BIOMÉRIEUX

BIONUMERICS®

version 8 - PLUGINS

E. coli functional genotyping 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

Starting and setting up BIONUMERICS

1.1 Introduction

The *E. coli functional genotyping plugin* allows you to screen *Escherichia coli* whole genome sequences to predict phenotypic traits such as antibiotic resistance, virulence, and serotype. It also allows you to detect phages and plasmids and to detect and extract PCR markers. The genome sequences can be imported in BIONUMERICS using one of the import routines available in the software or can be the result of a de novo assembly performed in BIONUMERICS on a sequence read set.

The *E. coli functional genotyping plugin* is supported in the **BIONUMERICS-SEQ** and **BIONUMERICS-SUITE** configurations.

1.2 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 1.1).

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.

1.3 Installation of the E. coli functional genotyping plugin

Proceed as follows to install the *E. coli functional genotyping plugin*:

3.1 Call the *Plugins and Scripts* dialog box from the *Main* window with ***File*** > ***Install / remove plugins...*** ().

3.2 Select the *E. coli functional genotyping plugin* from the list and press the <***Install***> button.

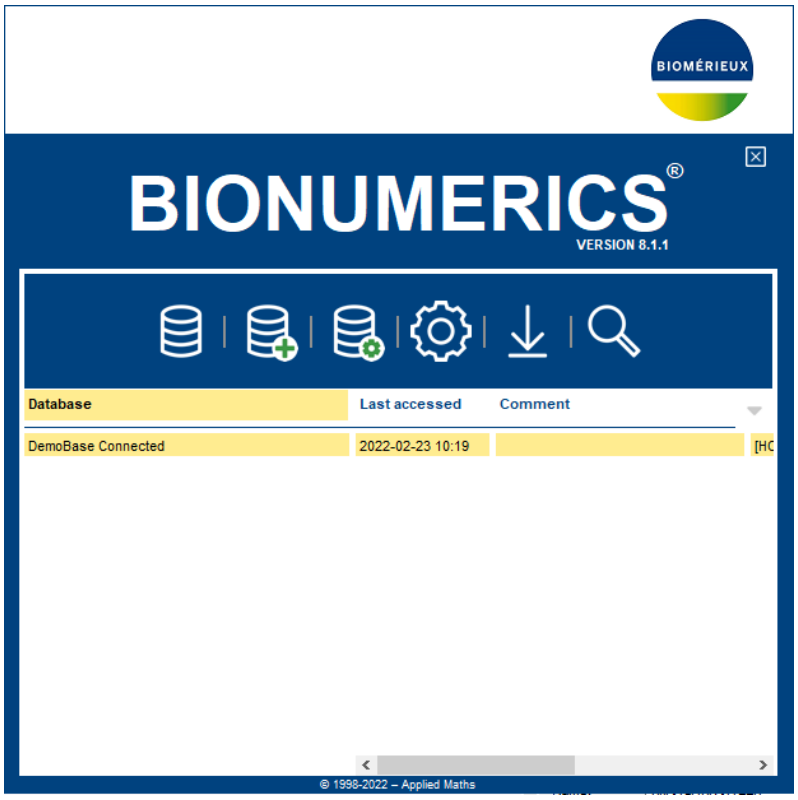

Figure 1.1: The *BIONUMERICS Startup* window.

3.3 Confirm the installation of the plugin.

During installation, the plugin downloads online knowledge bases (see 3) from `https://www.bionumerics.com`, which requires a connection to the internet. Depending on the bandwidth of your connection, this process can take up to several minutes. When the download is complete, the question 'Do you want to change the E. coli genotyping settings now?' pops up.

3.4 Press <***Yes***> to confirm that you want to change the plugin settings.

The *E. coli genotyping settings* dialog box appears, which will be discussed in detail in 2. For the plugin to work, at the very least we need to specify the sequence experiment containing the (de novo) genome sequences on which the screening will be performed.

3.5 If already available, select the sequence experiment in your database that contains genome sequences as ***Input sequence experiment*** and press <***OK***>.

A message appears, confirming the installation of the plugin and prompting you to restart BIONUMERICS.

3.6 Press <***OK***> in the confirmation message.

3.7 Press <***Close***> to close the *Plugins and Scripts* dialog box.

3.8 Close and re-open the database to complete the installation of the plugin.

The *E. coli functional genotyping plugin* installs menu items in the main menu of the software under ***E. coli*** (see Figure 1.2).

More information regarding the *E. coli functional genotyping plugin* settings is available under 4.

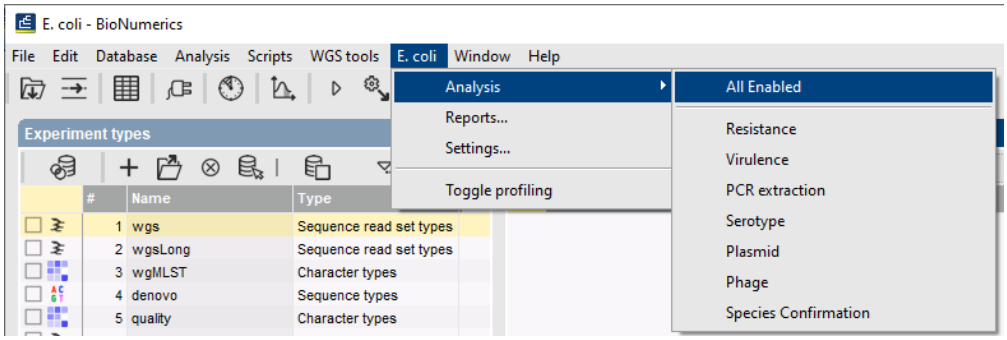

Figure 1.2: New menu items, available after installation of the *E. coli functional genotyping plugin*.

Chapter 2

Settings of the E. coli functional genotyping plugin

2.1 Accessing the genotyping settings

Settings for the *E. coli functional genotyping plugin* can be accessed via ***E. coli*** > ***Settings...*** in the *Main* window.

2.2 General settings

The *General* tab of the *E. coli genotyping settings* dialog box (see Figure 2.1) holds settings for the genotyping reports and for general processing.

Under ***Reporting***, the entry information fields that will be displayed in the genotyping reports can be specified in the ***Included info fields*** list. Simply check the ballot box next to an information field name to include the field in the report.

The ***Exports directory*** can be specified for all exports from the genotyping reports. By default, the exported files are stored in a subdirectory of the database directory, but a different location can be selected via the <***Browse***> button or entered directly in the text box.

The ***Input sequence experiment***, i.e. the sequence experiment containing the whole genome sequences to be screened, should be selected from the corresponding drop-down list. Select the <***Create***> option in case you wish to create a new sequence experiment type. In the latter case, make sure to import whole genome sequences in this experiment type before running the plugin.

It is crucial to specify at least the ***Input sequence experiment*** in the settings. If not specified, the error message ”The input sequence experiment must be set to process entries.” will be generated when the plugin is run.

In the ***Enabled features*** list, all features offered by the plugin are listed and enabled by default. If specific analyses are not required, you can uncheck them here to save on processing time and to omit the corresponding sections from the reports.

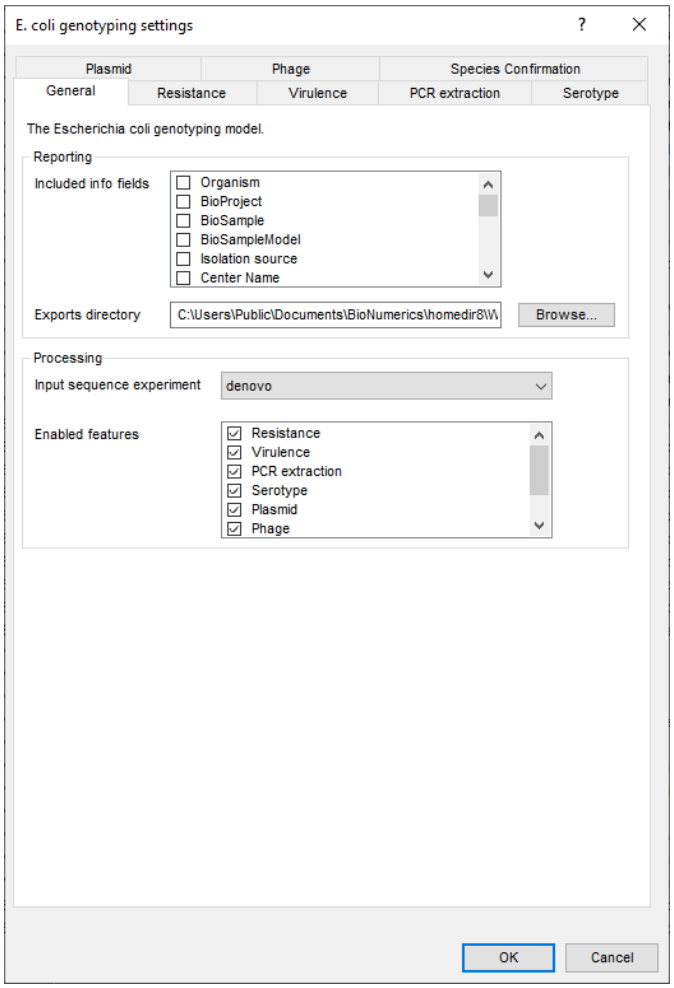

Figure 2.1: The *E. coli genotyping settings* dialog box, *General* tab.

2.3 Resistance settings

The *Resistance* tab in the *E. coli genotyping settings* dialog box (see Figure 2.2) groups all settings for the detection of antibiotics resistance and ESBL/CPE typing.

Resistance prediction is based on the absence or presence of certain antibiotic resistance related genes and alleles in the assembled *E. coli* genome. Presence or absence is detected based on a BLAST approach using the list of antibiotic resistance related genes as query and the de novo assembled genome as target. Additionally, the plugin detects mutations involved in resistance to antibiotics.

Under ***Knowledgebase***, the ***Name*** and ***Version*** of the specified knowledge base version for this feature is shown. When no knowledge base is specified yet, "<None>" will be indicated in both fields. When a specified knowledge base version cannot be found (e.g. because it is deleted), "<Missing>" is shown in both fields. A different knowledge base version can be selected by pressing the <***Change...***> button.

With ***Check for updates on startup*** checked, BIONUMERICS will check if a newer knowledge base version is available online for this feature each time the database is opened. This requires an internet connection.

See 3 for more information on *E. coli functional genotyping plugin* knowledge bases.

In the ***BLAST*** panel, two settings for the BLAST algorithm can be specified:

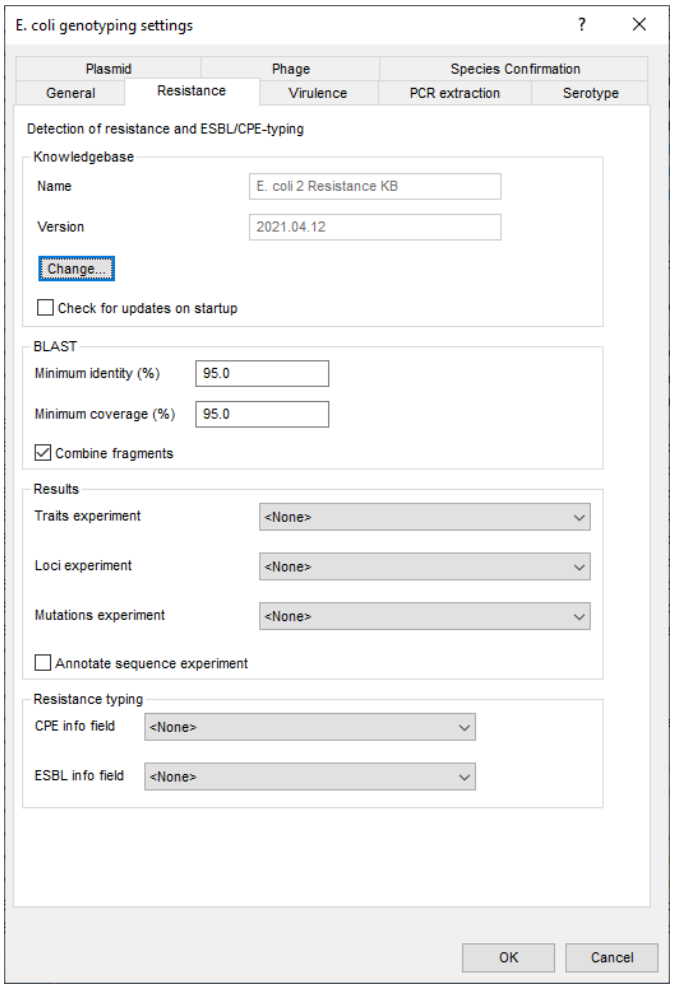

Figure 2.2: The *E. coli genotyping settings* dialog box, *Resistance* tab.

- ***Minimum identity (%)*** is the minimum sequence identity (as percentage) of the query sequence against the knowledge base's reference sequences.
- ***Minimum length for coverage*** specifies the minimum overlap (as percentage) between the subsequence found in the target assembly sequence and the reference sequence from the knowledge base.

If the option ***Combine fragments*** is checked, genes that occur fragmented in the genome (i.e. split over two contigs) can still be detected.

In the ***Results*** panel, the experiment types and entry information fields to which the screening results will be written can be dictated. Use the drop-down menu to choose an existing experiment type or information field or select the <***Create***> option to create a new experiment type or information field. A default name is suggested, but you can adjust this if you want to.

Following character experiments can optionally be specified for acquired resistance:

- ***Traits experiment***: contains the results for each antibiotic group: 0 = not detected (sensitive), 1 = detected (resistant). The default name is **Resistance_traits**.
- ***Loci experiment***: contains the results for each resistance gene: 0 = not detected (sensitive), when detected (resistant) the % identity of the best BLAST hit is shown. The default name is **Resistance_loci**.

Following character experiment can optionally be specified for mutational resistance:

- ***Mutations experiment***: contains the results for each resistance mutation: -2 = partially indecisive, -1 = fully indecisive, 0 = not detected (sensitive), 1 = detected (resistant). The default name is **Resistance_mutations**.

For mutational resistance analysis, the result ***-1/fully indecisive*** is given when one of the following requirements is true:

- The sequence identity of the hit is smaller than or equals 95%.
- The sequence coverage (after hit extension) is below 100%.
- The DNA sequence cannot be successfully translated into an amino acid sequence, for mutations defined at the protein level.

Some mutations present in the mutational resistance knowledge base only confer antibiotic resistance when another 'required' mutation is also present. The result ***-2/partially indecisive*** is given when the primary mutation is present but the status of the required mutation is uncertain, e.g. when the required locus was absent.

The characters in the characters experiments are displayed in the same order they are listed in their knowledge base. However, it might be more convenient for interpretation to have them displayed alphabetically. How to rearrange characters in a character type experiment is described in the Reference manual, Chapter Setting up character type experiments.

For the resistance character experiments, it can be convenient to map the character values (-1, 0, 1) to categorical names (indecisive (-), sensitive (S), resistant (R)) by creating a character mapping. How to do this is explained in the Reference manual, Chapter Setting up character type experiments.

Check ***Annotate sequence experiment*** to annotate the input sequence with the detected genotyping features.

In the ***Resistance typing*** panel, the entry information fields to which the resistance type results will be written can be dictated. Use the drop-down menu to choose an existing information field or select the <***Create***> option to create a new information field. A default name is suggested, but you can adjust this if you want to.

Following information fields can optionally be specified:

- ***CPE info field***: 'True' when one or more enzymes associated with the CPE resistance type were detected; 'False' if no such enzymes were detected. The default field name is **CPE**.
- ***ESBL info field***: 'True' when one or more enzymes associated with the ESBL resistance type were detected; 'False' if no such enzymes were detected. The default field name is **ESBL**.

2.4 Virulence settings

The *Virulence* tab of the *E. coli genotyping settings* dialog box (see Figure 2.3) groups all settings for the detection of virulence factors and virulence islands.

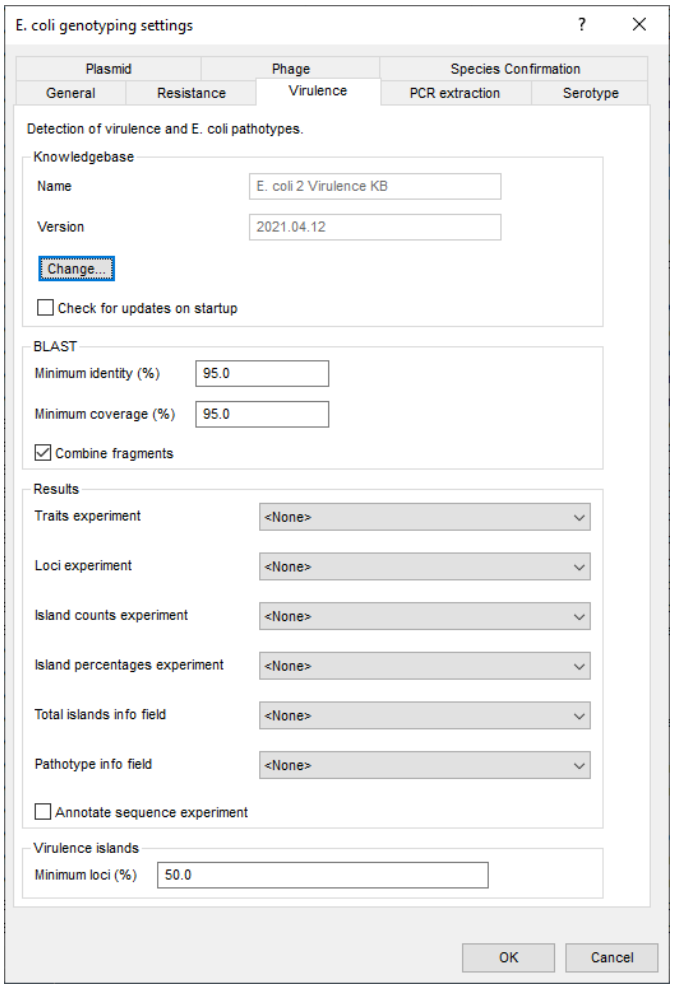

Figure 2.3: The *E. coli genotyping settings* dialog box, *Virulence* tab.

Virulence prediction is based on the absence or presence of certain virulence related genes and alleles in the assembled *E. coli* genome. Presence or absence is detected based on a BLAST approach using the list of virulence related genes as query and the de novo assembled genome as target.

Many of the *E. coli* virulence genes are found on 'virulence islands' (also called 'pathogenicity islands', PAIs) in the chromosome. A BLAST-based tool screens for loci associated to virulence islands and reports the number of found loci, the % of found loci, and the loci spread average minimum inter-loci distance (in bp), for each virulence island.

Under ***Knowledgebase***, the ***Name*** and ***Version*** of the specified knowledge base version for this feature is shown. When no knowledge base is specified yet, "<None>" will be indicated in both fields. When a specified knowledge base version cannot be found (e.g. because it is deleted), "<Missing>" is shown in both fields. A different knowledge base version can be selected by pressing the <***Change...***> button.

With ***Check for updates on startup*** checked, BIONUMERICS will check if a newer knowledge base version is available online for this feature each time the database is opened. This requires an internet connection.

See 3 for more information on *E. coli functional genotyping plugin* knowledge bases.

In the ***BLAST*** panel, two settings for the BLAST algorithm can be specified:

- ***Minimum identity (%)*** is the minimum sequence identity (as percentage) of the query sequence against the knowledge base's reference sequences.

- ***Minimum length for coverage*** specifies the minimum overlap (as percentage) between the subsequence found in the target assembly sequence and the reference sequence from the knowledge base.

If the option ***Combine fragments*** is checked, genes that occur fragmented in the genome (i.e. split over two contigs) can still be detected.

In the ***Results*** panel, the experiment types and entry information fields to which the screening results will be written can be dictated. Use the drop-down menu to choose an existing experiment type or information field or select the <***Create***> option to create a new experiment type or information field. A default name is suggested, but you can adjust this if you want to.

Following character experiments can optionally be specified:

- ***Traits experiment***: contains the results for each virulence trait: 0 = not detected, 1 = detected. The default name is **Virulence_traits**.

- ***Loci experiment***: contains the results for each virulence gene: 0 = not detected, when detected the % identity of the best BLAST hit is shown. The default name is **Virulence_loci**.

- ***Island counts experiment***: contains the number of detected loci associated to a pathogenicity island. The default name is **Island_counts**.

- ***Island percentages experiment***: contains the percentage of detected loci associated to a pathogenicity island. The default name is **Island_percentages**.

Following entry information fields can optionally be specified:

- ***Total islands info field***: contains the number of different virulence islands detected. The default field name is **Total islands**.

- ***Pathotype info field***: contains the different pathotypes that are detected. The plugin will detect the pathotype based on the presence of certain marker genes as defined by the by the EU Reference Laboratory at the Statens Serum Institute. If markers associated with different pathotypes are found, all pathotypes, separated by a comma (',') will be reported in this information field. Possible pathotypes are: enteropathogenic *E. coli* (EPEC), enterohemorrhagic *E. coli* (EHEC), enterotoxigenic *E. coli* (ETEC), enteroaggregative *E. coli* (EAEC), diffusely adherent *E. coli* (DAEC), enteroinvasive *E. coli* (EIEC), uropathogenic *E. coli* (UPEC), and neonatal meningitis *E. coli* (NMEC). The default field name is **Pathotype**.

Check ***Annotate sequence experiment*** to annotate the input sequence with the detected genotyping features.

Under ***Virulence islands***, you can specify the minimum percentage of virulence island loci (***Minimum loci (%)***) that needs to be detected before the presence of the virulence island is shown in the results. Please note that some virulence islands in the BIONUMERICS online knowledge base only contain two loci, for this reason we discourage setting the threshold above 50%.

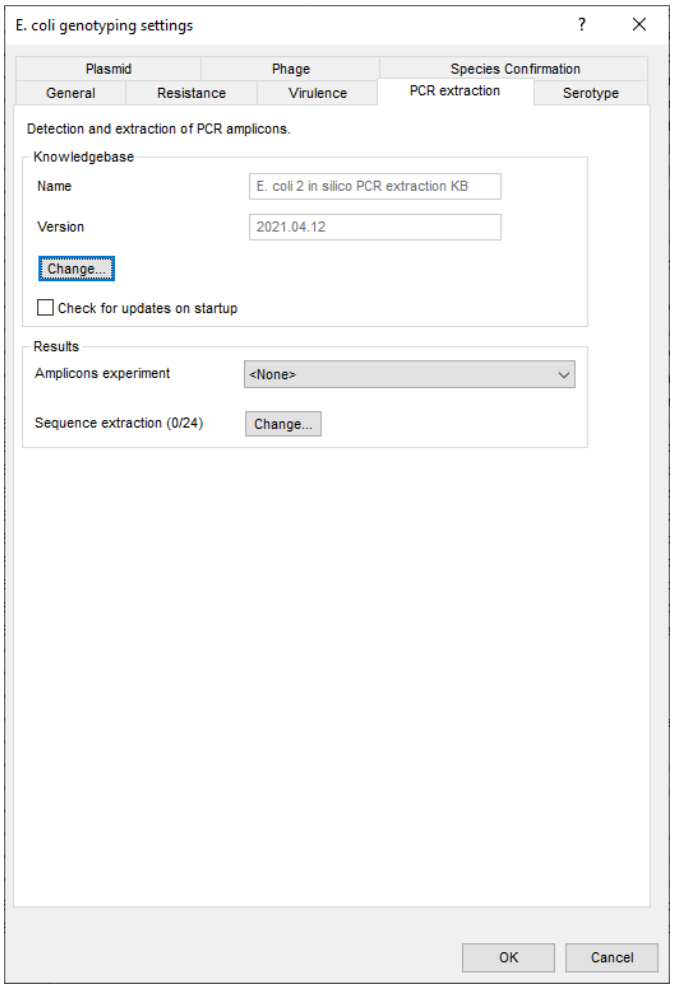

Figure 2.4: The *E. coli genotyping settings* dialog box, *PCR extraction* tab.

2.5 PCR extraction settings

The *PCR extraction* tab in the *E. coli genotyping settings* dialog box (see Figure 2.4) groups all settings for in silico PCR detection and extraction of amplicons.

The *E. coli functional genotyping plugin* offers an *in silico* variant of the classical PCR detection methods used in wet labs. The *in silico* PCR based extraction of marker sequences enables the detection and extraction of marker loci.

Under ***Knowledgebase***, the ***Name*** and ***Version*** of the specified knowledge base version for this feature is shown. When no knowledge base is specified yet, "<None>" will be indicated in both fields. When a specified knowledge base version cannot be found (e.g. because it is deleted), "<Missing>" is shown in both fields. A different knowledge base version can be selected by pressing the <***Change...***> button.

With ***Check for updates on startup*** checked, BIONUMERICS will check if a newer knowledge base version is available online for this feature each time the database is opened. This requires an internet connection.

See 3 for more information on *E. coli functional genotyping plugin* knowledge bases.

In the ***Results*** panel, the experiment types and entry information fields to which the screening results will be written can be dictated. Use the drop-down menu to choose an existing experiment type or information field or select the <***Create***> option to create a new experiment type or

information field. A default name is suggested, but you can adjust this if you want to.

The character experiment ***Amplicons experiment*** contains the results for each in silico PCR: 0 = no amplicon, 1 = amplicon generated. The default name is **PCR extraction_amplicons**.

Sequences are extracted based on an *in silico* PCR approach. That is, binding sites for primers are detected in the de novo assembly. The plugin allows for a maximum of two mismatches and the presence of one non-resolved (IUPAC) base (except for `N`) in the primer binding sites. Please note that if more than one unresolved bases are detected, these unresolved bases will also be considered as mismatch. If both the forward and reverse primer are able to bind resulting in an amplicon of the expected size, the sequence of the amplicon will be extracted and stored in the corresponding sequence experiment.

Under ***Sequence extraction***, press <***Change...***> to open the *Change sequence experiment* dialog box (see Figure 2.5).

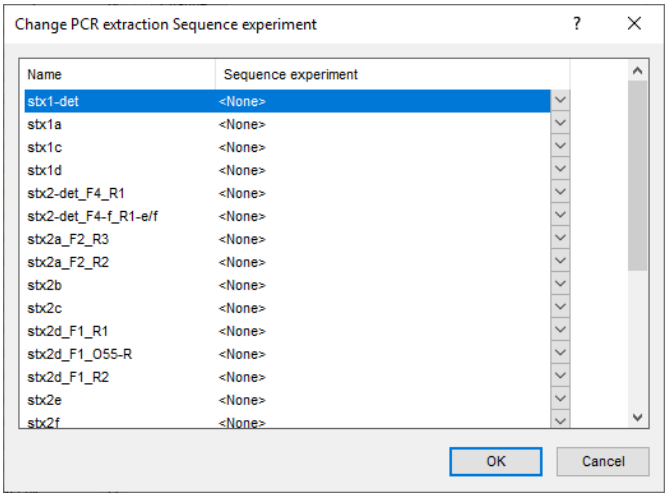

Figure 2.5: The *Change sequence experiment* dialog box.

The column 'Name' in the dialog contains all targets available in the knowledge base. Via the drop-down list in the 'Sequence experiment' column, an existing sequence experiment type can be specified in which the extracted sequence will be stored. When the <***Create***> option is selected, a dialog will pop up asking for the name of the sequence experiment type. By default, the name of the target will be suggested.

When all default names are accepted, following sequence experiment types will be created:

- **Stx1-det**, **stx1a**, **stx1c** and **stx1d**: contains the extracted amplicon sequences for stx1 toxin related markers.
- **Stx2-det**, **stx2a**, **stx2b**, **stx2c**, **stx2d**, **stx2e**, **stx2f** and **stx2g**: contains the extracted amplicon sequences for stx2 toxin related markers.
- **eaeA**: contains the extracted amplicon sequence for the *eaeA* marker.
- **ehxA**: contains the extracted amplicon sequence for the *ehxA* marker.
- **rpoB-AEM** and **rpoB-CDC**: contains the extracted amplicon sequences for the *rpoB* marker. Two different primers pairs are used for the extraction: one used by the EU Reference Laboratory at the Statens Serum Institute (rpoB-AEM) and one used by the US Centers for Disease Control (rpoB-CDC).
- **ipaH**: contains the extracted amplicon sequence for the *ipaH* marker.

- **e_coli-det**, **e_albertii-det** and **e_fergusonii-det**: contains the extracted amplicon sequences for the *Escherichia* species (*E. coli*, *E. albertii* and *E. fergusonii*) markers.

2.6 Serotype settings

The *Serotype* tab in the *E. coli genotyping settings* dialog box (see Figure 2.6) groups all settings for *E. coli* serotype determination.

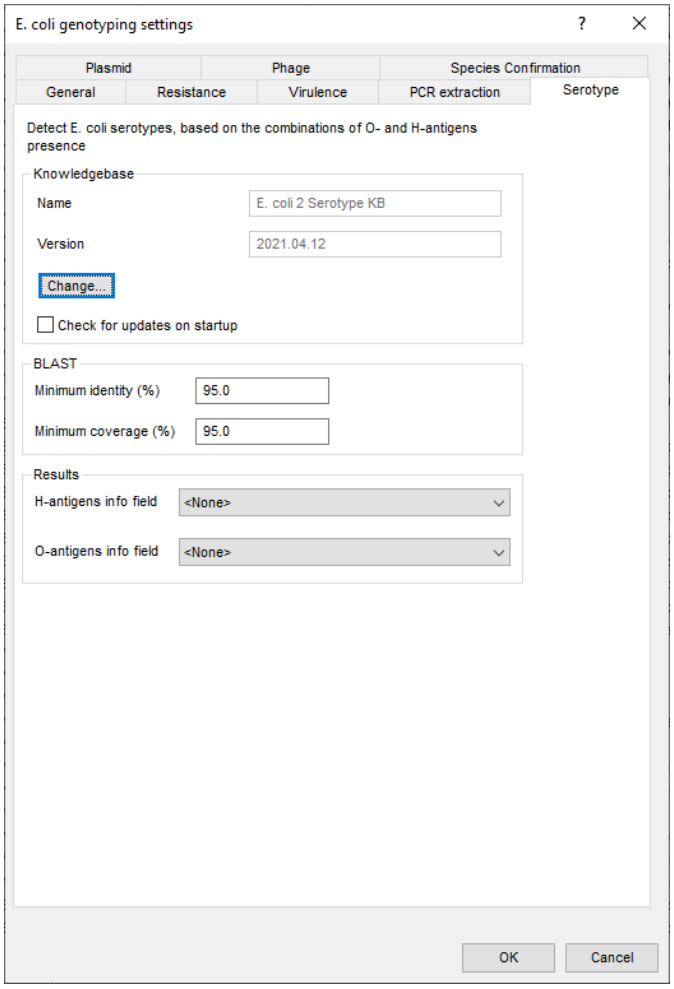

Figure 2.6: The *E. coli genotyping settings* dialog box, *Serotype* tab.

O:H serotyping is a standard method for characterization of pathogenic *E. coli*. The O group is determined primarily by analysis of *wzm*, *wzt*, *wzx* and *wzy* gene sequences in the assembled *E. coli* genome. The H group is determined through analysis of *fliC*, *flkA*, *fllA*, *flmA* and *flnA* gene sequences. Detection of both groups is based on a BLAST approach.

Under ***Knowledgebase***, the ***Name*** and ***Version*** of the specified knowledge base version for this feature is shown. When no knowledge base is specified yet, "<None>" will be indicated in both fields. When a specified knowledge base version cannot be found (e.g. because it is deleted), "<Missing>" is shown in both fields. A different knowledge base version can be selected by pressing the <***Change...***> button.

With ***Check for updates on startup*** checked, BIONUMERICS will check if a newer knowledge base version is available online for this feature each time the database is opened. This requires an internet connection.

See 3 for more information on *E. coli functional genotyping plugin* knowledge bases.

In the ***BLAST*** panel, two settings for the BLAST algorithm can be specified:

- ***Minimum identity (%)*** is the minimum sequence identity (as percentage) of the query sequence against the knowledge base's reference sequences.
- ***Minimum length for coverage*** specifies the minimum overlap (as percentage) between the subsequence found in the target assembly sequence and the reference sequence from the knowledge base.

If more than one BLAST hit that meets the minimum sequence identity and length for coverage criteria is found, a ratio of BLAST p-values between the two best BLAST hits is calculated as a discrimination score D:

$$D = 2^{b-s}$$

with b the bit score of the best hit and s the bit score of the second-best hit (runner-up). The bigger this ratio, the bigger the difference between the two best BLAST hits. In the ***Results*** panel, the experiment types and entry information fields to which the screening results will be written can be dictated. Use the drop-down menu to choose an existing experiment type or information field or select the <***Create***> option to create a new experiment type or information field. A default name is suggested, but you can adjust this if you want to.

Following entry information fields can optionally be specified:

- ***H-antigens info field***: the detected H-antigen group (H1 to H56) or "Unknown" if no H-antigen group could be assigned. The default field name is **H_antigens**.
- ***O-antigens info field***: the detected O-antigen group (O1 to O188) or "Unknown" if no O-antigen group could be assigned. The default field name is **O_antigens**.

2.7 Plasmid settings

The *Plasmid* tab in the *E. coli genotyping settings* dialog box (see Figure 2.7) groups all settings for plasmid detection.

Plasmids are detected at two different levels by the plugin. A first analysis screens for the presence of origins of replication (ori) using a BLAST approach. A second analysis screens for the presence of reference plasmids in the genome using sourmash [1].

Under ***Knowledgebase***, the ***Name*** and ***Version*** of the specified knowledge base version for this feature is shown. When no knowledge base is specified yet, "<None>" will be indicated in both fields. When a specified knowledge base version cannot be found (e.g. because it is deleted), "<Missing>" is shown in both fields. A different knowledge base version can be selected by pressing the <***Change...***> button.

With ***Check for updates on startup*** checked, BIONUMERICS will check if a newer knowledge base version is available online for this feature each time the database is opened. This requires an internet connection.

See 3 for more information on *E. coli functional genotyping plugin* knowledge bases.

In the ***Ori*** panel, two settings for the BLAST algorithm can be specified:

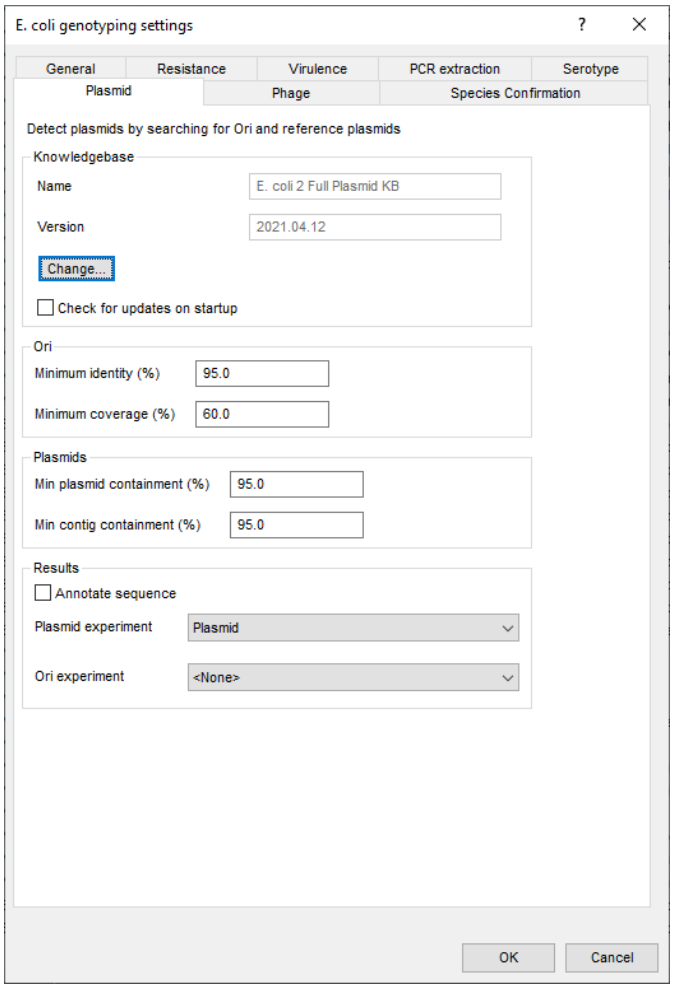

Figure 2.7: The *E. coli genotyping settings* dialog box, *Plasmid* tab.

- ***Minimum identity (%)*** is the minimum sequence identity (as percentage) of the query sequence against the knowledge base's reference sequences.
- ***Minimum coverage*** specifies the minimum overlap (as percentage) between the subsequence found in the target assembly sequence and the reference sequence from the knowledge base.

In the ***Plasmids*** panel, two settings for the sourmash algorithm can be specified:

- ***Min plasmid containment (%)*** is the minimum containment score (expressed as a percentage) of a plasmid sequence in the target assembly sequence. If the minimum plasmid containment score is set to e.g. 95% and less than 95% of the plasmid sequence is contained in the query sequence, the plasmid will not be reported.
- ***Min contig containment (%)*** is the minimum containment score (expressed as a percentage) of a contig sequence in the detected plasmid sequence. If the minimum contig containment score is set to e.g. 95% and less than 95% of the contig sequence is contained in the detected plasmid sequence, the contig will not be reported.

In the ***Results*** panel, the experiment types and entry information fields to which the screening results will be written can be dictated. Use the drop-down menu to choose an existing experiment type or information field or select the <***Create***> option to create a new experiment type or information field. A default name is suggested, but you can adjust this if you want to.

Following character experiments can optionally be specified:

- ***Plasmid experiment***: contains the results of the full plasmids detection: 0 = not detected, when detected the % containment of the detected plasmid is shown. The default name is **Plasmid**.
- ***Ori experiment***: contains the results of the plasmid ori detection: 0 = not detected, when detected the % BLAST identity with the ori reference sequence is shown. The default name is **Ori**.

2.8 Phage settings

The *Phage* tab in the *E. coli genotyping settings* dialog box (see Figure 2.8) groups all settings for phage detection.

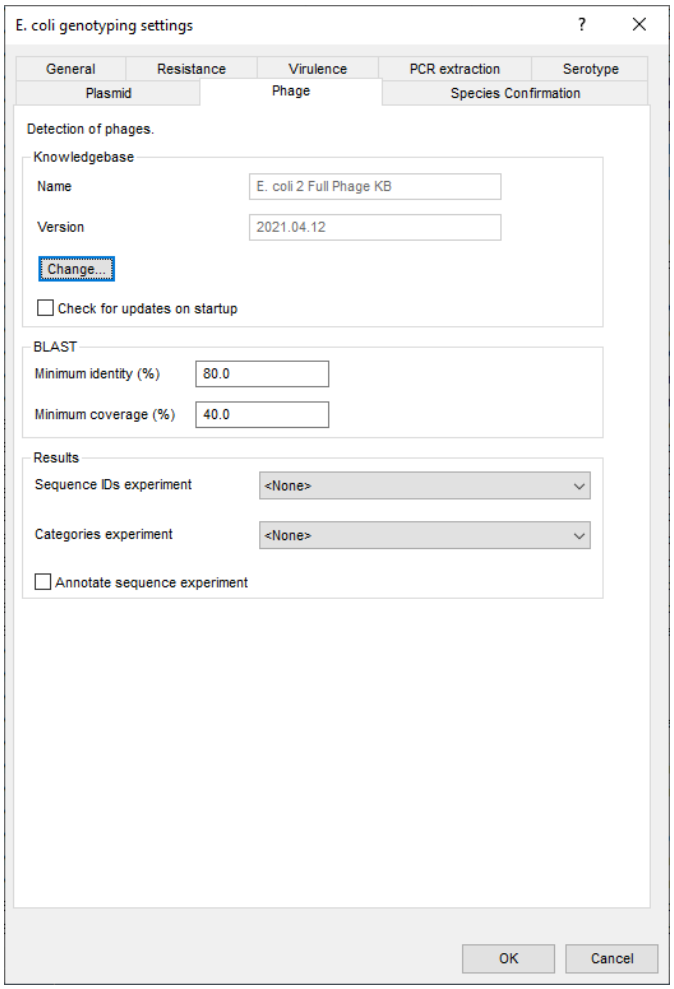

Figure 2.8: The *E. coli genotyping settings* dialog box, *Phage* tab.

Phages are detected in the assembled *E. coli* genome by a BLAST-based screening against a collection of known full length phages.

Under ***Knowledgebase***, the ***Name*** and ***Version*** of the specified knowledge base version for this feature is shown. When no knowledge base is specified yet, "<None>" will be indicated in both fields. When a specified knowledge base version cannot be found (e.g. because it is deleted),

"<Missing>" is shown in both fields. A different knowledge base version can be selected by pressing the <***Change...***> button.

With ***Check for updates on startup*** checked, BIONUMERICS will check if a newer knowledge base version is available online for this feature each time the database is opened. This requires an internet connection.

See 3 for more information on *E. coli functional genotyping plugin* knowledge bases.

In the ***BLAST*** panel, two settings for the BLAST algorithm can be specified:

- ***Minimum identity (%)*** is the minimum sequence identity (as percentage) of the query sequence against the knowledge base's reference sequences.
- ***Minimum length for coverage*** specifies the minimum overlap (as percentage) between the subsequence found in the target assembly sequence and the reference sequence from the knowledge base.

In the ***Results*** panel, the experiment types and entry information fields to which the screening results will be written can be dictated. Use the drop-down menu to choose an existing experiment type or information field or select the <***Create***> option to create a new experiment type or information field. A default name is suggested, but you can adjust this if you want to.

Following character experiments can optionally be specified:

- ***Sequence IDs experiment***: contains the results of the phages detection by sequence IDs: 0 = not detected, when detected the % of the detected full phage is shown. The default name is **Phage_seq_ids**.
- ***Categories experiment***: contains the results of the phages detection by phage categories: 0 = not detected, when detected the % of the detected full phage is shown. The default name is **Phage_categories**.

Check ***Annotate sequence experiment*** to annotate the input sequence with the detected genotyping features.

2.9 Species confirmation settings

The *Species Confirmation* tab in the *E. coli genotyping settings* dialog box (see Figure 2.9) groups all settings for species confirmation.

Species confirmation is achieved by comparing the input genome against a set of type strain and other reference genomes from the same genus (or highly related genera) using sourmash [1]. The species confirmation knowledge base specifies ascending thresholds for the hierarchical levels genus, species and subspecies (if subspecies are defined within the species). When the sourmash containment score is above the genus threshold but below the species threshold, only the genus name is returned as identification result. When the species threshold is met but not the subspecies threshold, the binomial name (genus + species) is returned. Finally, when subspecies are defined for the organism and the containment score exceeds the subspecies threshold, the species confirmation result consists of the genus, species and subspecies name.

Under ***Knowledgebase***, the ***Name*** and ***Version*** of the specified knowledge base version for this feature is shown. When no knowledge base is specified yet, "<None>" will be indicated in both fields. When a specified knowledge base version cannot be found (e.g. because it is deleted),

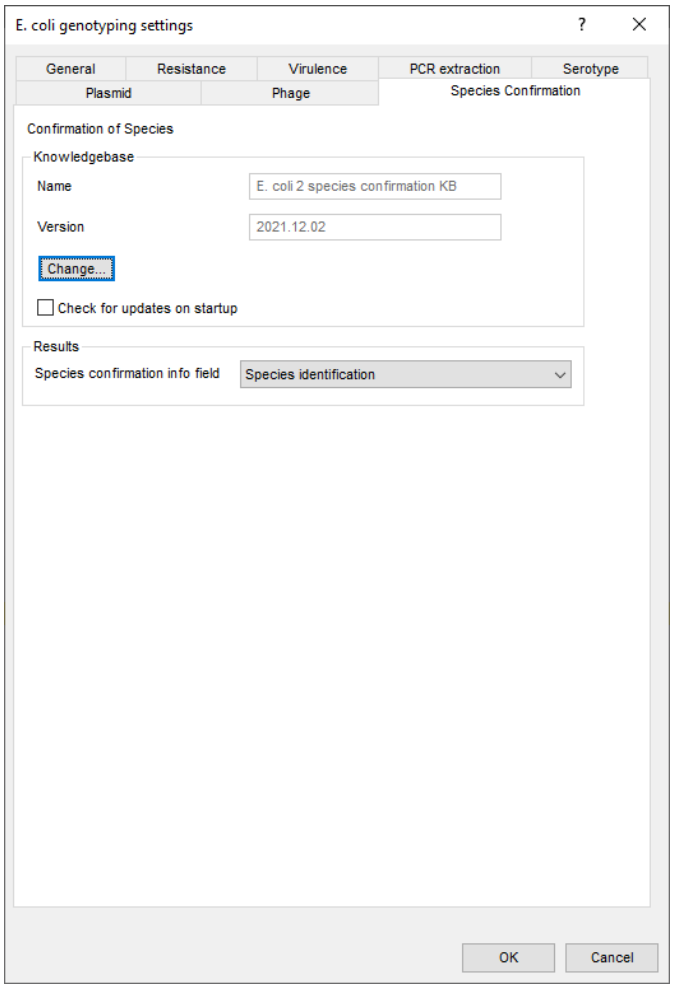

Figure 2.9: The *E. coli genotyping settings* dialog box, *Species Confirmation* tab.

”<Missing>” is shown in both fields. A different knowledge base version can be selected by pressing the <***Change...***> button.

With ***Check for updates on startup*** checked, BIONUMERICS will check if a newer knowledge base version is available online for this feature each time the database is opened. This requires an internet connection.

See 3 for more information on *E. coli functional genotyping plugin* knowledge bases.

In the ***Results*** panel, the experiment types and entry information fields to which the screening results will be written can be dictated. Use the drop-down menu to choose an existing experiment type or information field or select the <***Create***> option to create a new experiment type or information field. A default name is suggested, but you can adjust this if you want to.

A ***Species confirmation info field*** can optionally be specified. This will contain the identification result, i.e. the name of the best matching reference in the species confirmation knowledge base. Depending on the score, this may be the genus name, the genus and species name or genus, species and subspecies name.

Chapter 3

Genotyping plugin knowledge bases

3.1 Introduction

Nearly all genotyping features make use of a knowledge base of some kind. Knowledge bases are at the heart of functional genotyping because they literally contain the knowledge on how to interpret genome sequences in function of the feature they were designed for. By providing the knowledge bases online from https://www.bionumerics.com, they can easily be updated without the need to install a new plugin version.

All online knowledge bases are based on curated public repositories (e.g. ResFinder https://cge.cbs.dtu.dk/services/ResFinder/) and converted to the specific format required by the plugin. Each knowledge base has a change log with detailed information on the changes in each version.

Organism-specific functional genotyping plugins can only use *online* knowledge bases. The downloaded knowledge bases are encrypted and cannot be modified by the user. For using your own, custom knowledge bases we recommend the *Custom genotyping plugin*, which was specifically designed for this purpose.

By default, the plugin uses the most recent knowledge base version available at the time of installation. When more than one knowledge base version is available online, users can specify which version to use.

3.2 Specifying a different knowledge base version

The knowledge base for a genotyping feature is specified in the genotyping settings.

2.1 In the *Main* window, select ***E. coli*** > ***Settings...*** to open the *E. coli genotyping settings* dialog box.

In all tabs of the *E. coli genotyping settings* dialog box (except the *General tab*) a ***Knowledgebase*** section is available.

2.2 Click on the tab of the feature for which you want to change the knowledge base and press <***Change...***>.

This action opens the *Change knowledge base* dialog box (see Figure 3.1).

The *Change knowledge base* dialog box shows the downloaded knowledge bases for this feature. The currently used knowledge base (if specified) will be highlighted by default. Following information is displayed about the knowledge bases:

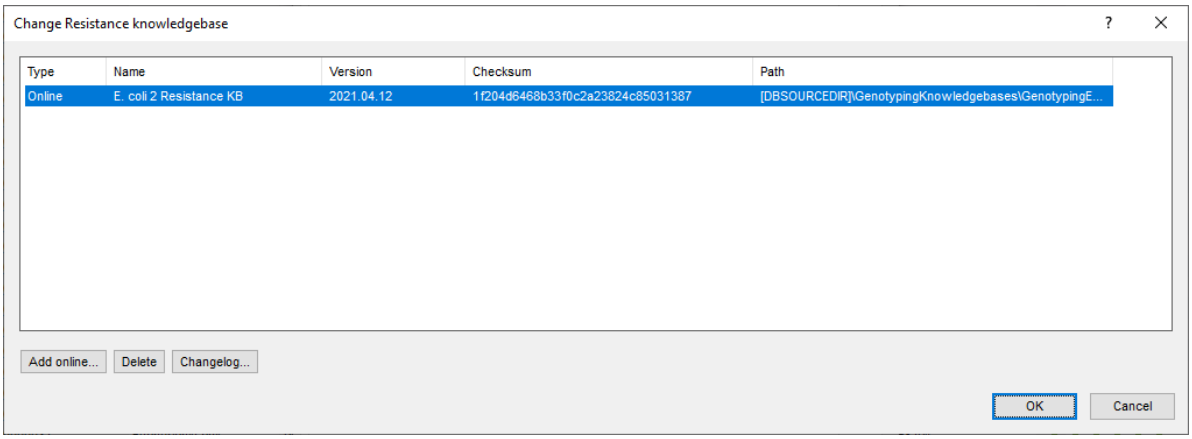

Figure 3.1: The *Change knowledge base* dialog box.

- 'Type': either Online or Local. The knowledge base type will always be Online for all organism-specific genotyping plugins and for the *Resistance detection plugin*.
- 'Name': the name of the knowledge base.
- 'Version': the knowledge base version, typically a last modified date formatted as YYYY.MM.DD.
- 'Checksum': MD5 checksum, used to verify the integrity of the downloaded knowledge base.
- 'Path': file path, relative to the source files directory (indicated with the token `DBSOURCEDIR`), where the knowledge base is downloaded to.

To manually check if other versions (newer or older) of the knowledge base are available online, press <***Add online...***>. This opens the *Download online knowledge base* dialog box (see Figure 3.2).

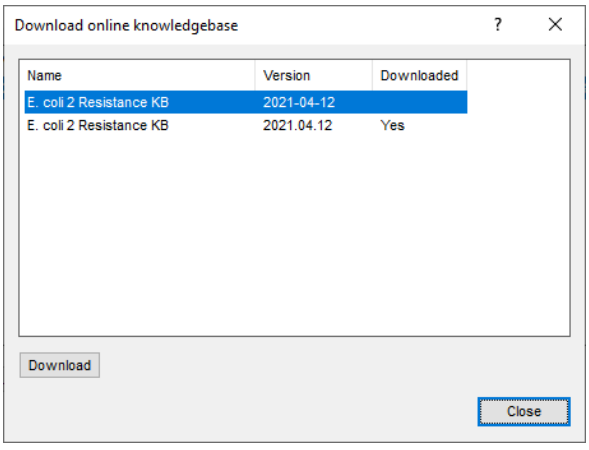

Figure 3.2: The *Download online knowledge base* dialog box.

Opening this dialog will check the BIONUMERICS website (`https://www.bionumerics.com`) for the latest knowledge base version for this feature. When the latest version is already downloaded (i.e. the knowledge base is up to date), the 'Downloaded' field will indicate "Yes". If this is not the case, you can press the <***Download***> button to download the latest version.

Close the dialog with <***Close***>.

A highlighted knowledge base can be deleted by pressing <***Delete***>. The software will ask for confirmation before actually removing the knowledge base.

By pressing <***Changelog...***> you are referred to the change log page on the BIONUMERICS website of the online knowledge base you have selected. Here you can find information regarding the source(s) used to create the knowledge base and the changes made in respect to prior versions of the knowledge base. This information should help you decide which knowledge base version to use.

To specify a different knowledge base for the feature, click on the preferred knowledge base to highlight it and press <***OK***>. This action will close the *Change knowledge base* dialog box and will show the genotyping settings again with the newly specified knowledge base.

3.3 Automated check for knowledge base updates

For each feature that uses an online knowledge base, there is a check box ***Check for updates on startup***. When this is checked, BIONUMERICS will automatically connect to `https://www.bionumerics.com` to check for knowledge base updates each time the database is opened. When updates are available, the *Update knowledge bases* dialog box pops up (see Figure 3.3).

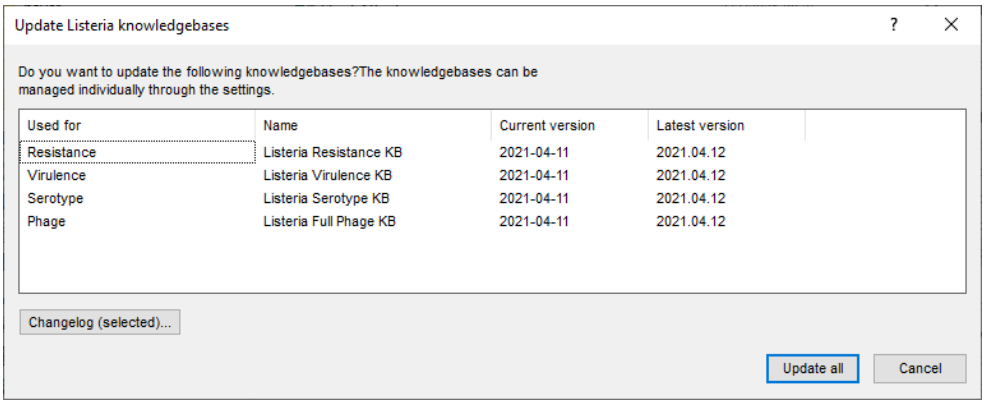

Figure 3.3: The *Update knowledge bases* dialog box pops up at startup when one or more new knowledge bases are available online.

By pressing <***Changelog (selected)***> you are referred to the change log page on the BIONUMERICS website of the highlighted online knowledge base.

Press <***Update all***> to apply all knowledge base updates. The updated knowledge bases will be downloaded and set as default in the genotyping settings, meaning that the updated knowledge base will be used the next time an analysis is run for the corresponding feature.

Knowledge base updates can only be done in bulk from this dialog box, i.e. for all available updates at once. If you need to update one or more knowledge bases selectively, press <***Cancel***> in the *Update knowledge bases* dialog box and update the knowledge bases one by one via the genotyping settings (see 3.2).

Chapter 4

E. coli genotyping analysis

4.1 Selecting entries

Once the plugin is installed and the settings have been specified, the actual screening of the genome sequences of the selected entries is an easy process.

Analyses are performed on the selected entries in the database. For example, to select a single entry, hold the **Ctrl**-key and click on the entry in the *Database entries* panel. Alternatively, use the **space bar** or click the ballot box next to the entry. In order to select a range of entries, hold the **Shift**-key and click on the last entry in the range.

More options for selecting entries can be found in the BIONUMERICS reference manual (see the Reference manual, Chapter Database entries).

4.2 Starting an analysis

Screening for the phenotypic traits can be done for all tools checked in the *E. coli genotyping settings* dialog box (using ***E. coli*** $>$ ***Analysis*** $>$ ***All Enabled***) or for each tool separately with the corresponding command (***E. coli*** $>$ ***Analysis***...).

The analysis time increases proportionally with the number of selected entries and the number of enabled genotyping features. It also depends on the type of feature and whether or not additional results are stored in character experiments and/or information fields. A complete analysis may take up to several minutes or even hours.

When the analysis is finished, the progress bar disappears. The screening results are stored in the database experiments and information fields which you have defined in the *E. coli genotyping settings* dialog box. The settings can always be consulted or adjusted using ***E. coli*** $>$ ***Settings...*** (see 2 for details).

Chapter 5

E. coli genotyping reports

5.1 Opening functional genotyping reports

A functional genotyping report (see Figure 5.1) can be opened for the selected entries with ***E. coli*** $>$ ***Reports...***.

Clicking on an entry in the *Entries* panel of the *Genotyping report* window (or using the up and down arrow keys on the keyboard) shows the report for the highlighted entry.

At the top of each report the creation date of the report (***Date***), the Key (***Name***), and information fields that were checked in the *General* tab of the genotyping settings are displayed, followed by a summary of the results of all analyzed traits.

Selecting ***File*** $>$ ***Exit*** closes the *Genotyping report* window.

5.2 Report styles

In the *Genotyping report* window, three different report styles can be applied from the drop-down list in the panel header or via the menu (***Report*** $>$ ***Report styles***):

1. ***Summary***: only a summary of the results is shown.
2. ***Default***: the summarized results and most details are shown in a tabular format. In this report style, all columns of the results tables can be sorted alphabetically or numerically by clicking on their headers.
3. ***Complete***: the summarized results and all available details are shown. More exhaustive information is presented in an additional row, for example descriptions of the detected genes, decision trees, etc.. Result tables cannot be sorted in this report style.

5.3 Details section

5.3.1 Introduction

The Details section in the genotyping report contains a detailed result table for each analyzed genotyping feature.

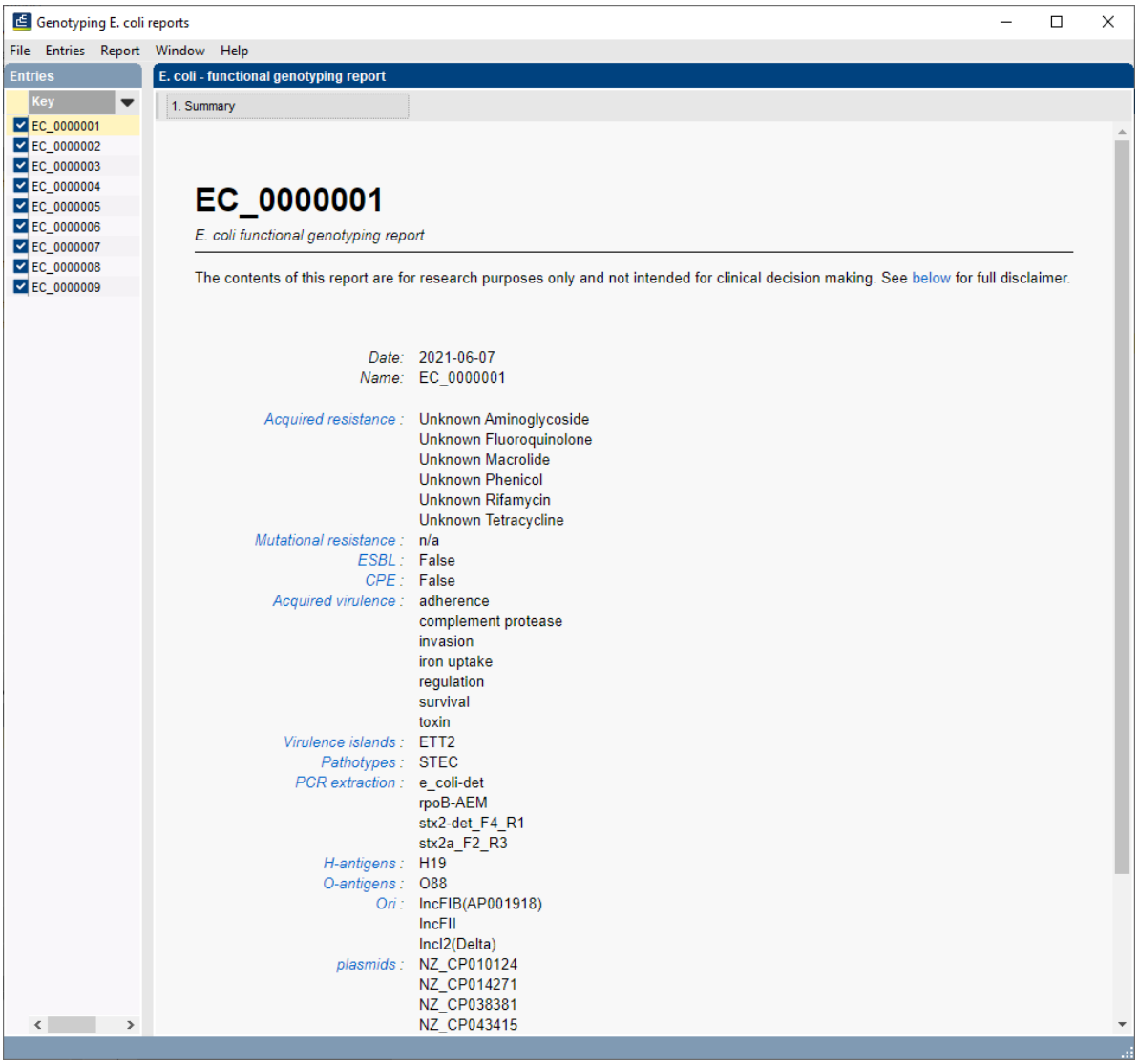

Figure 5.1: The *Genotyping report* window, showing a report for sample EC_0000001.

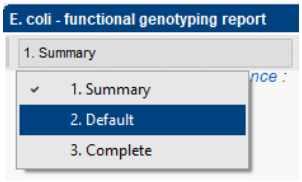

Figure 5.2: Drop-down list to select report styles in the *Genotyping report* window.

Some result tables contain hyperlinks. These hyperlinks open in the default web browser, except for the 'Position' field. The latter links open the *Sequence editor* window in BIONUMERICS with the positions highlighted on the sequence.

5.3.2 Resistance

All detected acquired resistance traits are listed in the table, with following fields:

- **Trait:** Name of the resistance trait.
- **Locus:** Detected locus that confers the trait.
- **Coverage (%):** Percent overlap between the query and the target sequence.
- **Identity (%):** BLAST identity, expressed as a percentage.
- **Position:** Position(s) of the BLAST match on the input genome. Multiple positions are possible when ***Combine fragments*** was checked in the genotyping settings.
- **Accession:** Link to the GenBank accession of the detected locus.
- **Description:** Description for the detected locus (in ***Complete*** template only).
- **Publication:** Link to the publication in PubMed describing the detected locus (in ***Complete*** template only).

All detected mutational resistance traits are listed in the table, with following fields:

- **Trait:** Name of the resistance trait.
- **Locus:** Name of the locus in which the mutation occurs.
- **Level:** The level on which the mutation was detected; either the amino acid (AA) or nucleic acid (NA) level. The latter is used mainly for 16S rDNA mutations.
- **Position:** Position(s) of the mutation in the locus.
- **Reference:** Amino acid in the reference sequence.
- **Mutation:** Amino acid in the mutated allele.
- **Requirements:** This field will be empty if the mutation confers resistance on its own (no requirements). In case the mutation should occur in combination with another mutation to confer resistance and the required mutation is present, ”Pass” is displayed. When the gene that contains the required mutation could not be (fully) detected, ”Uncertain” is shown.
- **Decision tree:** The decision tree leading to the result. Each mutation appears in a color: green when the mutation is detected, red when the locus is detected but the mutation is not present and yellow when its presence is indecisive due to locus absence (in ***Complete*** template only).

Exact matches with enzymes associated with respectively ESBL and CPE resistance types are listed in two separate tables, containing following fields:

- **Protein name:** Name of the detected enzyme, associated with the resistance type.
- **Query position:** Position on the input sequence where the gene for the enzyme is detected.
- **Reference accession:** Link to the GenBank accession of the reference sequence.
- **Reference position:** Position on the reference sequence. This always covers the full CDS, since a 100% overlap is required.

5.3.3 Virulence

All detected acquired virulence traits are listed in the table, with following fields:

- **Trait:** Name of the virulence trait.
- **Locus:** Detected locus that confers the trait.
- **Coverage (%):** Percent overlap between the query and the target sequence.
- **Identity (%):** BLAST identity, expressed as a percentage.
- **Position:** Position(s) of the BLAST match on the input genome. Multiple positions are possible when ***Combine fragments*** was checked in the genotyping settings.
- **Accession:** Link to the GenBank accession of the detected locus.
- **Description:** Description for the detected locus (in ***Complete*** template only).
- **Publication:** Link to the publication in PubMed describing the detected locus (in ***Complete*** template only).

All detected virulence islands are listed in the table, with following fields:

- **Virulence island:** Name of the virulence island.
- **Found:** Number of detected loci for this virulence island.
- **Total:** Total number of loci present in the knowledge base for this virulence island.
- **Percentage:** Percentage of detected loci, i.e. $100\frac{Loci_{Found}}{Loci_{Total}}$.
- **Loci:** Locus names of all detected loci (in ***Complete*** template only).

All detected loci that are associated with a certain pathotype are listed, with following fields:

- **Pathotype:** Name of the pathotype.
- **Locus:** Detected locus, associated with the pathotype.
- **Coverage (%):** Percent overlap between the query and the target sequence.
- **Identity (%):** BLAST identity, expressed as a percentage.
- **Position:** Position(s) of the BLAST match on the input genome. Multiple positions are possible when ***Combine fragments*** was checked in the genotyping settings.
- **Accession:** Link to the GenBank accession of the locus.

5.3.4 PCR extraction

All detected PCR targets are listed in the table, with following fields:

- **Identifier:** ID of the PCR target.
- **Position:** Position on the input sequence where the target was detected.

- **Strand:** Strand on which the PCR primers were detected: either the base strand ('base') or the reverse-complement ('revcomp') strand.
- **Length:** Length in nucleotides of the in silico PCR product.
- **Reference length:** Expected length of the target sequence.

5.3.5 Serotype

The *Serotype* section shows the predicted serotype as **H-antigens** and **O-antigens** detected. The p-value ratio or discrimination score (see 2.6) shown is for the best match against its runner-up. When the p-value ratio is smaller than ten, the result table will list the second-best matching antigen in addition to the best matching, with following fields:

- **Antigen:** Name of the antigen.
- **Locus:** Name of the detected locus that corresponds to the antigen.
- **Coverage (%):** Percent overlap between the query and the target sequence.
- **Identity (%):** BLAST identity, expressed as a percentage.
- **Position:** Position of the BLAST match on the input genome.
- **Accession:** Link to the GenBank accession of the detected locus.

5.3.6 Ori

All detected plasmid origins of replication are listed in the table, with following fields:

- **Ori:** Name of the detected ori.
- **Coverage (%):** Percent overlap between the query and the target sequence.
- **Identity (%):** BLAST identity, expressed as a percentage.
- **Length:** Length in nucleotides of the detected sequence.
- **Position:** Position of the BLAST match on the input genome.
- **Accession:** Link to the GenBank accession of the detected locus.

5.3.7 Plasmids

All detected plasmids are listed in the table, with following fields:

- **Plasmid accession**: Link to the GenBank accession of the detected plasmid sequence.
- **Overlap (bp):** Overlap between the detected plasmid sequence and the query sequence.
- **Largest overlapping contig**: Start and end position of the largest overlapping contig.
- **Contained contigs**: The number of contigs of which the contig containment score is higher than the user-defined minimum contig containment score.

- **Containment (%)**: The percentage of the plasmid sequence which is contained in the query sequence.
- **Associated ori**: Ori associated with the detected plasmid.

5.3.8 Phage

All detected phages are listed in the table, with following fields:

- **Category:** Category to which the detected phage belongs.
- **Sequence ID:** Name of the detected phage.
- **Coverage (%):** Percent overlap between the query and the target sequence.
- **Identity (%):** BLAST identity, expressed as a percentage.
- **Length:** Length in nucleotides of the detected sequence.
- **Accession:** Link to the GenBank accession of the detected phage sequence.

5.3.9 Species confirmation

The best matching identification with one of the reference genomes from the knowledge base is listed in the table, with following fields:

- **Confirmation:** The species confirmation (i.e. identification) result, at genus, species or subspecies level.
- **Containment score (%):** Sourmash containment score, expressed as a percentage.
- **Largest exceeded threshold (%):** The highest containment score threshold (as defined in the knowledge base) that was exceeded.
- **Ref. genome name:** Organism name of the reference genome as obtained from GenBank.
- **Accession:** GenBank accession of the reference genome sequence.
- **Taxid:** Taxonomy ID as used in NCBI's taxonomy browser (https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi).

5.4 Info section

At the bottom of each report, an *Info section* is shown which contains information regarding the analysis date, plugin version, knowledge base name and version, and settings of each analysis.

5.5 Exporting report information

The genotype information for all selected entries in the *Genotyping report* window can be exported with ***Entries*** > ***Export selected***. The results for the currently shown report can be exported with ***Report*** > ***Export current***.

A Tab Separated Values (`*.tsv`) file is created for each functionality and stored in the report export directory as specified in the genotyping settings. The location of the files is opened after the export.

The displayed report can be printed directly to a printer using ***Report*** > ***Print...***.

The dialog box that appears is the standard Windows Print dialog box, allowing you to choose a printer and change the properties. Depending on your system, this also allows printing to a PDF file to create an export of the displayed report. Please note that background colors, such as those in lists and mutational decision trees, may be lost in this step regardless of printer settings.

Bibliography

[1] C Titus Brown and Luiz Irber. sourmash: a library for minhash sketching of dna. *Journal of Open Source Software*, 1(5):27, 2016.

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