BIOMÉRIEUX

BIONUMERICS®

version 8 - PLUGINS

SARS-CoV-2 plugin

Contents

NOTES

SUPPORT BY APPLIED MATHS, A BIOMÉRIEUX COMPANY

While the best efforts have been made in preparing this manuscript, no liability is assumed by the authors with respect to the use of the information provided.

Applied Maths, a bioMérieux company, will provide support to research laboratories in developing new and highly specialized applications, as well as to diagnostic laboratories where speed, efficiency and continuity are of primary importance. Our software thanks its current status for a part to the response of many customers worldwide. Please contact us if you have any problems or questions concerning the use of BIONUMERICS®, or suggestions for improvement, refinement or extension of the software to your specific applications:

Applied Maths NV
Keistraat 120
9830 Sint-Martens-Latem
Belgium
PHONE: +32 9 2222 100
FAX: +32 9 2222 102
E-MAIL: BE-DAU-INFO@biomerieux.com
URL: https://www.bionumerics.com

Applied Maths, Inc.
11940 Jollyville Road, Suite 115N
Austin, Texas 78759
U.S.A.
PHONE: +1 512-482-9700
FAX: +1 512-482-9708
E-MAIL: US-DAU-INFO@biomerieux.com

LIMITATIONS ON USE

The BIONUMERICS® software, its plugin tools and their accompanying guides are subject to the terms and conditions outlined in the License Agreement. The support, entitlement to upgrades and the right to use the software automatically terminate if the user fails to comply with any of the statements of the License Agreement. No part of this guide may be reproduced by any means without prior written permission of the authors.

BIONUMERICS® uses following third-party software tools and libraries:

- Python 3.8 release from the Python Software Foundation, https://www.python.org/
- Xerces library for XML input and output from the Apache Software Foundation, https://xerces.apache.org/
- NCBI toolkit version 2.11.0, https://www.ncbi.nlm.nih.gov/BLAST/
- SRA Toolkit, https://ncbi.github.io/sra-tools/
- Boost c++ libraries, https://www.boost.org/
- Samtools for interacting with SAM / BAM files, https://www.htslib.org/download/
- 7-Zip (7za.exe), https://www.7-zip.org/
- Zlib library, https://zlib.net/
- Pigz for parallel gzip compression, https://zlib.net/pigz/
- Cairo 2D graphics library version 1.12.14, https://cairographics.org/
- Crypto++ library version 5.5.2, https://www.cryptopp.com/
- OpenSSL library, https://www.openssl.org/
- libSVM library for Support Vector Machines, https://www.csie.ntu.edu.tw/~cjlin/libsvm/
- SQLite version 3.7.17, https://www.sqlite.org/
- pymzML Python module version 2.4.7, https://github.com/pymzml/pymzML
- NumPy Python library version 1.19.1, https://www.numpy.org/
- BioPython Python library version 1.78, https://www.biopython.org/
- pyodbc Python module version 4.0.30, https://pypi.org/project/pyodbc/
- jinja2 Python library version 2.11.2, https://pypi.org/project/Jinja2/
- MarkupSafe Python library version 1.1.1, https://pypi.org/project/MarkupSafe/
- regex Python library version 2.5.91, https://pypi.org/project/regex/
- Chromium Embedded Framework, https://bitbucket.org/chromiumembedded/cef/wiki/Home
- SPAdes genome assembler version 3.15.3, https://bioinf.spbau.ru/spades *
- SKESA version 2.3.0, https://github.com/ncbi/SKESA/releases
- Unicycler version 0.5.0, https://github.com/rrwick/Unicycler/releases *
- Velvet for Windows, source code can be downloaded from https://www.bionumerics.com/download/open-source
- Bowtie2 version 2.2.5 (https://bowtie-bio.sourceforge.net/bowtie2/index.shtml)*
- SNAP version 2.0.0, https://www.microsoft.com/en-us/research/project/snap/
- RAxML version 8.2.11, https://github.com/stamatak/standard-RAxML/releases

- FastTree version 2.1.10, https://www.microbesonline.org/fasttree/
- CFSAN SNP pipeline version 2.2.0, https://github.com/CFSAN-Biostatistics/snp-pipeline *
- Prokka version 1.14.5, https://github.com/tseemann/prokka *
- sourmash version 4.1.0, https://github.com/dib-lab/sourmash **
- SeqSero2 for Windows, source code can be downloaded from https://www.bionumerics.com/download/open-source
- Fastp version 0.22.0, https://github.com/OpenGene/fastp

*: On Calculation Engine only **: See license conditions below

Sourmash license conditions:

Chapter 1

Introduction

The *SARS-CoV-2 plugin* facilitates the processing and analysis of SARS-CoV-2 genomic sequences, whether downloaded from a public data repository or generated locally. Each genomic sequence is separated (”extracted”) into subsequences, each of which is analyzed for SNPs relative to the reference sequence. All SNPs are stored together in an open (dynamic) character set, which allows for easy comparisons and strain typing based on the highest resolution available.

The *SARS-CoV-2 plugin* is a free add-on available in the **BIONUMERICS-SEQ** and **BIONUMERICS-SUITE** configurations.

Chapter 2

Starting and setting up BIONUMERICS

2.1 Startup program

Make sure the latest version of BIONUMERICS is installed (https://www.bionumerics.com/download/software). The installation manual can be downloaded from https://www.bionumerics.com/download/manuals.

When BIONUMERICS is launched from the Windows start panel or when the BIONUMERICS shortcut on your computer's desktop is double-clicked, the **Startup program** is run. This program shows the *BIONUMERICS Startup* window (see Figure 2.1).

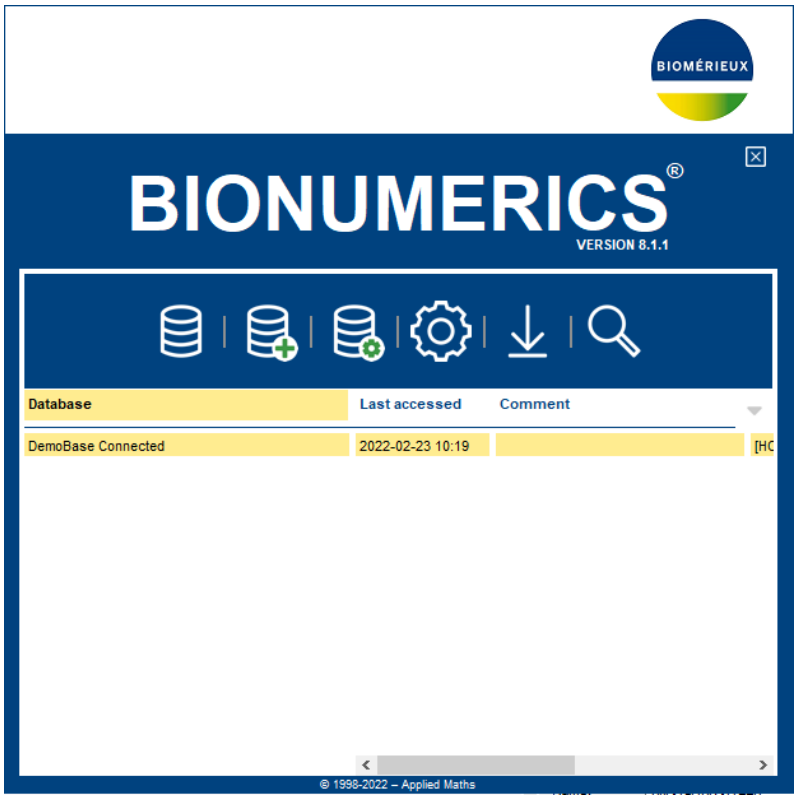

Figure 2.1: The *BIONUMERICS Startup* window.

A new BIONUMERICS database is created from the Startup program by pressing the button.

An existing database is opened in BIONUMERICS with or by simply double-clicking on a

database name in the list.

2.2 Creating a new database

2.1 Press the  button in the BIONUMERICS *BIONUMERICS Startup* window to enter the *New database* wizard.

2.2 Enter a name for the database, and press <***Next***>.

A new dialog box pops up, prompting for the type of database (see Figure 2.2).

2.3 Leave the default option selected and press <***Next***>.

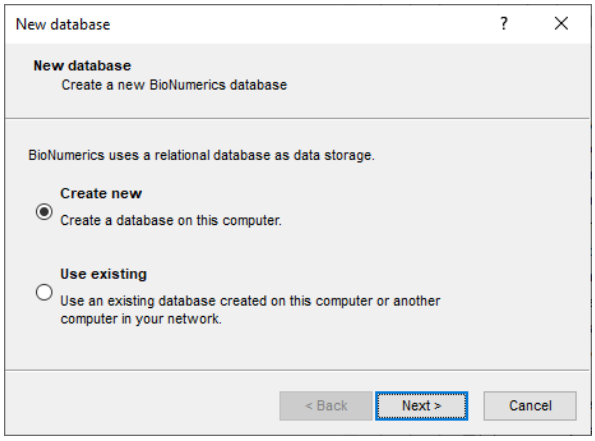

Figure 2.2: The *New database* wizard page.

A new dialog box pops up, prompting for the database engine (see Figure 2.3).

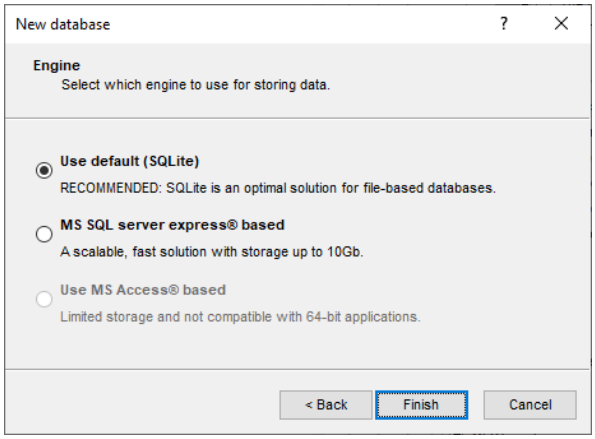

Figure 2.3: The *Engine* wizard page.

2.4 Leave the default option selected and press <***Finish***> to complete the setup of the new database.

2.3 Installing the Sars-CoV-2 plugin

The *Plugins and Scripts* dialog box can be called from the *Main* window by selecting ***File*** > ***Install / remove plugins...*** ( ) (see Figure 2.4).

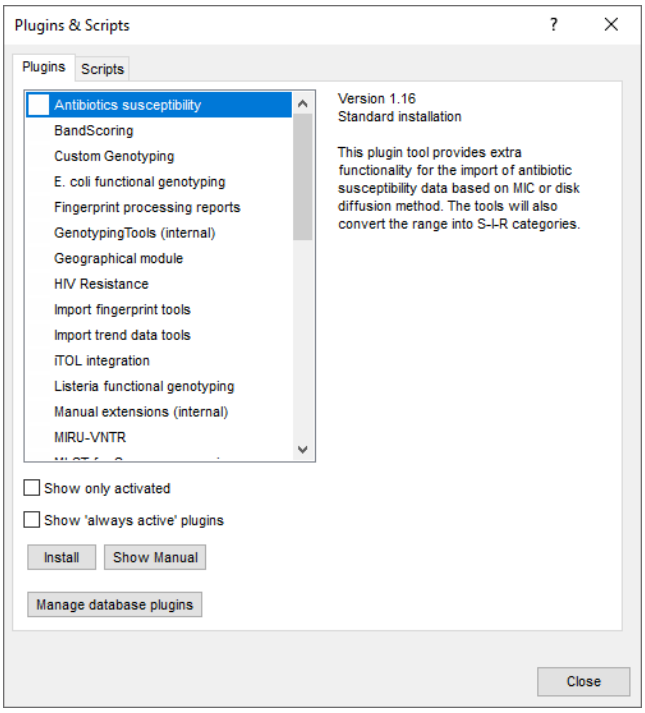

Figure 2.4: The *Plugins and Scripts* dialog box.

When a particular plugin is selected from the list of plugins, a short description appears in the right panel.

A selected plugin can be installed with the <***Install***> button. The software will ask for confirmation before installation. Some plugins are only supported in specific BIONUMERICS configurations. If the plugin is not supported by your BIONUMERICS configuration, it cannot be installed and an error message will be generated.

Once a plugin is installed, it is marked with a green V-sign. It can be removed again with the <***Uninstall***> button.

If the selected plugin is documented, pressing <***Show Manual***> will open its manual in the *Help* window.

An older version of the plugin (i.e. version 0.41) comes with the installation of BIONUMERICS and can be installed by selecting the plugin in the ***Plugins*** tab of the *Plugins and Scripts* dialog box and pressing the <***Install***> button. However, an updated version of the plugin is available online and can be added to the database by clicking the <***Manage database plugins***> button which will open the *Manage database plugins* dialog box.

3.1 Select ***File*** > ***Install / remove plugins...*** ( ) in the *Main* window to call the *Plugins and Scripts* dialog box.

3.2 Select the <***Manage database plugins***> button to open the *Manage database plugins* dialog box.

The *Manage database plugins* dialog box lists the plugins that are currently stored in the relational database. In a new BIONUMERICS database, this list will initially be empty. Adding or updating

database plugins can be done in the *Add database plugins* dialog box which can be accessed by clicking the <***Add/Update***> button.

3.3 Select the <***Add/Update***> button to open the *Add database plugins* dialog box.

The *Online plugins* panel lists all available online plugins which can be added by checking the check box in front of the plugins.

3.4 Check the check box in front of the *SARS-CoV-2 plugin* and click <***OK***>.

A message appears indicating that the plugin will be loaded after the database is restarted (see Figure 2.5).

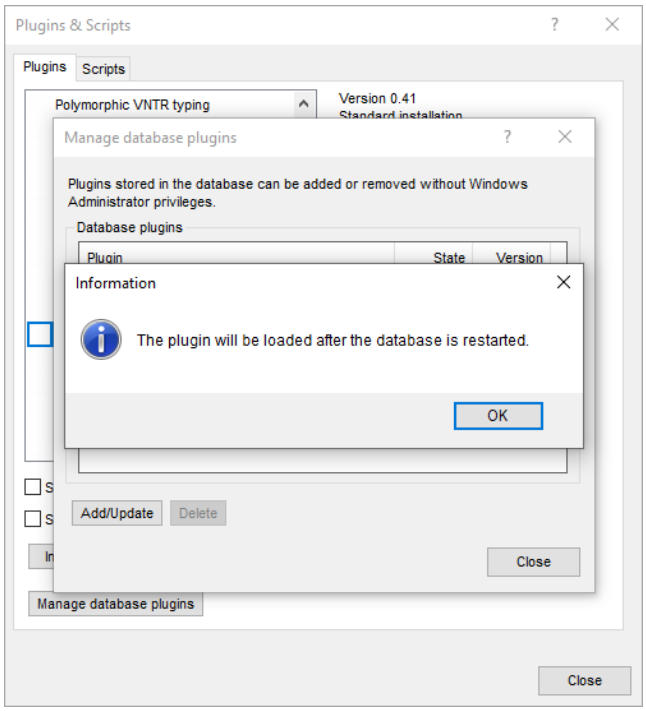

Figure 2.5: Information message indicating that the online plugin will be loaded after the database is restarted.

After adding a database plugin to the database, its status in the *Manage database plugins* dialog box will be inactive as it is not installed yet (see Figure 2.6).

3.5 Select the <***Close***> button two times and close and reopen your BIONUMERICS database.

3.6 Select ***File*** > ***Install / remove plugins...*** ( ) in the *Main* window to call the *Plugins and Scripts* dialog box.

After restarting the BIONUMERICS database, the added database plugins will be loaded into the database and can be recognized in the *Plugins tab* of the *Plugins and Scripts* dialog box as they are preceded by a database icon (see Figure 2.7).

3.7 Select the *SARS-CoV-2 plugin* in the list and press the <***Install***> button.

3.8 Confirm the installation of the plugin (see Figure 2.8).

The *Create database components* dialog pops up displaying all database components required by the plugin: entry fields, character type experiments and sequence type experiments (see Figure 2.9). The default suggested names can be changed if desired.

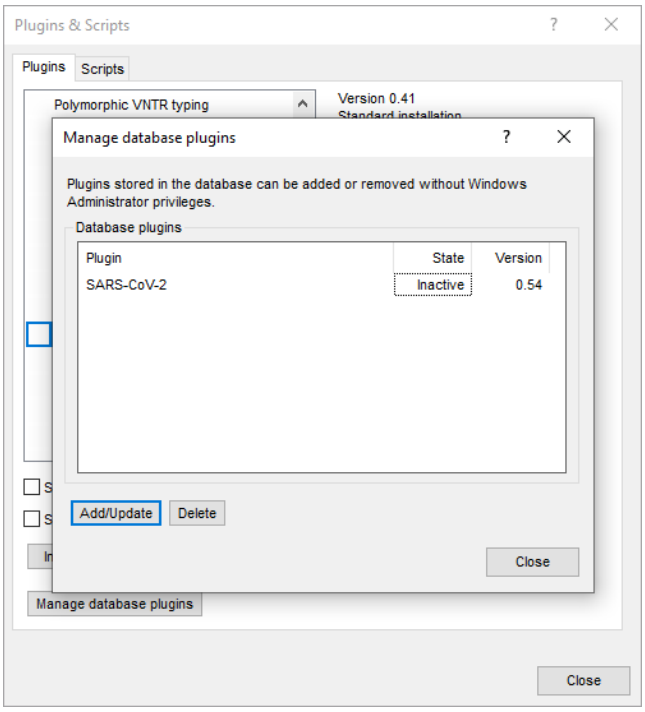

Figure 2.6: The online *SARS-CoV-2 plugin* added to the *Manage database plugins* dialog box.

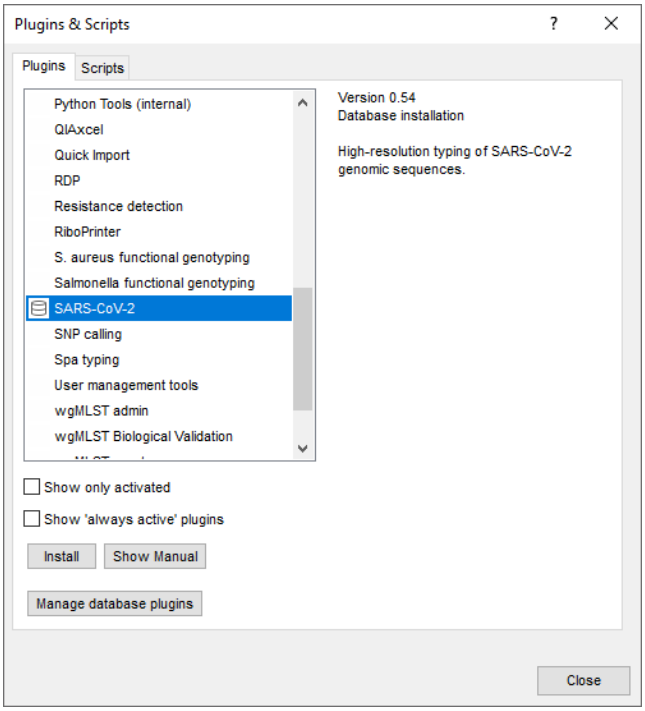

Figure 2.7: The online *SARS-CoV-2 plugin* added to the *Plugins* tab.

3.9 Press <***OK***> to confirm the creation of the database components.

A message pops up, displaying the successful installation of the plugin (see Figure 2.10).

3.10 Press <***OK***>.

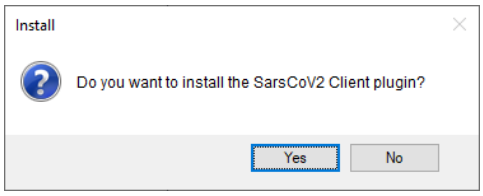

Figure 2.8: Confirm plugin installation.

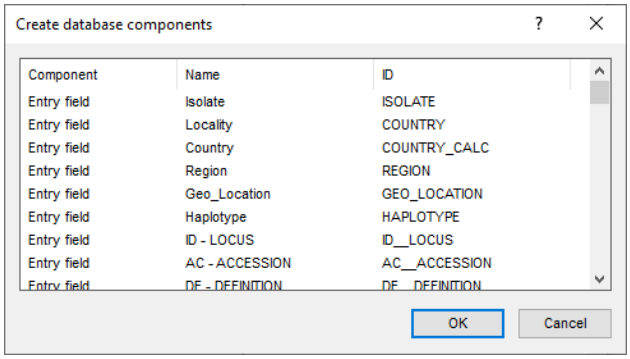

Figure 2.9: Create database components.

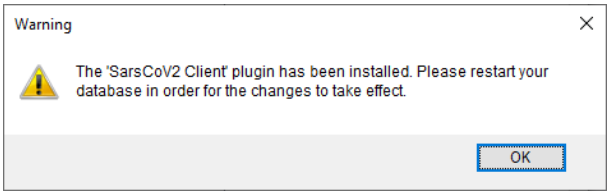

Figure 2.10: Installation confirmation.

The plugin is marked with a green V-sign in the *Plugins and Scripts* dialog box.

3.11 Close the *Plugins and Scripts* dialog box.

3.12 Close and reopen the database to activate the features of the *SARS-CoV-2 plugin*.

The *Main* window should now look like Figure 2.11.

The *SARS-CoV-2 plugin* installs menu items in the main menu of the software under ***SARSCoV2*** (see Figure 2.12) and following components:

- A character type called **SNP**, for the storage of the SNPs.
- A character type called **SNP_TRANSL**, for the storage of the translated SNPs.
- A sequence type called **genome**, for the storage of the (assembled) whole genome.
- 27 sequence types, for the storage of the extracted subsequences.
- 33 information fields, comprised of standard GenBank meta data fields and NCBI's SARS-CoV-2 data hub columns.

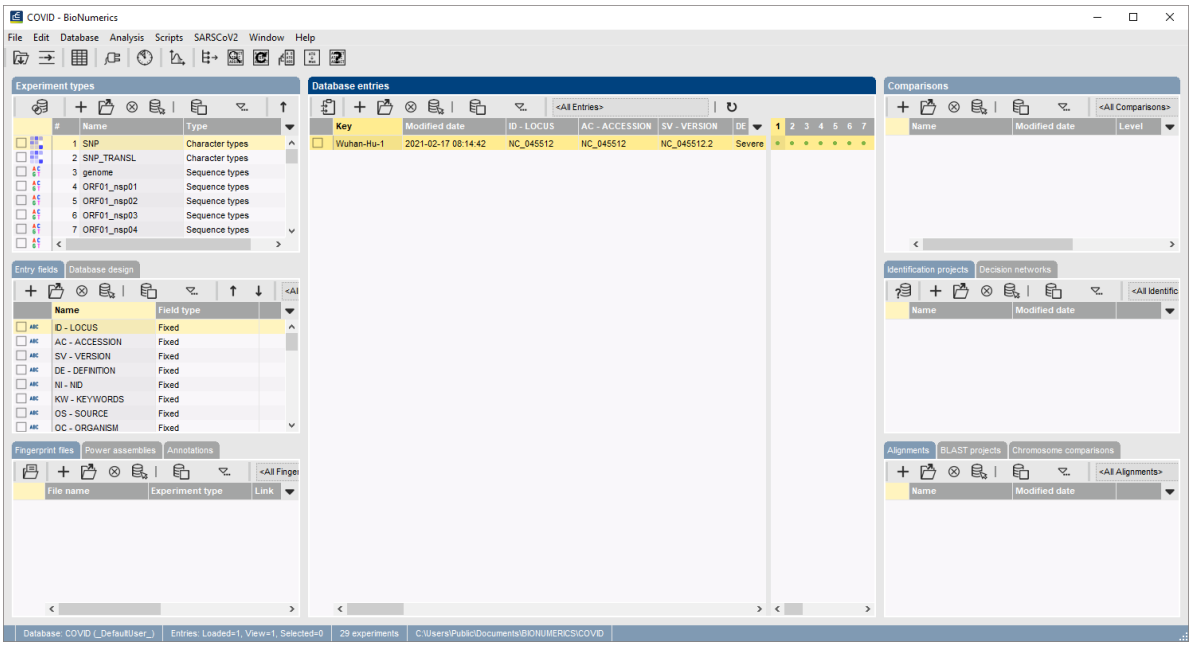

Figure 2.11: The *Main* window after installation of the plugin.

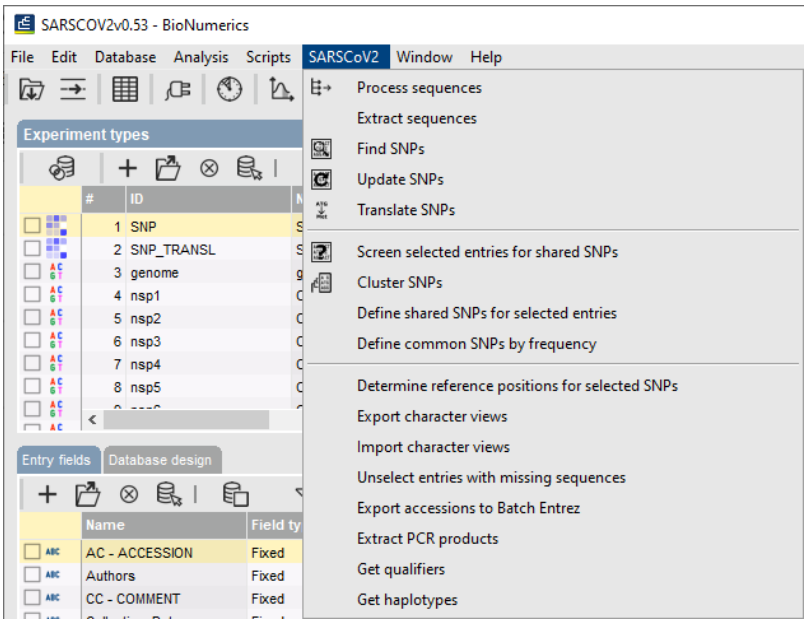

Figure 2.12: New menu items after installation of the plugin.

One entry is present in the database, with key **Wuhan-Hu-1**. The NCBI reference sequence for SARS-CoV-2, i.e. **NC_045512**, is stored in the sequence type **genome** for this entry.

3.13 Click on the green colored dot in the *Experiment presence* panel, corresponding to the **genome** experiment (i.e. the third column in default configuration) to open the *Sequence editor* window.

The sequence is displayed in the upper panel and a graphical representation of the sequence is displayed in the panel below (see Figure 2.13). The *Annotation* panel holds the NCBI features, and the header information is stored in the *Header* panel.

3.14 Close the *Sequence editor* window.

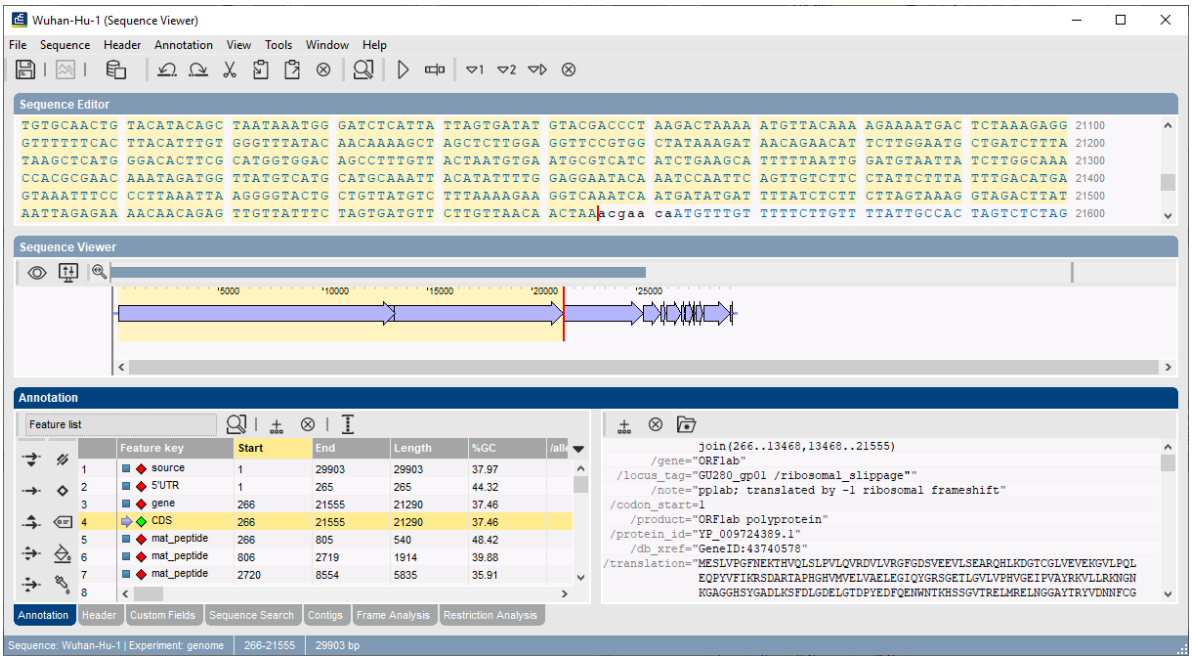

Figure 2.13: The *Sequence editor* window.

The subsequences found on the NCBI reference sequence for SARS-CoV (accession **NC_045512**) are stored in the corresponding destination sequence type experiments. These sequence types are composed of the tag **ORF** (Open Reading Frame) followed by a number and optionally a **nsp** (Nuclear Shuttle Protein) tag. For example: **ORF01_nsp01**. These subsequences are used as reference sequences for the BLAST search when sample sequences are screened (see 4.2).

3.15 To display the **ID** column next to the sequence type ***Name***, click on the column properties button in the header of the *Experiment types* panel and select ***Set active fields***.

3.16 Check **ID** and press <***OK***> (see Figure 2.14).

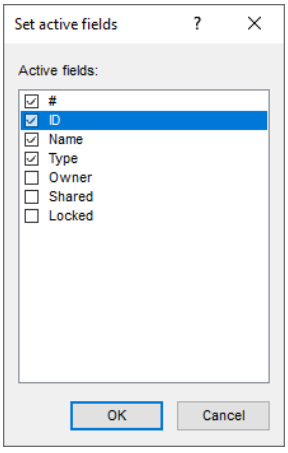

Figure 2.14: Show the ID column.

The **ID** column is now displayed in the *Experiment types* panel (see Figure 2.15).

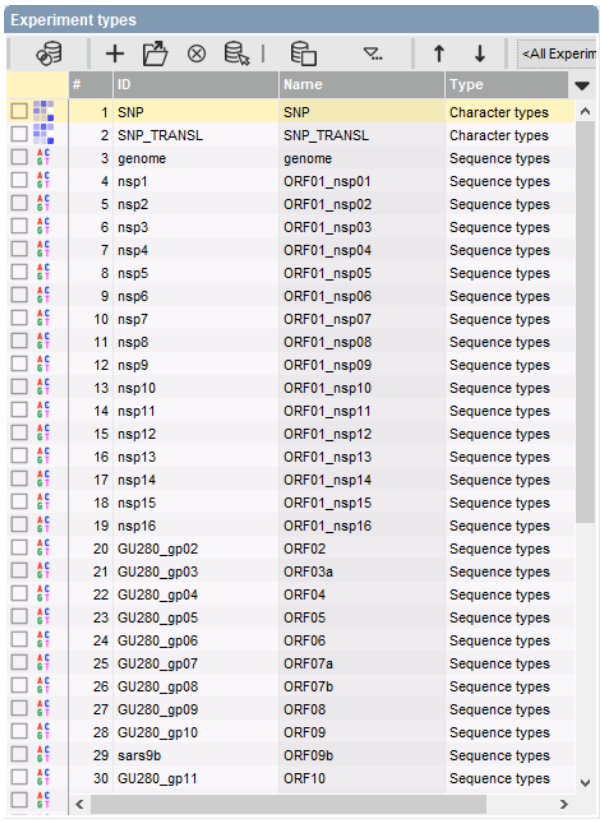

Figure 2.15: The *Experiment types* panel with the ID column displayed.

Chapter 3

Importing sequences

Genomic sequences can be imported into the database using the import routines available in BIONUMERICS.

0.1 Select ***File*** > ***Import...*** (, **Ctrl+I**) to call the *Import data* wizard.

0.2 Highlight the option <***Manual selection***> and press <***Next***>.

All import routines that import (assembled) genome sequences in BIONUMERICS are bundled under the ***Sequence type data*** topic.

0.3 To display all sequence import routines, expand the tree by pressing the "+" sign next to ***Sequence type data*** (see Figure 3.1).

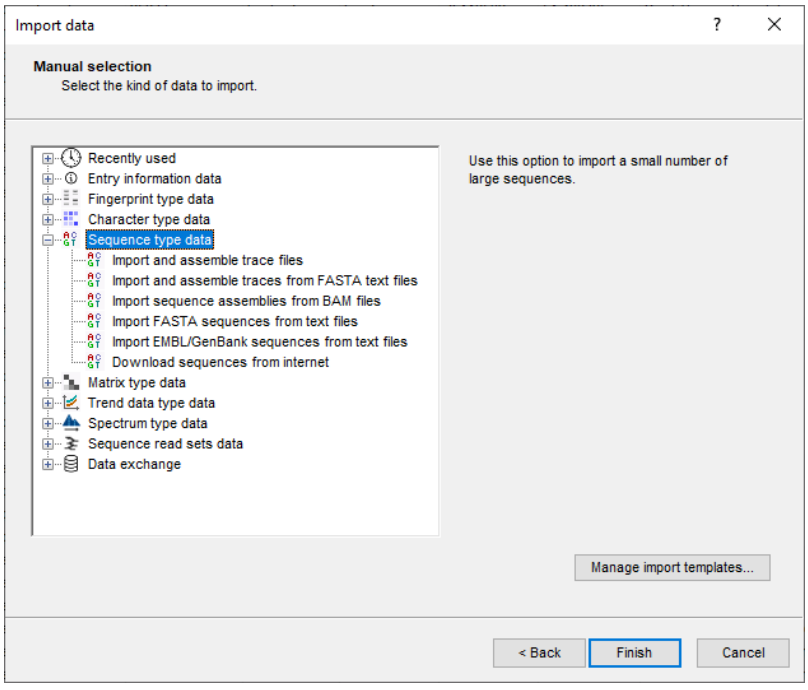

Figure 3.1: The Import tree.

As an example, we will fetch some sequences from EMBL/NCBI. More detailed information about the other sequence import routines can be found in the sequence tutorials available on our website.

A SarsCoV2 import template can be downloaded from the sample data download page on the BIONUMERICS website (https://www.bionumerics.com/download/sample-data, "COVID-19 im-

port template"). With this import template, NCBI/EMBL tags are mapped to entry fields created by the plugin.

0.4 In the *Import data* wizard, select <***Manage import templates***>.

0.5 Select <***Import from file***>, browse for the `SarsCoV2 template.xml` file and press <***OK***> (see Figure 3.2).

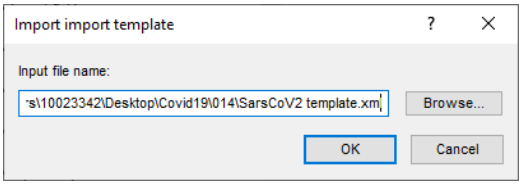

Figure 3.2: Import template.

The import template maps the EMBL/NCBI tags to the entry fields created by the *SARS-CoV-2 plugin*.

0.6 Press <***OK***> to add the import template to the database and close the dialog.

0.7 In the *Import data* wizard, choose the option ***Download sequences from internet*** under the ***Sequence type data*** item in the tree and click <***Finish***>.

0.8 Enter the accession codes (e.g. "MT385458,MT385436,MT385431") in the ***Accession codes*** input field, separated by the separation character ",".

0.9 Specify "," as the ***Separation character*** and choose one of the available download sites from the list, e.g. ***EBI*** (see Figure 3.3).

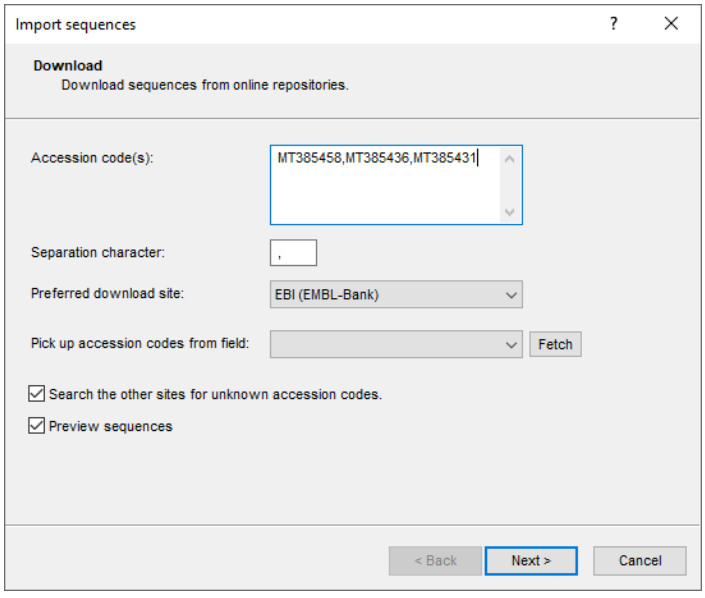

Figure 3.3: Download sequences.

0.10 With the option ***Preview sequences*** checked, press <***Next***>.

The import routine fetches the sequences from the selected database and shows detailed information in the next step (see Figure 3.4).

0.11 Press <***Next***>.

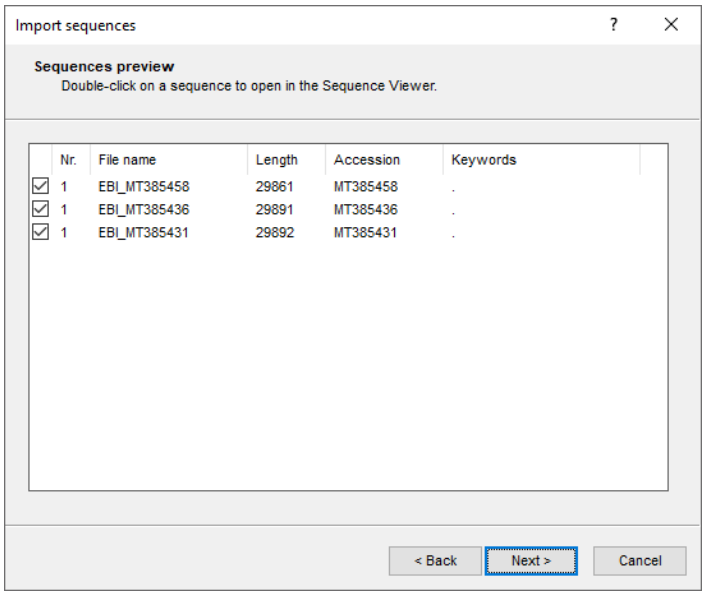

Figure 3.4: Fetched sequences.

The next step of the import wizard lists the templates that are present to import sequence information in the database. The predefined import template that was imported in the database in one of the previous steps is listed (see Figure 3.5).

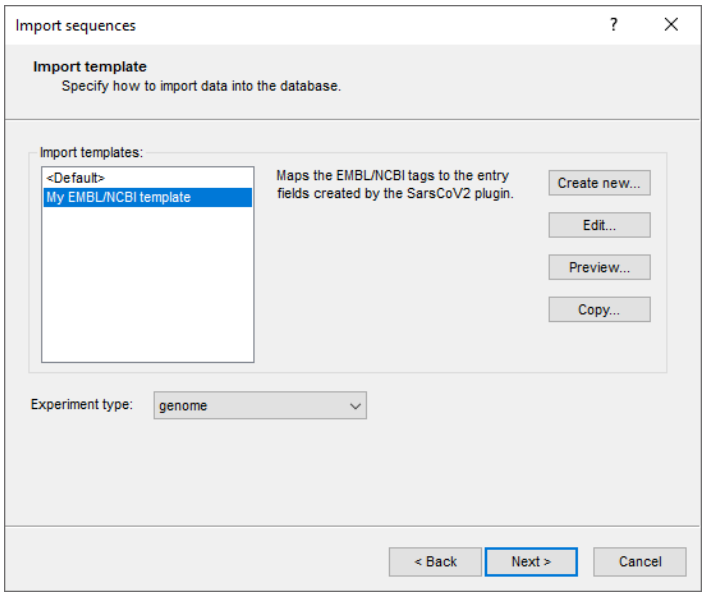

Figure 3.5: Import template.

0.12 Make sure the ***My EMBL/NCBI template*** is selected and press the <***Preview***> button to check the mapping (see Figure 3.6).

0.13 Close the preview.

0.14 Make sure ***My EMBL/NCBI template*** and **genome** are selected and press <***Next***>.

0.15 Press <***Finish***>. Confirm the import.

The entries are created and are automatically selected. The entry fields are updated and the

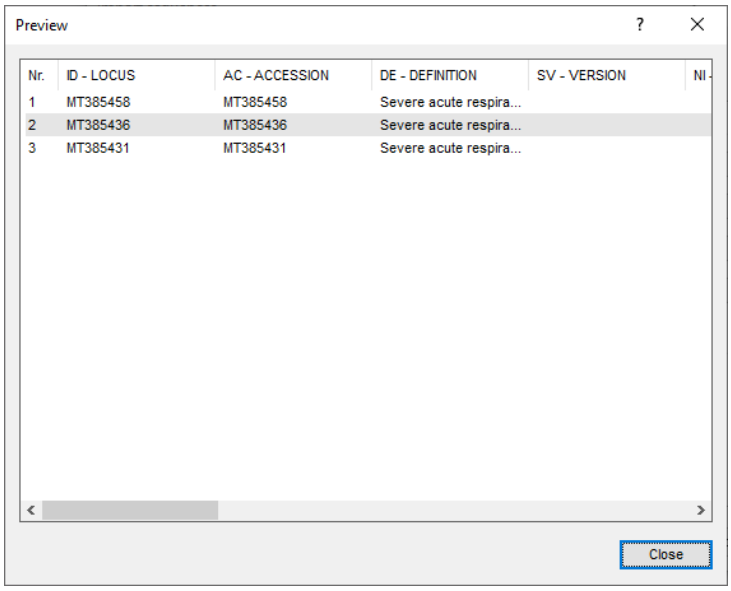

Figure 3.6: Preview.

sequences are stored in the **genome** experiment (see Figure 3.7).

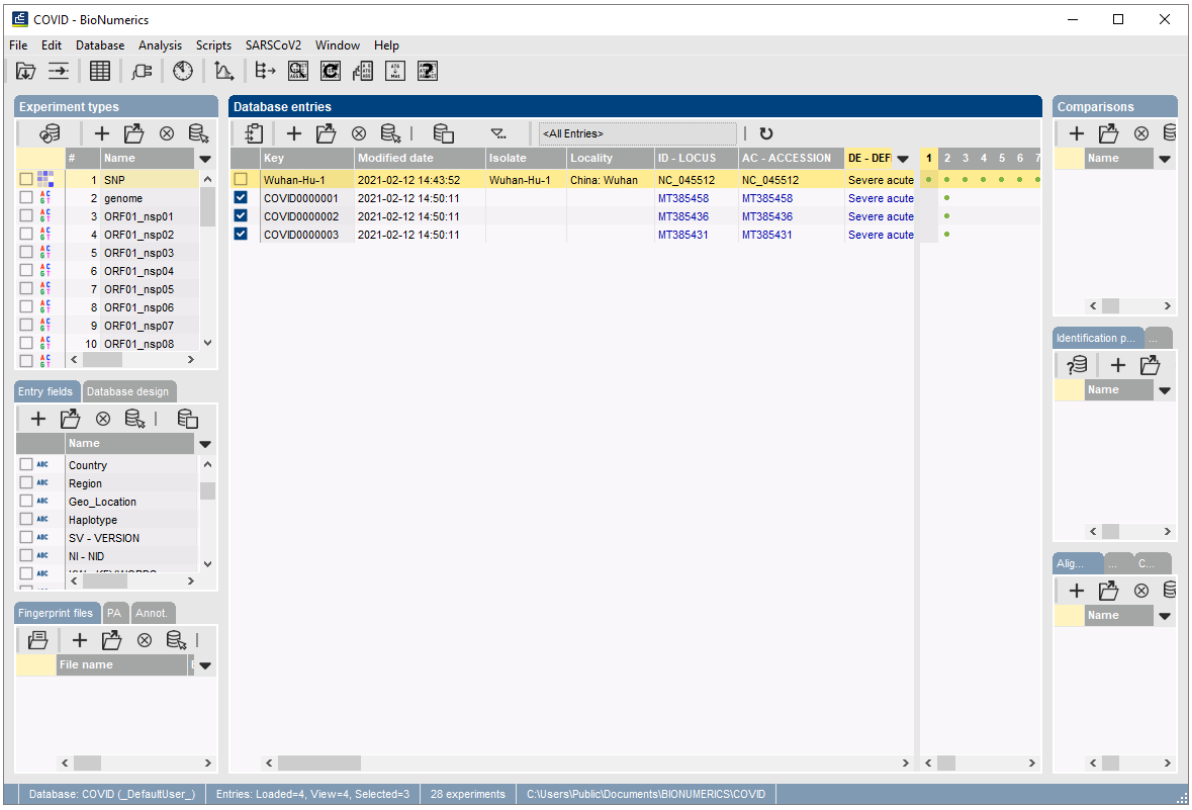

Figure 3.7: The *Main* window after import of three sample sequences.

Chapter 4

Processing sequences

4.1 Procedure

Genomic sequences, imported and stored in the ***genome*** experiment (see 3) can now be processed with the *SARS-CoV-2 plugin*.

1.1 In the *Database entries* panel of the *Main* window, select the entries you wish to process using the **Ctrl**-key. Alternatively, use the check box next to the entries in the *Database entries* panel.

1.2 Select ***SARSCoV2*** > ***Process sequences*** or click the  button to start the processing of the sequences.

The processing includes the following actions:

1. Extract 27 subsequences from the genomic sequence stored in the ***genome*** experiment (where possible) and save these sequences in the corresponding destination experiments (see 4.2).
2. Screen the 27 subsequences for SNPs with the reference sequence (see 4.3).
3. Add the new SNP positions to the character set of previously processed entries (see 4.3).
4. Translate the detected SNPs (see 4.4).
5. Screen the selected entries for shared SNPs (see 4.5).

These five actions can also be performed independently by selecting the corresponding menu-items (i.e. ***SARSCoV2*** > ***Extract sequences***, ***SARSCoV2*** > ***Find SNPs***, ***SARSCoV2*** > ***Update SNPs***, ***SARSCoV2*** > ***Translate SNPs*** and ***SARSCoV2*** > ***Screen selected entries for shared SNPs***).

We highly recommend to re-process all entries analyzed with a *SARS-CoV-2 plugin* version lower than 0.54 for consistent results.

4.2 Sequence extraction

The *SARS-CoV-2 plugin* uses a BLAST approach to extract subsequences from the sequence stored in the **genome** experiment. The subsequences of entry **Wuhan-Hu-1** are used as reference sequences for the BLAST search.

The subsequences found on the genome sequences of the selected entries, are stored in the corresponding destination sequence type experiments. These sequence types are composed of the tag **ORF** (Open Reading Frame) followed by a number and optionally a **nsp** (Nuclear Shuttle Protein) tag. For example: **ORF01_nsp01**.

2.1 In the *Database entries* panel of the *Main* window, select the entries you wish to process using the **Ctrl**-key. Alternatively, use the check box next to the entries in the *Database entries* panel.

2.2 Select ***SARSCoV2*** > ***Extract sequences*** to open the *Extract sequences* dialog box (see Figure 4.1.

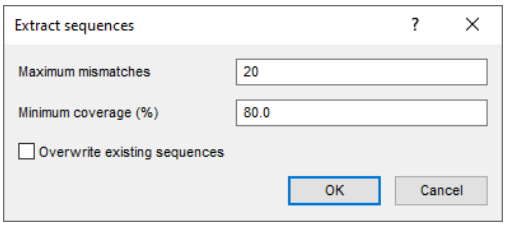

Figure 4.1: The *Extract sequences* dialog box.

The *Extract sequences* dialog box allows you to adjust the maximum number of allowed mismatches and the minimum sequence coverage).

2.3 Click on <***OK***>.

The *Select experiment(s)* dialog box pops up and allows you to select ORFs of interest to extract from the genomes of the selected entries (see Figure 4.2).

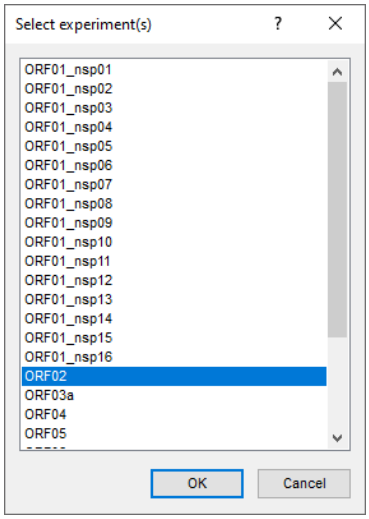

Figure 4.2: The *Select experiment(s)* dialog box.

2.4 In the *Select experiment(s)* dialog box, select the ORFs which you wish to extract from the genomes of the selected entries using the **Ctrl**-key. Alternatively, select all ORFs by using the **Ctrl + A** shortcut.

2.5 Click on <***OK***> to start the sequence extraction.

A message box will pop up indicating how many sequences were extracted and saved in the database (see Figure 4.3).

2.6 Click on <***OK***>.

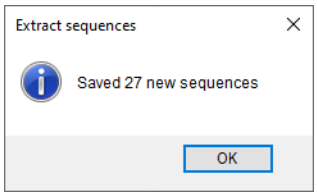

Figure 4.3: Extracted sequences.

2.7 Click on a green colored dot in the *Experiment presence* panel for one of the **ORF** sequence type experiments of the selected entries.

This action opens the *Sequence editor* window, containing the extracted sequence (see Figure 4.4 for an example).

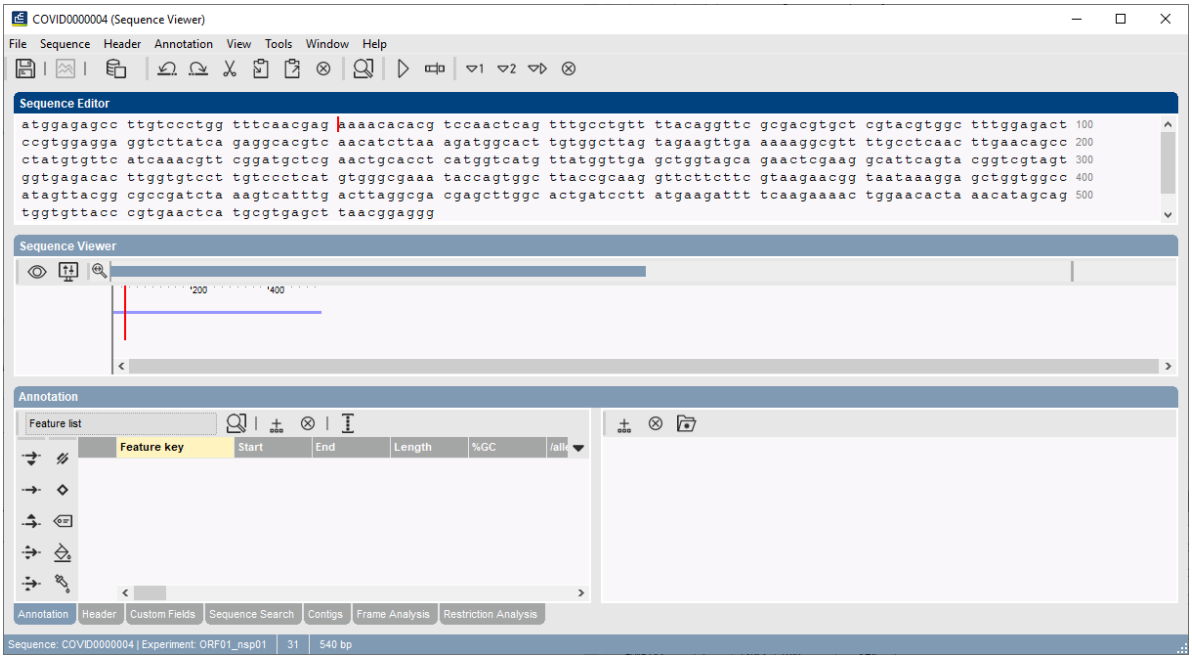

Figure 4.4: The *Sequence editor* window containing an extracted sequence.

2.8 Close the *Sequence editor* window.

4.3 Calculating SNPs

After the subsequence extraction (see 4.2), the subsequences can be screened for SNPs:

3.1 Make a selection of entries in the *Database entries* panel of the *Main* window using the **Ctrl**-key. Alternatively, use the check box next to the entries in the *Database entries* panel.

3.2 Select ***SARSCoV2*** > ***Find SNPs*** or press the button.

After the SNP screening, a message box pops up displaying the number of detected SNPs (see Figure 4.5). If new SNP positions are detected, this is also reported.

3.3 Click on <***OK***> to close the confirmation dialog.

Figure 4.5: SNP information.

If new SNPs were detected, a dialog box pops up asking if you would like to add the new SNPs to the entries already processed (see Figure 4.6).

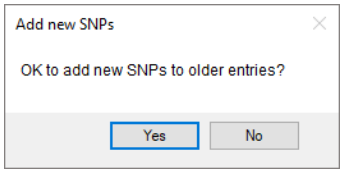

Figure 4.6: Update SNPs.

3.4 Click on <***OK***> to add the new SNP positions to the SNP character set of previously processed entries.

The SNP character sets of processed entries can be searched for missing SNPs at any time by selecting the entries of interest in the *Main* window and selecting ***SARSCoV2*** > ***Update SNPs*** or pressing the button.

3.5 Click on a green colored dot in the *Experiment presence* panel corresponding to the **SNP** character experiment of one the selected entries to open the character experiment card.

The character experiment card lists all SNPs detected for the sample. The bases are listed in the 'Mapping' column (see Figure 4.7).

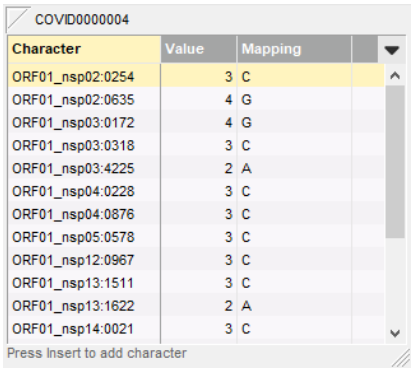

Figure 4.7: SNP character card.

3.6 Close the experiment card by clicking in the left upper corner of the card.

When new SNPs are detected the reference position is automatically determined and saved in the SNP character experiment type. This makes it easier to select SNPs that correspond to a published variant and to create a character view that can be used for entry screening.

3.7 In the *Main* window double-click on the character experiment type **SNP** in the *Experiment types* panel to call the *Character type* window.

The reference position for each SNP is indicated in the **RefPos** information field in the character experiment type (see Figure 4.8).

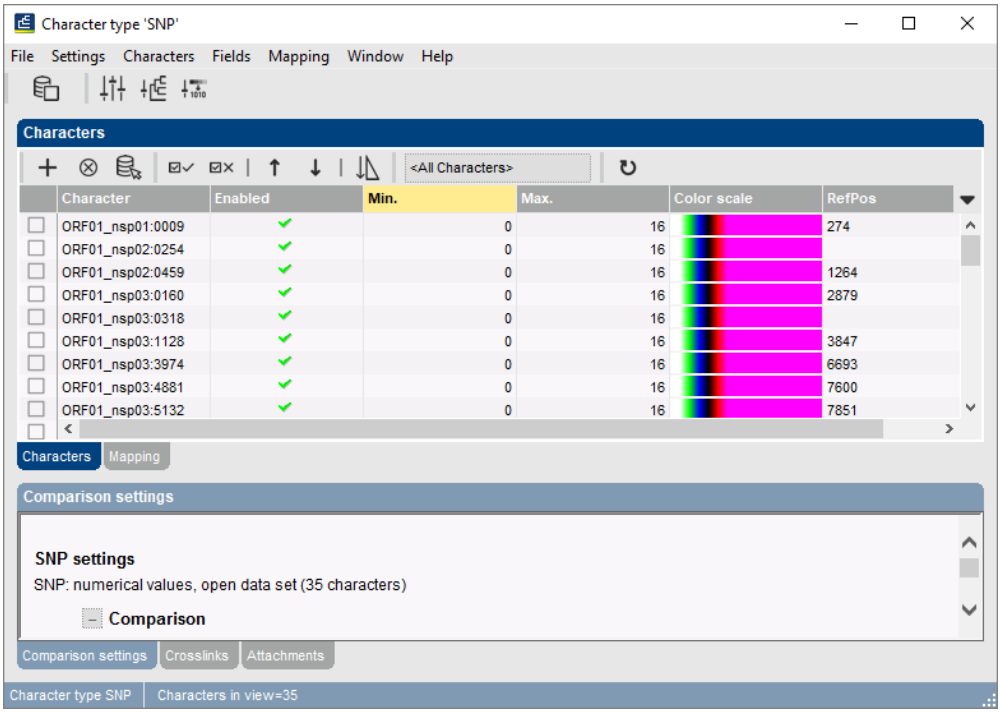

Figure 4.8: The SNP character experiment type with the **RefPos** information field .

The reference position of SNPs already present in the database can be determined by selecting the SNPs in the **SNP** character experiment type or *Comparison* window and selecting ***SARSCoV2*** > ***Determine reference positions for selected SNPs*** in the *Main* window.

3.8 Close the *Character type* window.

4.4 Translating SNPs

With ***SARSCoV2*** > ***Translate SNPs***, SNPs stored in the **SNP** character experiment, are translated.

4.1 Make a selection of entries in the *Database entries* panel of the *Main* window for which a SNP character experiment is available.

4.2 Select ***SARSCoV2*** > ***Translate SNPs*** or click on the button.

The *Select 'SNP' views* dialog box appears. The user can choose between the following character views:

- ***All characters***: All SNP positions present in the **SNP** character experiment.
- ***Selected characters***: All selected SNP positions in the **SNP** character experiment.
- ***common***: All common SNP positions (see 6.1).
- All user-defined character views in the SNP character experiment, e.g. character views which contain shared SNPs (see 4.5).

4.3 Select a character view from the list and click on <***OK***>.

The SNPs stored in the **SNP** experiment of the selected entries are translated and the amino acids are stored in the **SNP_TRANSL** experiment.

4.4 Click on a green colored dot in the *Experiment presence* panel corresponding to the **SNP_TRANSL** character experiment of one the selected entries to open the character experiment card.

The amino acids are listed in the **Mapping** column (see Figure 4.9).

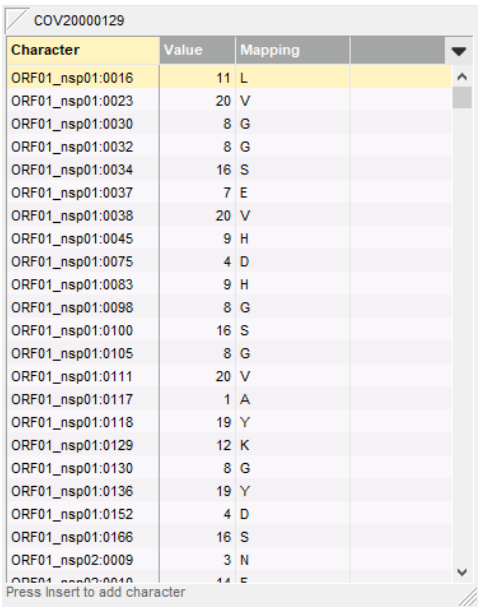

Figure 4.9: SNP_TRANSL character card.

4.5 Close the experiment card by clicking in the left upper corner of the card.

This translation tool assumes that the frame for each sequence starts at position 1.

4.5 Defining shared SNPs and screening for shared SNPs

The plugin allows to identify SNPs that are shared by selected entries. The shared SNPs can be saved as a character view in the SNP character experiment type and this character view can then be used to screen other entries in the database for the presence of these defined SNPs. This serves as a rapid screener for variants of concern, such as B.1.1.7, B.1.1.351, and P.1.

5.1 In the *Database entries* panel of the *Main* window, select the entries you wish to include in the SNP screening.

5.2 Select ***SARSCoV2*** > ***Define shared SNPs for selected entries***.

The *Select experiment(s)* dialog box appears. The user can choose between the character experiment which contains the detected SNPs (**SNP**) and the character experiment which contains the translated SNPs (**SNP_TRANSL**).

5.3 Select a character experiment from the list and click on <***OK***>.

5.4 Specify a name for the new SNP character view which will contain the SNPs shared between the selected entries (see Figure 4.10) and press <***OK***>.

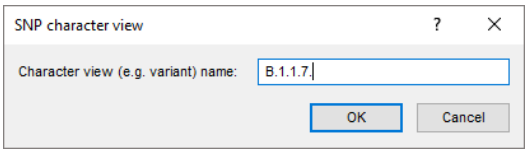

Figure 4.10: SNP character view.

5.5 Press <***OK***> to close the dialog.

The shared SNPs are saved to the newly created character view of the **SNP** or the **SNP_TRANSL** experiment:

5.6 In the *Main* window, double-click the character experiment type **SNP** or the **SNP_TRANSL** in the *Experiment types* panel to call the *Character type* window.

5.7 Click on the drop-down bar in the toolbar and select the newly created character view from the list.

The shared SNPs identified with the command ***SARSCoV2*** > ***Define shared SNPs for selected entries*** are listed (see Figure 4.11 for an example).

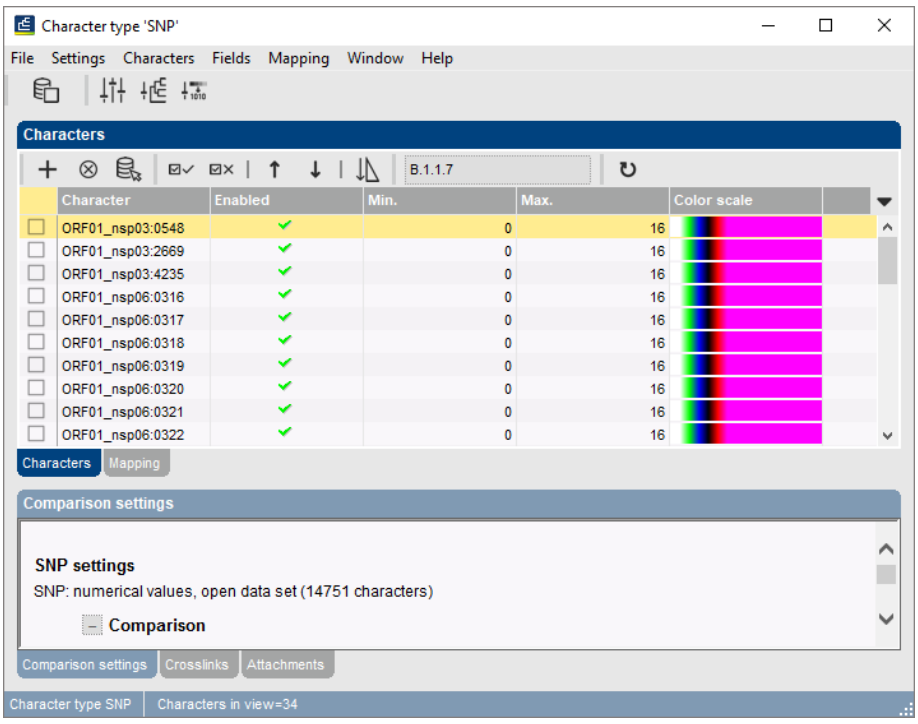

Figure 4.11: The **SNP** character experiment type with the shared SNPs listed in the newly created character view **B.1.1.7**.

5.8 Close the *Character type* window.

This character view can now be used to screen other entries for the presence of these defined SNPs. The screening will not exclude entries with additional SNPs.

5.9 In the *Database entries* panel of the *Main* window, select the entries you wish to include in the SNP screening.

5.10 Select ***SARSCoV2*** > ***Screen selected entries for shared SNPs*** or click on the button.

The *Select experiment(s)* dialog box appears. The user can choose between the character experiment which contains the detected SNPs (**SNP**) and the character experiment which contains the translated SNPs (**SNP_TRANSL**).

5.11 Select a character experiment from the list and click on <***OK***>.

The *Select character view* dialog box appears and the user can choose a character view which contains the SNPs of interest. Entries which contain the defined SNPs are automatically added to an entry view with the same name as the selected character view (see Figure 4.12 for an example).

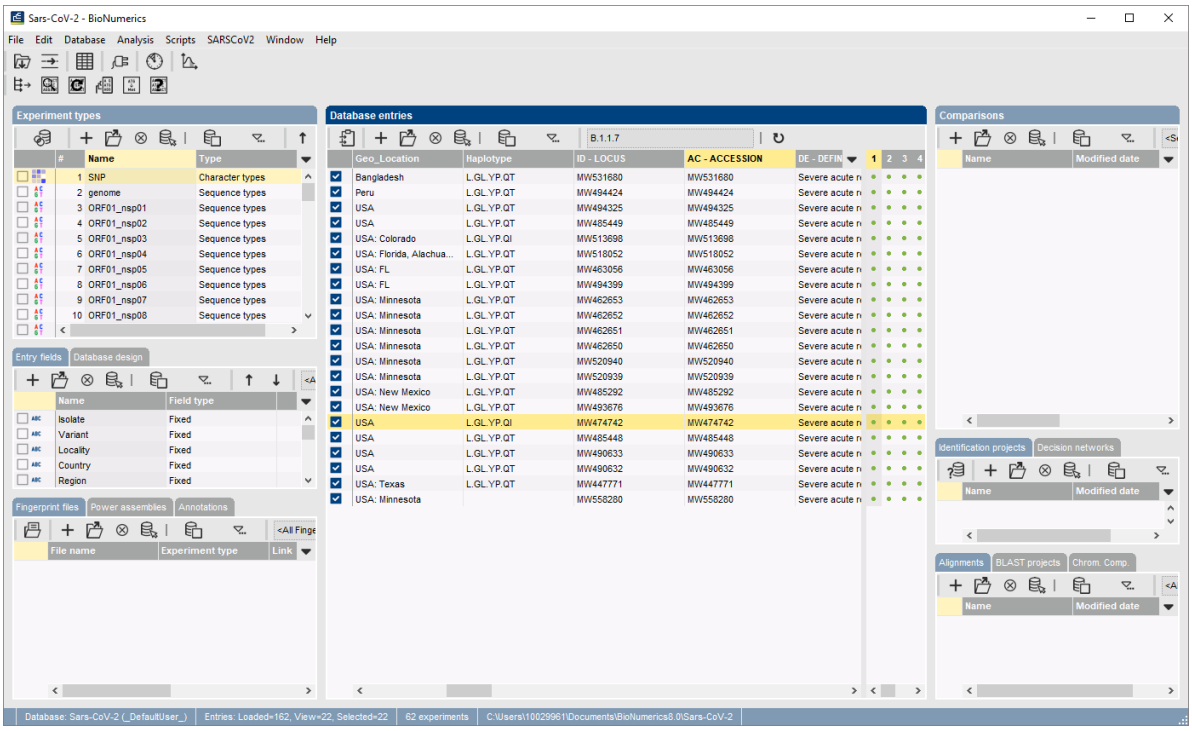

Figure 4.12: Screening of selected entries for the SNPs present in the **B.1.1.7** character view.

Chapter 5

Clustering SNP data

0.1 In the *Database entries* panel of the *Main* window, select the entries you wish to cluster.

Entries for which one ore more subsequences are missing, have an incomplete SNP character set and can be excluded from the comparison by selecting ***SARSCoV2*** > ***Unselect entries with missing sequences***.

0.2 Select ***SARSCoV2*** > ***Unselect entries with missing sequences*** if you do not want to include entries with missing sequences in your cluster analysis.

0.3 Select ***SARSCoV2*** > ***Cluster SNPs*** or click on the  button to cluster the selected entries.

The *Select experiment(s)* dialog box appears. The user can choose between the character experiment which contains the detected SNPs (i.e. **SNP**) and the character experiment which contains the translated SNPs (i.e. **SNP_TRANSL**).

0.4 Select a character experiment from the list and click on <***OK***>.

The *Select character view* dialog box appears (see Figure 5.1). The user can choose between the following character views:

- ***All characters***: All SNP positions present in the SNP character experiment will be included in the comparison.
- ***Selected characters***: All selected SNP positions in the SNP character experiment will be included in the comparison.
- ***common***: All common SNP positions (see 6.1) will be included in the comparison.
- All user-defined character views in the SNP character experiment, e.g. character views which contain shared SNPs (see 4.5).

If the option ***Select polymorphic characters*** is selected, the polymorphic SNP positions for the selected entries will be selected in the *Comparison* window.

When a character view is selected, the *Comparison* window appears and should look like Figure 5.2 with following settings:

- A similarity matrix is calculated based on the **SNP** or **SNP_TRANSL** experiment, using the ***Categorical (differences)*** similarity coefficient and is displayed in the *Similarities* panel.
- The dendrogram is calculated based on the ***Complete linkage*** clustering algorithm and is displayed in the *Dendrogram* panel.

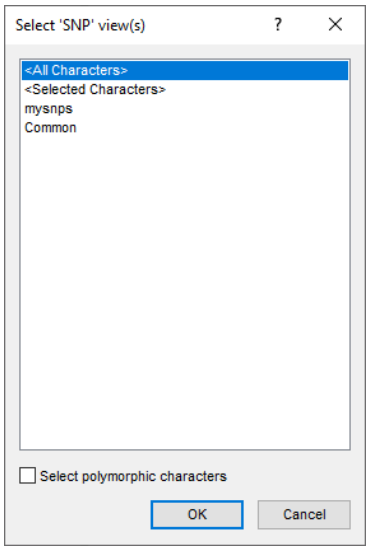

Figure 5.1: *Select character view* dialog box.

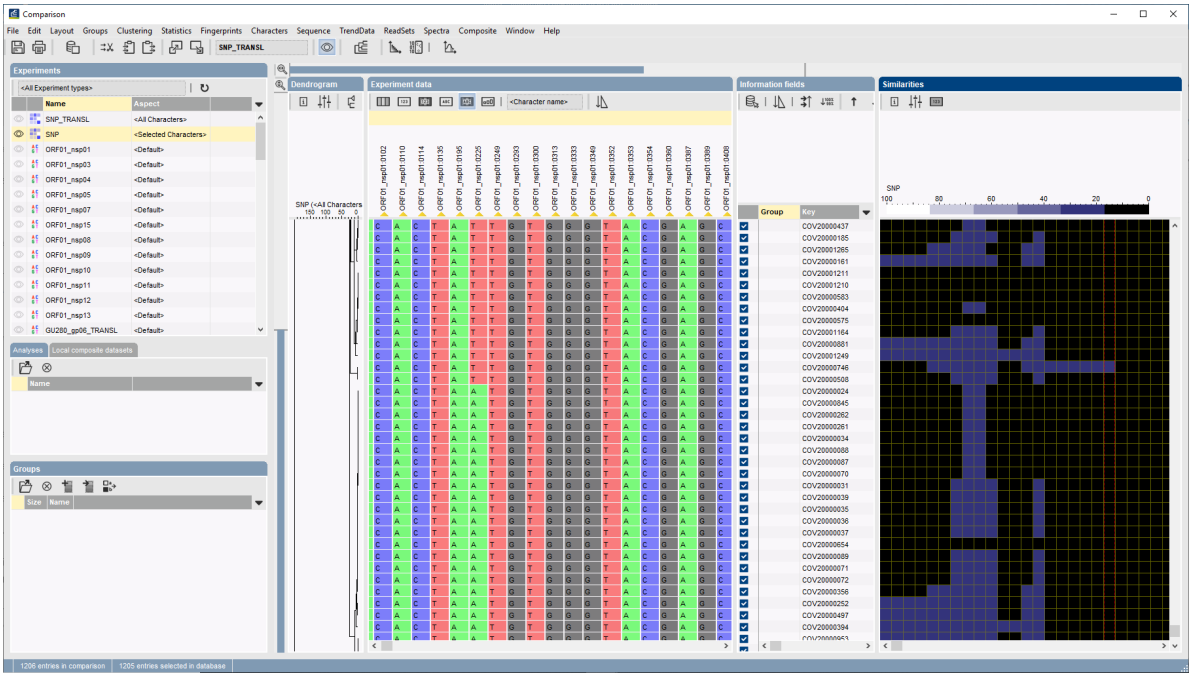

Figure 5.2: The *Comparison* window.

0.5 The settings used to calculate the dendrogram that is displayed in the *Dendrogram* panel can be called with ***Clustering*** > ***Show information*** ( ).

0.6 To view the number of SNP differences on the nodes, select ***Clustering*** > ***Dendrogram display settings...*** ( ), and tick the option ***Show node information***.

In the *Comparison* window, groups can be defined from clusters, from database fields (e.g. based on the geographic location or haplotype), or just from any selection.

0.7 To create groups based on a database field, right-click on the field name in the *Information fields* panel, and select ***Create groups from database field***. To create groups based on a selection of entries in the *Comparison* window, use the commands under the ***Groups*** menu item.

A minimum spanning tree in BIONUMERICS is calculated in the *Cluster analysis* window. This

window can be launched from the *Comparison* window:

0.8 Make sure **SNP** is selected in the *Experiments* panel of the *Comparison* window.

0.9 Select ***Clustering*** > ***Calculate*** > ***Advanced cluster analysis...*** or press the button and select ***Advanced cluster analysis*** to launch the *Network wizard dialog* wizard.

0.10 Select ***MST for categorical data***, and press <***Next***>.

The minimum spanning tree is calculated and displayed in the *Cluster analysis* window (see Figure 5.3 for an example).

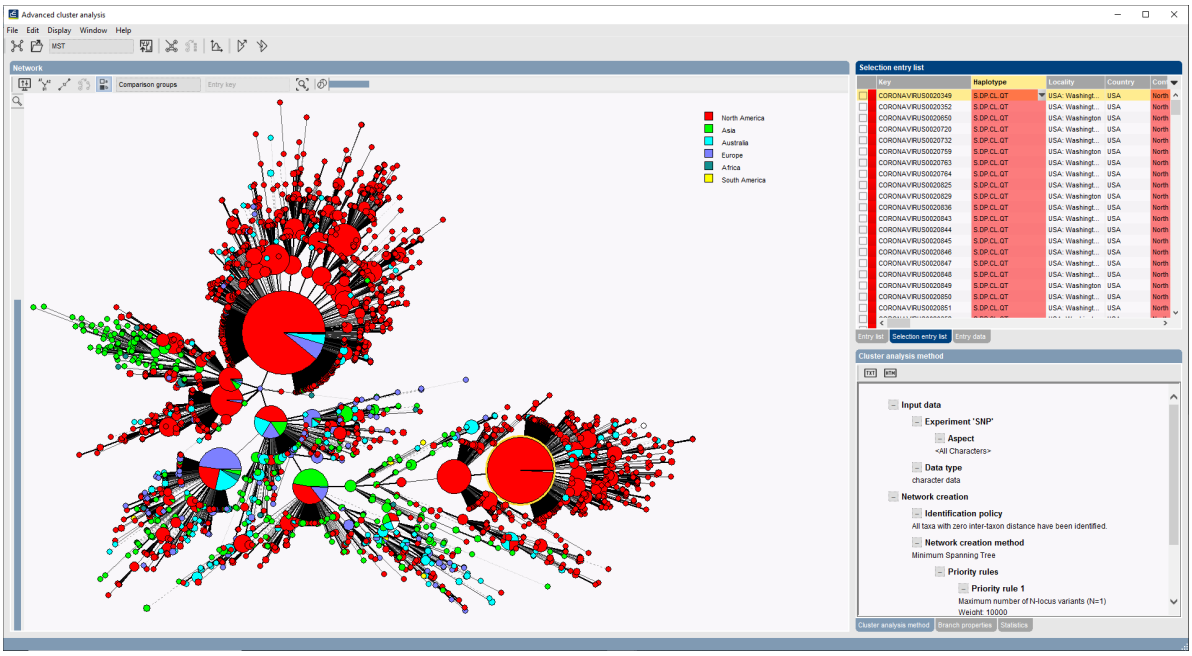

Figure 5.3: The *Cluster analysis* window.

0.11 Close the *Cluster analysis* window.

0.12 Save the comparison with ***File*** > ***Save as...*** and close the comparison with ***File*** > ***Exit***.

Chapter 6

Miscellaneous tools

6.1 Defining common SNPs

The plugin allows to define common SNPs, i.e. polymorphisms with a minimum frequency above a specified threshold.

1.1 In the *Database entries* panel of the *Main* window, select the entries you wish to include in the SNP screening.

1.2 Select ***SARSCoV2*** > ***Define common SNPs by frequency***.

The *Select experiment(s)* dialog box appears. The user can choose between the character experiment which contains the detected SNPs (**SNP**) and the character experiment which contains the translated SNPs (**SNP_TRANSL**).

1.3 Select a character experiment from the list and click on <***OK***>.

1.4 Specify the minimum frequency in the dialog (see Figure 6.1) and press <***OK***>.

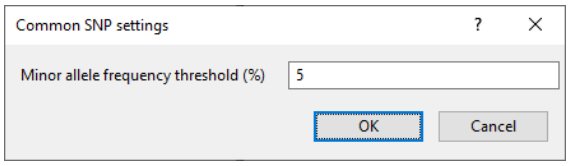

Figure 6.1: Specify threshold.

The number of common SNPs – identified based on the provided frequency – is displayed (see Figure 6.2 for an example).

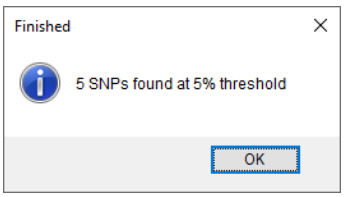

Figure 6.2: Result.

1.5 Press <***OK***> to close the dialog.

The common SNPs are saved to the **Common** character view of the **SNP** or the **SNP_TRANSL** experiment:

1.6 In the *Main* window double-click the character experiment type **SNP** or **SNP_TRANSL** in the *Experiment types* panel to call the *Character type* window.

1.7 Click on the drop-down bar in the toolbar and select the **common** character view from the list (see Figure 6.3).

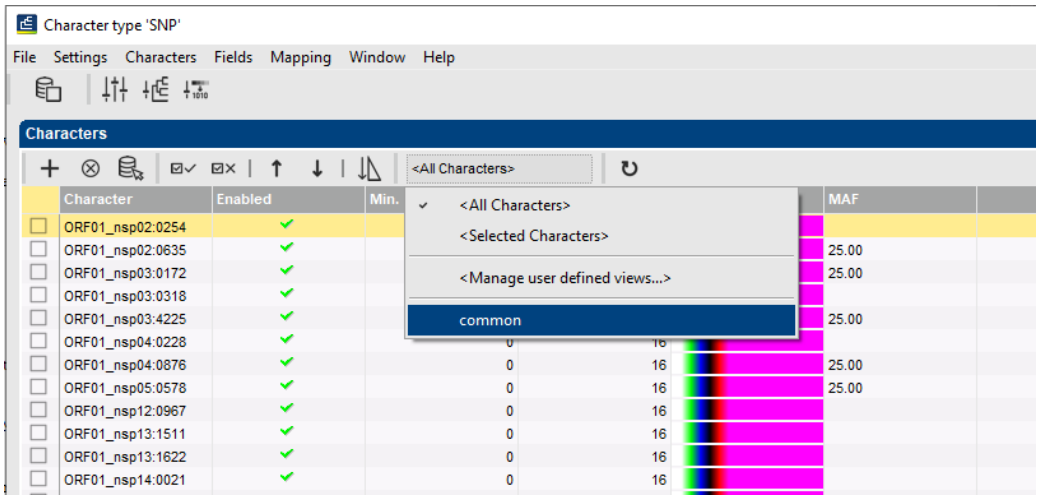

Figure 6.3: Character view.

The common SNPs identified with the command ***SARSCoV2*** > ***Define common SNPs*** are listed. The **MAF** character field displays the allele frequency.

1.8 Close the *Character type* window.

6.2 Exporting accessions to BLAST Entrez

There is a standard BIONUMERICS import tool available in the *Import data* wizard to download GenBank sequences from NCBI, but new sequences might not yet be available for import.

To retrieve GenBank-formatted sequences in bulk follow these steps:

2.1 In the *Database entries* panel of the *Main* window, select the entries you wish to export the accessions for.

2.2 Select ***SARSCoV2*** > ***Export accessions to Batch Entrez***.

2.3 Browse for an existing folder and press <***OK***>.

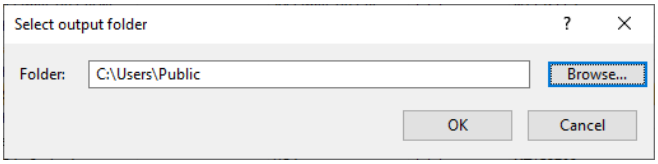

Figure 6.4: Browse for folder.

This command exports the accessions (stored in the **AC - ACCESSION** entry field) for selected entries to a space-delimited text file in the selected folder, and opens the NCBI BLAST Entrez

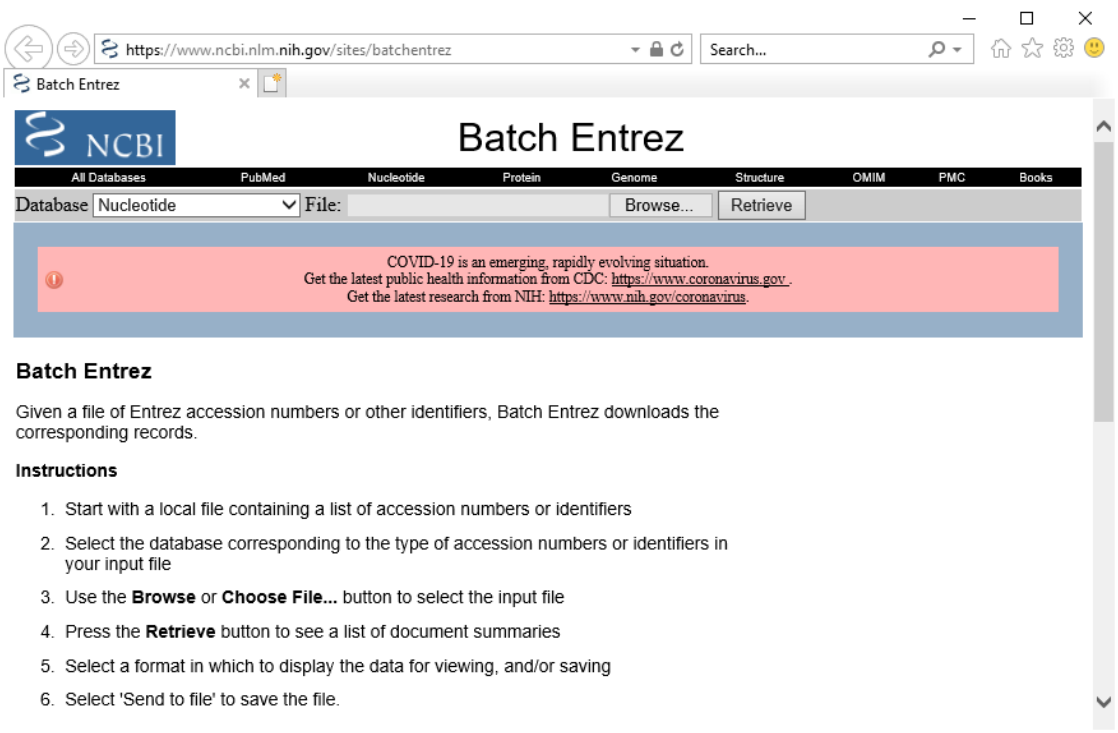

Figure 6.5: Batch Entrez.

website in the default browser (see Figure 6.5), from which the accessions file can be selected (with the <***Browse***> button) to retrieve GenBank-formatted sequences in bulk.

6.3 Extracting PCR products

With the *SARS-CoV-2 plugin*, PCR products can be extracted based on the WHO-standard primer sequences (https://www.who.int/publications/m/item/molecular-assays-to-diagnose-covid-19-sum

3.1 In the *Database entries* panel of the *Main* window, select the entries you wish to export the PCR products from.

3.2 Select ***SARSCoV2*** > ***Extract PCR products***.

The first time this menu item is selected, the sequence types are created and added to the *Experiment types* panel.

The extracted PCR products are stored in the corresponding PCR sequence type experiments (see Figure 6.6).

In the *Comparison* window, one can further analyze specific PCR products:

3.3 In the *Database entries* panel of the *Main* window, select the entries you wish to analyze.

3.4 Highlight the *Comparisons* panel in the *Main* window and select ***Edit*** > ***Create new object...*** (+) to create a new comparison for the selected entries.

3.5 Click on the icon next to the sequence type in the *Experiments* panel to display the sequences in the *Experiment data* panel (see Figure 6.7 for an example).

The sequences can be further analyzed using the sequence analysis tools available in BIONUMERICS.

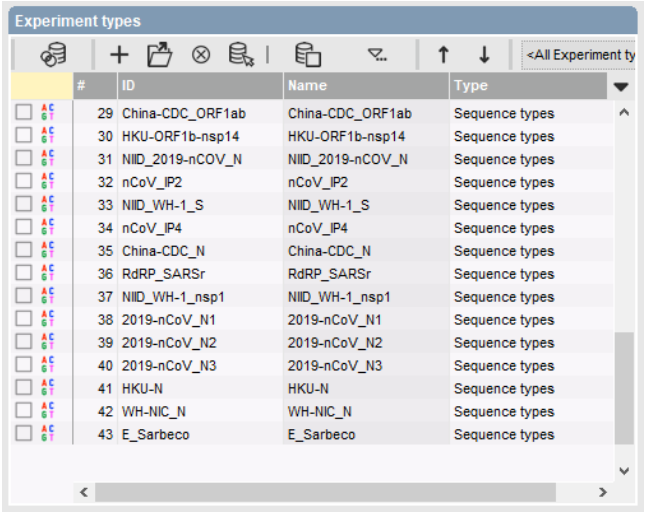

Figure 6.6: PCR sequence type experiments.

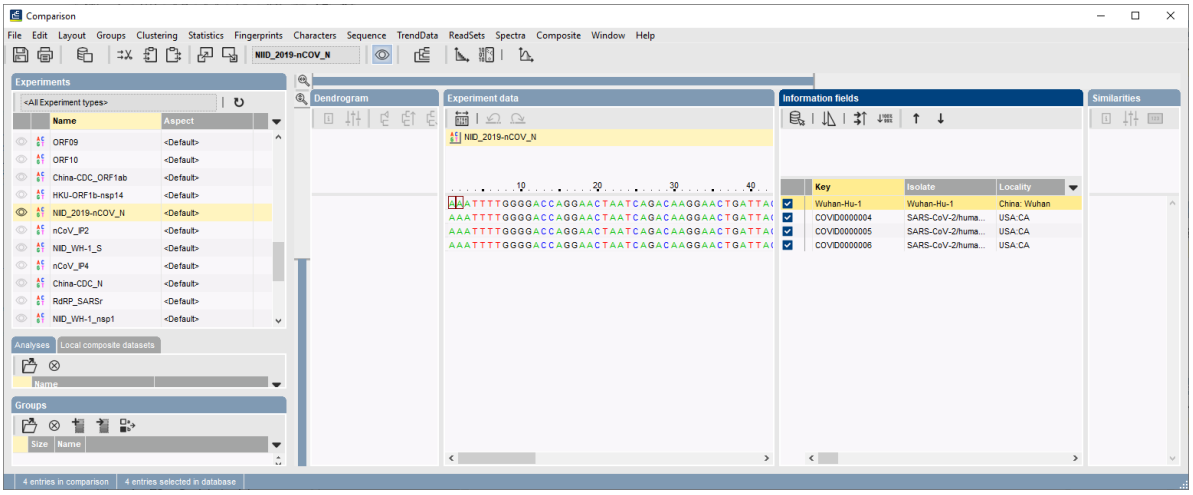

Figure 6.7: PCR products.

6.4 Get qualifiers

Meta data can be parsed from the sequence annotations – if present – and stored in the **Isolate** and **Locality** entry fields (see Figure 6.8).

4.1 In the *Database entries* panel of the *Main* window, select the entries for which you want to extract the isolate and locality information.

4.2 Select ***SARSCoV2*** > ***Get qualifiers***.

Using the *calculated field* option in BIONUMERICS, information stored in the ***Locality*** entry field (e.g. China:Wuhan or USA:CA) can be parsed to only contain the country information (e.g. China and USA respectively). More information on how to create calculated fields can be found in the reference manual.

Figure 6.8: Locality and isolate information extraction.

6.5 Haplotype determination

The plugin allows to determine haplotypes. The haplotype as defined in the *SARS-CoV-2 plugin* is a set of high-frequency amino acid substitutions, organized by linkage groups. They are ordered from left to right by the date on which they first appeared.

Position (genome)	ORF8: 251	ORF2: 1841	ORF1nsp12: 941	ORF1nsp13: 1622	ORF1nsp13: 1511	ORF3a: 171	ORF1nsp2: 254
Ancestral allele	S	D	P	Y	P	Q	T
Derived allele	L	G	L	C	L	H	I

Table 6.1: High-frequency amino acid substitutions.

5.1 In the *Database entries* panel of the *Main* window, select the entries for which you want to determine the haplotype.

5.2 Select ***SARSCoV2*** > ***Get haplotypes***.

The result is displayed in the **Haplotype** entry field (see Figure 6.9).

Figure 6.9: Haplotype determination.

Optionally colors can be assigned to every haplotype:

5.3 Right-click on the **Haplotype** information field in the *Database entries* panel and choose ***Field properties*** from the floating menu (see Figure 6.10).

The *Database field properties* dialog box appears.

5.4 Press <***Add all***> to create all existing states for the **Haplotype** field. Confirm the action.

5.5 Check ***Use colors*** to display a specific color code for each field state.

5.6 Press <***OK***> to accept the new settings.

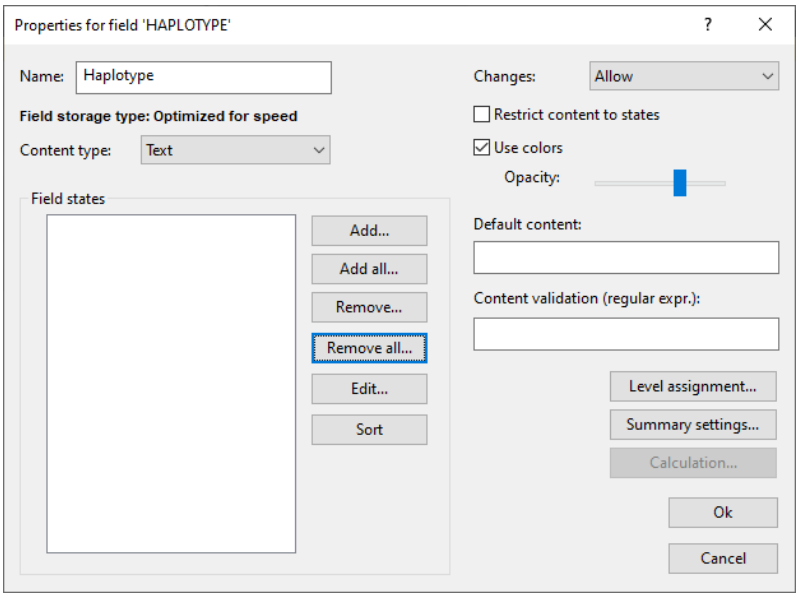

Figure 6.10: Haplotype field properties.

The *Database entries* panel is updated (see Figure 6.11).

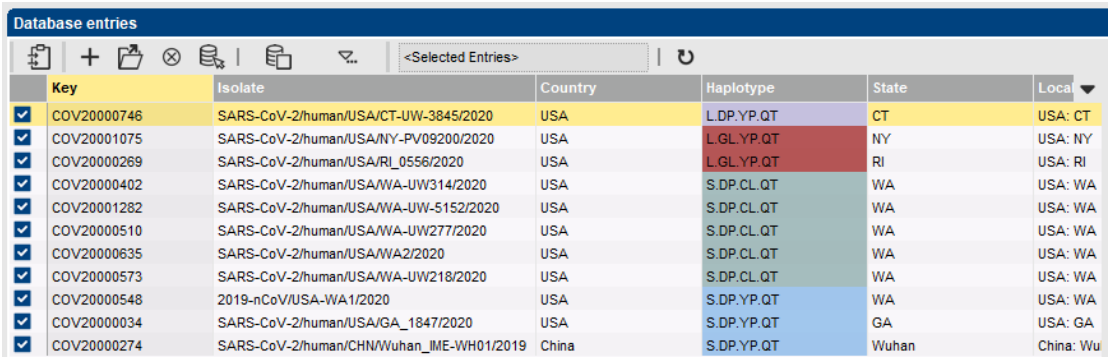

Figure 6.11: Haplotype information field in the *Database entries* panel.

6.6 Exporting and importing character views

The default and user-created character views of the **SNP** and **SNP_TRANSL** character experiments can be exported and imported into the BIONUMERICS database. This allows users to easily share character views between BIONUMERICS databases.

6.1 Select ***SARSCoV2*** > ***Export character views***.

6.2 In the *File selection* dialog box that appears, accept the default proposed file or click the <***Browse...***> button to browse for a preferred export file (see Figure 6.12). Click <***OK***>.

The *Select experiment(s)* dialog box appears. The user can choose between the character experiment which contains the detected SNPs (**SNP**) and the character experiment which contains the translated SNPs (**SNP_TRANSL**).

6.3 Select a character experiment from the list and click <***OK***>.

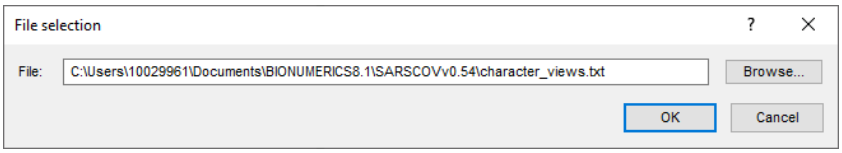

Figure 6.12: The *File selection* dialog box.

The *Select 'SNP' views* dialog box appears. The user can choose between the following character views:

- ***All characters***: All SNP positions present in the SNP character experiment.
- ***Selected characters***: All selected SNP positions in the SNP character experiment.
- ***common***: All common SNP positions (see 6.1).
- All user-defined character views in the SNP character experiment, e.g. character views which contain shared SNPs (see 4.5).

6.4 Select one or multiple character views from the list and click on <***OK***>.

The exported text file will contain the SNP positions of the selected character views. The exported character views can now be imported into another BIONUMERICS database. This can be done by selecting ***SARSCoV2*** > ***Import character views***.

6.5 In another database, select ***SARSCoV2*** > ***Import character views*** and browse for the exported text file. Click <***OK***>.

An information dialog box pops up indicating the amount of imported characters and character views (see Figure 6.13).

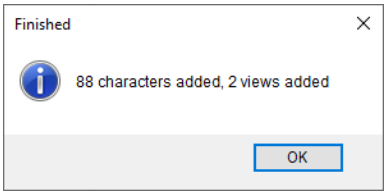

Figure 6.13: Information dialog box indicating the amount of imported characters and character views.

bioMérieux S.A. • 69280 Marcy l'Étoile • France • Tel.: + 33 (0)4 78 87 20 00 • Fax: +33 (0)4 78 87 20 90
www.biomerieux.com