BIONUMERICS Tutorial:

Creation of custom knowledgebases

1 Aim

The *Custom genotyping* plugin allows the creation and usage of custom knowledgebases for the detection and extraction of sequences using a BLAST or in silico PCR approach, for the detection of mutations using a BLAST approach and for the confirmation of species identity using sourmash. Depending on the type of knowledgebase (BLAST-based, PCR-based or MinHash-based) and depending on the genotyping feature, the knowledgebases need to be in a specific format.

The *Custom genotyping* plugin itself offers functionality to guide and help the user in the creation of custom knowledgebases. A typical workflow for the creation of custom knowledgebases therefore consists of the following steps:

- Collect data to populate the knowledgebase.
- Install the *Custom genotyping* plugin in a BIONUMERICS database.
- Create example knowledgebases.
- Put data into the correct knowledgebase format.

This tutorial will show you how to create a custom BLAST-based knowledgebase for acquired and mutational resistance detection, an in-silico PCR based knowledgebase and a MinHash-based knowledgebase.

2 Collect data to populate the knowledgebase

All *Custom genotyping* plugin features make use of a knowledgebase of some kind. Knowledgebases 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. Both online repositories as data obtained during your own research can serve as input for the generation of a custom knowledgebase.

For this tutorial data was already extracted from the Resfinder database version 2022-02-04 and the PointFinder database version 2021-02-01 (see https://cge.cbs.dtu.dk/services/ResFinder/) for the custom BLAST-based knowledgebases and from NCBI for the custom MinHash-based knowledgebase. The example data that will be used in this tutorial can be downloaded from the BIONUMERICS website (https://www.bionumerics.com/download/sample-data, click on "Custom knowledgebase creation data").

3 Install the custom genotyping plugin in a BIONUMERICS database

1. Create a new database (see tutorial ”Creating a new database”) or open an existing database.

Proceed as follows to install the *Custom genotyping plugin*:

2. Call the *Plugins and Scripts* dialog box from the *Main* window with ***File*** > ***Install / remove plugins...*** ( ).
3. Select the *Custom genotyping plugin* from the list and press the <***Install***> button.
4. Confirm the installation of the plugin.

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

5. Press <***OK***> in the confirmation message.
6. Press <***Close***> to close the *Plugins and Scripts* dialog box.
7. Close and re-open the database to complete the installation of the plugin.

The *Custom genotyping plugin* installs menu items in the main menu of the software under ***Genotyping*** (see Figure 1).

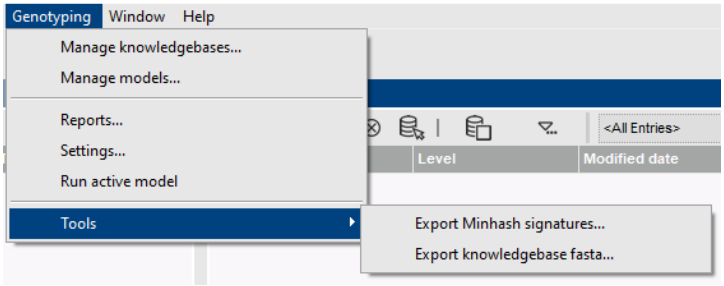


Figure 1: New menu items, available after installation of the *Custom genotyping plugin*.

4 Create example knowledgebases

BLAST-, PCR- and MinHash-based example knowledgebases can be created by the *Custom genotyping* plugin to illustrate the required knowledgebase format.

1. Select ***Genotyping*** > ***Manage knowledgebases...*** to open the *Manage knowledge bases* dialog box (see Figure 2).
2. Press the <***Create example...***> button.

This opens the *Create example knowledge base* dialog box (see Figure 3).

3. Press <***Browse...***> and specify a directory in which you want to create the example knowledgebases.

Three example knowledgebases are available in the drop-down list: ***BLAST based***, ***PCR based*** and ***minhash based***. As we want to create custom knowledgebases of these three types in the

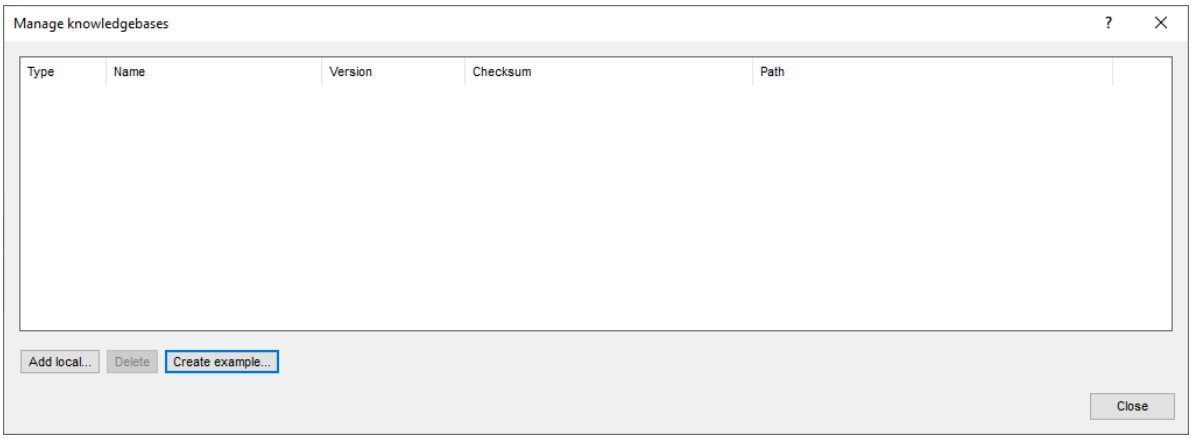


Figure 2: The *Manage knowledge bases* dialog box.

current tutorial, we will create an example knowledgebase for each knowledgebase type.

4. With the ***BLAST based*** type selected in the drop-down list, press <***Create***> to create the selected example knowledgebase.

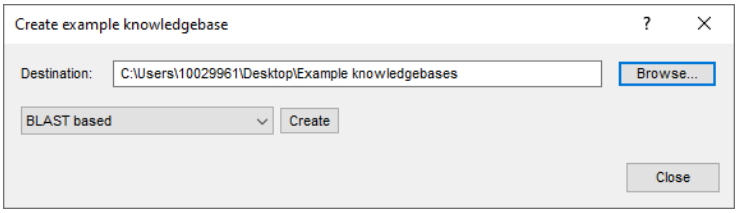


Figure 3: The *Create example knowledge base* dialog box.

The software automatically opens the knowledgebase in Windows Explorer.

5. Select the ***PCR based*** type in the drop-down list and press <***Create***> to create the selected example knowledgebase. Do the same for the ***minhash based*** type.

The example knowledgebase folders contain examples of the required files and contain a `README.md` file which further explains the format. We will use these files and the information provided in the `README.md` file to generate custom knowledgebases for the open-source data we collected from ResFinder and NCBI.

6. Press <***Close***> twice to return to the *Main* window.

5 Put data into the correct knowledgebase format

5.1 Acquired resistance BLAST-based knowledgebase

We will first create a BLAST-based knowledgebase for the detection of disinfectant resistance genes. The disinfectant resistance genes and the associated metadata (e.g. gene accession, resistance gene class, PubMed ID and the disinfectant to which the gene confers resistance to) which will be used in this tutorial have been extracted from the ResFinder database version 2022-02-04 (see https://cge.cbs.dtu.dk/services/ResFinder/).

Open the `README.md` file in the exported BLAST-based example knowledgebase folder with e.g. Notepad. The `README.md` file explains that not all files in the example knowledgebase folder are required for all features. Aside from the `info file` required by all knowledgebases, additional files required for the sequence detection, sequence extraction and acquired traits detection features are the following:

- A `sequences.fasta` file: a file containing the sequences you want to detect. If you want to detect acquired traits this file should also contain at least one occurence of the ”@trait” key in each header.
- A `trimming_patterns.tsv` file: a file which is only required if you intend to use trimming patterns for sequence correction.

For each required or optional file the `README.md` file provides detailed information on the required format. We will first create a folder for our disinfectant resistance knowledgebase and include an appropriate info.json file.

1. Create a new folder (e.g. `Disinfectant_resistance`) on your computer for the disinfectant resistance knowledgebase.
2. Copy the `info.json` file from the BLAST-based example knowledgebase folder into this new folder and open the json file in e.g. Notepad.
3. Optionally adapt the `info.json` file by changing the version, name, description and changelog and save your changes (see Figure 4).

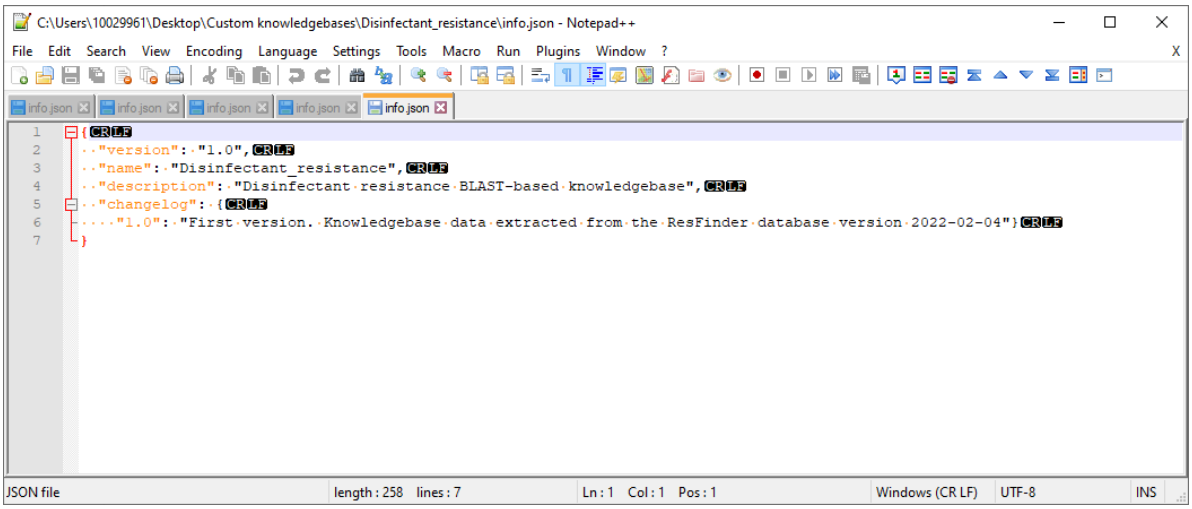


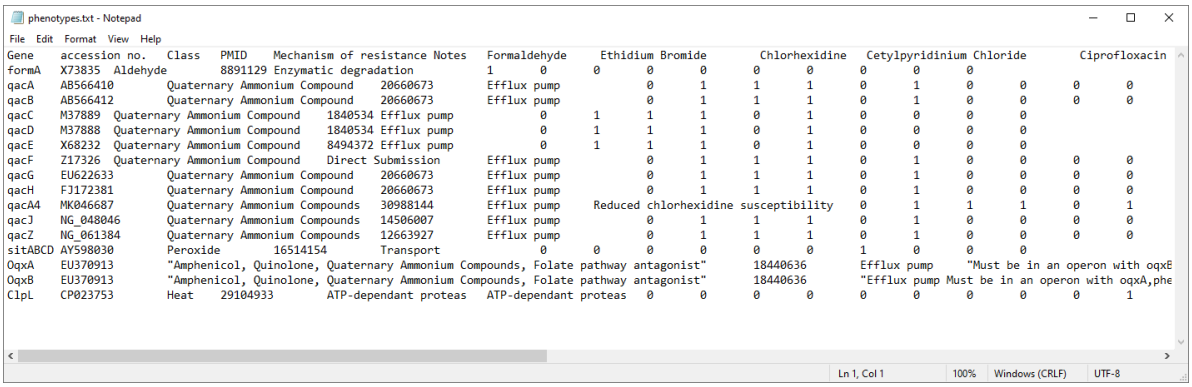
Figure 4: The `info.json` file.

The `sequences.fasta` file can easily be created from the data obtained from ResFinder by using the ***export knowledgebase fasta*** tool provided by the *Custom genotyping* plugin. To be able to use the ***export knowledgebase fasta*** tool we first need to import our disinfectant resistance data into the BIONUMERICS database.

First we will import the metadata and character data present in the `phenotypes.txt` file (see Figure 5) which is present in the ”Custom knowledgebase creation data” folder previously downloaded from our website.

Since we will import character data (i.e. the disinfectants to which the resistance genes confer resistance to), we will first create a character type to hold this data.

4. In the *Main* window, click on + in the toolbar of the *Experiment types* panel and select ***Character type*** from the list. Press <***OK***>.



phenotypes.txt - Notepad

File Edit Format View Help

Gene	accession no.	Class	PMID	Mechanism of resistance	Notes	Formaldehyde	Ethidium Bromide		Chlorhexidine		Cetylpyridinium Chloride			Ciprofloxacin
formA	X73835	Aldehyde	8891129	Enzymatic degradation		1	0	0	0	0	0	0	0	
qacA	AB566410	Quaternary Ammonium Compound	20660673	Efflux pump		0	1	1	1	0	1	0	0	0	0
qacB	AB566412	Quaternary Ammonium Compound	20660673	Efflux pump		0	1	1	1	0	1	0	0	0	0
qacC	M37889	Quaternary Ammonium Compound	1840534	Efflux pump		0	1	1	1	0	1	0	0	0	0
qacD	M37888	Quaternary Ammonium Compound	1840534	Efflux pump		0	1	1	1	0	1	0	0	0	0
qacE	X68232	Quaternary Ammonium Compound	8494372	Efflux pump		0	1	1	1	0	1	0	0	0	0
qacF	Z17326	Quaternary Ammonium Compound	Direct Submission		Efflux pump		0	1	1	1	0	1	0	0	0
qacG	EU622633	Quaternary Ammonium Compound	20660673		Efflux pump		0	1	1	1	0	1	0	0	0
qacH	FJ172381	Quaternary Ammonium Compound	20660673		Efflux pump		0	1	1	1	0	1	0	0	0
qacA4	MK046687	Quaternary Ammonium Compounds	30988144		Efflux pump	Reduced chlorhexidine susceptibility					0	1	1	1	0	1
qacJ	NG_048046	Quaternary Ammonium Compounds	14506007		Efflux pump		0	1	1	1	0	1	0	0	0	0
qacZ	NG_061384	Quaternary Ammonium Compounds	12663927		Efflux pump		0	1	1	1	0	1	0	0	0	0
sitABCD	AY598030	Peroxide	16514154		Transport		0	0	0	0	0	1	0	0		
OqxA	EU370913	"Amphenicol, Quinolone, Quaternary Ammonium Compounds, Folate pathway antagonist"									18440636		Efflux pump	"Must be in an operon with oqxB		
OqxB	EU370913	"Amphenicol, Quinolone, Quaternary Ammonium Compounds, Folate pathway antagonist"									18440636		"Efflux pump Must be in an operon with oqxA,phe			
ClpL	CP023753	Heat	29104933		ATP-dependant proteas	ATP-dependant proteas	0	0	0	0	0	0	0	0	0	1

Ln 1, Col 1 100% Windows (CRLF) UTF-8

Figure 5: The `phenotypes.txt` file.

The *New character type* dialog box prompts you to enter a name for the new character type.

5. Enter a name, for example “Disinfectants” and press <***Next***>.

In the next step of the wizard, the choice is offered between ***Numerical values*** and ***Binary data***.

6. Choose ***Binary data*** since only two possible states are present in our dataset: ”0” and ”1” (see Figure 5). Press <***Next***>.

The wizard asks if the character type has an open (***Yes***) or closed (***No***) character set.

7. Answer ***No*** and press the <***Finish***> button to complete the setup of the new character type.

The *Experiment types* panel now lists the new character type **Disinfectants**.

We will now import the metadata and disinfectants character data.

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

9. Press <***Browse***>, navigate to the ”Custom knowledgebase creation data” folder previously downloaded from our website, select the `phenotypes.txt` file and press <***Open***>.

10. With the ***Import fields and characters (txt file)*** option highlighted, press <***Finish***> and press <***Next***>.

11. Select ”Gene”, ”accession no.”, ”Class”, ”PMID”, ”Mechanism of resistance” and ”Notes” from the list by holding down the **Ctrl**-key. Click on <***Edit destination***>, select ***Entry information field*** and click <***OK***> twice and then <***Yes***> to confirm the creation of new information fields.

12. Select all remaining file fields from the list by holding down the **Ctrl**-key. Click on <***Edit destination***>, select ***Disinfectants*** under the ***Character value*** option and click <***OK***> and then <***Yes***> to confirm the creation of new characters.

The grid panel is updated (see Figure 6).

13. Press <***Preview***> to see what you are about to import.

14. Press the <***Close***> button to close the preview.

15. Press <***Next***> and <***Finish***>, optionally save the import template and press <***OK***>.

In the *Import template* dialog box, the newly created template is automatically selected.

16. Click <***Next***> and <***Finish***> to start the actual import.

The character data is stored in the character type **Disinfectants**.

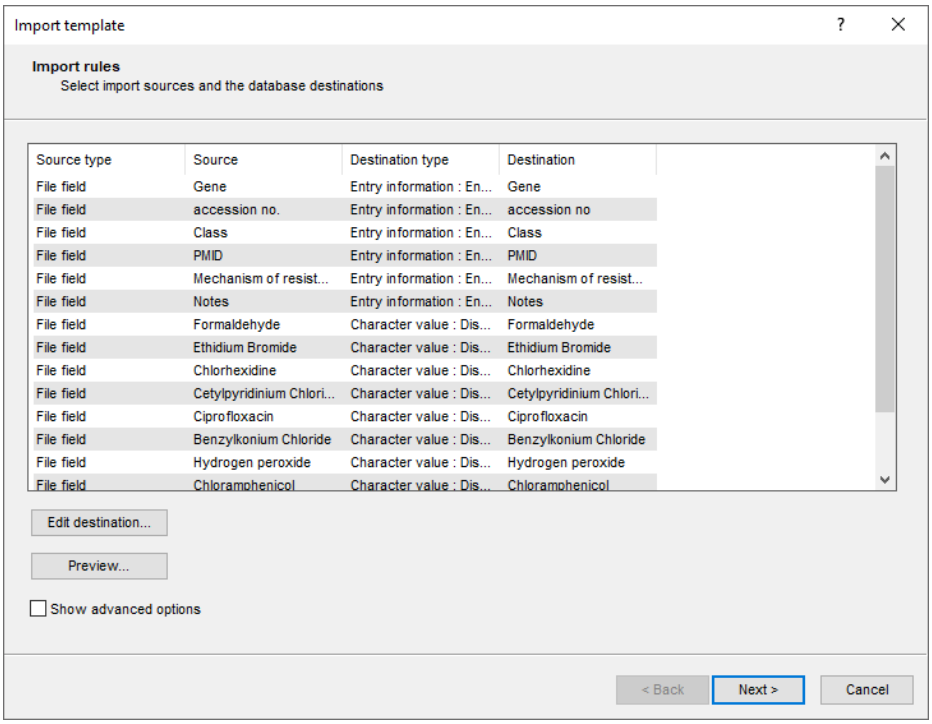


Figure 6: Import rules.

We will now import the resistance genes into our database and link the gene sequences to the entries with the disinfectant metadata and character data.

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

18. Press <***Browse***>, navigate to the ”Custom knowledgebase creation data” folder previously downloaded from our website, select the `disinfectant.fasta` file and press <***Open***>.

19. With the ***Import FASTA sequences from text files*** option highlighted, press <***Finish***>.

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

The import wizard now displays a preview of the sequence data in the FASTA file. From this preview, it is clear that the first FASTA field contains the gene and gene accession number.

21. Press <***Next***>.

22. Click <***Create new***> to create a new import template.

23. Select ***Field 1*** in the list and click <***Edit destination***> or simply double-click on ***Field 1***. Under ***Entry info field***, select ***Gene*** and press <***OK***>.

24. Check the checkbox next to ***Show advanced options*** and click the <***Edit parsing...***> button. As data parsing string use [DATA]_* to parse the gene name from the header information (see Figure 7) and click <***OK***>.

The grid is updated.

25. Optionally, you can press <***Preview***> to obtain a preview of the data you are about to import.

26. Click <***Next***>.

27. Select ***Gene*** as ***Entry link field***. Press <***Finish***>.

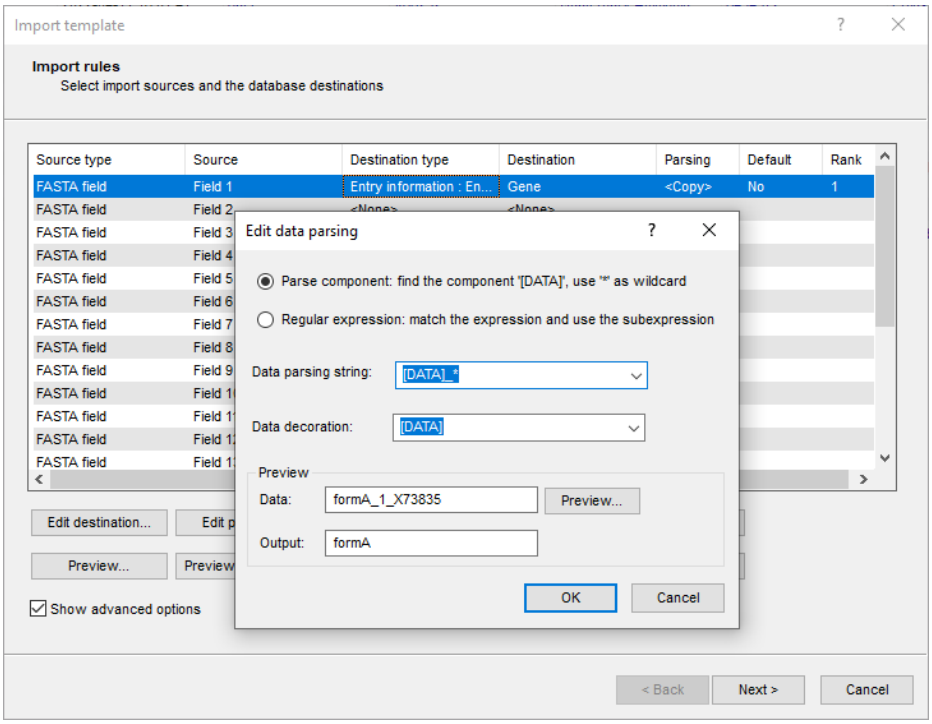


Figure 7: Specifying a data parsing string.

28. Optionally save the template and press <***OK***>.

29. Highlight the newly created template and select ***Create new*** as ***Experiment type***.

30. Press <***Next***>.

31. Specify a sequence type name (e.g. **Disinfectant resistance genes**) and press <***OK***> and confirm the action.

32. Press <***Finish***> to start the import into the database.

The *Main* window now looks like Figure 8.

Now the metadata, character data and sequence data is available in our BIONUMERICS database, the `sequences.fasta` file can easily be created by using the ***export knowledgebase fasta*** tool of the *Custom genotyping* plugin.

33. Select all entries in the database and select ***Genotyping*** > ***Tools*** > ***Export knowledgebase fasta...***.

This action opens the *Export knowledge base fasta* dialog box (see Figure 9).

34. In the ***Configuration*** drop-down list select the "Generic sequence file (standard + traits)" configuration and in the ***Sequence experiment*** drop-down list select the "Disinfectant resistance genes" experiment.

35. Select the "name" target tag and click the <***Edit mapping...***> button. In the "name" drop-down list select "Gene" and click <***OK***>. Repeat this action to map the other target tags to the appropriate information fields (i.e. "accession" to "accession no", "publication" to "PMID" and "description" to "mechanism of resistance").

36. Select the "trait" target tag and click the <***Edit mapping...***> button. In the "trait" drop-down list select the "Disinfectants" character experiment and click <***OK***>.

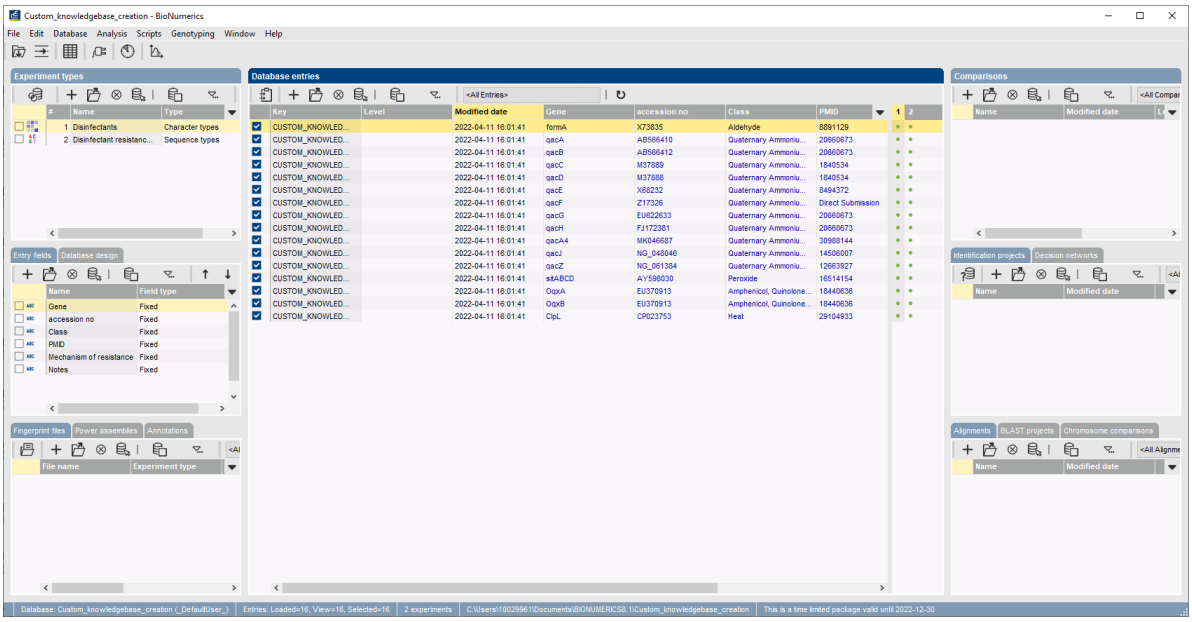


Figure 8: The *Main* window.

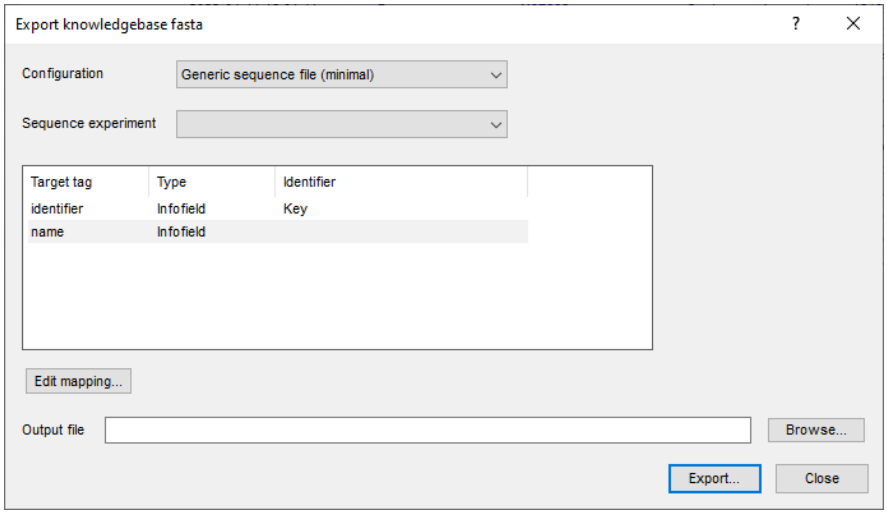


Figure 9: The *Export knowledge base fasta* dialog box.

The sequences.fasta file can now be exported to the respective knowledgebase folder (see Figure 10).

37. Click <***Browse...***> and browse for the `Disinfectant_resistance` knowledgebase folder. As file name enter `sequences.fasta` and click <***Open***>. In the *Export knowledge base fasta* dialog box click <***Export...***> and <***Yes***> to export the fasta file.

The BLAST-based disinfectants resistance knowledgebase (see Figure 11) is now ready to be used by the custom genotyping plugin for the features sequence detection, sequence extraction and acquired traits detection.

This knowledgebase can also be downloaded from the BIONUMERICS website (`https://www.bionumerics.com/download/sample-data`, click on ”Custom knowledgebases”).

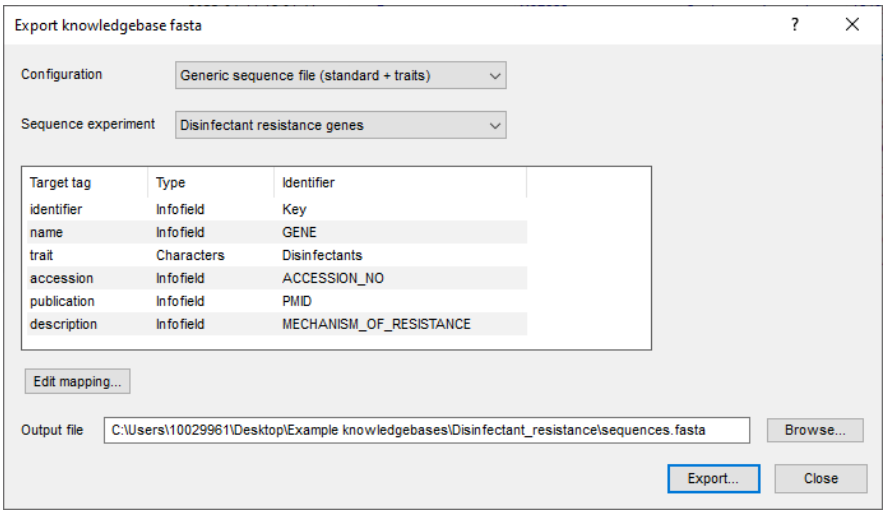


Figure 10: The *Export knowledge base fasta* dialog box with the adapted configuration.

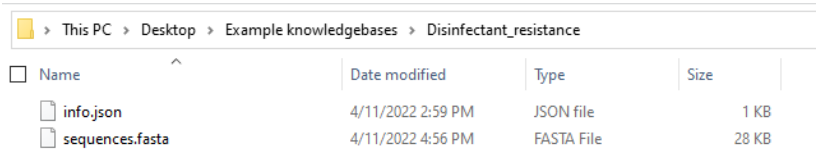


Figure 11: The disinfectant_resistance knowledgebase.

5.2 Mutational resistance BLAST-based knowledgebase

We will now create a BLAST-based knowledgebase for the detection of mutational resistance in *Salmonella*. The mutational resistance genes for *Salmonella* and the associated metadata (e.g. gene, codon position, amino acid in the reference, amino acid in resistant phenotype etc.) which will be used in this tutorial have been extracted from the PointFinder database version 2021-02-01 (see https://cge.cbs.dtu.dk/services/ResFinder/). Open the `README.md` file in the exported BLAST-based example knowledgebase folder with e.g. Notepad. The `README.md` file explains that not all files in the example knowledgebase folder are required for all features. Aside from the `info file` required by all knowledgebases, additional files required for the mutation scanning and mutational traits detection features are the following:

- A `mutational_loci.fasta` file: a file containing the reference sequences of the loci in which mutations will be sought.
- A `mutations.tsv` file: a file containing known mutations about loci in the `mutational_loci.fasta` file. The file is only required if you intend to detect known mutations and/or their associated traits.

For each required or optional file the `README.md` file provides detailed information on the required format.

We will first create a folder for our mutational resistance knowledgebase and include an appropriate info.json file.

38. Create a new folder (e.g. `Mutational_resistance_Salmonella`) on your computer for the mutational resistance knowledgebase.

39. Copy the `info.json` file from the BLAST-based example knowledgebase folder into this new folder and open the json file in e.g. Notepad.

40. Optionally adapt the `info.json` file by changing the version, name, description and changelog and save your changes (see Figure 12).

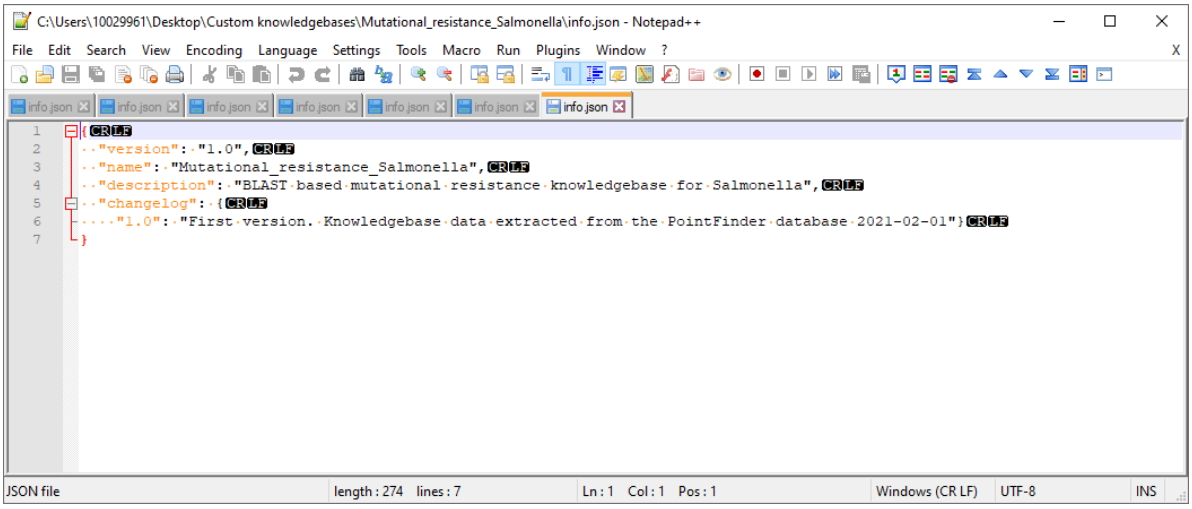


Figure 12: The `info.json` file.

The `mutational_loci.fasta` file can easily be created from the data obtained from PointFinder by using the ***export knowledgebase fasta*** tool provided by the *Custom genotyping* plugin. To be able to use the ***export knowledgebase fasta*** tool we first need to import our mutational resistance data into the BIONUMERICS database.

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

42. Press <***Browse***>, navigate to the "Custom knowledgebase creation data" folder previously downloaded from our website, select all the fasta files in the `Mutational_resistance_Salmonella` folder and press <***Open***>.

43. With the ***Import FASTA sequences from text files*** option highlighted, press <***Finish***>.

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

The import wizard now displays a preview of the sequence data in the FASTA file. From this preview, it is clear that the first FASTA field contains the gene name.

45. Press <***Next***>.

46. Click <***Create new***> to create a new import template.

47. Select ***Field 1*** in the list and click <***Edit destination***> or simply double-click on ***Field 1***. Under ***Entry info field***, select ***Gene*** and press <***OK***>.

The grid is updated.

48. Optionally, you can press <***Preview***> to obtain a preview of the data you are about to import.

49. Click <***Next***>.

50. Do not select an information field as ***Entry link field***. Press <***Finish***>.

51. Optionally save the template and press <***OK***>.

52. Highlight the newly created template and select ***Create new*** as ***Experiment type***.

53. Press <***Next***>.

54. Specify a sequence type name (e.g. **Mutational resistance genes**) and press <***OK***> and confirm the action.

55. Press <***Finish***> to start the import into the database.

The *Main* window now looks like Figure 13.

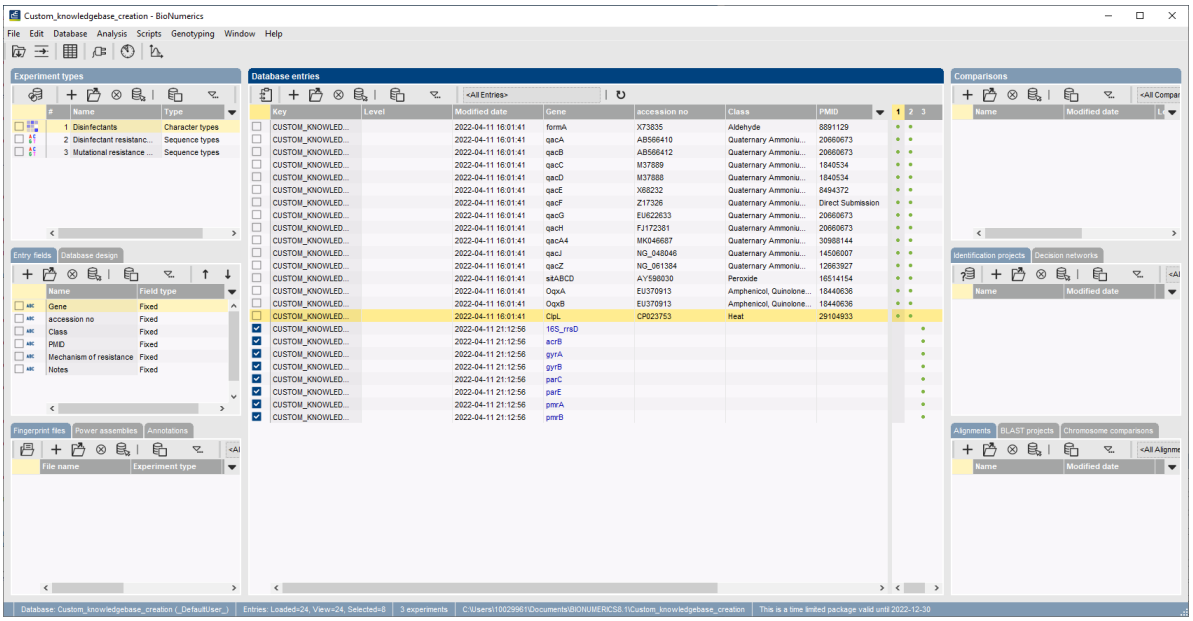


Figure 13: The *Main* window.

Now the sequence data is available in our BIONUMERICS database, the `mutational_loci.fasta` file can easily be created by using the ***export knowledgebase fasta*** tool of the *Custom genotyping* plugin.

56. Select all entries in the database which have data in the ***Mutational resistance genes*** sequence experiment and select ***Genotyping*** > ***Tools*** > ***Export knowledgebase fasta...***.

This action opens the *Export knowledge base fasta* dialog box.

57. In the ***Configuration*** drop-down list select the ”Generic sequence file (minimal)” configuration and in the ***Sequence experiment*** drop-down list select the ”Mutational resistance genes” experiment.

58. Select the ”name” target tag and click the <***Edit mapping...***> button. In the ”name” drop-down list select ”Gene” and click <***OK***>.

The mutational_loci.fasta file can now be exported to the respective knowledgebase folder (see Figure 14).

59. Click <***Browse...***> and browse for the `Mutational_resistance_Salmonella` knowledgebase folder. As file name enter `mutational_loci.fasta` and click <***Open***>. In the *Export knowledge base fasta* dialog box click <***Export...***> and <***Yes***> to export the fasta file.

As the header of the second column in the `mutational_loci.fasta` should be ”locus” instead of ”name” we should still replace all occurrences of ”name” in the exported fasta file with ”locus”.

60. Open the exported fasta file and replace all occurrences of the word ”name” with the word ”locus”. Save the changes to the file.

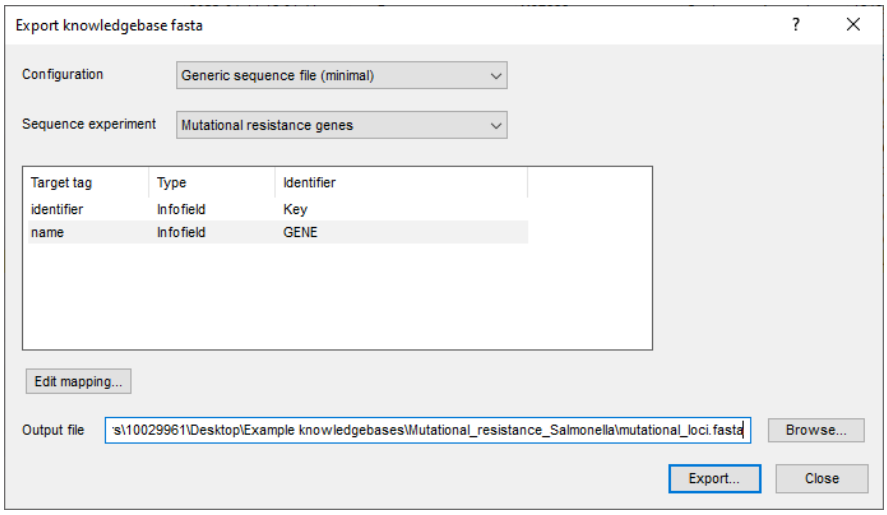


Figure 14: The *Export knowledge base fasta* dialog box with the adapted configuration.

The `resistens-overview.txt` file which was extracted from the PointFinder database includes known mutations and associated resistance traits. This file is present in the "Mutational_resistance_Salmonella" folder in the "Custom knowledgebase creation data" folder which was previously downloaded from our website. We will now create the `mutations.tsv` file based on the information present in the `resistens-overview.txt` file.

61. Open an excel file and import the data which is present in the `resistens-overview.txt` file.

62. Open the `README.md` file in the exported BLAST-based example knowledgebase folder with e.g. Notepad.

The `README.md` file indicates that the following columns are required in the `mutations.tsv` file:

- identifier: The unique identifier in the specific BIONUMERICS mutational identifier format.
- required: A logical expression (optional). A logical combination of related mutations that must be fulfilled for the trait associated with this mutation to be present.
- trait: The trait resulting from the presence of the mutation and, if applicable, the 'required' expression evaluating to True.

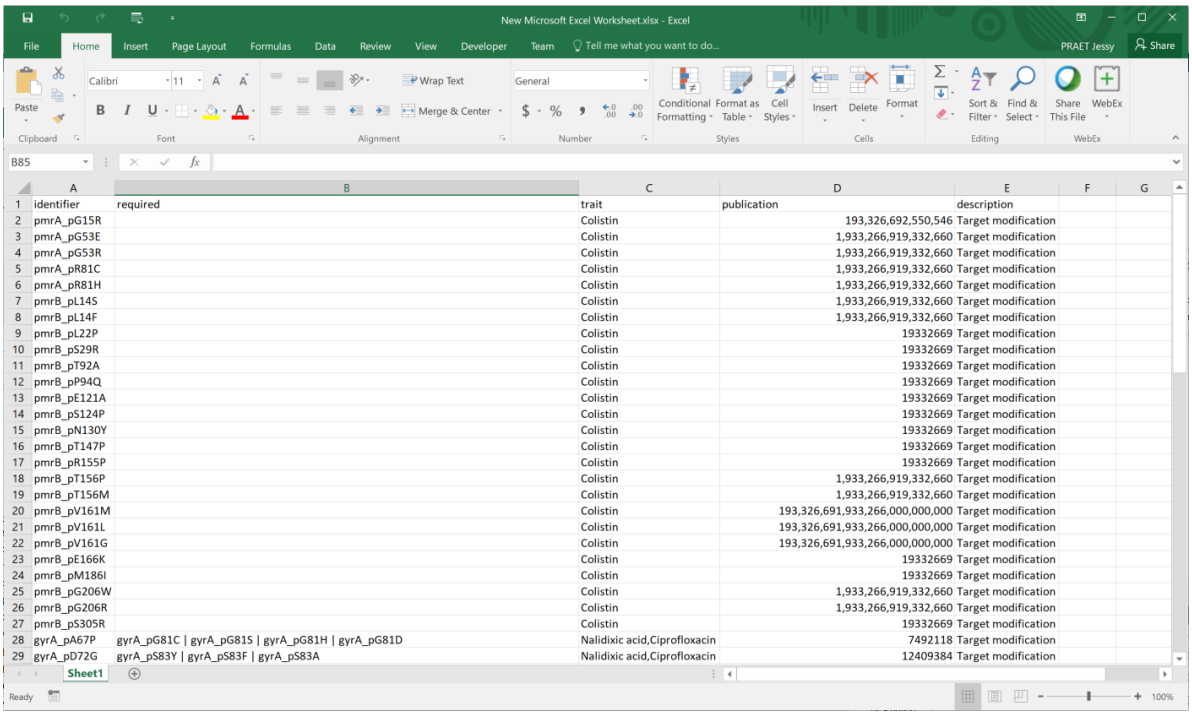
The following columns are optional:

- publication: The publication source of the sequence.
- description

Note that the `README.md` file provides more extensive information on the mutational identifiers and logical expressions.

63. In a new excel sheet create the columns which are required and optional in the `mutations.tsv` file based on the information present in the `resistens-overview.txt` file and based on the required format indicated in the `README.md` file (see Figure 15). This requires some manual copying and pasting of the information present in the `resistens-overview.txt` file.

64. Save the created excel sheet as a tsv file with the name `mutations.tsv` in the `Mutational_resistance_Salmonella` knowledgebase folder.



	A	B	C	D	E	F	G
1	identifier	required	trait	publication	description		
2	pmrA_pG15R		Colistin	193,326,692,550,546	Target modification		
3	pmrA_pG53E		Colistin	1,933,266,919,332,660	Target modification		
4	pmrA_pG53R		Colistin	1,933,266,919,332,660	Target modification		
5	pmrA_pR81C		Colistin	1,933,266,919,332,660	Target modification		
6	pmrA_pR81H		Colistin	1,933,266,919,332,660	Target modification		
7	pmrB_pL14S		Colistin	1,933,266,919,332,660	Target modification		
8	pmrB_pL14F		Colistin	1,933,266,919,332,660	Target modification		
9	pmrB_pL22P		Colistin	19332669	Target modification		
10	pmrB_pS29R		Colistin	19332669	Target modification		
11	pmrB_pT92A		Colistin	19332669	Target modification		
12	pmrB_pP94Q		Colistin	19332669	Target modification		
13	pmrB_pE121A		Colistin	19332669	Target modification		
14	pmrB_pS124P		Colistin	19332669	Target modification		
15	pmrB_pN130Y		Colistin	19332669	Target modification		
16	pmrB_pT147P		Colistin	19332669	Target modification		
17	pmrB_pR155P		Colistin	19332669	Target modification		
18	pmrB_pT156P		Colistin	1,933,266,919,332,660	Target modification		
19	pmrB_pT156M		Colistin	1,933,266,919,332,660	Target modification		
20	pmrB_pV161M		Colistin	193,326,691,933,266,000,000,000	Target modification		
21	pmrB_pV161L		Colistin	193,326,691,933,266,000,000,000	Target modification		
22	pmrB_pV161G		Colistin	193,326,691,933,266,000,000,000	Target modification		
23	pmrB_pE166K		Colistin	19332669	Target modification		
24	pmrB_pM186I		Colistin	19332669	Target modification		
25	pmrB_pG206W		Colistin	1,933,266,919,332,660	Target modification		
26	pmrB_pG206R		Colistin	1,933,266,919,332,660	Target modification		
27	pmrB_pS305R		Colistin	19332669	Target modification		
28	gyrA_pA67P	gyrA_pG81C | gyrA_pG81S | gyrA_pG81H | gyrA_pG81D	Nalidixic acid,Ciprofloxacin	7492118	Target modification		
29	gyrA_pD72G	gyrA_pS83Y | gyrA_pS83F | gyrA_pS83A	Nalidixic acid,Ciprofloxacin	12409384	Target modification		

Figure 15: Excel sheet with the columns which are required or optional in the `mutations.tsv` file.

The BLAST-based mutational resistance knowledgebase (see Figure 16) is now ready to be used by the custom genotyping plugin for the mutation scanning and mutational traits detection features.

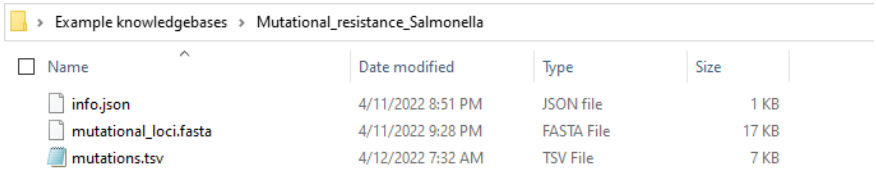


Figure 16: The Mutational_resistance_Salmonella knowledgebase.

This knowledgebase can also be downloaded from the BIONUMERICS website (`https://www.bionumerics.com/download/sample-data`, click on "Custom knowledgebases").

5.3 In-silico PCR-based knowledgebase

A selective amplification of the *tyv* (rfbE), *prt* (rfbS), *viaB*, and *fliC* genes by multiplex PCR for the identification of *Salmonella enterica* serovars Typhi and Paratyphi A was described in Hirose *et al.* [2] (see Figure 17). To be able to perform this multiplex PCR in BIONUMERICS, we will create an in-silico PCR based knowledgebase which we can use in the *custom genotyping* plugin.

Open the `README.md` file in the exported PCR-based example knowledgebase folder with e.g. Notepad. Aside from the `info file` required by all knowledgebases, the `README.md` file explains that a `primers.tsv` file is also required for the in-silico PCR detection and in-silico PCR extraction features.

For each required file the `README.md` file provides detailed information on the required format. We will first create a folder for our in-silico PCR-based knowledgebase and include an appropriate

TABLE 1.

Primers for multiplex PCR amplification of *Salmonella enterica* serovars Typhi and Paratyphi A

Gene and primer (oligonucleotide sequence)	Length (bp)	Amplified fragment size (bp)	Source[b]
tyv (rfbE)			
tyv-s (5"-GAG GAA GGG AAA TGA AGC TTT T-3")	22	615	M29682
tyv-as (5"-TAG CAA ACT GTC TCC CAC CAT AC-3")	23		M29682
prt (rfbS)			
parat-s (5"-CTT GCT ATG GAA GAC ATA ACG AAC C-3")	25	258	M29682
parat-as, (5"-CGT CTC CAT CAA AAG CTC CAT AGA-3")	24		M29682
viaB			
vi-s (5"-GTT ATT TCA GCA TAA GGA G-3")	19	439	D14156
vi-as (5"-CTT CCA TAC CAC TTT CCG-3")	18		D14156
fliC			
fliCcom-s (5"-AAT CAA CAA CAA CCT GCA GCG-3")	21		L21912
fliCd-as (5"-GCA TAG CCA CCA TCA ATA ACC-3")	21		L21912
fliCa-as (5"-TAG TGC TTA ATG TAG CCG AAG G-3")	22		X03393
fliCcom-fliCd-as		750 (489)[a]	
fliCcom-fliCa-as		329	

Figure 17: Primers for multiplex PCR amplification of *Salmonella enterica* serovars Typhi and Paratyphi A (see [2]).

info.json file.

65. Create a new folder (e.g. `In-silico PCR`) on your computer for the PCR-based knowledgebase.

66. Copy the `info.json` file from the PCR-based example knowledgebase folder into this new folder and open the json file in e.g. Notepad.

67. Optionally adapt the `info.json` file by changing the version, name, description and changelog and save your changes (see Figure 18).

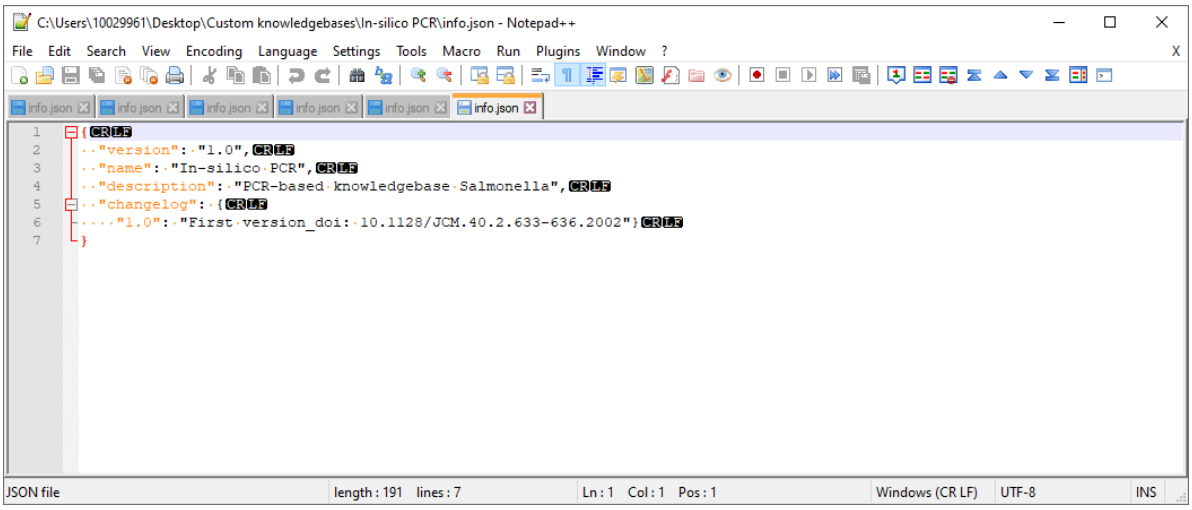


Figure 18: The `info.json` file.

We will now create the `primers.tsv` file based on the information present in the cited article (see Figure 17).

68. Copy the `primers.tsv` file from the PCR-based example knowledgebase folder into the `In-silico PCR` knowledgebase folder and open the tsv file in e.g. Notepad.

69. Open the `README.md` in the exported PCR-based example knowledgebase folder with e.g. Notepad.

The `README.md` file indicates that the following columns are required in the `primers.tsv` file:

- identifier: The unique identifier of the primer pair definition.
- primer_fwd: The sequence of the forward primer, based on the sense strand of the reference.
- primer_rev: The sequence of the reverse primer, based on the antisense strand of the reference.
- reference_length: The length of the expected reference amplicon, including the primers themselves.
- max_length_offset: The maximum allowed size difference between an amplicon length and the reference amplicon length.
- max_iupac: The maximum ambiguous bases allowed in the alignment of each primer with the query (default 0).
- max_mismatch: The maximum number of mismatches allowed in the alignment of each primer with the query (default 0).

Note that the `README.md` file provides more extensive information on the required format.

70. Add the primer information from the article (see Figure 17) to the `primers.tsv` file in the required format as indicated in the `README.md` file (see Figure 19). This requires some manual copying and pasting of the information.

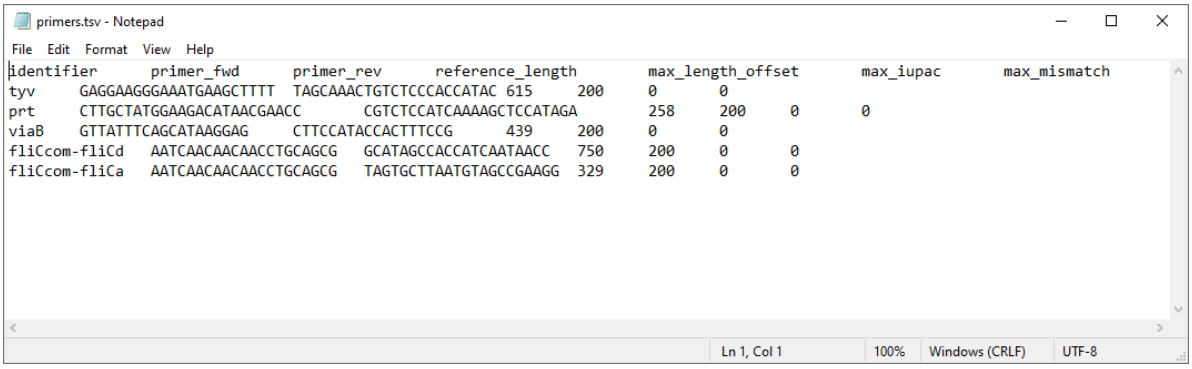


Figure 19: The `primers.tsv` file.

71. Save the adapted tsv file.

The PCR-based knowledgebase (see Figure 20) is now ready to be used by the custom genotyping plugin for the in-silico PCR detection and in-silico PCR extraction features.

This knowledgebase can also be downloaded from the BIONUMERICS website (https://www.bionumerics.com/download/sample-data, click on "Custom knowledgebases").

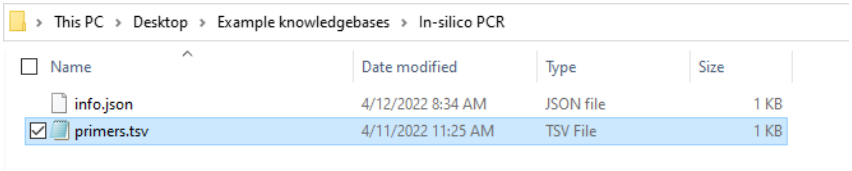


Figure 20: The In-silico PCR knowlegdebase.

5.4 Minhash-based knowledgebase

The custom genotyping plugin allows species confirmation based on minhashing with sourmash ([1]). In this tutorial we will generate a minhash-based knowledgebase for the species confirmation of *Salmonella* subspecies. Open the `README.md` file in the exported minhash-based example knowledgebase folder with e.g. Notepad. Aside from the `info file` required by all knowledgebases, the `README.md` file explains that the following files are required for the species confirmation feature:

- A `sourmash_params.json` file: A JSON formatted file that includes the sourmash kmer size and scaling factor.
- A `genomes.sig` file: A JSON formatted file that includes the minhash signatures. It can be generated with the sourmash sketch function or the ***Export Minhash Signatures*** menu item in BIONUMERICS.
- A `genome_info.tsv` file: A TSV formatted file that includes information on the reference genomes and includes the applied mash containment thresholds for genus, species and subspecies.

For each required file the `README.md` file provides detailed information on the required format. We will first create a folder for our minhash knowledgebase and include an appropriate info.json file and sourmash_params.json file.

72. Create a new folder (e.g. `Species_confirmation_Salmonella`) on your computer for the minhash-based knowledgebase.
73. Copy the `info.json` file and the `sourmash_params.json` file from the minhash-based example knowledgebase folder into this new folder and open the json file in e.g. Notepad.
74. Optionally adapt the `info.json` file by changing the version, name, description and changelog and save your changes (see Figure 21).
75. Open the `sourmash_params.json` file and optionally change the sourmash k-mer size and scaling factor (see Figure 22).

To allow species confirmation of *Salmonella* subspecies the `genomes.sig` file should contain mash signatures of *Salmonella* subspecies reference genomes. Reference genome sequences of *Salmonella* subspecies have been downloaded from NCBI and are available in the ”Minhash_Salmonella” folder in the ”Custom knowledgebase creation data” folder which was previously downloaded from our website.

The `genomes.sig` file can easily be created from the data obtained from NCBI by using the ***export Minhash signatures*** tool provided by the *Custom genotyping* plugin. To be able to use the ***export Minhash signatures*** tool we first need to import the reference genomes into the BIONUMERICS database.

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

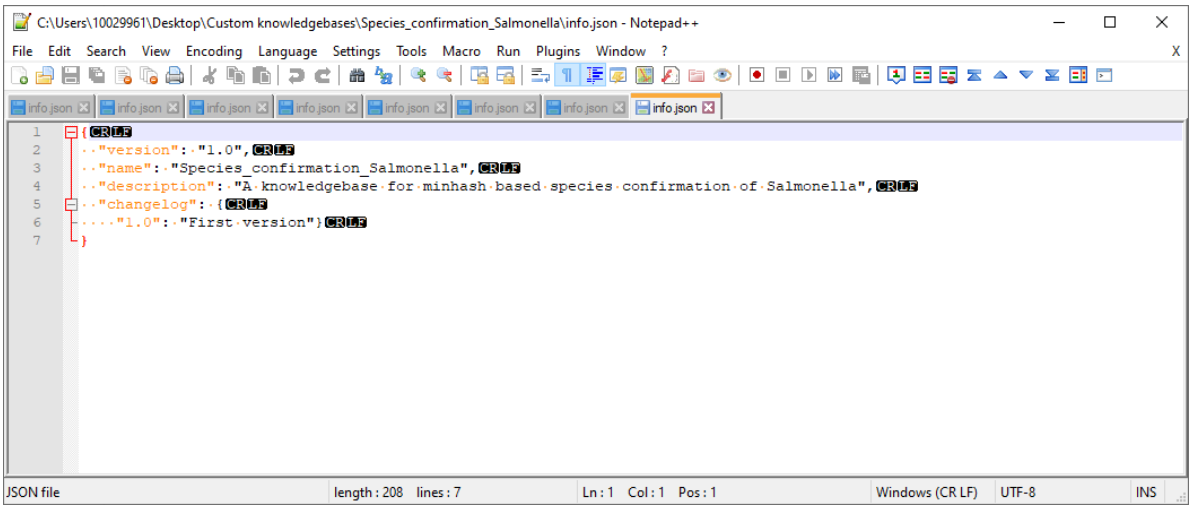


Figure 21: The `info.json` file.

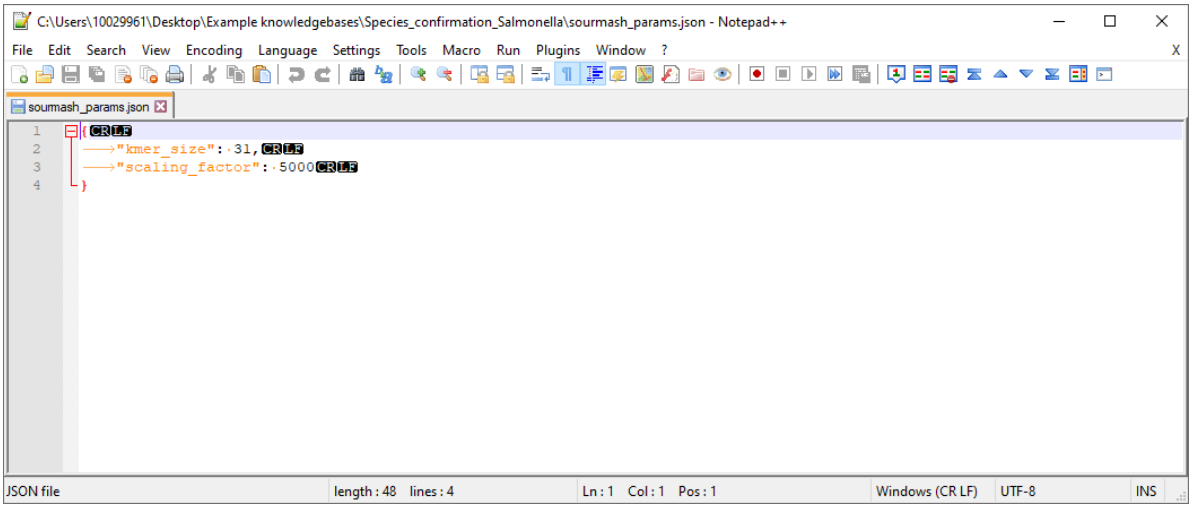


Figure 22: The `sourmash_params.json` file

77. Press <***Browse***>, navigate to the ”Custom knowledgebase creation data” folder previously downloaded from our website, open the ”Minhash_Salmonella” folder, select all fasta files and press <***Open***>.

78. With the ***Import FASTA sequences from text files*** option highlighted, press <***Finish***>.

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

The import wizard now displays a preview of the sequence data in the FASTA file.

80. Press <***Next***>.

81. Click <***Create new***> to create a new import template.

82. Select ***Name*** in the list and click <***Edit destination***> or simply double-click on ***Name***. Select ***Key*** and press <***OK***>.

The grid is updated.

83. Optionally, you can press <***Preview***> to obtain a preview of the data you are about to import.

84. Click <***Next***>.

85. Select ***Key*** as ***Entry link field***. Press <***Finish***>.

86. Optionally save the template and press <***OK***>.

87. Highlight the newly created template and select ***Create new*** as ***Experiment type***.

88. Press <***Next***>.

89. Specify a sequence type name (e.g. **Reference genomes**) and press <***OK***> and confirm the action.

90. Press <***Finish***> to start the import into the database.

The *Main* window now looks like Figure 23.

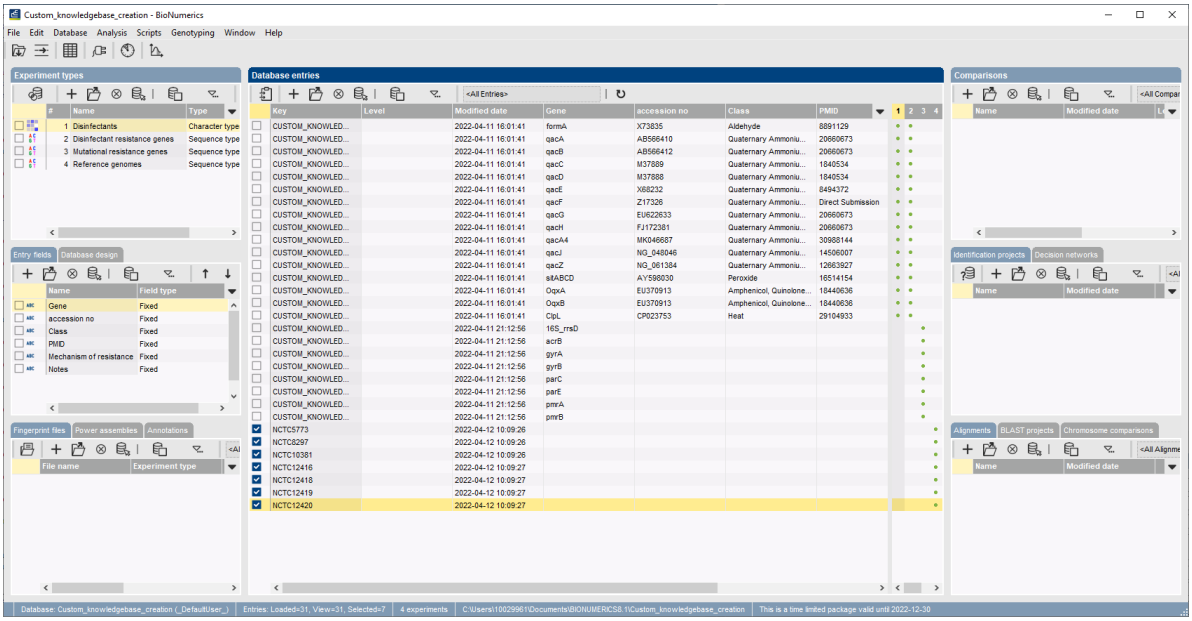


Figure 23: The *Main* window.

Now the sequence data is available in our BIONUMERICS database, the `genomes.sig` file can easily be created by using the ***export Minhash signatures*** tool of the *Custom genotyping* plugin.

91. Select all entries in the database which have data in the ***Reference genomes*** sequence experiment and select ***Genotyping*** > ***Tools*** > ***Export Minhash signatures...***.

This action opens the *Export Minhash signatures* dialog box (see Figure 24).

92. Click <***Browse...***> and browse for the Species_confirmation_Salmonella knowledgebase folder. As file name enter `genomes.sig` and click <***Open***>.

93. In the ***Sequence experiment*** drop-down list select the ”Reference genomes” experiment.

94. Optionally change the sourmash k-mer size and scaling factor according to what was specified in the `sourmash_params.json file`.

95. Click <***OK***> to export the `genomes.sig` file to the respective knowledgebase folder (see Figure 25).

The `genome_info.tsv` file should include information on the reference genomes and should include the applied mash containment thresholds for genus, species and subspecies. The metadata

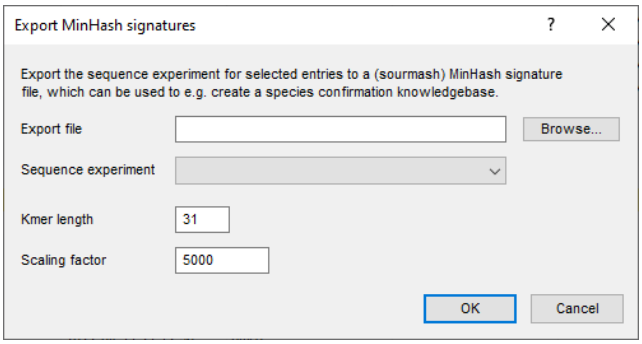


Figure 24: The *Export Minhash signatures* dialog box.

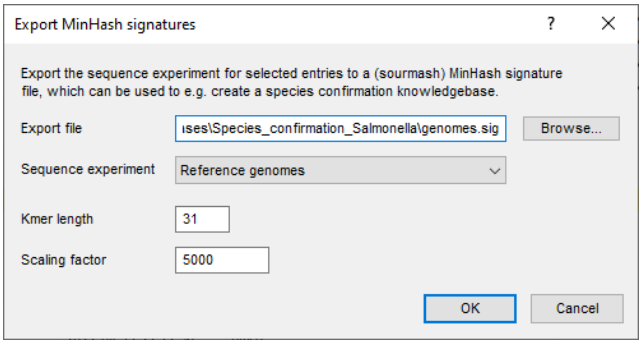


Figure 25: The *Export Minhash signatures* dialog box with the adapted configuration.

of the reference genomes and the appropriate mash containment thresholds have already been filled in in the `Salmonella.csv` file which was downloaded from our website.

96. Open the `README.md` file in the exported minhash-based example knowledgebase folder with e.g. Notepad.

The `README.md` file indicates that the following columns are required or optional in the `genome_info.tsv` file:

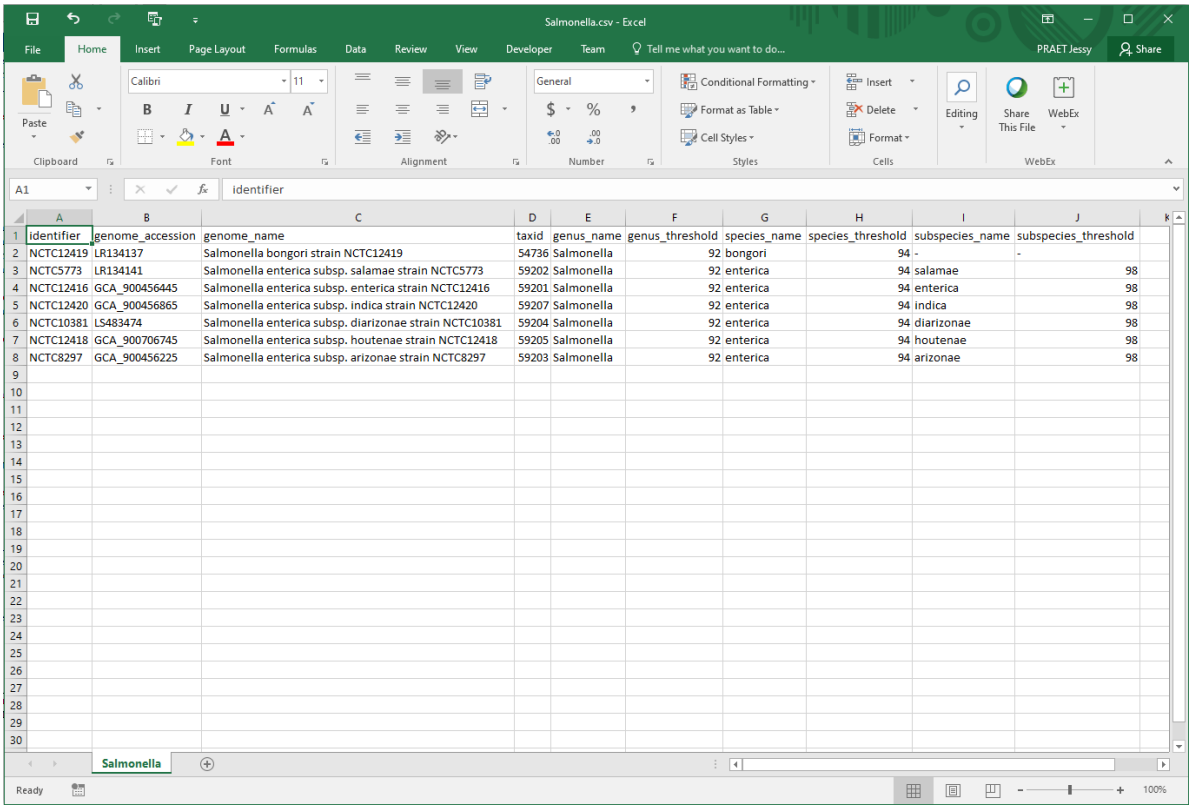
- identifier: The unique identifier of the reference sequence. This is the entry key from the BIONUMERICS database when you create the signatures with the ***Export Minhash Signatures*** menu item.
- genome_accession: Optional. The accession number corresponding to the reference sequence.
- genome_name: A description of the genome.
- taxid: The taxid number.
- genus_name: The genus name.
- species_name: The species name.
- subspecies_name: The subspecies name.
- genus_threshold: A mash containment threshold between ”0.0” and ”100.0”.
- species_threshold: A mash containment threshold between ”0.0” and ”100.0”.

- subspecies_threshold: A mash containment threshold between ”0.0” and ”100.0”.

Note that the `README.md` file provides more extensive information on the required format.

97. Open the `Salmonella.csv` file in the ”Minhash_Salmonella” folder.

The metadata of the reference genomes and the appropriate mash containment thresholds have already been filled in in the required columns (see Figure 26).



identifier	genome_accession	genome_name	taxid	genus_name	genus_threshold	species_name	species_threshold	subspecies_name	subspecies_threshold
NCTC12419	LR134137	Salmonella bongori strain NCTC12419	54736	Salmonella	92	bongori	94	-	-
NCTC5773	LR134141	Salmonella enterica subsp. salamae strain NCTC5773	59202	Salmonella	92	enterica	94	salamae	98
NCTC12416	GCA_900456445	Salmonella enterica subsp. enterica strain NCTC12416	59201	Salmonella	92	enterica	94	enterica	98
NCTC12420	GCA_900456865	Salmonella enterica subsp. indica strain NCTC12420	59207	Salmonella	92	enterica	94	indica	98
NCTC10381	LS483474	Salmonella enterica subsp. diarizonae strain NCTC10381	59204	Salmonella	92	enterica	94	diarizonae	98
NCTC12418	GCA_900706745	Salmonella enterica subsp. houtenae strain NCTC12418	59205	Salmonella	92	enterica	94	houtenae	98
NCTC8297	GCA_900456225	Salmonella enterica subsp. arizonae strain NCTC8297	59203	Salmonella	92	enterica	94	arizonae	98

Figure 26: The `Salmonella.csv` file.

98. Save the `Salmonella.csv` file as a tsv file in the ”Species_confirmation_Salmonella” knowledgebase folder with the name `genome_info.tsv`.

The minhash-based knowledgebase (see Figure 27) is now ready to be used by the custom genotyping plugin for the species confirmation feature.

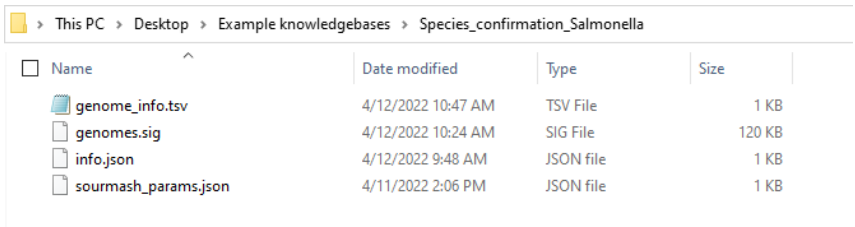


Figure 27: The ”Species_confirmation_Salmonella” knowledgebase.

This knowledgebase can also be downloaded from the BIONUMERICS website (`https://www.bionumerics.com/download/sample-data`, click on ”Custom knowledgebases”).

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.

[2] Hirose Kenji, Itoh Ken-Ichiro, Nakajima Hiroshi, Kurazono Takayuki, Yamaguchi Masanori, Moriya Kazuo, Ezaki Takayuki, Kawamura Yoshiaki, Tamura Kazumichi, and Watanabel Haruo. Selective amplification of tyv (rfbe), prt (rfbs), viab, and flic genes by multiplex pcr for identification of salmonella enterica serovars typhi and paratyphi a. *Journal of Clinical Microbiology*, 40(2):633–636, 2002.