BIONUMERICS Tutorial:

E. coli functional genotyping: predicting phenotypic traits from whole genome sequences

1 Aim

In this tutorial we will screen whole genome sequences of *Escherichia coli* samples for phenotypic traits using the *E. coli functional genotyping plugin*. This plugin contains knowledgebases for serotype, virulence and antibiotic resistance prediction, as well as plasmid and phage detection. An *in silico* PCR tool is also implemented, making it possible to detect Shiga toxin gene subtypes and virulence genes, mimicking the wet lab PCR.

The different steps are illustrated using the whole genome demonstration database of *Escherichia coli*. This database is available for download on our website (see 2) and contains 60 publicly available sequence read sets of *Escherichia coli* with already calculated de novo assemblies.

2 Preparing the database

2.1 Introduction to the demonstration database

We provide a **WGS demo database** for *Escherichia coli* containing sequence read set data links for 60 samples, calculated de novo assemblies and wgMLST results (allele calls and quality information).

The wgMLST workflow and results will not be discussed in this tutorial.

The **WGS_demo_database_for_Escherichia_coli** can be downloaded directly from the *BIONUMERICS Startup* window (see 2.2), or restored from the back-up file available on our website (see 2.3).

Installation of the *E. coli functional genotyping plugin* is only possible when no spaces are present in the BIONUMERICS home directory and in the name of the database. Before downloading or restoring the **WGS demo database** for *Escherichia coli*, please check if your BIONUMERICS home directory does not contain any spaces:

1. Click the button, located in the toolbar in the *BIONUMERICS Startup* window and select ***Change home directory...*** to call the *Home directory* dialog box.

2. In case the currently specified home directory contains spaces, update the path to a path containing no spaces and close the *Home directory* dialog box.

2.2 Option 1: Download demo database from the Startup Screen

3. Click the  button, located in the toolbar in the *BIONUMERICS Startup* window.

This calls the *Tutorial databases* window (see Figure 1).

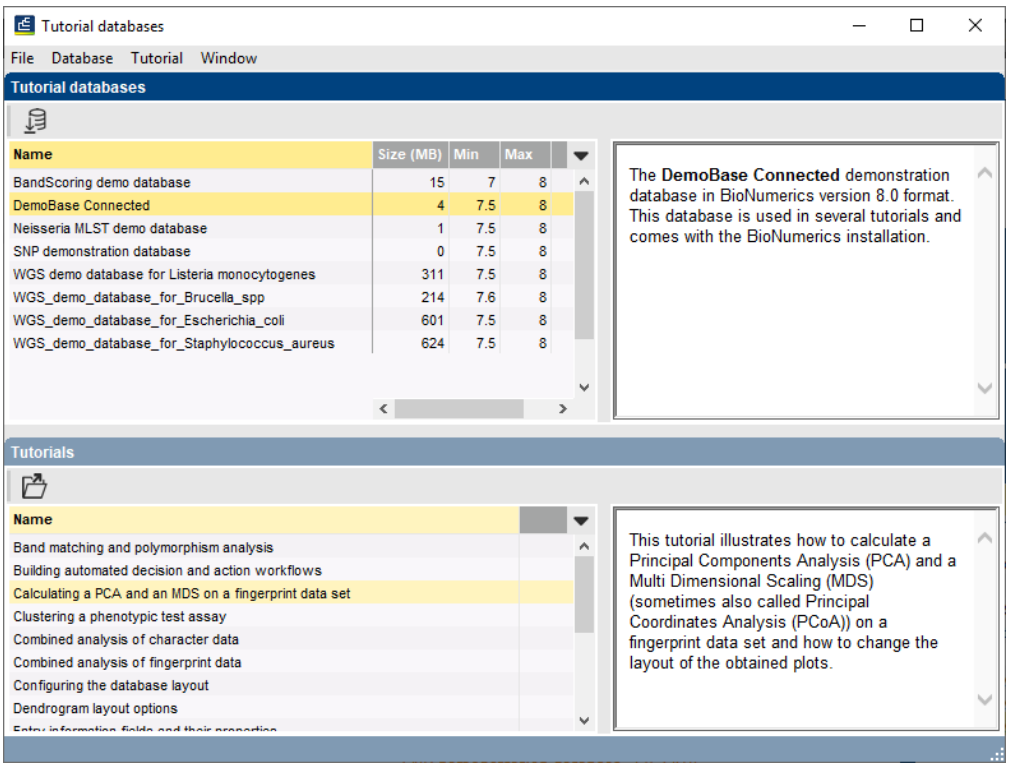


Figure 1: The *Tutorial databases* window, used to download the demonstration database.

4. Select **WGS_demo_database_for_Escherichia_coli** from the list and select ***Database*** > ***Download*** ( ).

5. Confirm the installation of the database and press <***OK***> after successful installation of the database.

6. Close the *Tutorial databases* window with ***File*** > ***Exit***.

The **WGS_demo_database_for_Escherichia_coli** appears in the *BIONUMERICS Startup* window.

7. Double-click the **WGS_demo_database_for_Escherichia_coli** in the *BIONUMERICS Startup* window to open the database.

2.3 Option 2: Restore demo database from back-up file

A BIONUMERICS back-up file of the demo database for *Escherichia coli* is also available on our website. This backup can be restored to a functional database in BIONUMERICS.

8. Download the file `WGS_EC.bnbk` file from https://www.applied-maths.com/download/sample-data, under 'WGS_demo_database_for_Escherichia_coli'.

In contrast to other browsers, some versions of Internet Explorer rename the `WGS_EC.bnbk` database backup file into `WGS_EC.zip`. If this happens, you should manually remove the `.zip` file extension and replace with `.bnbk`. A warning will appear ("If you change a file name extension, the file might become unusable."), but you can safely confirm this action. Keep in mind that Windows might not display the `.zip` file extension if the option "Hide extensions for known file types" is checked in your Windows folder options.

9. In the *BIONUMERICS Startup* window, press the  button. From the menu that appears, select ***Restore database...***.

10. Browse for the downloaded file and select ***Create copy***. Note that, if ***Overwrite*** is selected, an existing database will be overwritten.

11. Specify a new name for this demonstration database, e.g. "WGS_Ecoli_demobase".

12. Click <***OK***> to start restoring the database from the backup file.

13. Once the process is complete, click <***Yes***> to open the database.

The *Main* window is displayed.

3 About the demonstration database

The **WGS_demo_database_for_Escherichia_coli** contains data for a set of 60 samples. The sample information, stored in entry info fields (Isolation source, Center Name, etc.) was collected from the publications. Seven experiments are present in the demo database and are listed in the *Experiment types* panel (see Figure 2).

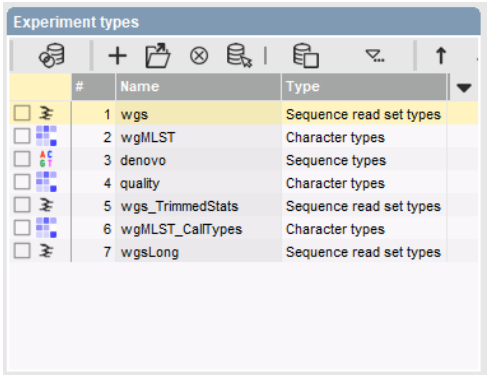


Figure 2: The *Experiment types* panel in the *Main* window.

1. Click on the green colored dot for one of the entries in the first column in the *Experiment presence* panel. Column 1 corresponds to the first experiment type listed in the *Experiment types* panel, which is **wgs**.

In the *Sequence read set experiment* window, the link to the sequence read set data on NCBI (SRA) with a summary of the characteristics of the sequence read set is displayed: *Read set size*, *Sequence length statistics*, *Quality statistics*, *Base statistics* (see Figure 3).

2. Close the *Sequence read set experiment* window.

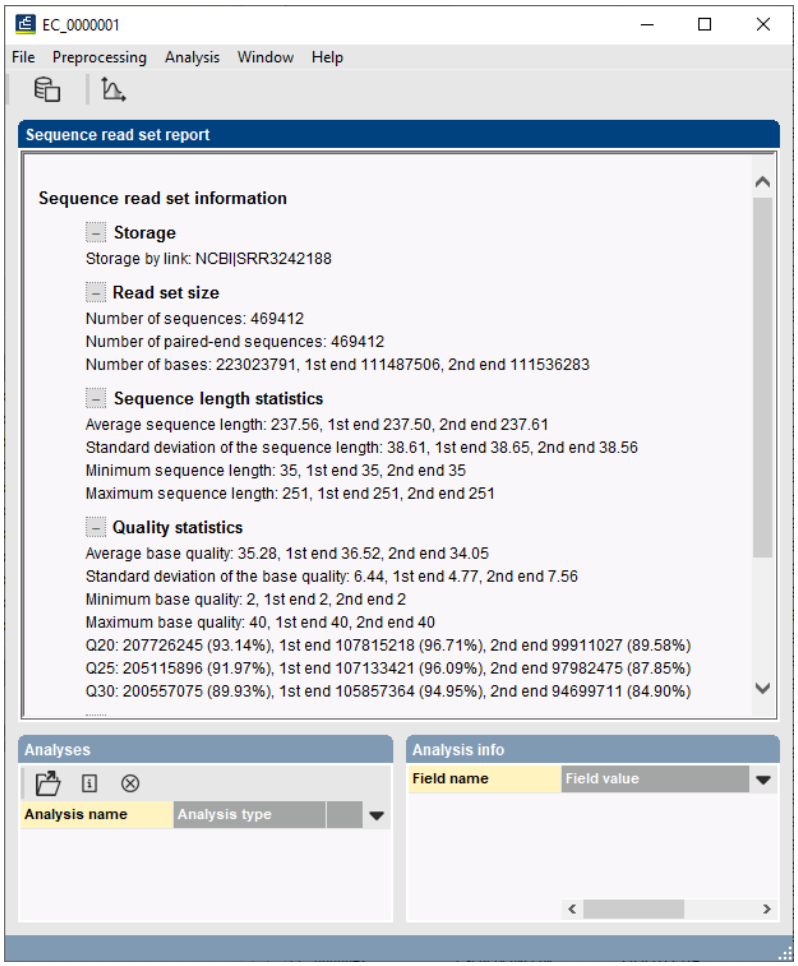


Figure 3: The sequence read set experiment card for an entry.

3. Click on the green colored dot for one of the entries in the third column in the *Experiment presence* panel. Column 3 corresponds to the third experiment type listed in the *Experiment types* panel, which is **denovo**.

The *Sequence editor* window opens, containing the results from the de novo assembly algorithm, i.e. concatenated de novo contig sequences (see Figure 4).

4. Close the *Sequence editor* window.

The sequence read set experiment type **wgs_TrimmedStats** contains some data statistics about the reads retained after trimming, used for the de novo assembly.

The sequence read set experiment type **wgsLong** contains the links to long read sequence read data (typically PacBio or MinION datasets). In this demo database, no links are defined for this experiment.

The other three experiments contain data related to the wgMLST analysis performed on the samples:

- Character experiment type **wgMLST** contains the allele calls for detected loci in each sample, where the consensus from assembly-based and assembly-free calling resulted in a single allele ID.
- Character experiment type **quality** contains quality statistics for the raw data, the de novo assembly and the different allele identification algorithms.

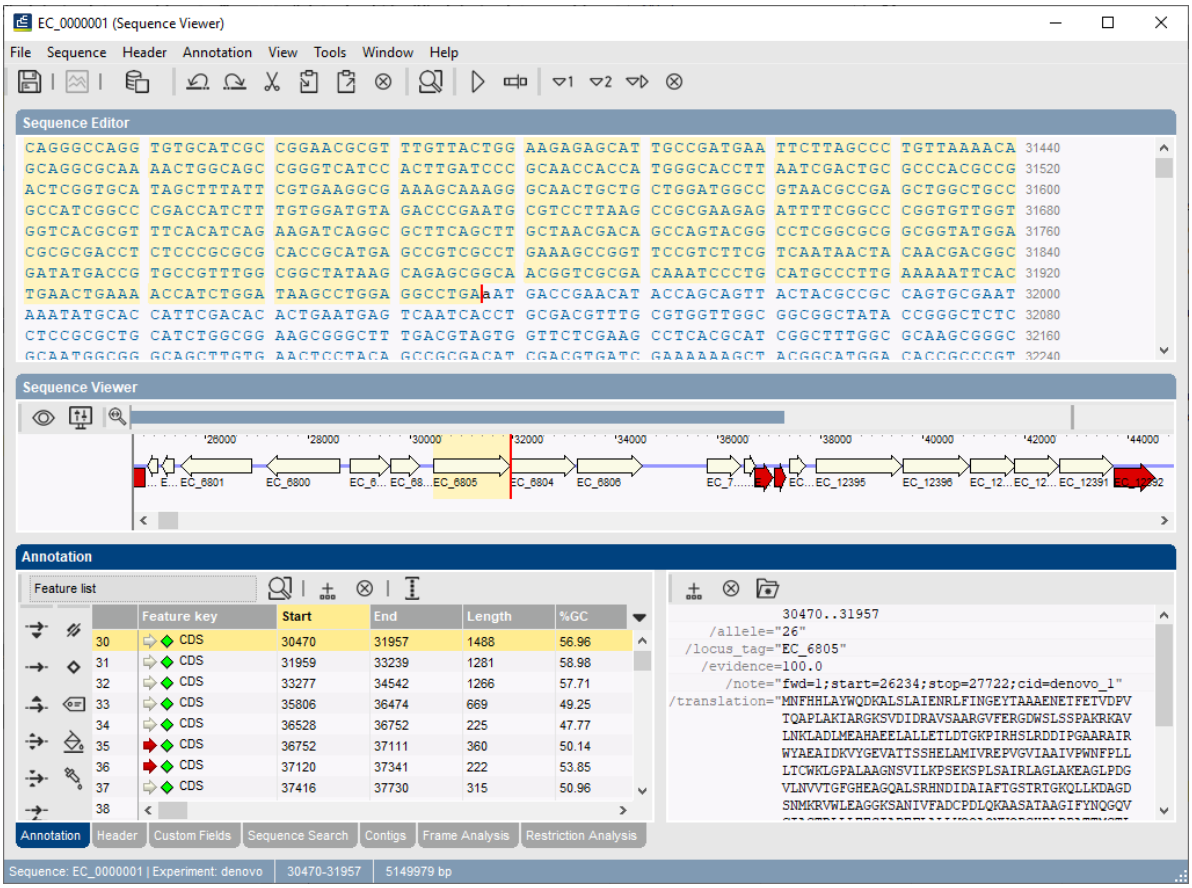


Figure 4: The *Sequence editor* window.

- Character experiment type **wgMLST_CallTypes**: contains details on the call types.

4 Installing the *E. coli* functional genotyping plugin

1. Call the *Plugins* dialog box from the *Main* window by selecting ***File*** > ***Install / remove plugins...*** ( ).
2. Select the *E. coli functional genotyping plugin* in the *Application tab* and press the <***Activate***> button (see Figure 5).
3. Confirm the installation of the plugin.
4. Click <***Yes***> to review the settings.

The *Settings* dialog box pops up, consisting of 9 tabs (see Figure 6). In the *General tab* following general settings need to be specified:

- The ***Information fields*** that will appear in the report (see 6).
- The ***Exports directory*** for the export of the reports (see 6).
- The ***Input Sequence experiment*** that holds the (whole) genome sequences that will be screened.

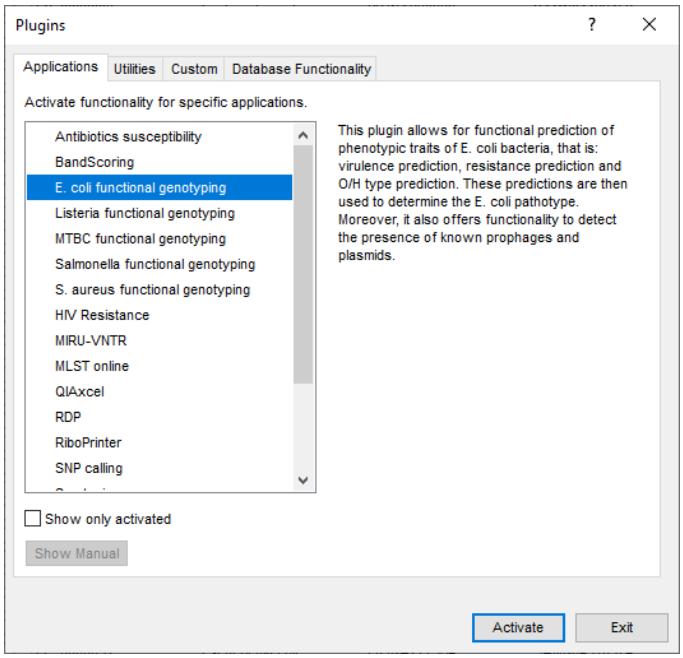


Figure 5: Install the plugin.

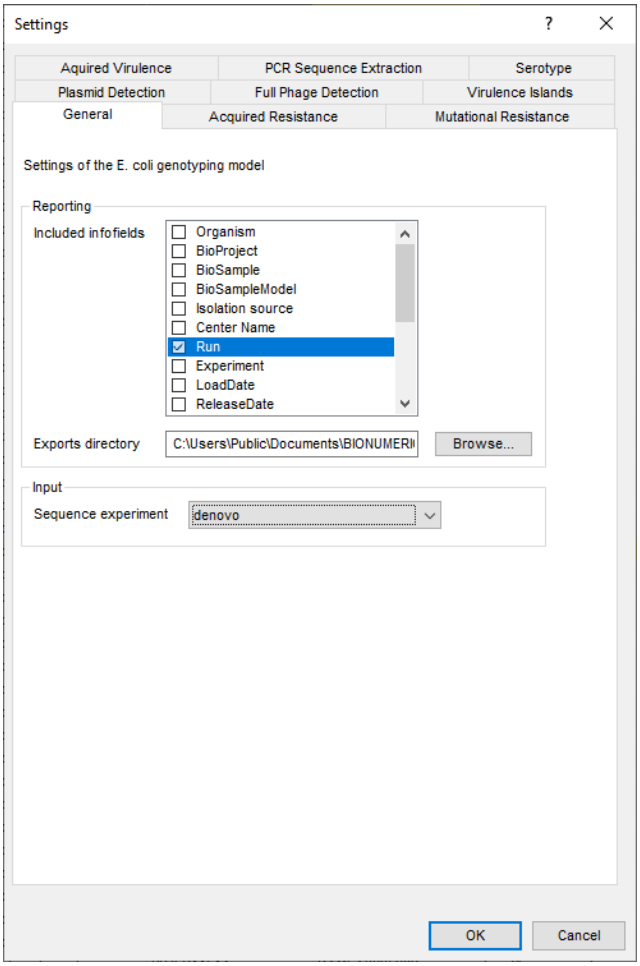


Figure 6: The *Settings dialog*: *General tab*.

5. In our demonstration database, the assembled sequences are stored in the ***denovo*** sequence experiment. Make sure this experiment is selected from the drop-down list and check the ***Run*** number to include in the report (see Figure 6).

The other tabs group the settings for each possible search: Serotype, Mutational/Acquired Resistance, Virulence (Acquired/Island), Plasmid, Full Phage, and PCR products. By default, all searches are enabled. Except for the *PCR Sequence Extraction* tab, all tabs consist of maximum three separate panels:

1. *Knowledgebase*: in this panel the knowledgebase against which you want to screen can be specified.
2. *Blast*: in this panel two settings for the BLAST algorithm can be specified; the ***Minimum percent identity (%)*** and the ***Minimum coverage (%)*** of your query sequence against the knowledgebase's reference sequences. If the option ***Combine fragments*** is checked, genes that occur fragmented in the genome (i.e. split over two contigs) can still be detected. Please note that this option is not available for the *Mutational Resistance* detection.
3. *Results*: in this panel the output database information fields and experiments to which the screening results will be written can be dictated. Use the drop-down list to choose an existing experiment type or field, or the <***Create***> option to create new experiments and fields. A default name for the experiment or information field is suggested, but you can adjust this if you want to. In the *Virulence Islands* tab you can specify the minimum percentage of virulence island loci that needs to be detected (***Minimum loci (%)***) before the presence of the virulence island is shown in the results.

Next to a *Knowledgebase* and *Results* panel, the *PCR Sequence Extraction* tab also contains the *Extraction* panel. In this panel you can set the output database experiments for each of the different *in silico* PCR amplicons by double-clicking the amplicon's identifier.

6. In this tutorial, specify the experiment types and information fields in all tabs by selecting the <***Create***> option in the drop-down lists and accepting the default names. Leave the other settings unaltered.
7. Click <***OK***> in the *Settings* dialog box.
8. When the *E. coli functional genotyping plugin* is successfully installed, a confirmation message pops up. Press <***OK***>.
9. Press <***Exit***> to close the *Plugins* dialog box.
10. Close and reopen the database to activate the features of the *E. coli functional genotyping plugin*.

The *E. coli functional genotyping plugin* installs menu items in the main menu of the software under ***E. coli*** (see Figure 7).

The settings specified during installation of the plugin can be called again at any time with ***E. coli*** > ***Settings...***.

5 Screening of entries

The screening can be done on any selection of entries in the database.

1. Select a single entry in the *Database entries* panel by holding the **Ctrl**-key and left-clicking on the entry. Alternatively, use the **space bar** to select a highlighted entry or click the ballot box next to the entry.

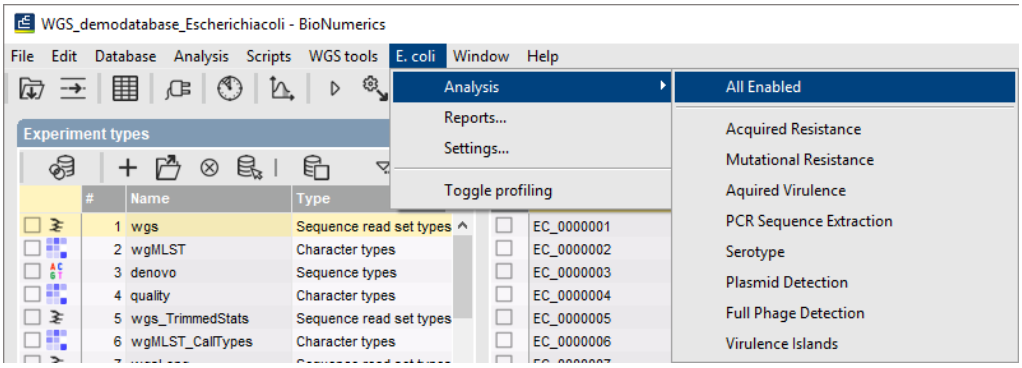


Figure 7: New menu-items after installation of the *E. coli functional genotyping plugin*.

Selected entries are marked by a checked ballot box (☑) and can be unselected in the same way.

2. In order to select a group of entries, hold the **Shift**-key and click on another entry.

A group of entries can be unselected the same way.

3. Make sure a few entries are selected in the *Database entries* panel of the demonstration database.

Screening for the phenotypic traits can be done for all tools checked in the *Settings* dialog box (***E. coli*** > ***Analysis*** > ***All Enabled***) or for each tool separately (***E. coli*** > ***Analyze*** > ***...***).

4. Select ***E. coli*** > ***Analysis*** > ***All Enabled*** to screen the selected entries for all enabled traits.

A progress bar appears. The analysis time depends on the number of selected entries. When the analysis is finished, the progress bar disappears. The detected traits for the screened entries are stored in the database.

The predicted **pathotypes**, **H and O serotypes** and **total islands** are written to the information fields in the *Database entries* panel (see Figure 8). Please note that the shown names of the information fields are those created per default, but can be different in your case depending whether you choose an alternative name during installation.

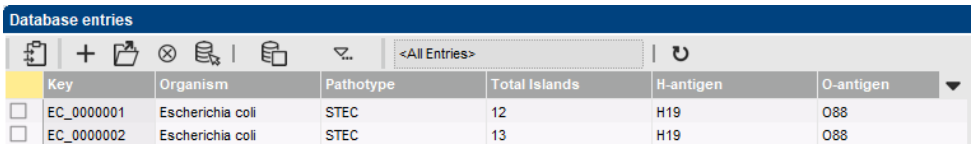


Figure 8: Example output of the Pathotype, Total Islands, H-antigen and O-antigen information fields.

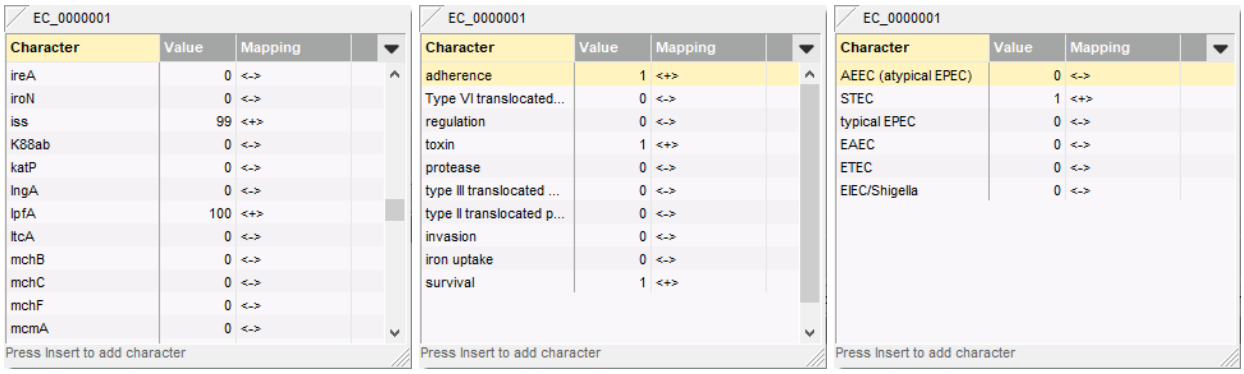
The character experiment types for **Acquired virulence**, **Virulence islands**, **Acquired resistance**, **Mutational resistance**, **Plasmid detection** and **Full phage detection** are created and updated with the predicted traits. Please note that the shown names of the experiment types are those created per default, but can be different in your case depending whether you choose an alternative name during installation.

5. Open a character card for one of the analyzed entries by clicking on the corresponding green colored dot in the *Experiment presence* panel.

Below, the interpretation of the results gathered in the character experiment types is given.

1. Acquired virulence (see Figure 9):

- ***Acquired_Virulence_loci_identity***: contains the results for each virulence gene: 0 = not detected, when detected the % identity of the best hit is shown.
- ***Acquired_Virulence_traits_identity***: contains the results for each virulence type: 0 = not detected, 1 = detected.
- ***E_coli_pathotype***: contains the results for each pathotype: 0 = not detected, 1 = detected.



EC_0000001

Character	Value	Mapping
ireA	0	<->
iroN	0	<->
iss	99	<+>
K88ab	0	<->
katP	0	<->
lngA	0	<->
lpfA	100	<+>
ltcA	0	<->
mchB	0	<->
mchC	0	<->
mchF	0	<->
mcmA	0	<->

Press Insert to add character

EC_0000001

Character	Value	Mapping
adherence	1	<+>
Type VI translocated...	0	<->
regulation	0	<->
toxin	1	<+>
protease	0	<->
type III translocated ...	0	<->
type II translocated p...	0	<->
invasion	0	<->
iron uptake	0	<->
survival	1	<+>

Press Insert to add character

EC_0000001

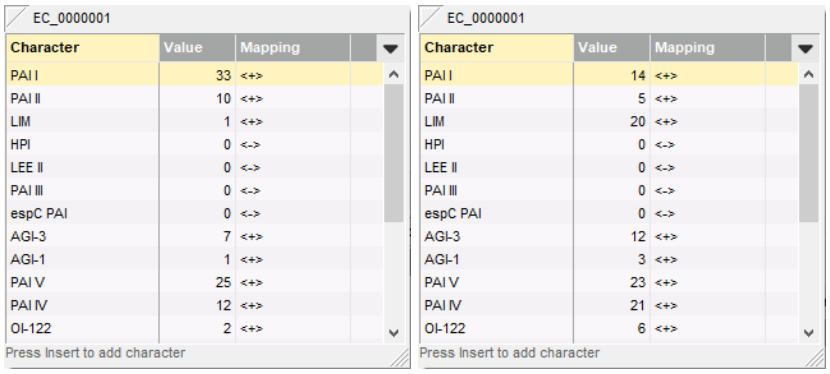
Character	Value	Mapping
AEEC (atypical EPEC)	0	<->
STEC	1	<+>
typical EPEC	0	<->
EAEC	0	<->
ETEC	0	<->
EIEC/Shigella	0	<->

Press Insert to add character

Figure 9: Example output of the ***Acquired_Virulence_loci_identity***, the ***Acquired_Virulence _traits_identity*** and the ***E_coli_pathotype*** experiment types for sample EC_0000001.

2. Virulence islands (see Figure 10):

- ***Island_Counts***: contains the number of detected loci associated to a pathogenicity island.
- ***Island_Percentages***: contains the percentage of detected loci associated to a pathogenicity island.



EC_0000001

Character	Value	Mapping
PAI I	33	<+>
PAI II	10	<+>
LIM	1	<+>
HPI	0	<->
LEE II	0	<->
PAI III	0	<->
espC PAI	0	<->
AGI-3	7	<+>
AGI-1	1	<+>
PAI V	25	<+>
PAI IV	12	<+>
OI-122	2	<+>

Press Insert to add character

EC_0000001

Character	Value	Mapping
PAI I	14	<+>
PAI II	5	<+>
LIM	20	<+>
HPI	0	<->
LEE II	0	<->
PAI III	0	<->
espC PAI	0	<->
AGI-3	12	<+>
AGI-1	3	<+>
PAI V	23	<+>
PAI IV	21	<+>
OI-122	6	<+>

Press Insert to add character

Figure 10: Example output of the ***Island_Counts*** and the ***Island_Percentages*** experiment types for sample EC_0000001.

3. Acquired resistance (see Figure 11):

- ***Acquired_Resistance_loci_identity***: contains the results for each resistance gene: 0 = not detected (sensitive), when detected (resistant) the % identity of the best hit is shown.
- ***Acquired_Resistance_traits_identity***: contains the results for each antibiotic: 0 = not detected (sensitive), 1 = detected (resistant).

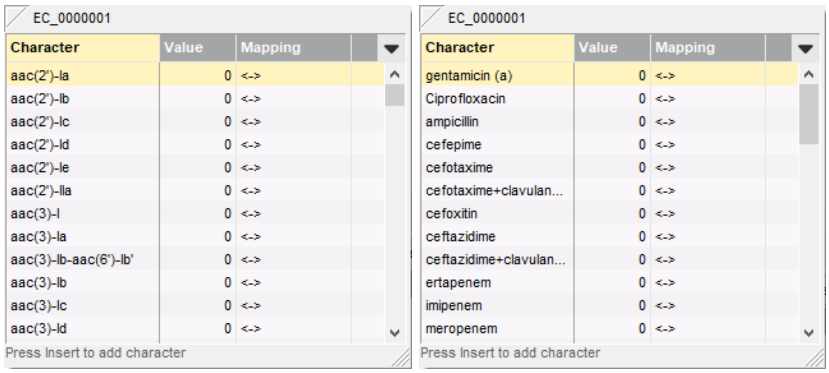


Figure 11: Example output of the ***Acquired_Resistance_loci_identity*** and the ***Acquired_Resistance _traits_identity*** experiment types for sample EC_0000001.

4. Mutational resistance (see Figure 12):

   - ***Mutational_Resistance_identifiers***: contains the results for each resistance mutation: -2 = partially indecisive, -1 = fully indecisive, 0 = not detected (sensitive), 1 = detected (resistant).
   - ***Mutational_Resistance_traits***: contains the results for each antibiotic group: -2 = partially indecisive, -1 = fully indecisive, 0 = not detected (sensitive), 1 = detected (resistant).

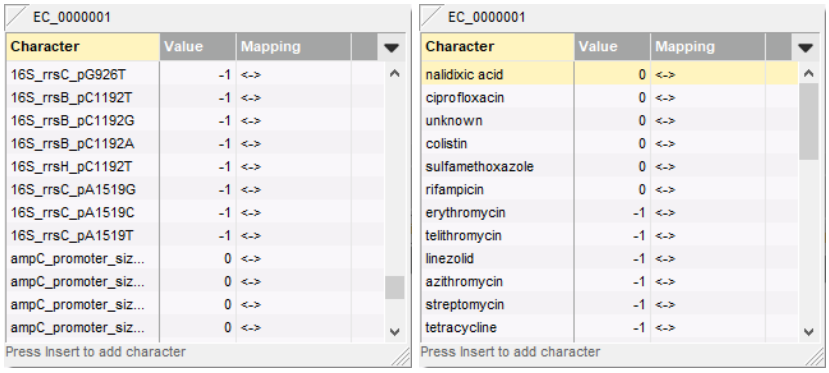


Figure 12: Example output of the ***Mutational_Resistance_identifiers*** and the ***Mutational_Resistance _traits*** experiment types for sample EC_0000001.

5. Plasmid detection (see Figure 13):

   - ***Plasmid_Ori_identity***: contains the results of the plasmid ORI detection: 0 = not detected, when detected the % BLAST identity with the ORI reference sequence is shown.
   - ***Plasmid_Full_coverage***: contains the results of the full plasmids detection: 0 = not detected, when detected the % coverage of the detected full plasmid is shown.

6. Full phage detection (see Figure 14):

   - ***Phage_Full_coverage***: contains the results of the full phages detection: 0 = not detected, when detected the % of the detected full phage is shown.

6. Close the character card(s) by clicking in the top left corner of the card.

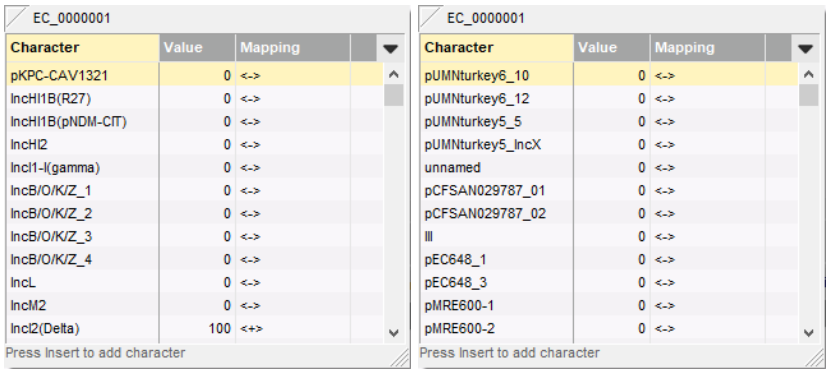


Figure 13: Example output of the ***Plasmid_Ori_identity*** and the ***Plasmid_Full_coverage*** experiment types for sample EC_0000001.

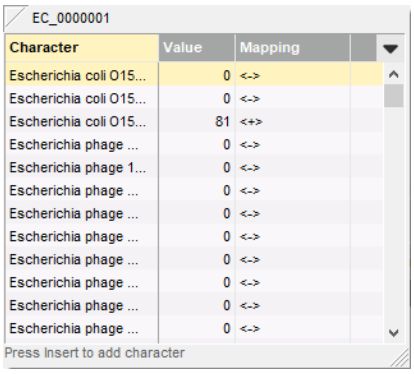


Figure 14: Example output of the ***Phage_Full_coverage*** experiment type for sample EC_0000001.

7. Open the **Amplicon** character card for one of the analyzed entries by clicking on the corresponding green colored dot in the *Experiment presence* panel.

The **Amplicons** character card (see Figure 15) lists all *in silico* PCR sequences that passed the search criteria.

8. Close the character card by clicking in the top left corner of the card.

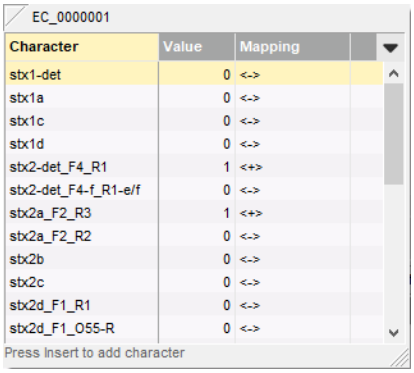


Figure 15: Example output of the ***Amplicons*** experiment type for sample EC_0000001.

The predicted *In silico* PCR sequences are stored in the corresponding sequence type experi-

ments.

9. Clicking on a green colored dot for an *in silico* experiment opens the *Sequence editor* window displaying the sequence (see Figure 16).

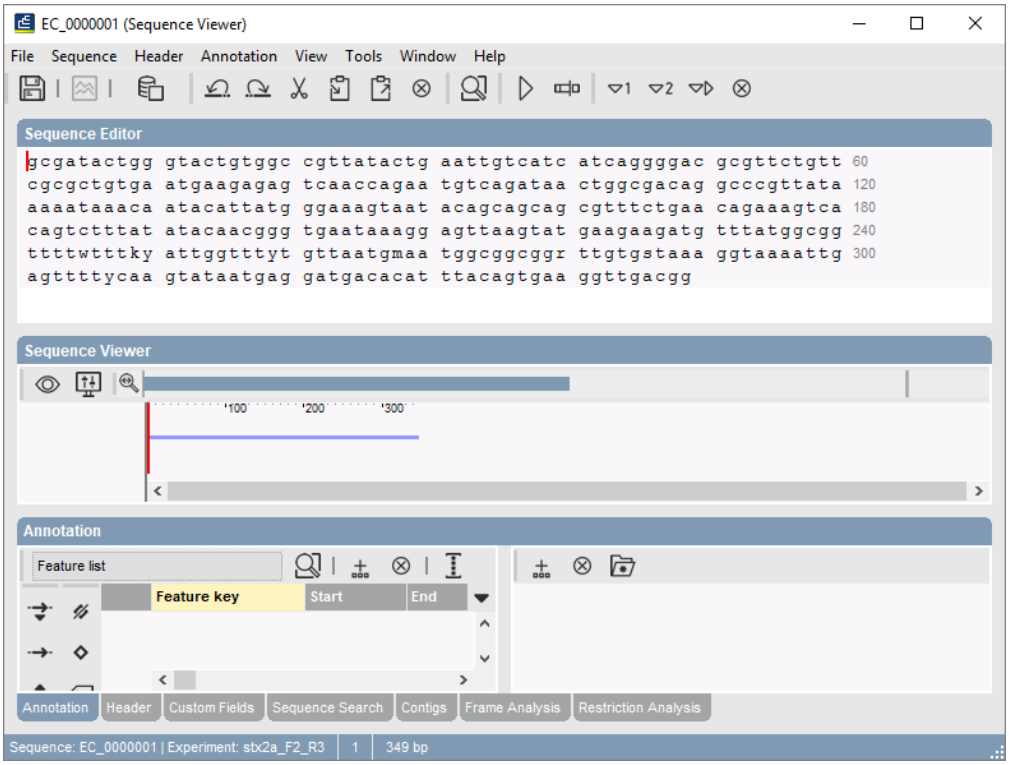


Figure 16: Example output of the ***stx2a_F2_R3*** experiment type for sample EC_0000001.

10. Close the *Sequence editor* window.

6 Reports

1. Open the genotype report for the selected entries with ***E. coli*** > ***Reports...***.

The *Report* window contains a genotype report for each of the selected entries (see Figure 17).

2. Select another entry in the *Entries* panel to update the results in the *Genotype report* panel.

The creation date of the report (***Date***), the Key (***Name***), and information fields checked in the *Settings* dialog box are displayed in the *Genotype report* panel.

3. Select ***Report*** > ***Report templates*** in the *Report* window and make sure the option ***Summary*** is selected.

A summary of the results of all analyzed traits is displayed in the *Report* window.

4. Select ***Report*** > ***Report templates*** in the *Report* window and select the option ***Complete*** (see Figure 18).

In the ***Complete*** view, the summarized results as well as all details are shown, including the serotype antigens, descriptions of the detected genes, the mutational resistance decision trees, the detected loci of the virulence islands, the genome positions of the full phages, ...

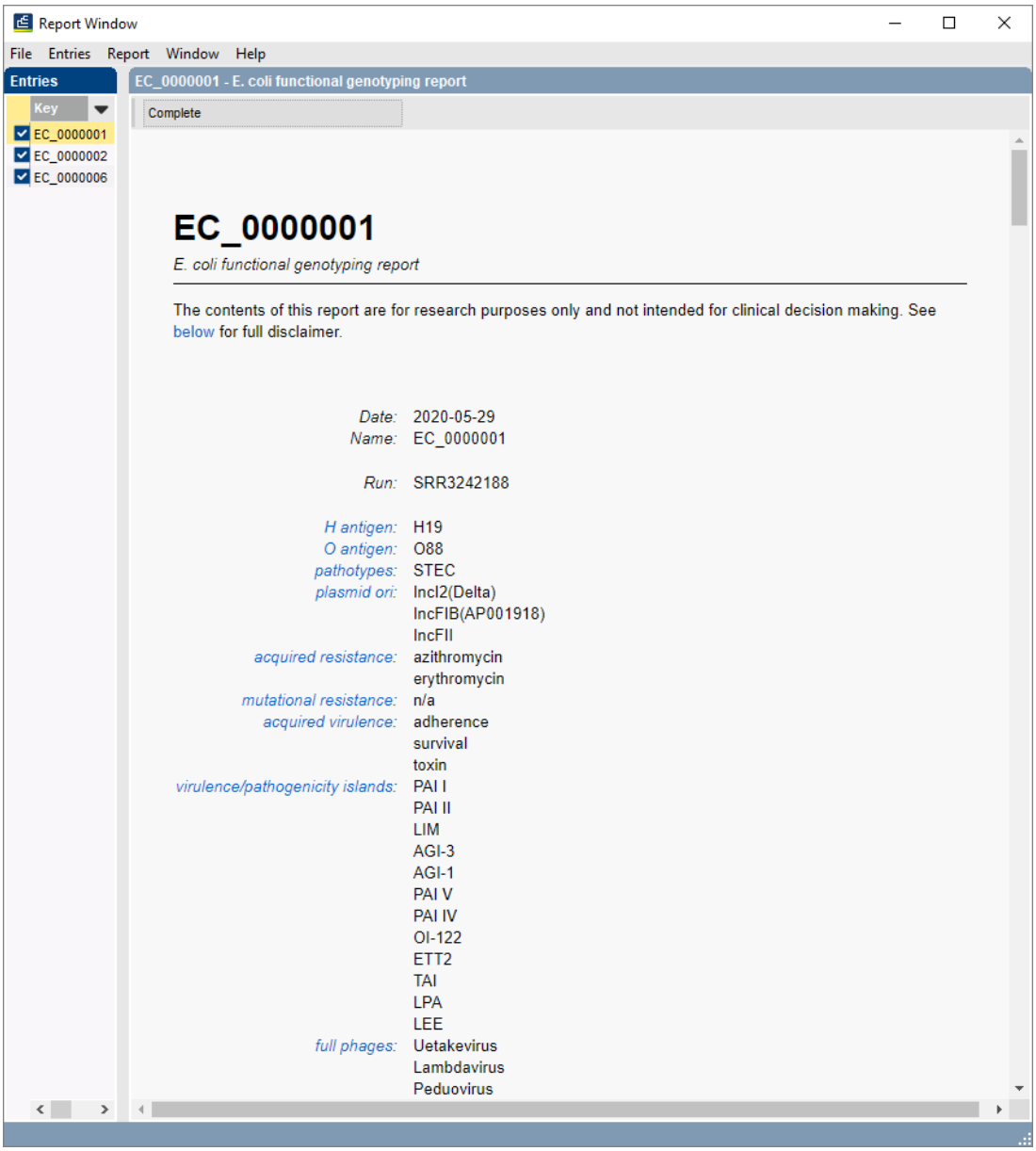


Figure 17: Functional genotyping report.

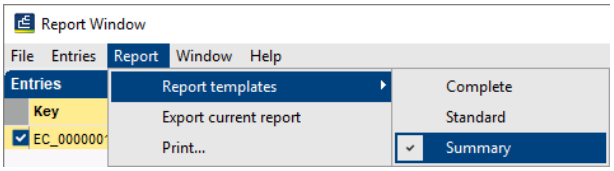


Figure 18: Report templates in the *Report window*.

5. Click on a hyperlink of one of the predicted traits to display the detailed BLAST results in the *Genotype report* panel (see Figure 19).

All hits that passed the settings for ***Acquired Resistance***, ***Mutational Resistance***, ***Acquired Virulence***, ***Virulence Islands***, ***Serotype***, ***Pathotypes***, ***Plasmid ORI***, ***Full Phages*** and ***Full Plasmids*** screening are listed. Detailed BLAST results include trait, locus, BLAST similarity scores (***Coverage*** (%), ***Identity*** (%)), the position in the assembly the locus was found (***Position***) and

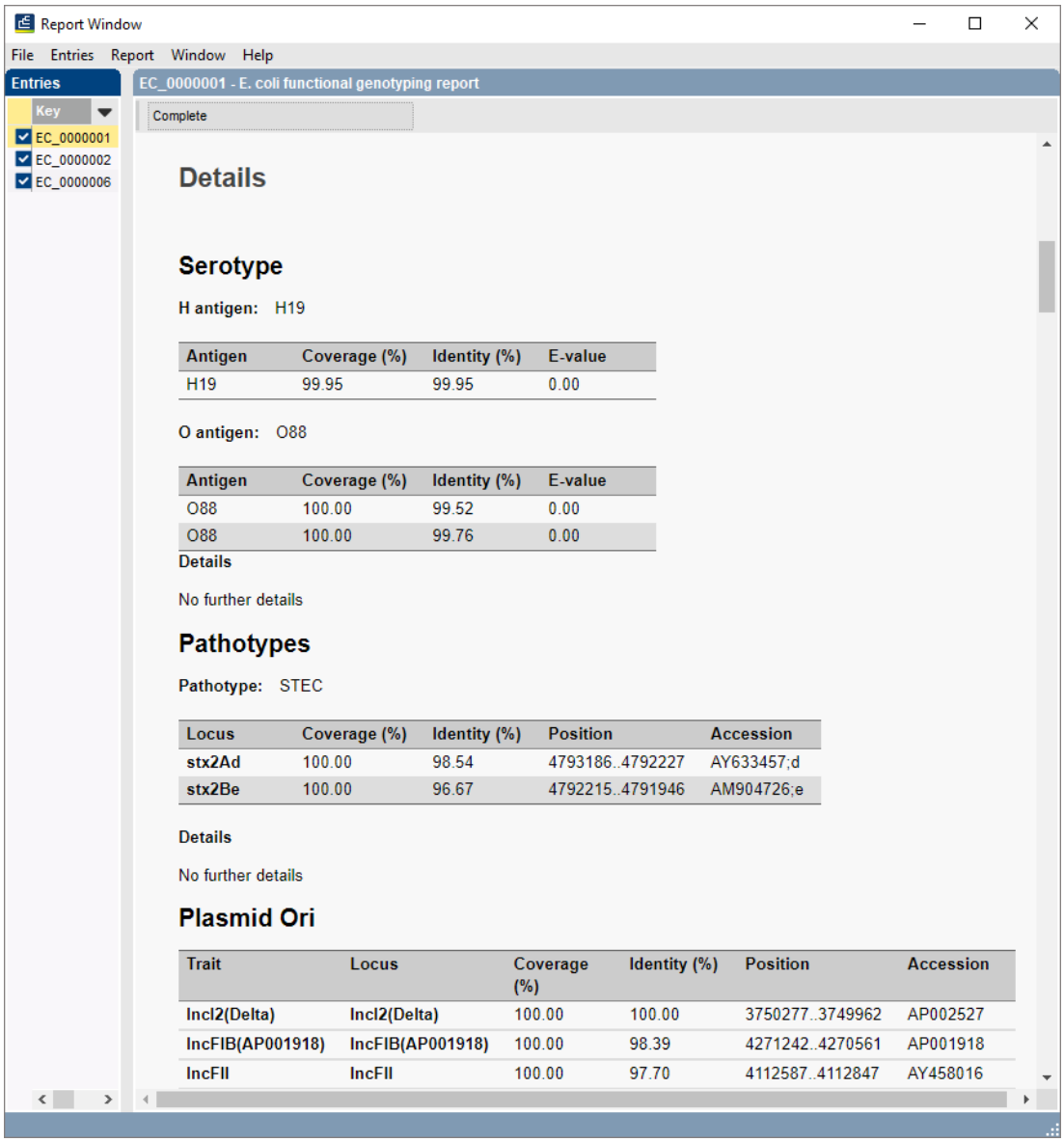


Figure 19: Report details.

the accession number of the detected locus (***Accession***).

6. Select ***File*** > ***Exit*** to close the *Report* window.