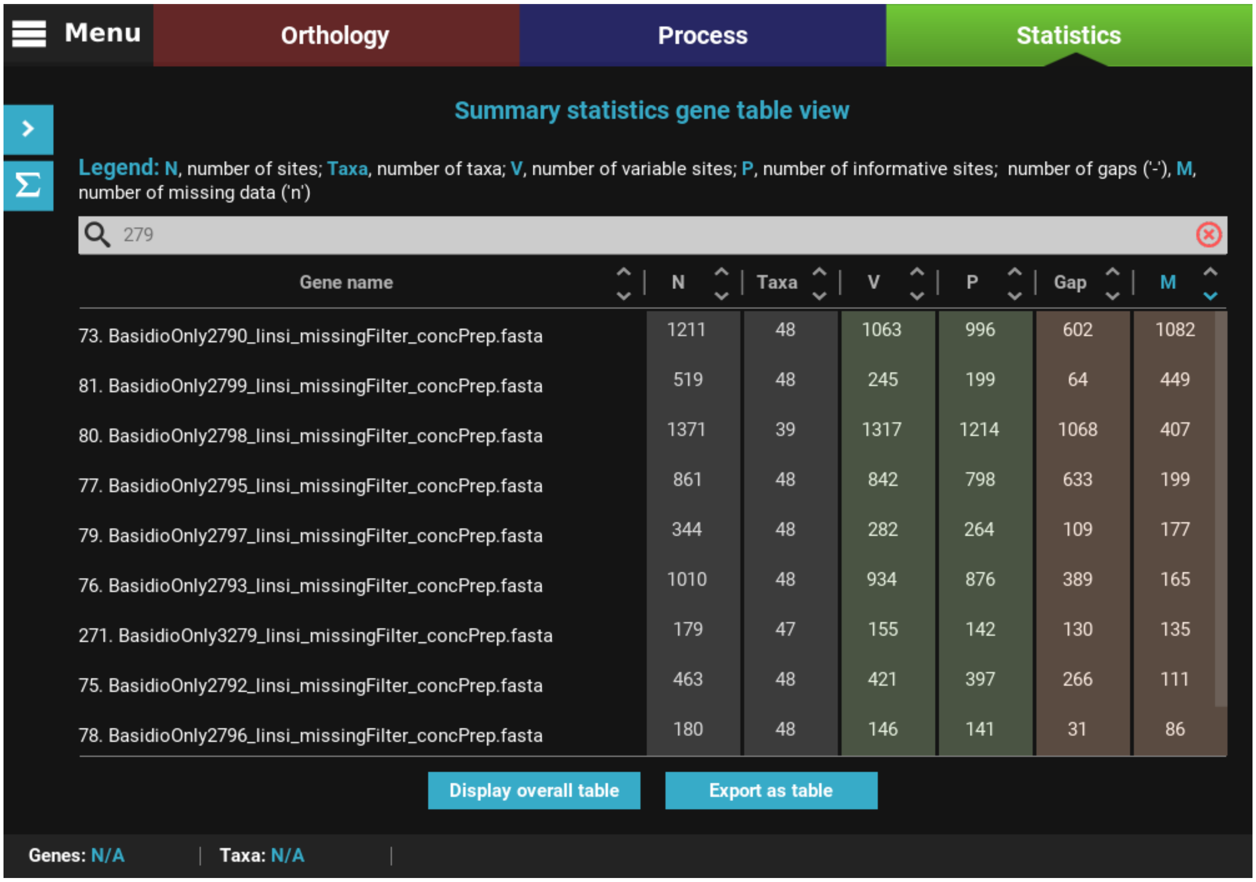
Summary statistics gene table view

Legend: N, number of sites; Taxa, number of taxa; V, number of variable sites; P, number of informative sites; number of gaps ('-'), M, number of missing data ('n')

Gene name	N	Taxa	V	P	Gap	M
333. BasidioOnly3345_linsi_missingFilter_concPrep.fasta	2253	47	2174	2027	1373	1691
558. BasidioOnly3595_linsi_missingFilter_concPrep.fasta	1739	47	1709	1644	1248	1509
60. BasidioOnly2765_linsi_missingFilter_concPrep.fasta	1439	48	1421	1371	781	1439
197. BasidioOnly3201_linsi_missingFilter_concPrep.fasta	1408	47	1306	1125	916	1408
560. BasidioOnly3597_linsi_missingFilter_concPrep.fasta	1867	47	1831	1712	1263	1370
259. BasidioOnly3267_linsi_missingFilter_concPrep.fasta	1332	47	1268	1224	525	1219
131. BasidioOnly2873_linsi_missingFilter_concPrep.fasta	1812	48	1775	1683	1343	1208
581. BasidioOnly3620_linsi_missingFilter_concPrep.fasta	1427	47	1370	1306	1289	1207
233. BasidioOnly3238_linsi_missingFilter_concPrep.fasta	1382	47	1354	1302	555	1186

Genes: N/A | Taxa: N/A

The table now displays the alignments with higher amount of missing data. If you want, you can filter alignments using the **Search** field above the table. We could search for alignment names containing the string *279* by typing it in the search field and pressing `Enter`.



Summary statistics gene table view

Legend: N, number of sites; Taxa, number of taxa; V, number of variable sites; P, number of informative sites; number of gaps ('-'), M, number of missing data ('n')

279

Gene name	N	Taxa	V	P	Gap	M
73. BasidioOnly2790_linsi_missingFilter_concPrep.fasta	1211	48	1063	996	602	1082
81. BasidioOnly2799_linsi_missingFilter_concPrep.fasta	519	48	245	199	64	449
80. BasidioOnly2798_linsi_missingFilter_concPrep.fasta	1371	39	1317	1214	1068	407
77. BasidioOnly2795_linsi_missingFilter_concPrep.fasta	861	48	842	798	633	199
79. BasidioOnly2797_linsi_missingFilter_concPrep.fasta	344	48	282	264	109	177
76. BasidioOnly2793_linsi_missingFilter_concPrep.fasta	1010	48	934	876	389	165
271. BasidioOnly3279_linsi_missingFilter_concPrep.fasta	179	47	155	142	130	135
75. BasidioOnly2792_linsi_missingFilter_concPrep.fasta	463	48	421	397	266	111
78. BasidioOnly2796_linsi_missingFilter_concPrep.fasta	180	48	146	141	31	86

Genes: N/A | Taxa: N/A

As you can see, the table is still sorting the alignments by the missing data (**M**) column, but only for alignment names containing '*279*'. You can play quite a bit with the sorting and filters to obtain more information about your data.

To switch to the **overall summary statistics** view, click the `Display overall table` button.

1.15.3 Displaying summary statistics

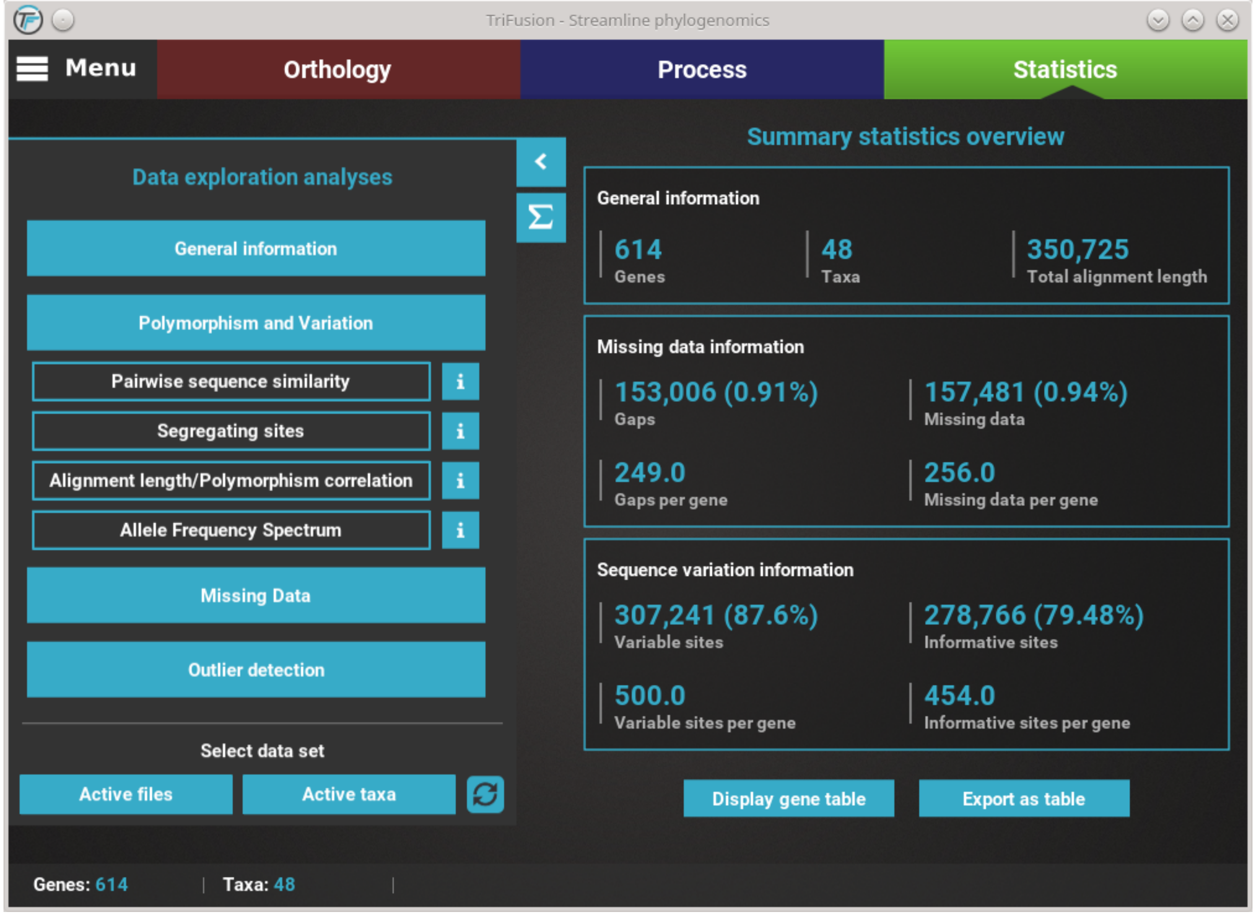
At any time, you can return to the **summary statistics** display by clicking the **Summary statistics** icon button at the edge of the Statistics' side panel.

1.16 Data exploration analyses

Note: Data availability for this tutorial: the medium sized data set of 614 genes and 48 taxa that will be used can be downloaded here.

All data exploration analyses are contained within the four main category buttons that are found in **Statistics**' side panel. Clicking any of these buttons will expand all available analyses under that category. For example, clicking the **Polymorphism and Variation** button, will show four individual analyses.

Note: This tutorial is not meant to be an exhaustive description of all plot types and analyses. For such a description please refer to TriFusion's user guide


TriFusion - Streamline phylogenomics
Menu
Orthology
Process
Statistics
Data exploration analyses
General information
Polymorphism and Variation
Pairwise sequence similarity
Segregating sites
Alignment length/Polymorphism correlation
Allele Frequency Spectrum
Missing Data
Outlier detection
Select data set
Active files
Active taxa
Display gene table
Export as table

Summary statistics overview

General information

614 Genes | 48 Taxa | 350,725 Total alignment length

Missing data information

153,006 (0.91%) Gaps | 157,481 (0.94%) Missing data

249.0 Gaps per gene | 256.0 Missing data per gene

Sequence variation information

307,241 (87.6%) Variable sites | 278,766 (79.48%) Informative sites

500.0 Variable sites per gene | 454.0 Informative sites per gene

Genes: 614 | Taxa: 48

1.16.1 How to view analysis specific information

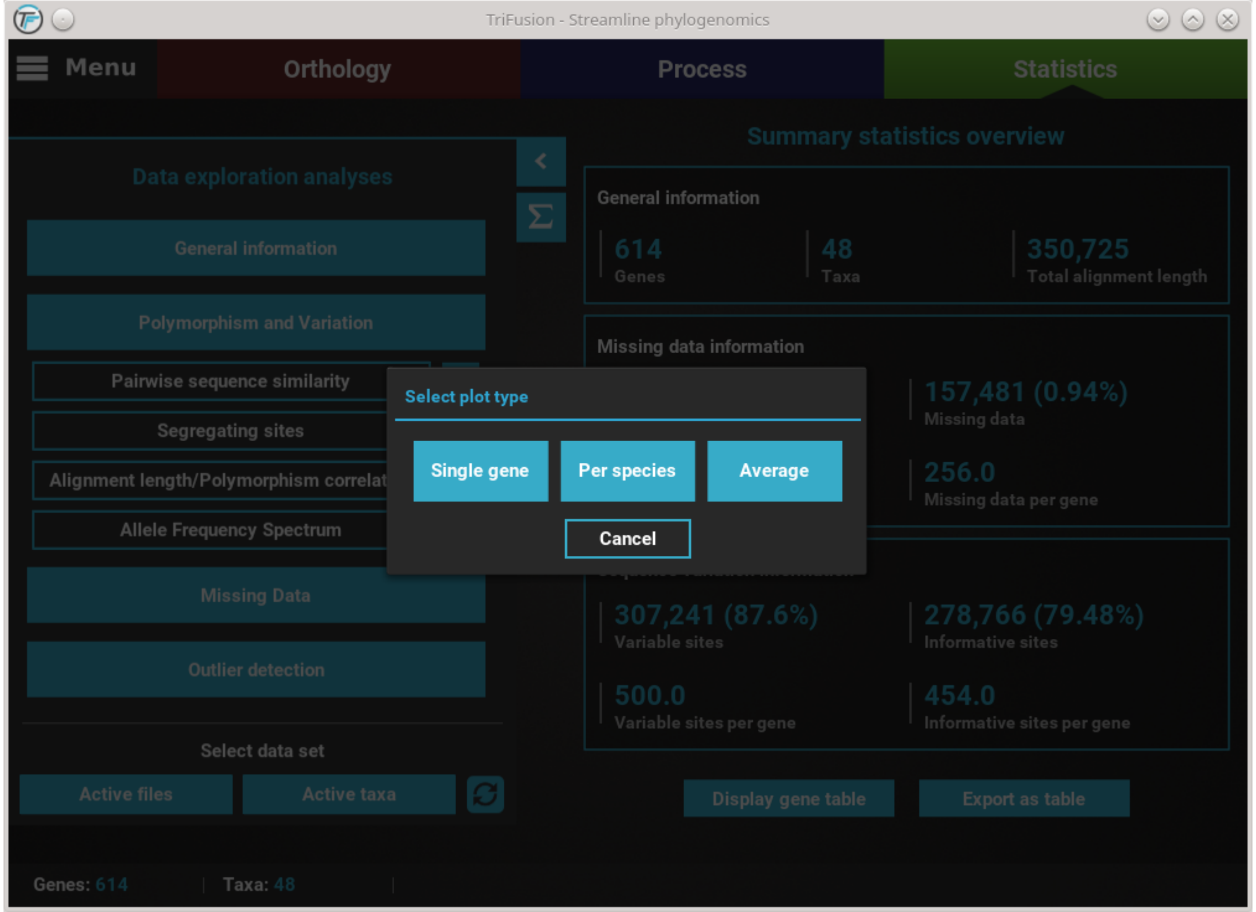
A detailed description of each analysis is provided in TriFusion's user guide, but you can also click the information buttons (**i**) that are coupled with every analysis button. For instance, clicking the **information button** of the **Pairwise sequence similarity** analysis shows a pop-up with a short description of the analysis, the available plot types and what the axis represent.

1.16.2 Plot types

In the majority of the individual analysis, there are up to **three plot types** available that represent different perspectives of the same analysis:

- *Single gene*: You choose a single a gene from the data set and the analysis is performed on that gene (usually a sliding window plot).
- *Per species*: The analyses will be focused on gathering information for each taxa or discriminates it by taxa in some way.
- *Average*: The analyses will produce an average distribution/result across the whole data set.

For example, clicking the Pairwise sequence similarity button will ask you which plot type you wish to produce.



General information

614 Genes | 48 Taxa | 350,725 Total alignment length

Missing data information

157,481 (0.94%) Missing data

256.0 Missing data per gene

307,241 (87.6%) Variable sites

278,766 (79.48%) Informative sites

500.0 Variable sites per gene

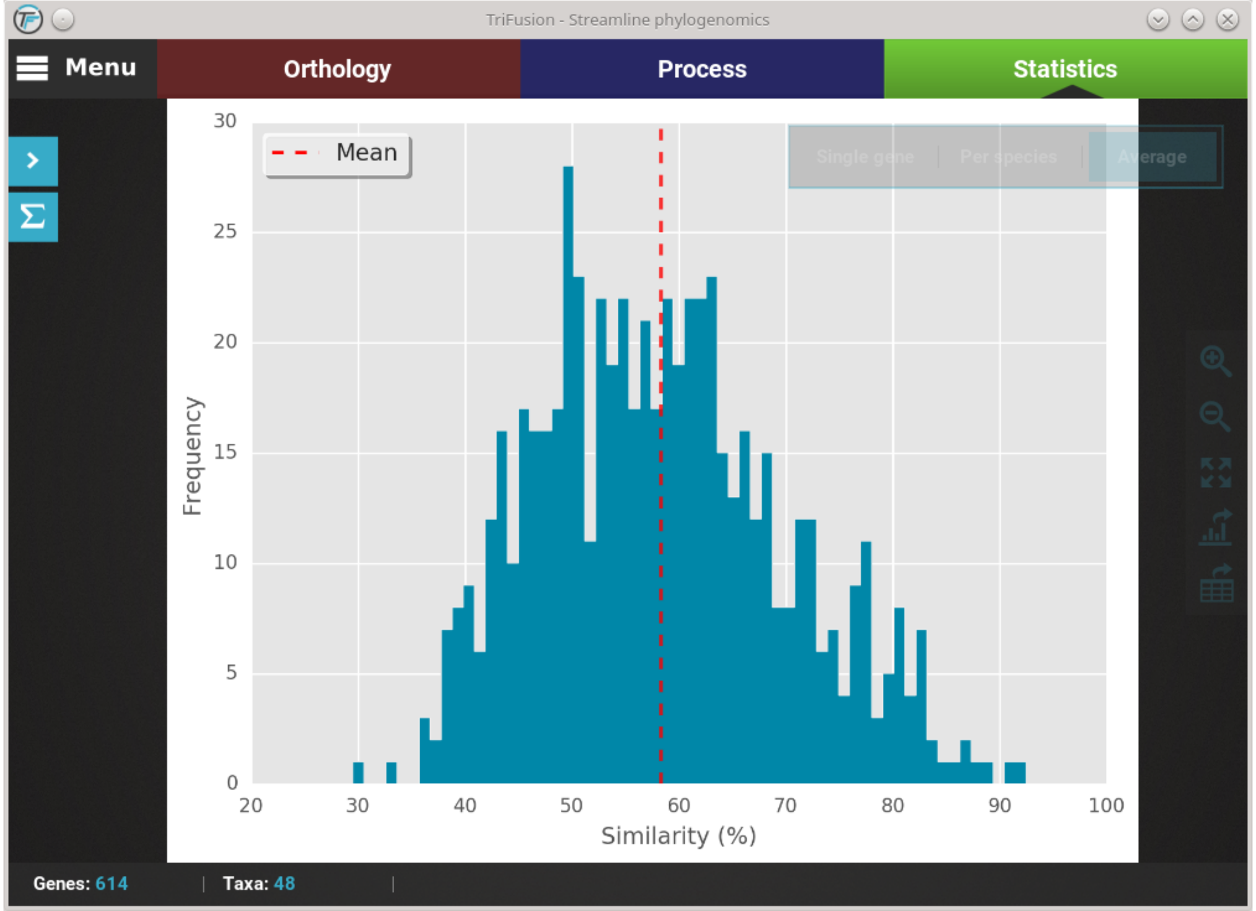
454.0 Informative sites per gene

Genes: 614 | Taxa: 48

In this case, all three plot types are available. However, some options will have only two plot types available, and others only one. It will depend on the analysis.

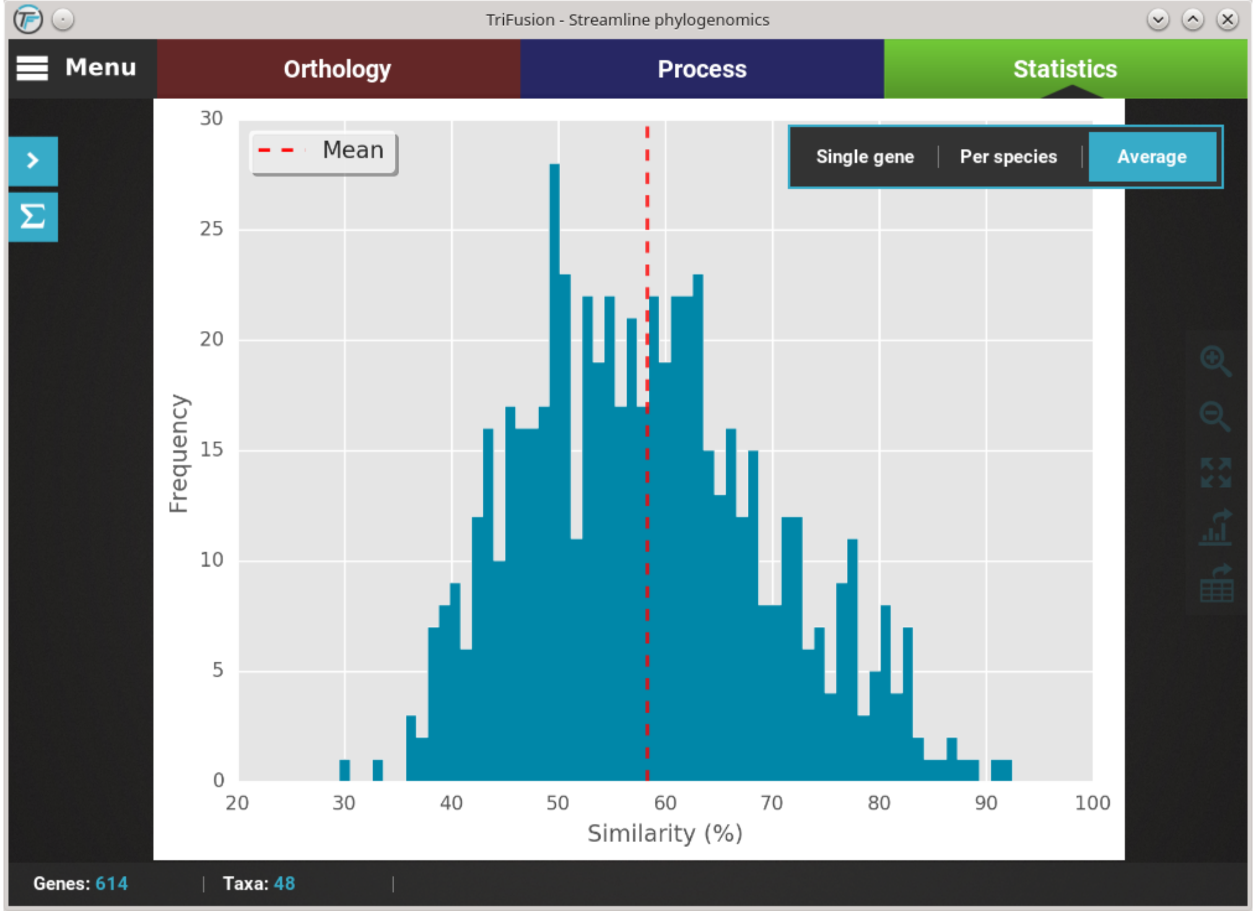
1.16.3 Executing an analysis

Let's explore the **distribution of sequence similarity** across our entire data set. Since we are interested in an average of the data set , click on the **Average** button. The computation of sequence similarity and segregating sites are some of the most computationally intensive in TriFusion, so this may take some time the first time. However, TriFusion uses a hash look-up table technique which considerably speeds up future computations of these analyses in the same session. Once complete, you should see a bar plot with the distribution and mean of the pairwise sequence similarity across the data set.

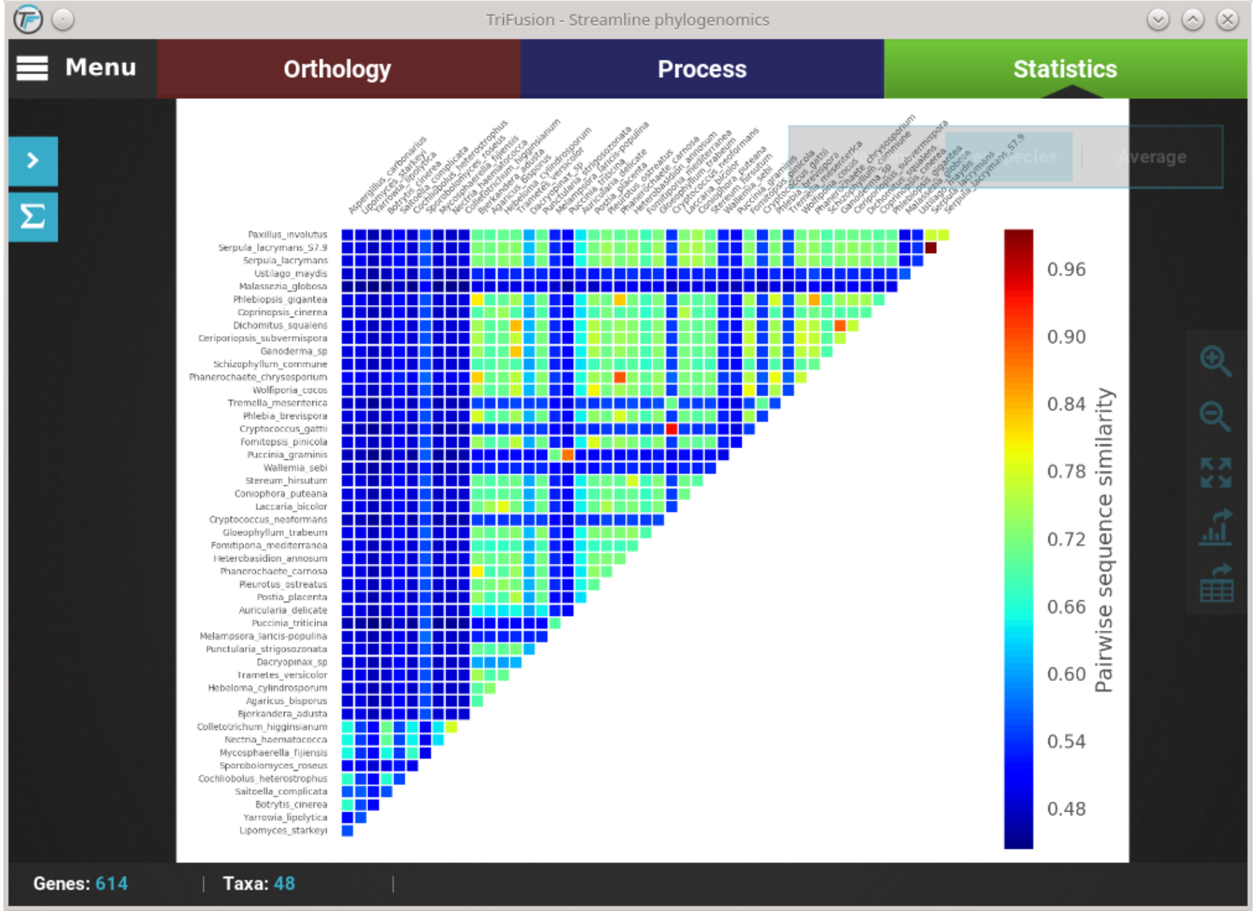


1.16.4 Changing plot type

If you want to change the plot type of the current analysis, there is a floating box in the top right of the screen.



The current plot type appears with a filled blue background (**Average** in this case). To change to the **Per species** plot type, simply click the corresponding button and a new analyses should be started. At the end of the analysis, you should see a triangular heat map matrix with the sequence similarity between every species pair in the data set.



1.16.5 Fast plot switching

While the active data set remains the same, all generate plots are stored locally. This means that if you need to visualize an analysis that you already performed in your current session, you do not have to repeat the entire computation. For instance, we are currently visualizing the **Per species** plot type of the **Pairwise sequence similarity** analysis. If you click the **Average** button in the floating box to change the plot type, you'll notice that the switch will be almost instantaneous.

1.16.6 Single gene analyses

Some analyses can be performed for single genes in the form of a sliding window analysis that contain additional features. Let's investigate the averaged pairwise sequence similarity for a single gene in our data set. Click the **Pairwise sequence similarity** analysis and then the `Single gene` plot type.

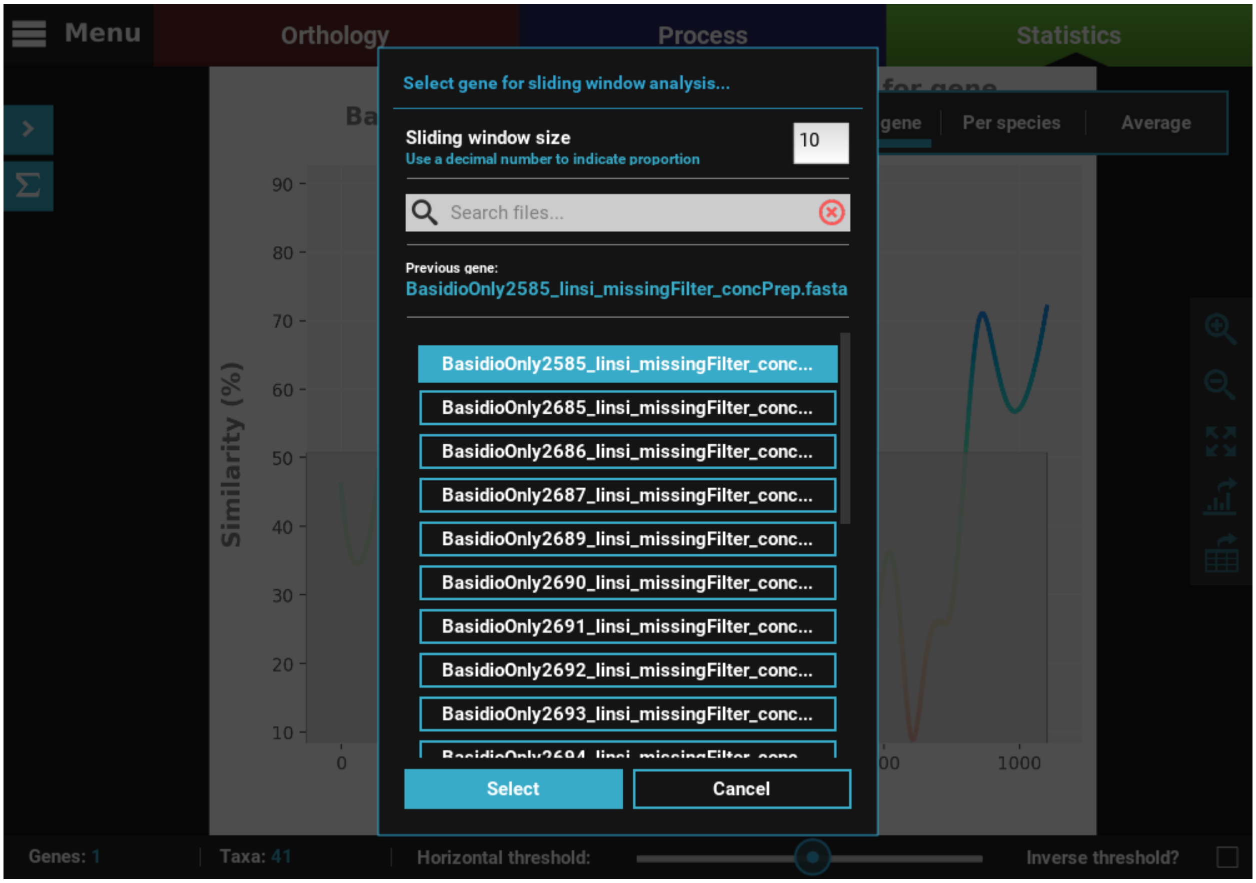
Here you can select any loaded alignment along with the size of the sliding window. The value of the **sliding window** may be:

- An **absolute** value will set the window size to exactly that value (e.g. a value of 20 will calculate the sequence similarity for every stretch of 20 alignment columns).
- A **decimal** value will set the window size to a proportion of the total alignment (e.g. a value of 0.1 will calculate the sequence similarity for stretches equivalent to 10% of the alignment size).

Let's choose the first alignment in the list with a window size of *20*.

Note: If the specified window size results in a very high number of sliding windows (>500), a warning will be raised where you can cancel, update the sliding window to a more sensible value or continue anyway.

If you want to calculate the sequence similarity for another single gene, you can click on the `Change gene` button on the plot type floating box.



Select gene for sliding window analysis...

Sliding window size: 10
Use a decimal number to indicate proportion

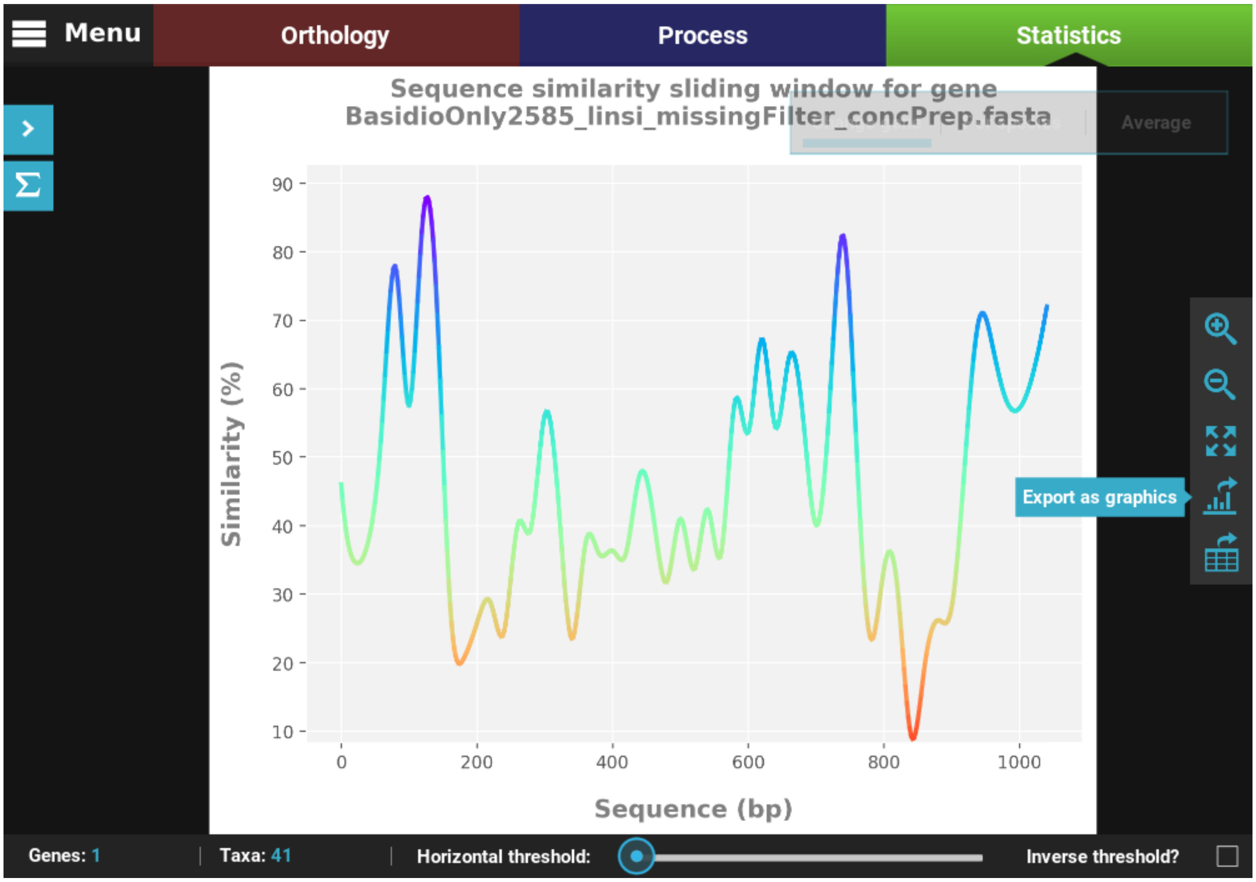
Previous gene:
BasidioOnly2585_linsi_missingFilter_concPrep.fasta

BasidioOnly2585_linsi_missingFilter_conc...
BasidioOnly2685_linsi_missingFilter_conc...
BasidioOnly2686_linsi_missingFilter_conc...
BasidioOnly2687_linsi_missingFilter_conc...
BasidioOnly2689_linsi_missingFilter_conc...
BasidioOnly2690_linsi_missingFilter_conc...
BasidioOnly2691_linsi_missingFilter_conc...
BasidioOnly2692_linsi_missingFilter_conc...
BasidioOnly2693_linsi_missingFilter_conc...
BasidioOnly2694_linsi_missingFilter_conc...

Notice that the previously selected gene will appear under the **Previous gene** section and will be already selected in the alignment list. Here you can select another alignment and window size, using the search field if you like.

1.16.7 Export figures and tables

All plots generated in TriFusion can be exported as a graphics file and almost all can be exported in table format. These functions are available in the plot screen bar at the right of the screen.



Export a figure

Click the `Export as graphics` button in the plot screen right bar. This will open a file browser where you can choose where to export the figure, its name and graphics format.

Here we provided some name to our figure, and set the image format to **svg**. Finally, click `Save` and the figure will be exported.

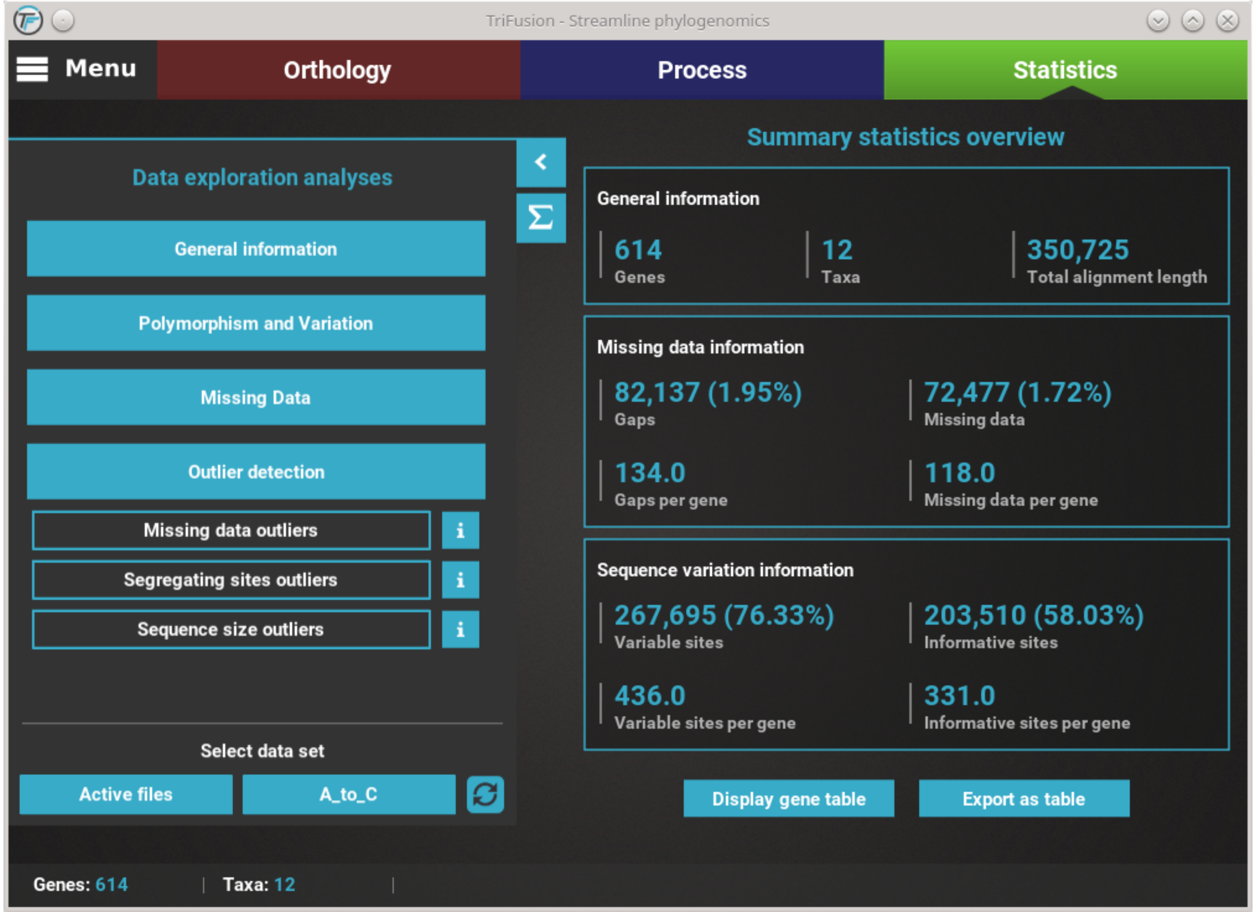
Export a table

Click the `Export as table` button in the plot screen right bar. As in the previous section, this will open a file browser where you can choose where to export the table and its name.

Then click `Save` to export the table. The generated table will be in **csv** format, which can be readily imported by **LibreOffice** or **MS Excel** or viewed as a plain text file.

1.16.8 Dealing with outliers

Outlier analyses in TriFusion are a bit different because they offer you the option to **remove files and/or taxa** that may have an outlier behaviour for some statistics. If you click on the **Outlier Dectection** category in Statistic's sidepanel you'll see three outlier detection analyses: by **missing data**, **segregating sites** and **sequence size**.



Summary statistics overview

General information

614 Genes

12 Taxa

350,725 Total alignment length

Missing data information

82,137 (1.95%) Gaps

72,477 (1.72%) Missing data

134.0 Gaps per gene

118.0 Missing data per gene

Sequence variation information

267,695 (76.33%) Variable sites

203,510 (58.03%) Informative sites

436.0 Variable sites per gene

331.0 Informative sites per gene

Genes: 614 | Taxa: 12 |

Let’s exemplify outlier handling by checking for outlier taxa for missing data, that is, taxa that contain unusual amounts of missing data. Click on the **Missing data outliers** button, and then the **Per species** plot type.

You can see that the missing data distribution is bimodal (two peaks) and that one taxa outlier was found (see the footer of the screen). In the footer of the screen are three functions to handle potential outliers:

- **Remove**: Clicking the Remove button will remove the outlier taxa from the current TriFusion session. This is equivalent to manually remove the taxa in TriFusion’s side panel.
- **Export**: Clicking the Export button will save the outlier taxa to a csv file, where each line will contain a taxon name. This can be used to change the active data set in TriFusion using a text file
- **View**: Clicking the View will display a list of the outlier taxa.

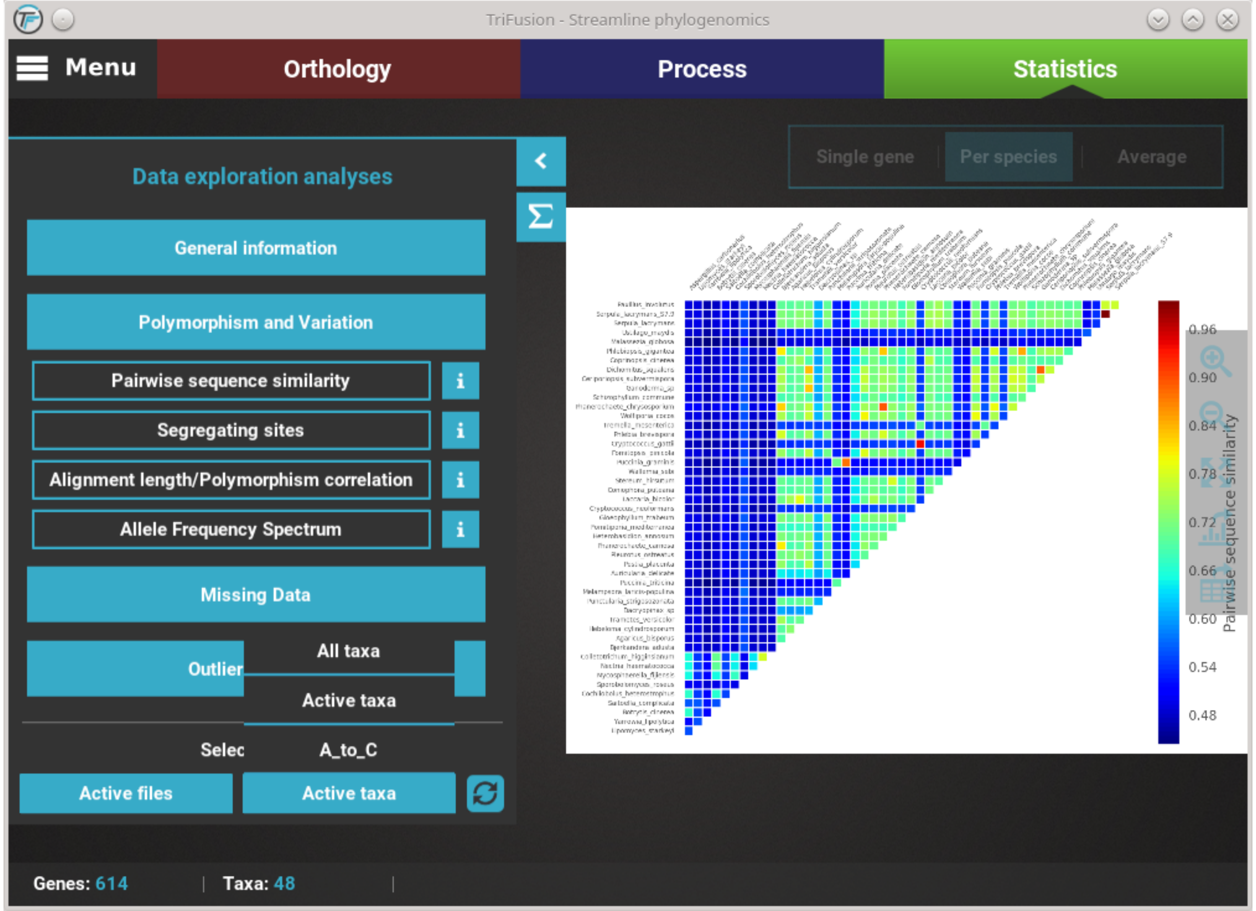
1.17 Update the active data set

Note: **Data availability for this tutorial**: the medium sized data set of 614 genes and 48 taxa that will be used can be downloaded here.

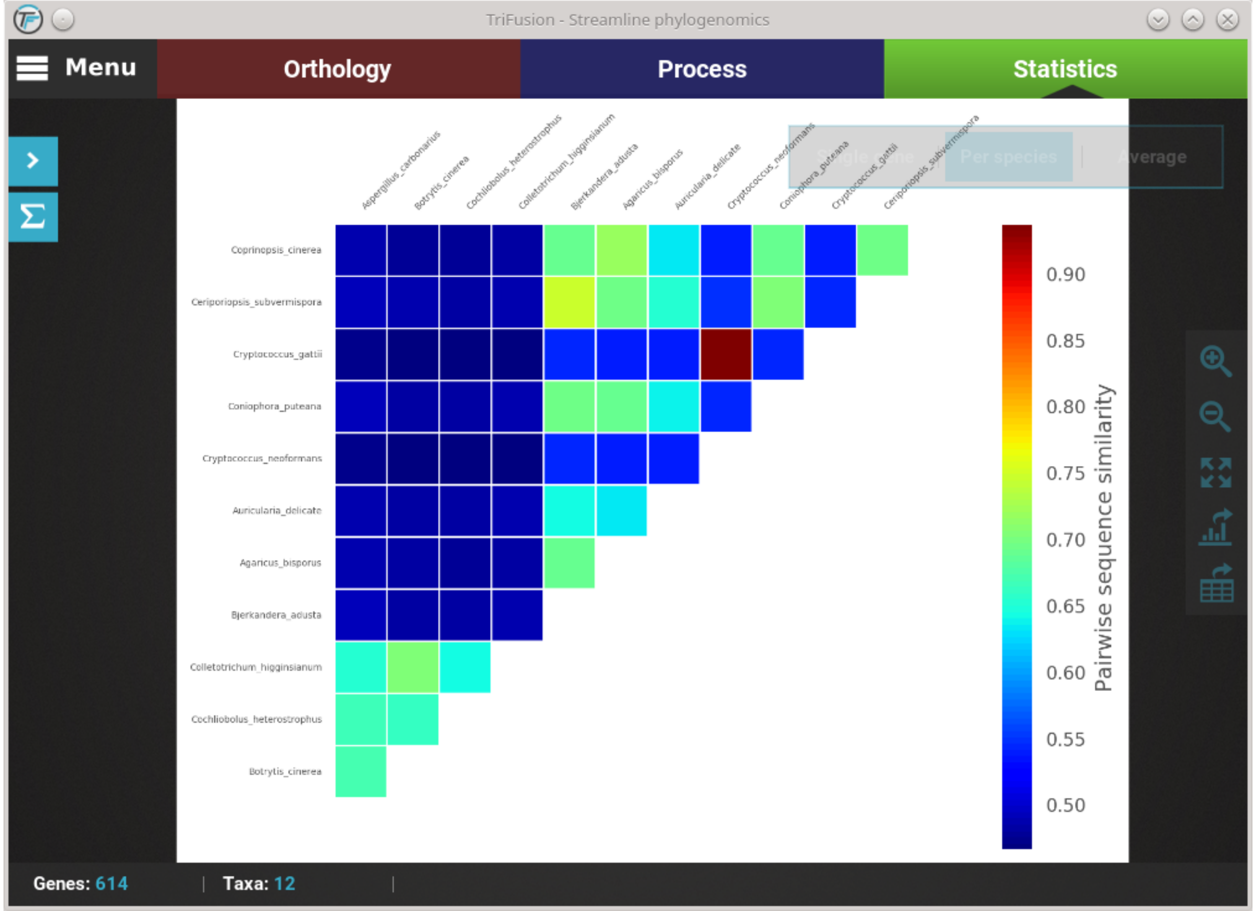
1.17.1 Data exploration analyses

The analyses in the **Statistics** module are not limited to the **total data** set loaded into TriFusion. You can **modify the active file/taxa data sets** or create data set groups in TriFusion (see tutorial *Data set groups*), and then select them in the bottom of the Statistics side panel.

Following the guidelines in the *Data set groups* tutorial, we created a taxa group of 12 elements that contains taxa whose name starts with an "A", "B" or "C", named **A_to_C**. To change the taxa data set to the newly define group, click in the drop down menu for the taxa data set and select the **A_to_C** option.



Now, all selected analyses will use this set of 12 taxa instead of the full 48 taxa data set. If you want to update the currently displayed analyses, click the **refresh** button next to the data set selection drop down menus.



Genes: 614 | Taxa: 12

1.17.2 Summary statistics

It is also possible to change the active data set when visualizing the summary statistics of your data set and it can be particularly useful. For example, if you suspect that a group of taxa or alignment may be responsible for a particular large share of variability of missing data, you could create