BioNumerics®

MICROBIAL DATA ANALYSIS SOFTWARE

PLUGINS

VERSION 7.6

www.applied-maths.com

Contents

NOTES

SUPPORT BY APPLIED MATHS

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 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.applied-maths.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 2.7.4 release from the Python Software Foundation, http://www.python.org/
- Xerces library for XML input and output from the Apache Software Foundation, https://xerces.apache.org/
- NCBI toolkit version 2.2.28, http://www.ncbi.nlm.nih.gov/BLAST/
- SRA Toolkit, https://ncbi.github.io/sra-tools/
- Boost c++ libraries, http://www.boost.org/
- Samtools for interacting with SAM / BAM files, http://www.htslib.org/download/
- 7-Zip (7za.exe), http://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, http://cairographics.org/
- Crypto++ library version 5.5.2, http://www.cryptopp.com/
- OpenSSL library, https://www.openssl.org/
- libSVM library for Support Vector Machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm/
- SQLite version 3.7.17, http://www.sqlite.org/
- pymzML Python module for high throughput bioinformatics on mass spectrometry data, https://github.com/pymzml/pymzML
- Numpy Python library version 1.8.1, http://www.numpy.org/
- BioPython Python library version 1.64, http://www.biopython.org/
- PIL Python library version 1.1.7, http://www.pythonware.com/products/pil/
- Chromium Embedded Framework, https://bitbucket.org/chromiumembedded/cef/wiki/Home
- SPAdes genome assembler version 3.13.1, http://bioinf.spbau.ru/spades *
- SKESA version 2.3.0, https://github.com/ncbi/SKESA/releases
- Unicycler version 0.4.8, https://github.com/rrwick/Unicycler/releases *
- Velvet for Windows, source code can be downloaded from https://www.applied-maths.com/download/open-source
- Ray for Windows, source code can be downloaded from https://www.applied-maths.com/download/open-source
- Bowtie2 version 2.2.5 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml)*
- SNAP version 1.0.18, http://snap.cs.berkeley.edu/
- RAxML version 8.2.11, https://github.com/stamatak/standard-RAxML/releases
- FastTree version 2.1.10, http://www.microbesonline.org/fasttree/

- CFSAN SNP pipeline version 0.8.2, https://github.com/CFSAN-Biostatistics/snp-pipeline *
- Prokka version 1.12, https://github.com/tseemann/prokka *

*: On Calculation Engine only

Chapter 1

Introduction

The *SarsCoV2 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 *SarsCoV2 plugin* is a free add-on. The minimum configuration for installation of the plugin includes the Character data, Sequence data, Tree and Network inference and Genome Analysis Tools modules.

Chapter 2

Starting and setting up BIONUMERICS

2.1 Startup program

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

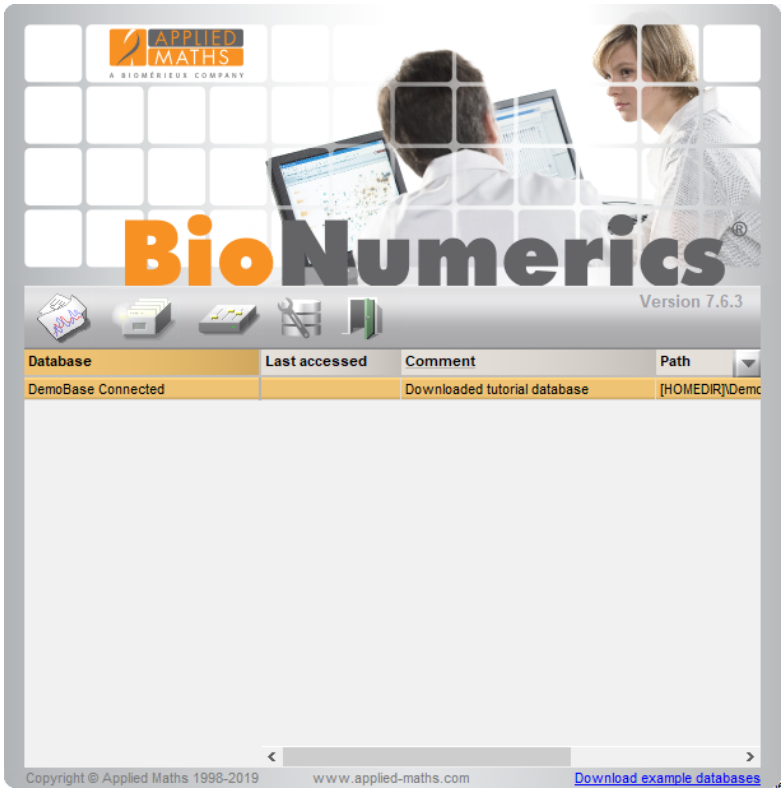


Figure 2.1: The *BIONUMERICS Startup* window.

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 Startup* window to enter the *New database* wizard.

2.2 Enter a name for the database (e.g. **My database**), and press <***Next***>.

A new dialog box pops up, prompting for the type of database.

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

A new dialog box pops up, prompting for the database engine.

2.4 Leave the default option selected (see Figure 2.2) and press <***Finish***> to complete the setup of the new database.

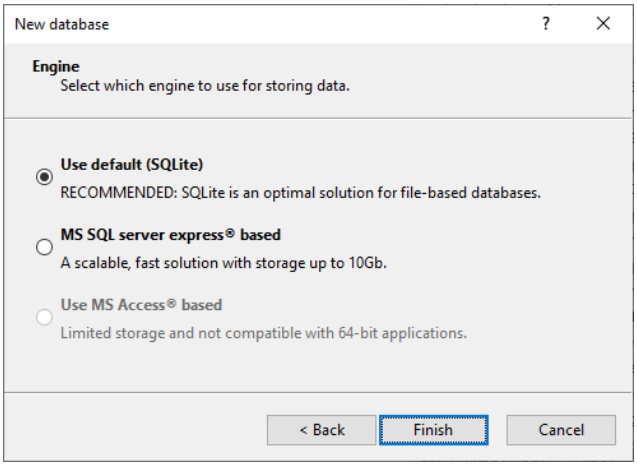


Figure 2.2: Select database engine.

The *Plugins* dialog box pops up which allows you to install additional functionality.

2.5 Press <***Proceed***> to start BIONUMERICS.

2.3 Installing the Sequence Extraction plugin

Before installing the *SarsCoV2 plugin* (see 2.4), please make sure you install the *Sequence extraction plugin* first:

3.1 Select ***File*** > ***Install / remove plugins...*** from the *Main* window to call the *Plugins* dialog box.

3.2 Select the *Utilities tab* in the *Plugins* dialog box, select the *Sequence extraction plugin* from the list and press the <***Activate***> button.

3.3 Confirm the installation of the plugin (see Figure 2.3).

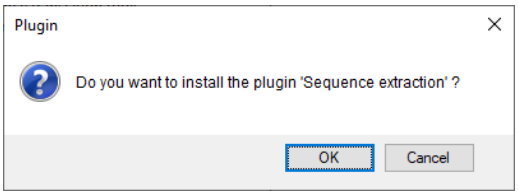


Figure 2.3: Confirm installation of plugin.

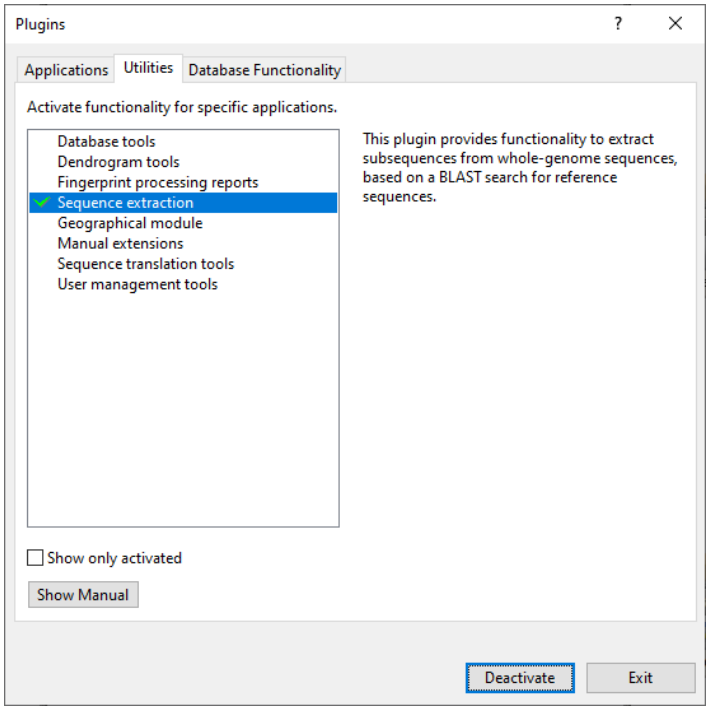


Figure 2.4: Installed plugin.

Once the plugin is successfully installed, it is marked with a green V-sign in the *Plugins* dialog box (see Figure 2.4).

3.4 Close the *Plugins* dialog box.

2.4 Installing the SarsCoV2 plugin

4.1 Select ***File*** > ***Install / remove plugins...*** in the *Main* window to call the *Plugins* dialog box again.

4.2 Select the *Database Functionality tab* in the *Plugins* dialog box and press the <***Add / Update...***> button.

4.3 Check the *SarsCoV2 plugin* in the list of online plugins (see Figure 2.5).

4.4 Press <***OK***> to download the plugin.

4.5 Confirm the installation of the plugin (see Figure 2.6).

The *Create database components* dialog pops up displaying all database components required by the plugin: entry fields, a character type experiment and sequence type experiments (see Figure

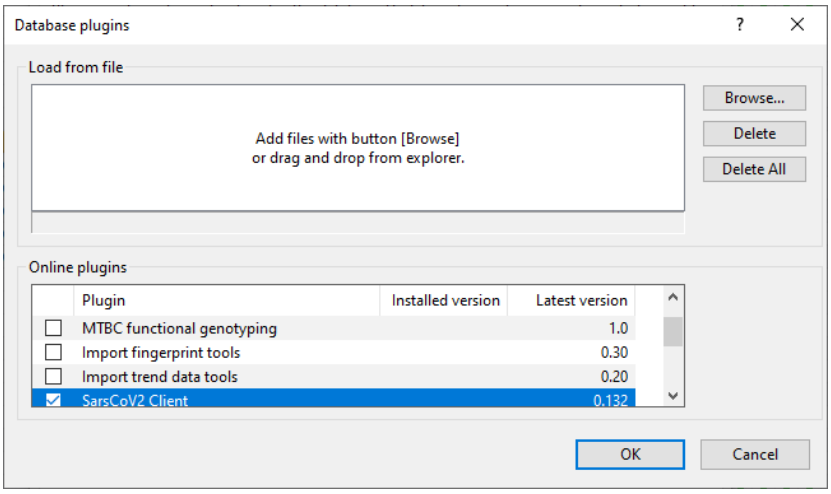


Figure 2.5: Select online plugin.

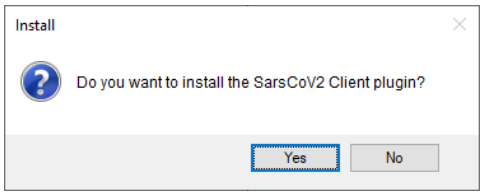


Figure 2.6: Confirm installation of the plugin.

2.7). The default suggested names can be changed if desired.

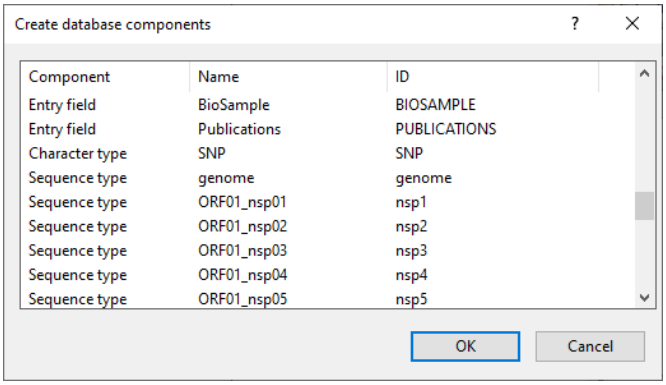


Figure 2.7: New database components.

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

A message pops up, displaying the successful installation of the plugin (see Figure 2.8).

4.7 Press <***OK***>.

The plugin is marked with a green V-sign in the *Plugins* dialog box (see Figure 2.9).

4.8 Close the *Plugins* dialog box.

4.9 Close and reopen the database to activate the features of the *SarsCoV2 plugin*.

The *Main* window should now look like Figure 2.10.

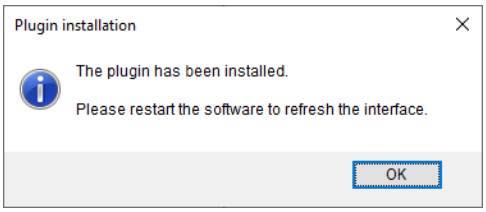


Figure 2.8: Installed plugin.

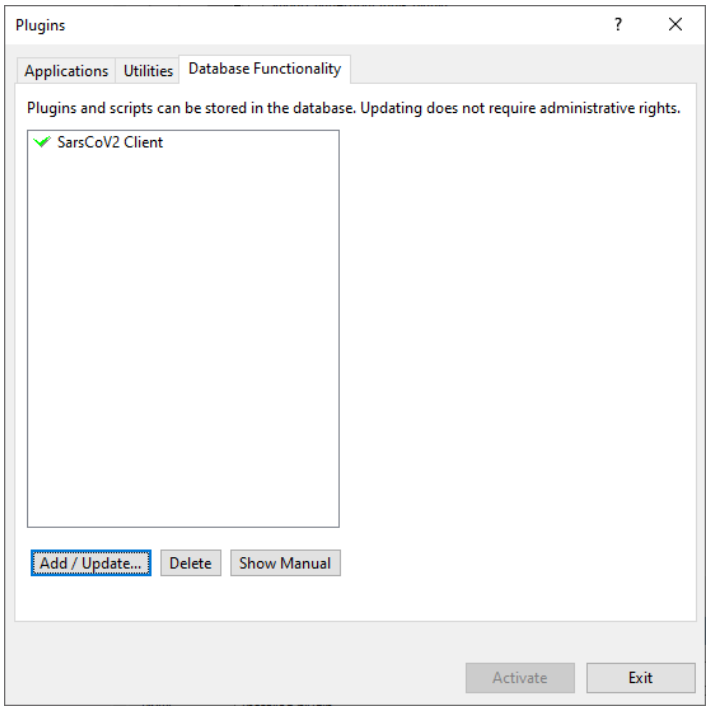


Figure 2.9: Installed plugin.

The *SarsCoV2 plugin* installs menu items in the main menu of the software under ***SARSCoV2*** (see Figure 2.11) and following components (see Figure 2.10):

- 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.
- 26 sequence types, for the storage of the extracted subsequences.
- 27 information fields, comprised of standard GenBank metadata fields and NCBI's SARS-CoV-2 data hub columns.

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.

4.10 Click on the green colored dot in the *Experiment presence* panel, corresponding to the ***genome*** experiment (i.e. the second 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 2.12). The *Annotation* panel holds the NCBI features, and the header information is stored in the *Header* panel.

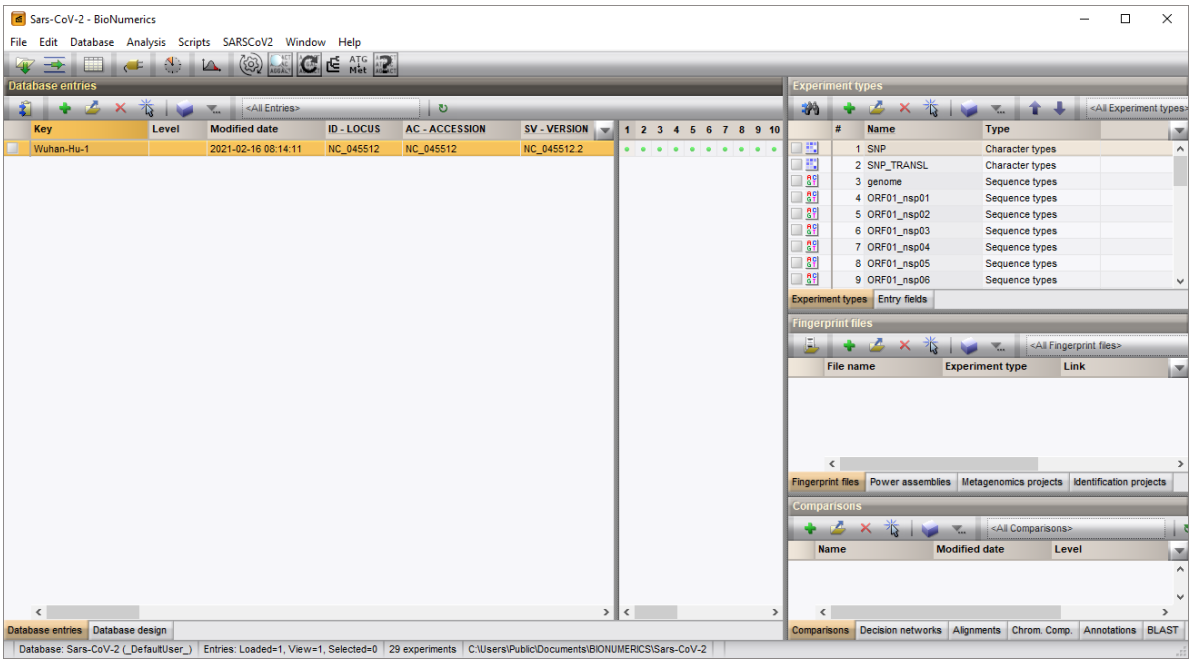


Figure 2.10: The *Main* window after installation of the *SarsCoV2 plugin*.

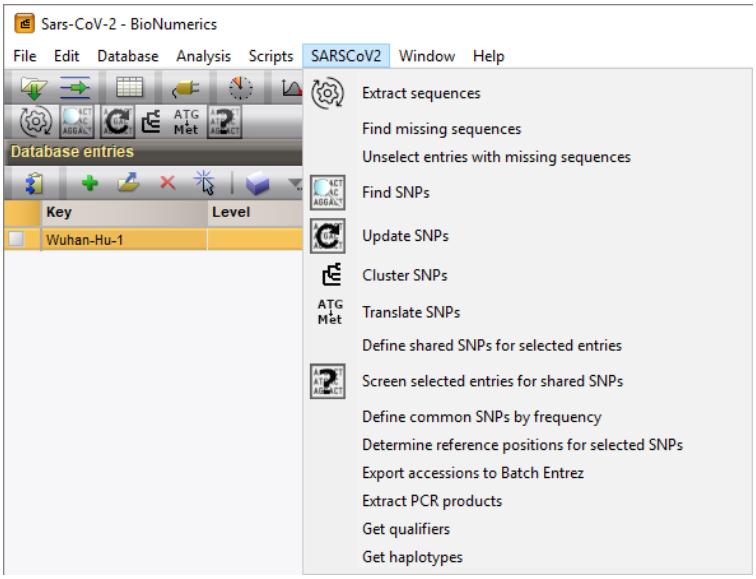


Figure 2.11: Menu items.

4.11 Close the *Sequence editor* window.

The subsequences found on the NCBI reference sequence for SARS-CoV i.e. **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 4.1).

4.12 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*** (see Figure 2.13).

4.13 Check ***ID*** and press <***OK***>.

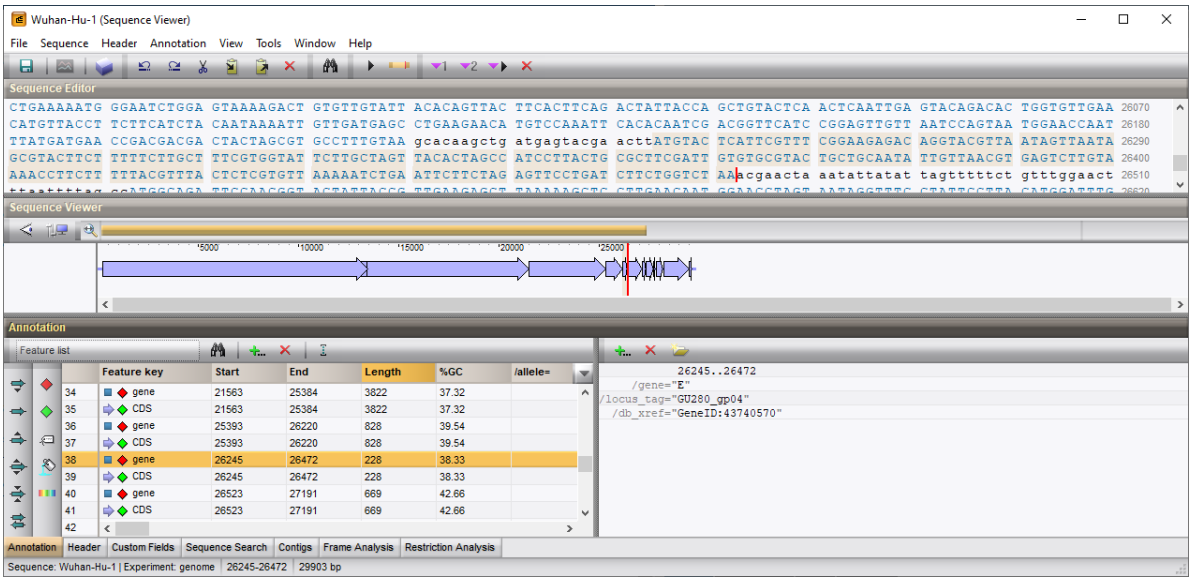


Figure 2.12: The *Sequence editor* window.

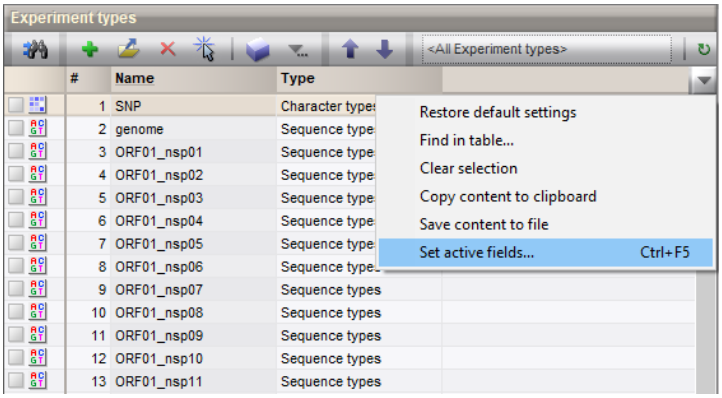


Figure 2.13: Set active fields.

The ***ID*** column is now displayed in the *Experiment types* panel (see Figure 2.14).

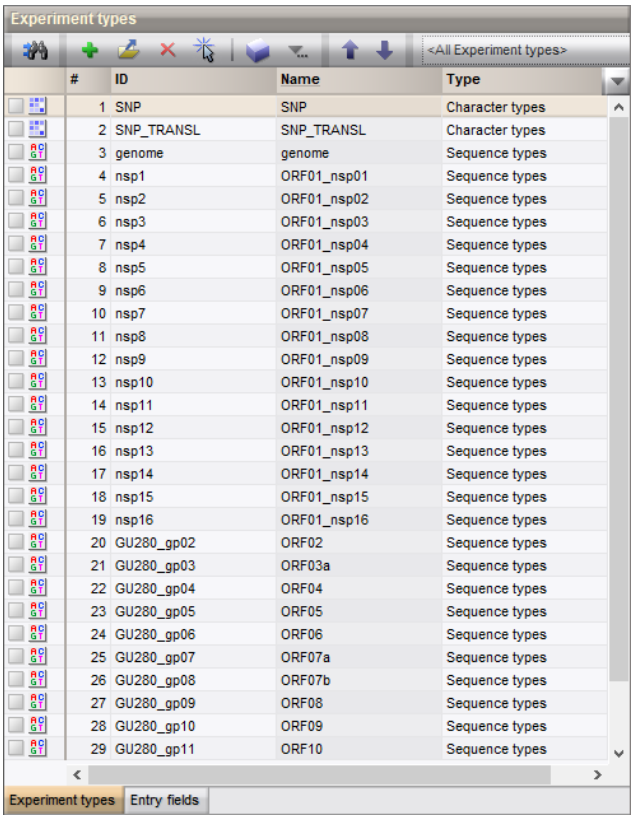


Figure 2.14: ID column displayed.

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* dialog box.

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

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

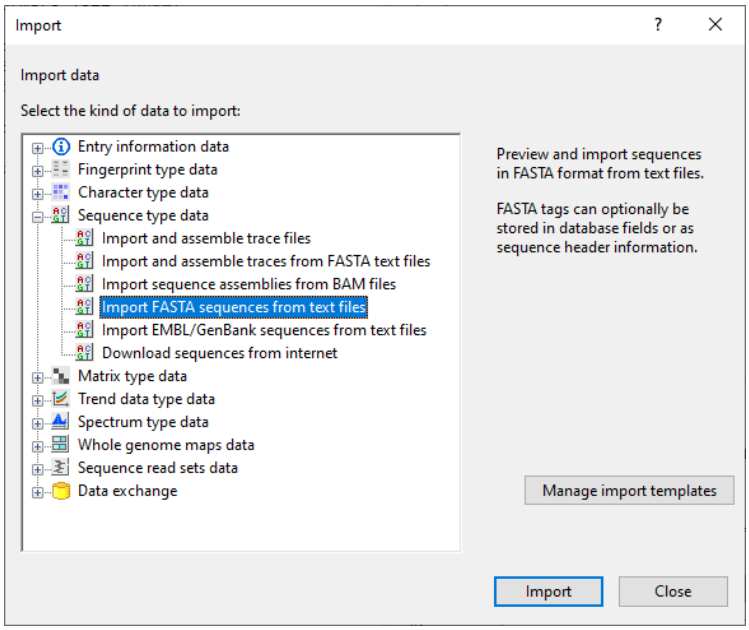


Figure 3.1: The Import tree.

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 Applied Maths website (https://www.applied-maths.com/download/sample-data, "COVID-19 import template"). With this import template, NCBI/EMBL tags are mapped to entry fields created by the plugin.

0.3 In the *Import* dialog box, select <***Manage import templates***>.

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

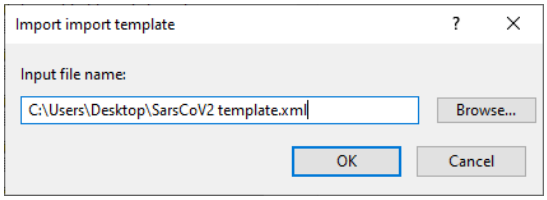


Figure 3.2: XML template.

The import template maps the EMBL/NCBI tags to the entry fields created by the *SarsCoV2 plugin*.

0.5 Press <***OK***> to add the import template to the database and close the dialog (see Figure 3.3).

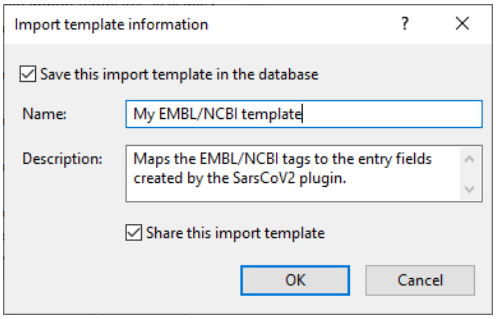


Figure 3.3: Import xml template.

0.6 In the *Import* dialog box, choose the option ***Download sequences from internet*** under the ***Sequence type data*** item in the tree and click <***Import***>.

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

0.8 Specify "," as the ***Separation character*** and choose one of the available download sites from the list, e.g. ***EBI***.

0.9 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).

0.10 Press <***Next***>.

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 3.6).

0.11 Make sure the ***My EMBL/NCBI template*** is selected and press the <***Preview***> button to check the mapping (see Figure 3.5).

0.12 Close the preview.

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

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

The entries are created and are automatically selected. The entry fields are updated and the sequences are stored in the ***genome*** experiment (see Figure 3.7).

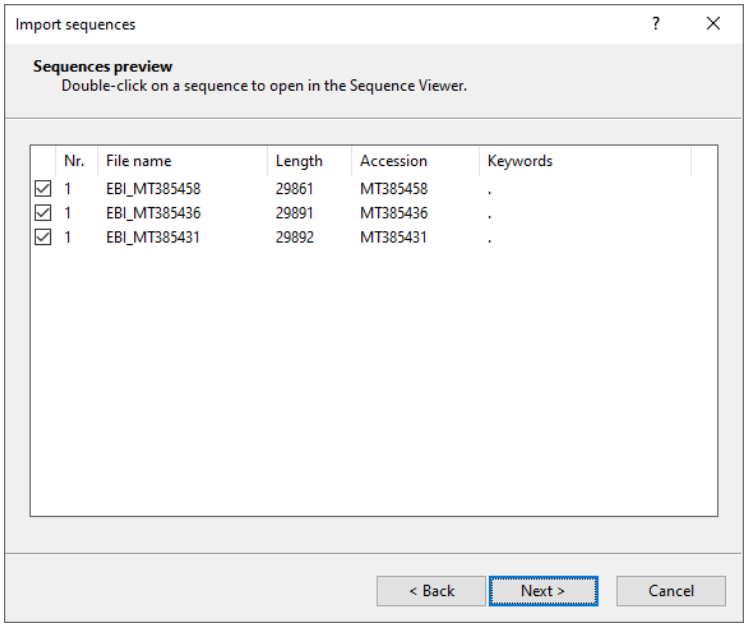


Figure 3.4: Fetched information.

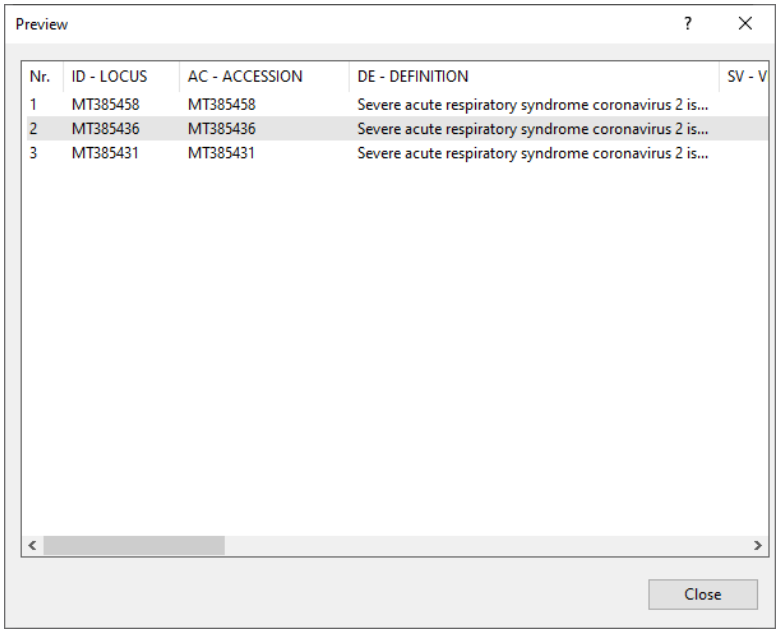


Figure 3.5: Preview.

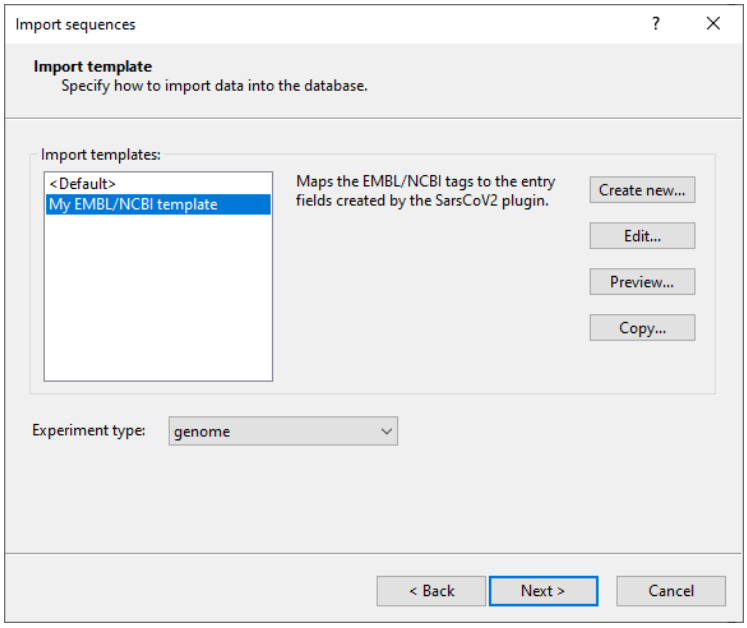


Figure 3.6: Import template.

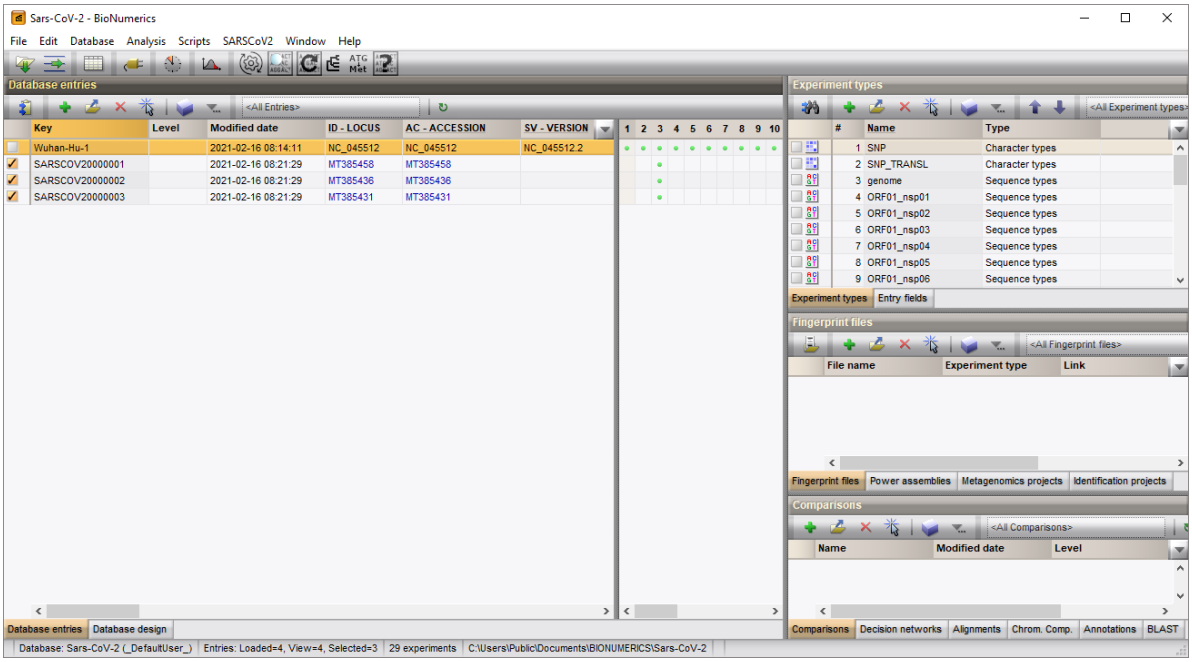


Figure 3.7: The *Main* window after import of some genome sequences.

Chapter 4

Processing sequences

Genomic sequences, imported and stored in the ***genome*** experiment (see 3) can now be processed with the *SarsCoV2 plugin*:

The process includes the following:

1. Extract 26 subsequences from the genomic sequence stored in the ***genome*** experiment (where possible) and save these sequences in the corresponding destination experiments (see 4.1).
2. Screen the 26 subsequences for SNPs with the reference sequence (see 4.2).

We highly recommend to re-process all entries analyzed with a Sars-CoV-2 plugin version lower than 0.32 for consistent results.

4.1 Sequence extraction

The *SarsCoV2 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***.

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*** > ***Extract sequences*** or click the  button to start the sequence extraction.

After the BLAST screening, a message box pops up asking to display a report with the BLAST results (see Figure 4.1).

1.3 Click on <***Yes***> to open the *Report* window.

If all subsequences could be extracted from the genomic sequences, the *Report* window opens and can be consulted immediately. However, if sequences could not be extracted because the length of the detected sequence is not the same as the length of the reference sequence a message box pops up on top of the *Report* window. When detected subsequences do not have the

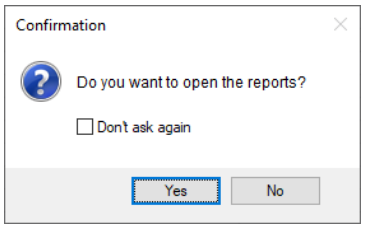


Figure 4.1: Confirmation dialog.

correct length or when subsequences could not be detected based on the specified BLAST criteria (often due to the presence of deletions in the subsequences) the ***Find missing sequences*** dialog box automatically pops up (see Figure 4.2) allowing you to perform a second sequence search which allows to extract sequences which contain deletions.

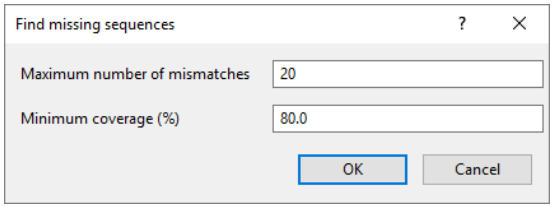


Figure 4.2: Find missing sequences dialog box.

The maximum number of allowed mismatches and the minimum sequence coverage for the second sequence search can be adjusted.

1.4 Click on <***OK***> to search for missing sequences.

Searching for missing sequences can also be performed independently for entries that have already been processed by selecting ***SARSCoV2*** > ***Find missing sequences*** in the *Main* window.

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

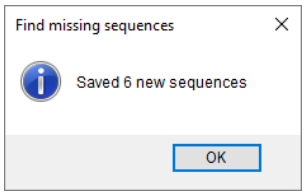


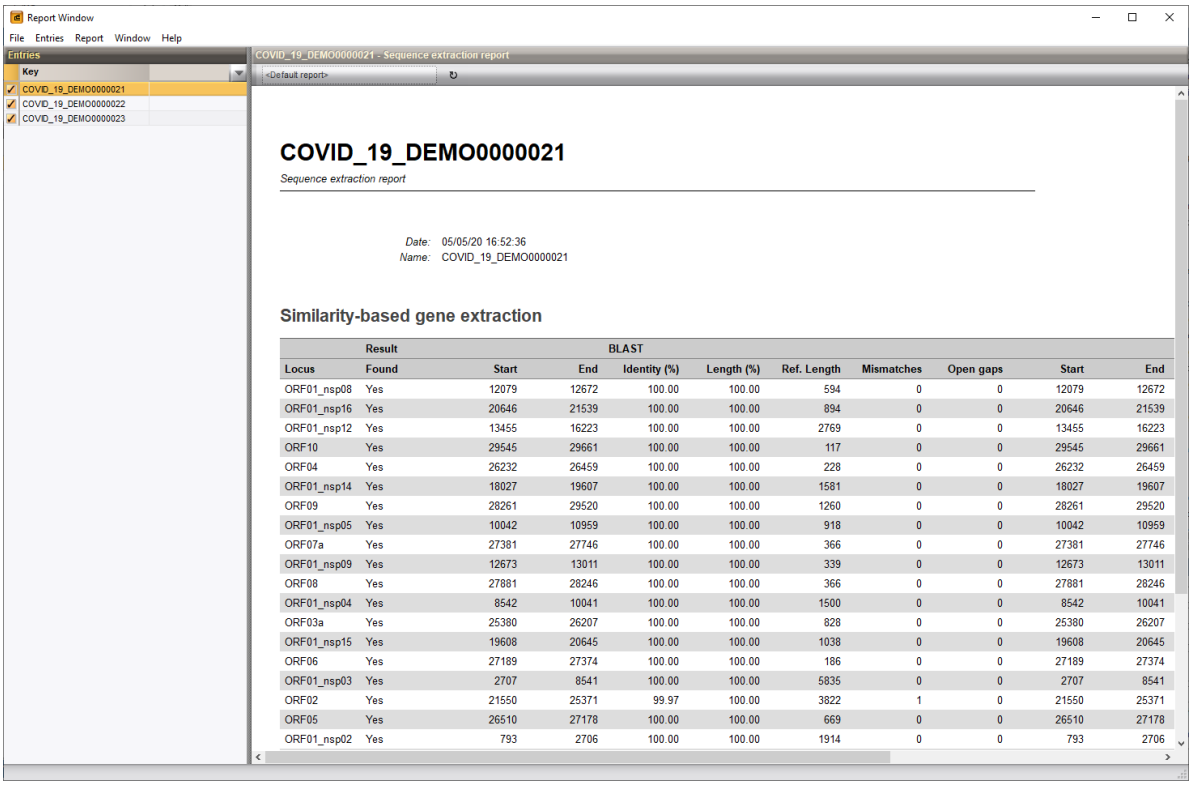
Figure 4.3: Detected missing sequences.

1.5 Click on <***OK***>.

The *Report* window can now be consulted and contains a report for each of the processed entries (see Figure 4.4). The processed entries are grouped in the *Entries* panel.

The *Report* window only lists the subsequences which were detected by the ***Extract sequences*** action, not the subsequences detected by the ***Find missing sequences*** action.

1.6 Select an entry in the *Entries* panel.



	Result			BLAST						
Locus	Found	Start	End	Identity (%)	Length (%)	Ref. Length	Mismatches	Open gaps	Start	End
ORF01_nsp08	Yes	12079	12672	100.00	100.00	594	0	0	12079	12672
ORF01_nsp16	Yes	20646	21539	100.00	100.00	894	0	0	20646	21539
ORF01_nsp12	Yes	13455	16223	100.00	100.00	2769	0	0	13455	16223
ORF10	Yes	29545	29661	100.00	100.00	117	0	0	29545	29661
ORF04	Yes	26232	26459	100.00	100.00	228	0	0	26232	26459
ORF01_nsp14	Yes	18027	19607	100.00	100.00	1581	0	0	18027	19607
ORF09	Yes	28261	29520	100.00	100.00	1260	0	0	28261	29520
ORF01_nsp05	Yes	10042	10959	100.00	100.00	918	0	0	10042	10959
ORF07a	Yes	27381	27746	100.00	100.00	366	0	0	27381	27746
ORF01_nsp09	Yes	12673	13011	100.00	100.00	339	0	0	12673	13011
ORF08	Yes	27881	28246	100.00	100.00	366	0	0	27881	28246
ORF01_nsp04	Yes	8542	10041	100.00	100.00	1500	0	0	8542	10041
ORF03a	Yes	25380	26207	100.00	100.00	828	0	0	25380	26207
ORF01_nsp15	Yes	19608	20645	100.00	100.00	1038	0	0	19608	20645
ORF06	Yes	27189	27374	100.00	100.00	186	0	0	27189	27374
ORF01_nsp03	Yes	2707	8541	100.00	100.00	5835	0	0	2707	8541
ORF02	Yes	21550	25371	99.97	100.00	3822	1	0	21550	25371
ORF05	Yes	26510	27178	100.00	100.00	669	0	0	26510	27178
ORF01_nsp02	Yes	793	2706	100.00	100.00	1914	0	0	793	2706

Figure 4.4: The Report window.

The results of the selected entry are displayed in the *Report* panel. The date of processing (***Date***) and the entry Key (***Name***) are displayed.

For each destination sequence type (***Locus*** column), it is indicated whether or not a BLAST hit was ***Found***, its position on the screened genome sequence (***Start*** and ***Stop***), sequence identity (***Identity (%)***) and sequence overlap (***Length (%)***).

Furthermore, the length of the retrieved subsequence is reported (***Ref length***), the number of mismatches with the reference sequence (***Mismatches***), number of gaps (***Open gaps***) and length correction (if applied).

1.7 Close the *Report window*.

The *Report* window can be consulted again by selecting the entries in the *Main* window and selecting ***Analysis*** > ***Sequence types*** > ***Extract sequences*** > ***Show reports***.

1.8 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 4.5 for an example).

1.9 Close the *Sequence editor* window.

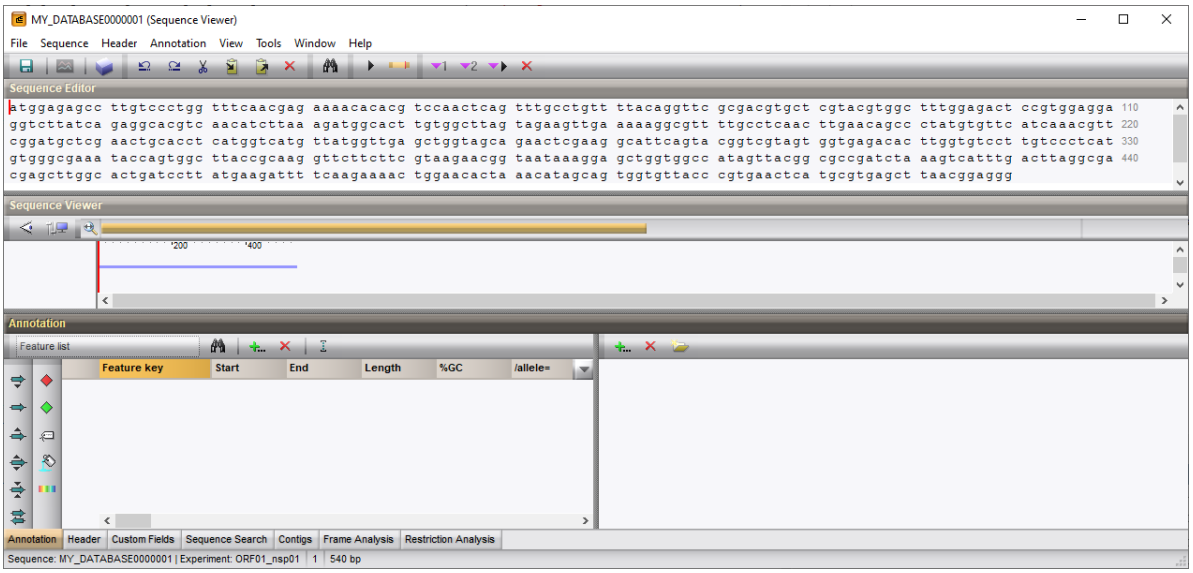


Figure 4.5: The ORF01_nsp01 sequence.

4.2 Calculating SNPs

After the subsequence extraction (see 4.1), the subsequences can be screenend for SNPs:

2.1 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.

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

The subsequences are screened for SNPs using the built-in BIONUMERICS SNP analysis tool (accessible via ***Analysis*** > ***Sequence types*** > ***Start SNP analysis***).

The resulting SNP set is filtered based on the *Relaxed SNP filtering* template and the retained SNPs are stored in the ***SNP*** character experiment.

With the *Relaxed SNP filtering* template non-ACGT bases are included in the analysis. However, non-ACGT bases are not actually saved in the SNP character set, leaving an absent value for that position instead.

After the SNP screening, a message box pops up displaying the number of detected SNPs (see Figure 4.6). If new SNP positions are detected, this is also reported.

Figure 4.6: SNP information.

2.3 Click <***OK***> to close the confirmation dialog.

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 4.7).

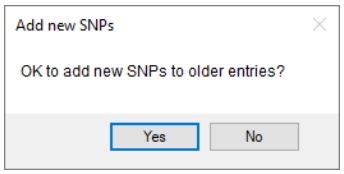


Figure 4.7: Update SNPs.

2.4 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.

2.5 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 characer experiment card.

The character experiment card lists all SNPs detected for the sample. The bases are listed in the ***Mapping*** column (see Figure 4.8).

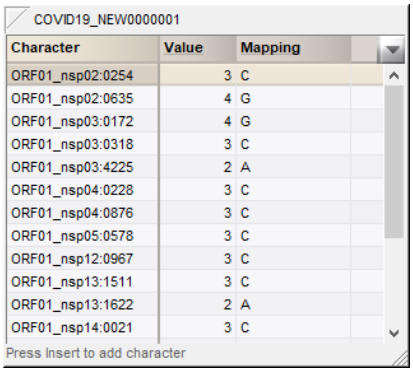


Figure 4.8: SNP character card.

2.6 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.

2.7 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 4.9).

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.

2.8 Close the *Character type* window.

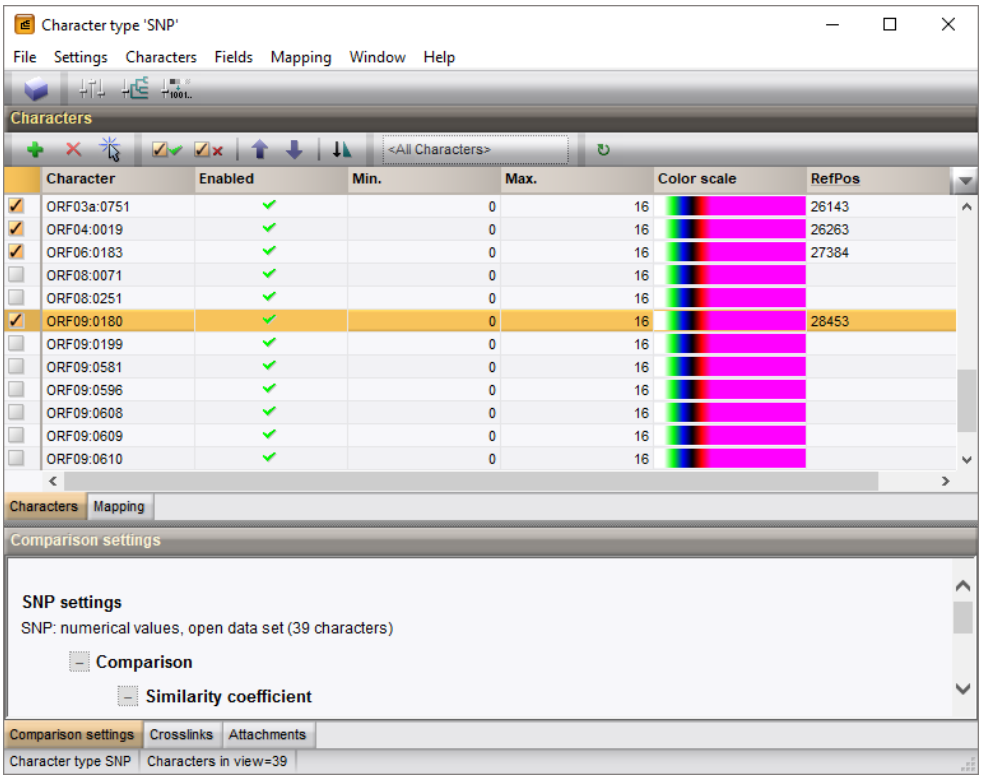


Figure 4.9: The SNP character experiment type with the ***RefPos*** information field.

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

- ***All***: All SNP positions present in the SNP character experiment will be included in the comparison.
- ***Selected***: All selected SNP positions in the SNP character experiment will be included in the comparison.
- ***common***: All common SNP positions (see 6.3) 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 6.2).

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

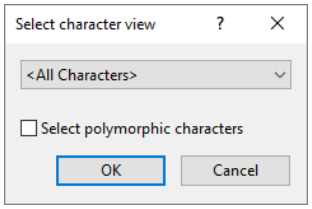


Figure 5.1: The ***Select character view*** dialog box.

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*** 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.

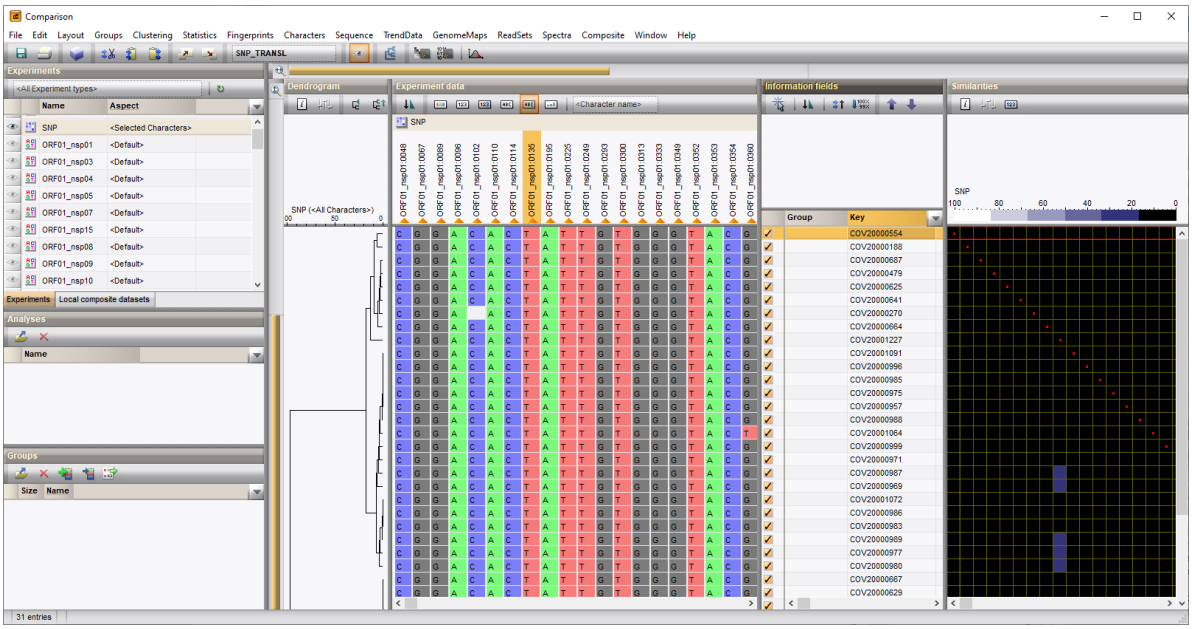


Figure 5.2: The *Comparison* window.

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

0.5 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.

0.6 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 *Advanced cluster analysis* window. This window can be launched from the *Comparison* window:

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

0.8 Select ***Clustering*** > ***Calculate*** > ***Advanced cluster analysis...*** or press the button and select ***Advanced cluster analysis*** to launch the *Create network* wizard.

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

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

0.10 Close the *Advanced cluster analysis* window.

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

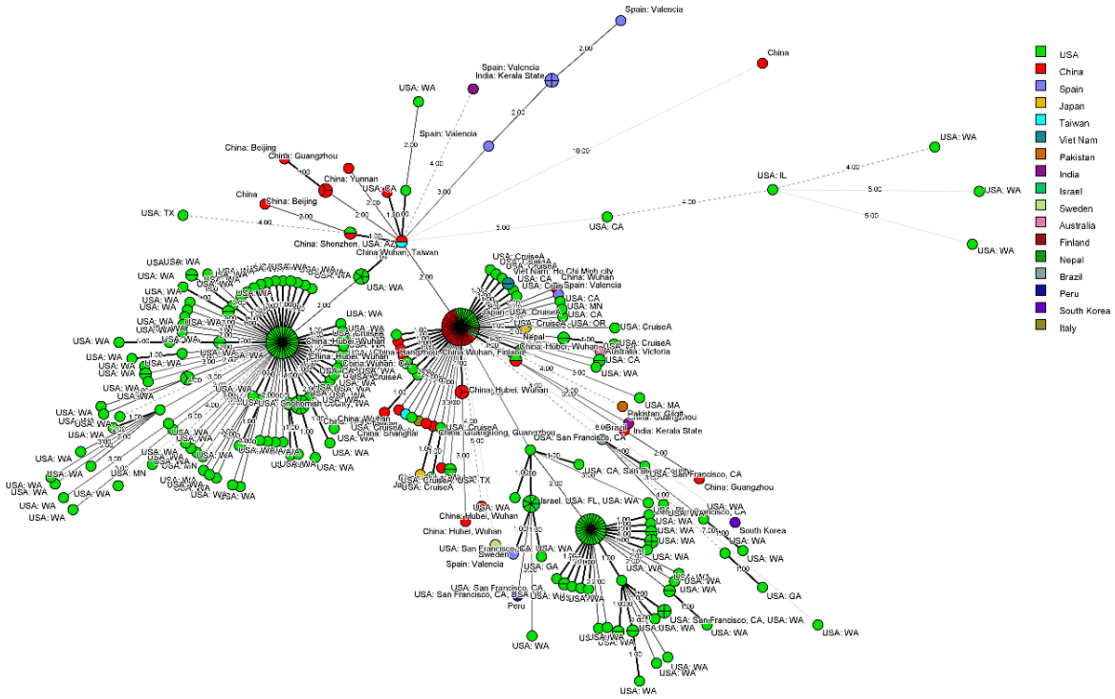


Figure 5.3: Minimum Spanning Tree with groups defined.

Chapter 6

Miscellaneous tools

6.1 Translating SNPs

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

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

1.2 Select ***SARSCoV2*** $>$ ***Translate SNPs*** or click on the ATG Met button.

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

- ***_All_***: All SNP positions present in the SNP character experiment.
- ***_Selected_***: All selected SNP positions in the SNP character experiment.
- ***common***: All common SNP positions (see 6.3).
- All user-defined character views in the SNP character experiment e.g. character views which contain shared SNPs (see 6.2).

1.3 Select a character view from the drop-down 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.

1.4 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 characer experiment card.

The amino acids are listed in the ***Mapping*** column (see Figure 6.1).

1.5 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.

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

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

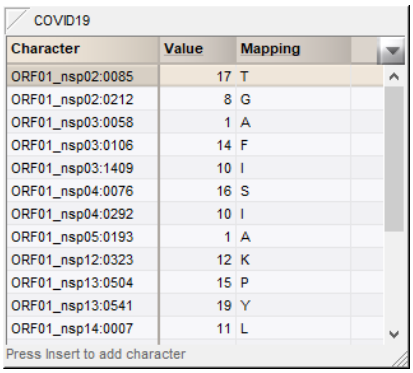


Figure 6.1: Character card.

1.8 Click on the next to the experiment name ***SNP_TRANSL*** in the *Experiments* panel to display the data in the *Experiment data* panel.

Initially, the character values are displayed as colors.

1.9 Select ***Characters*** > ***Show mappings*** ( ) to show the mapping values.

1.10 Select ***Characters*** > ***Show mappings+colors*** ( ) to show the mapping values and colors in overlay (see Figure 6.2).

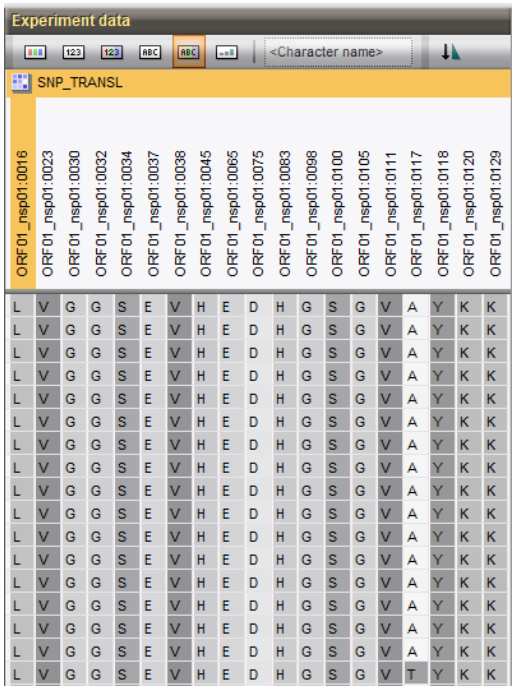


Figure 6.2: The *Comparison* window.

6.2 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.

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

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

2.3 Specify a name for the new SNP character view which will contain the SNPs shared between the selected entries (see Figure 6.3) and press <***OK***>.

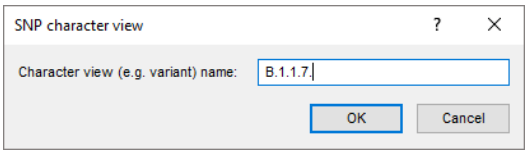


Figure 6.3: SNP character view.

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

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

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

2.6 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 6.4 for an example).

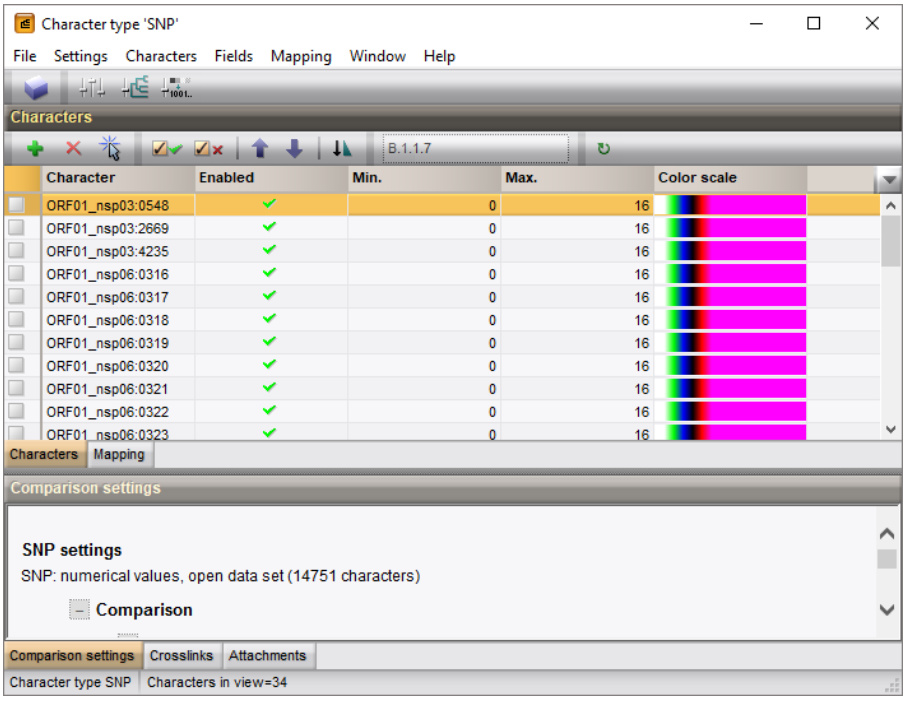


Figure 6.4: The SNP character experiment type with the shared SNPs listed in the newly created character view 'B.1.1.7'.

2.7 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.

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

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

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 and this entry view will automatically be selected (see Figure 6.5 for an example).

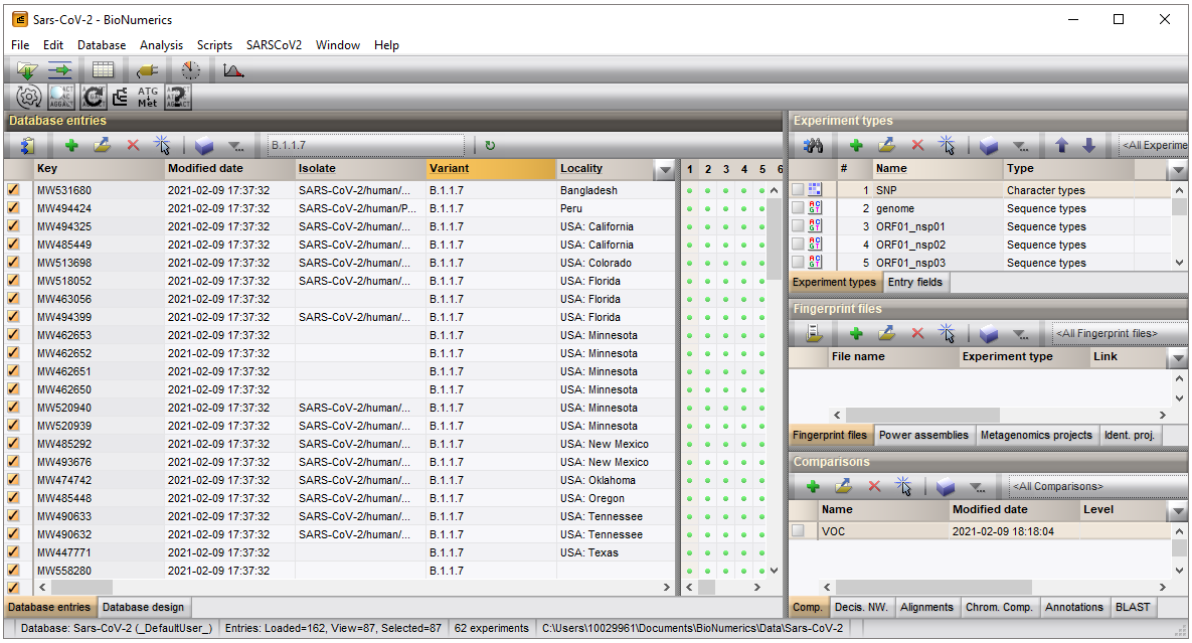


Figure 6.5: Screening of selected entries for the SNPs present in the 'B.1.1.7' character view.

6.3 Defining common SNPs

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

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

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

3.3 Specify the minimum frequency in the dialog (see Figure 6.6) and press <***OK***>.

The number of common SNPs - identified based on the provided frequency - is displayed (see Figure 6.7 for an example).

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

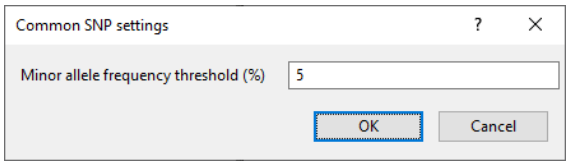


Figure 6.6: Specify threshold.

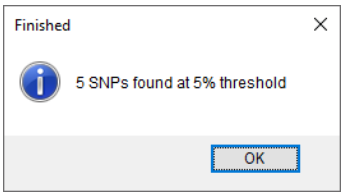


Figure 6.7: Result.

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

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

3.6 Click on the drop-down bar in the toolbar and select the ***common*** character view from the list (see Figure 6.8).

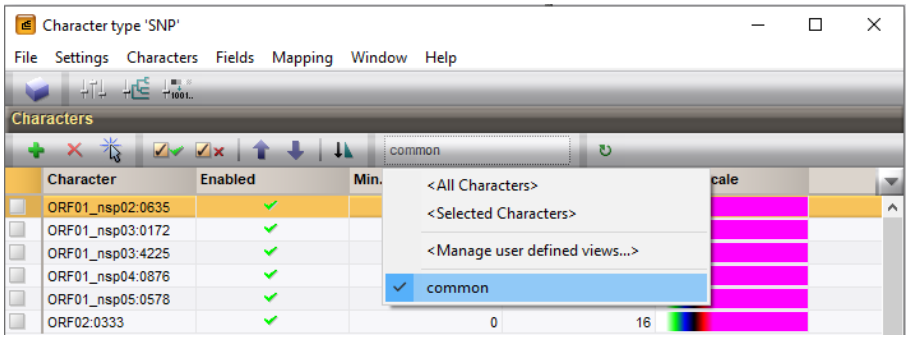


Figure 6.8: Character view.

The common SNPs identified with the command ***SARSCoV2*** > ***Define common SNPs*** are listed. The ***MAF*** character field displays the allele frequency (see Figure 6.9).

3.7 Close the *Character type* window.

6.4 Exporting accessions to BLAST Entrez

There is a standard BIONUMERICS import tool available in the *Import* dialog box 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:

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

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

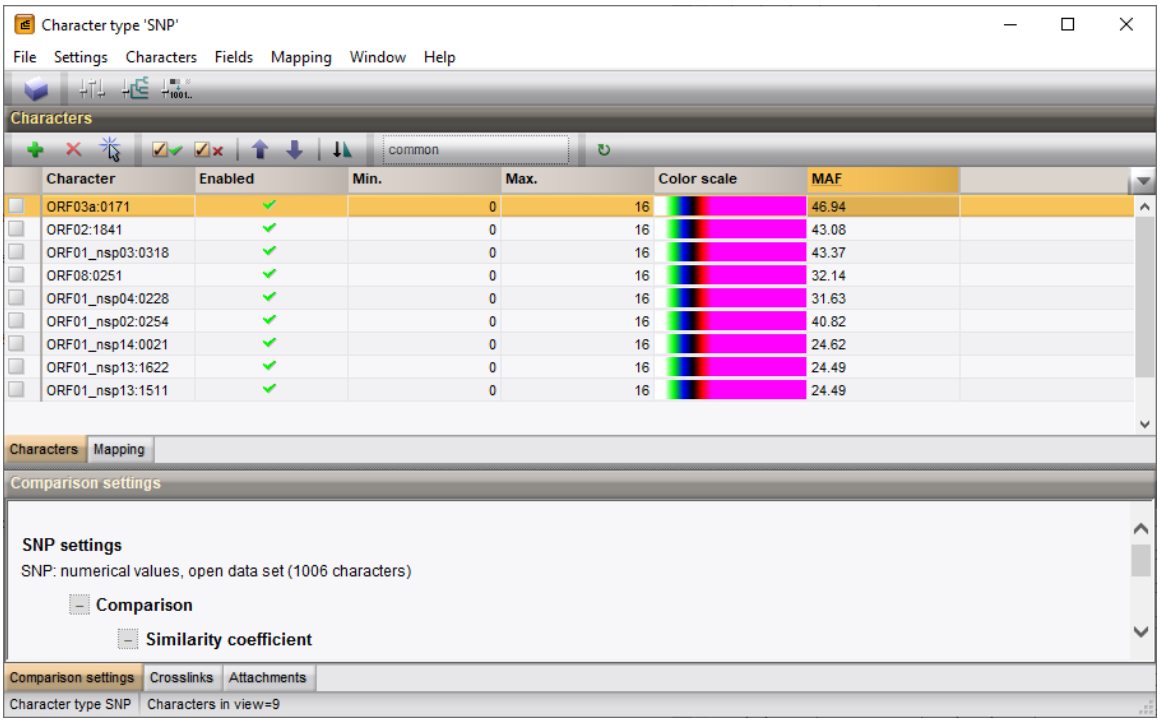


Figure 6.9: The ***MAF*** character field.

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

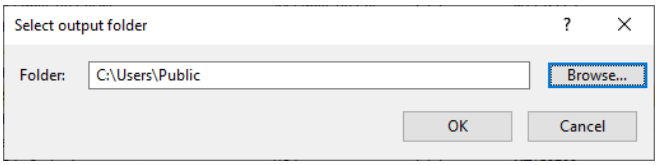


Figure 6.10: Browse for folder.

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 website in the default browser (see Figure 6.11), from which the accessions file can be selected (with the <***Browse***> button) to retrieve GenBank-formatted sequences in bulk.

6.5 Extracting PCR products

With the *SarsCoV2 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-summary-

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

5.2 Select ***SARSCoV2*** > ***Exctract PCR products***.

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

The extracted PCR products are stored in the corresponding PCR sequence type experiments

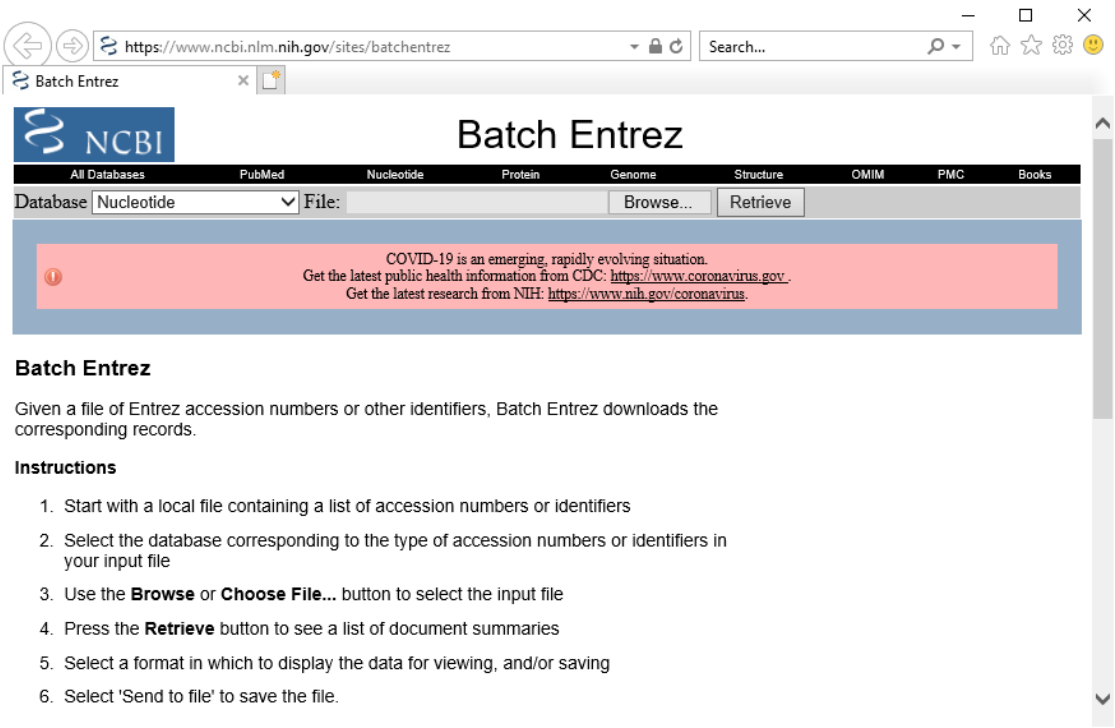


Figure 6.11: Batch Entrez.

(see Figure 6.12).

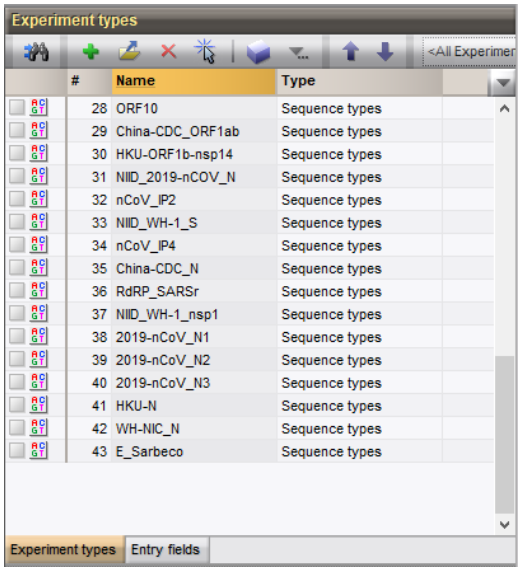


Figure 6.12: Sequence types for the storage of PCR products.

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

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

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

5.5 Click on the next to the sequence type in the *Experiments* panel to display the sequences in the *Experiment data* panel (see Figure 6.13 for an example).

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

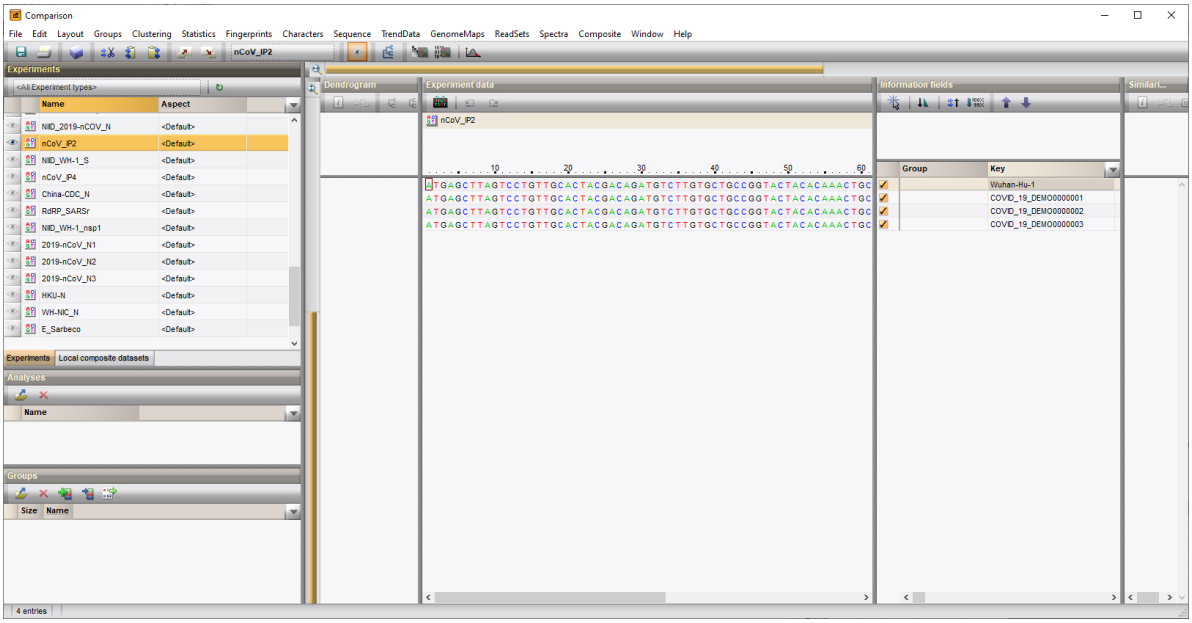


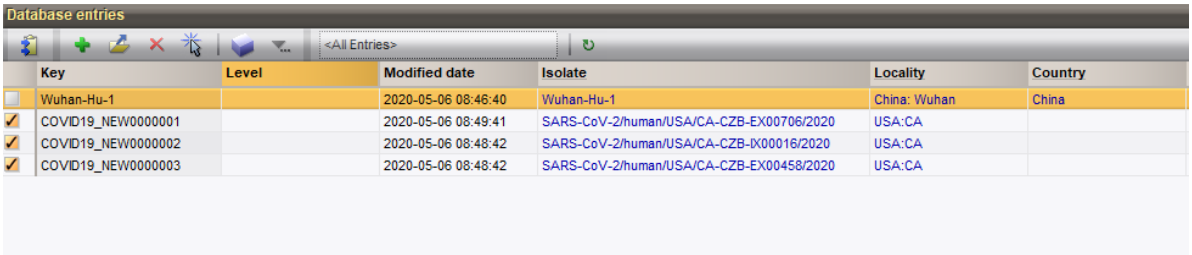
Figure 6.13: PCR products.

6.6 Get qualifiers

Metadata can be parsed from the sequence annotations - if present - and stored in the ***Isolate*** and ***Locality*** entry fields (see Figure 6.14).

6.1 In the *Database entries* panel of the *Main* window, select the entries for which you want to extract the Isolate and Locality information.

6.2 Select ***SARSCoV2*** > ***Get qualifiers***.



Key	Level	Modified date	Isolate	Locality	Country
Wuhan-Hu-1		2020-05-06 08:46:40	Wuhan-Hu-1	China: Wuhan	China
COVID19_NEW0000001		2020-05-06 08:49:41	SARS-CoV-2/human/USA/CA-CZB-EX00706/2020	USA:CA	
COVID19_NEW0000002		2020-05-06 08:48:42	SARS-CoV-2/human/USA/CA-CZB-IX00016/2020	USA:CA	
COVID19_NEW0000003		2020-05-06 08:48:42	SARS-CoV-2/human/USA/CA-CZB-EX00458/2020	USA:CA	

Figure 6.14: Locality and Isolate information extraction.

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.

6.7 Haplotype determination

The plugin allows to determine haplotypes. The haplotype as defined in the *SarsCoV2 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.

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

7.2 Select ***SARSCoV2*** > ***Get haplotypes***.

The result is displayed in the ***Haplotype*** entry field.

Optionally colors can be assigned to every haplotype:

7.3 Right-click on the ***Haplotype*** information field in the *Database entries* panel and choose ***Field properties*** from the floating menu (see Figure 6.15).

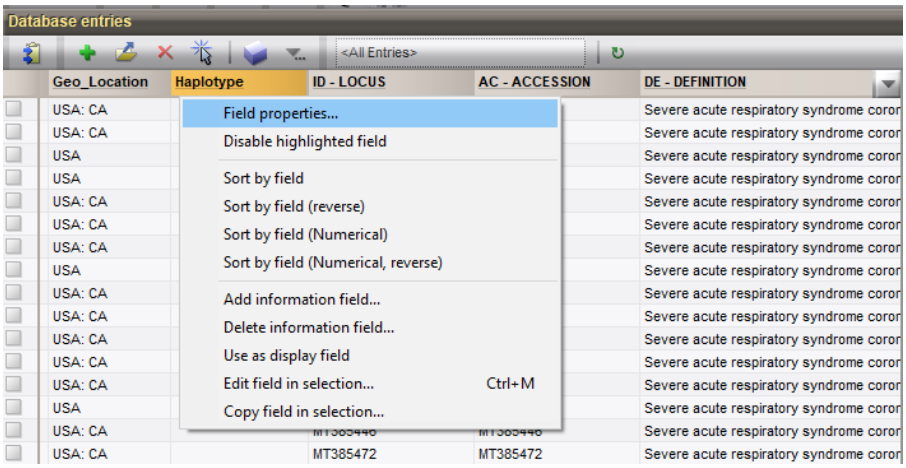


Figure 6.15: Field properties.

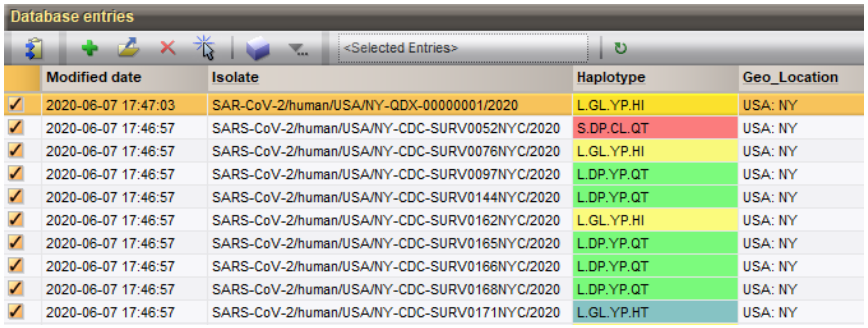
The *Database field properties* dialog box appears.

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

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

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

The *Database entries* panel is updated (see Figure 6.16).



	Modified date	Isolate	Haplotype	Geo_Location
✓	2020-06-07 17:47:03	SAR-CoV-2/human/USA/NY-QDX-00000001/2020	L.GL.YP.HI	USA: NY
✓	2020-06-07 17:46:57	SARS-CoV-2/human/USA/NY-CDC-SURV0052NYC/2020	S.DP.CL.QT	USA: NY
✓	2020-06-07 17:46:57	SARS-CoV-2/human/USA/NY-CDC-SURV0076NYC/2020	L.GL.YP.HI	USA: NY
✓	2020-06-07 17:46:57	SARS-CoV-2/human/USA/NY-CDC-SURV0097NYC/2020	L.DP.YP.QT	USA: NY
✓	2020-06-07 17:46:57	SARS-CoV-2/human/USA/NY-CDC-SURV0144NYC/2020	L.DP.YP.QT	USA: NY
✓	2020-06-07 17:46:57	SARS-CoV-2/human/USA/NY-CDC-SURV0162NYC/2020	L.GL.YP.HI	USA: NY
✓	2020-06-07 17:46:57	SARS-CoV-2/human/USA/NY-CDC-SURV0165NYC/2020	L.DP.YP.QT	USA: NY
✓	2020-06-07 17:46:57	SARS-CoV-2/human/USA/NY-CDC-SURV0166NYC/2020	L.DP.YP.QT	USA: NY
✓	2020-06-07 17:46:57	SARS-CoV-2/human/USA/NY-CDC-SURV0168NYC/2020	L.DP.YP.QT	USA: NY
✓	2020-06-07 17:46:57	SARS-CoV-2/human/USA/NY-CDC-SURV0171NYC/2020	L.GL.YP.HT	USA: NY

Figure 6.16: Haplotype information field.

APPLIED MATHS

A BIOMÉRIEUX COMPANY

Please contact us for any additional information you might require, we will gladly help you!

Headquarters
Keistraat 120 • 9830 Sint-Martens-Latem • Belgium
+32 922 22 100 info@applied-maths.com

USA and Canada
11940 Jollyville Rd., Suite 115N • Austin, TX 78750 USA
+1 512 482 9700 info-us@applied-maths.com