BIONUMERICS Tutorial:

Importing FASTQ files and FASTQ file links

1 Aim

Essentially, there are two ways to import FASTQ files in your BIONUMERICS database: the default import method stores the sequence reads in the BIONUMERICS database (either in the source files directory or optionally in the relational database (see 3)) and the second import method only imports the links to the location of the FASTQ files (see 4).

In this tutorial both options are described but please keep in mind that the second method, the storage by link method, is recommended since this keeps the BIONUMERICS database lightweight and avoids duplication of data.

2 Example data

Example data that will be used in this tutorial can be downloaded from the Applied Maths website: https://www.applied-maths.com/download/sample-data, "FASTQ files").

The data set contains 10 gzipped fastq files of 5 paired end read data file pairs coming from *Staphylococcus aureus* and an Excel file `Strain information.xlsx` containing some meta data on the sequence read sets.

3 Importing FASTQ files in the BioNumerics database

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

2. Select **File** > **Import...** (, **Ctrl+I**) to open the *Import* dialog box.

3. Select the option **Import sequence read set files** under **Sequence read sets data** (see Figure 1).

Using this import functionality, sequence read sets can be imported from the following formatted files:

- Roche/454® sequence files, with extensions .fna (sequence information) and .qual (quality information).

- FASTA files, with extensions .fasta, .fna, .ffn, .faa or .txt.
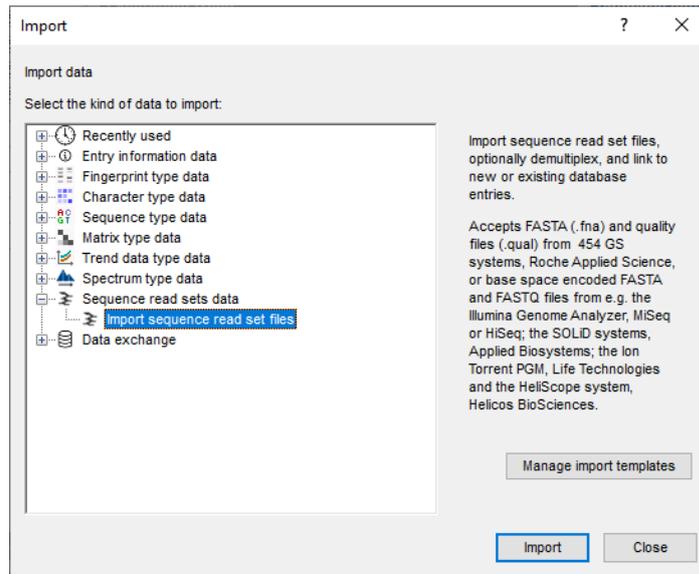
Figure 1: Import sequence read set files.

- FASTQ files, with extensions .fq, .fastq or .txt.

   4. Press <***Import***> to start the *Import sequence read sets* wizard.

   5. Press <***Browse***>, navigate to the correct location, select all 10 files in the `FastQ files` folder and press <***Open***> to add the selected files to the import dialog.

The *Import sequence read sets* wizard has detected that the ten gzipped fastq files form five paired end read data file pairs, because they have the same name apart from the _1 or _2 suffix.

   6. Press <***Next***> to proceed.

   7. No demultiplexing is needed so press <***Next***> to continue.

Now you need to define how the data should be stored in the database. The default template **Example import** can be applied to most file names. This template will only retain the SRA run accession numbers from the file names and store this information in the BIONUMERICS ***Key*** field.

   8. Select the ***Example import*** template and press the <***Preview***> button to check the outcome of the parsing. Close the preview.

✎ If the default template is not applicable to your files, press the <***Create new***> button to create your own template and rules.

   9. Make sure <***Create new***> is selected from the ***Experiment type*** list or select an existing experiment and press <***Next***> (see Figure 7).

   10. Specify a sequence type name when prompted for, e.g. "wgs" (see Figure 8). Click <***OK***> and confirm the creation of the experiment.

   11. Leave the option **create 5 entries** checked and press <***Finish***> to start the import of the sequence read sets.

The sequences are linked to new entries in the database.

   12. Once the import is completed, click on one of the green dots in the *Experiment presence* panel in the ***wgs*** column to visualize some basic statistics on the imported sequence read sets (see 6 for more detailed information about these statistics).
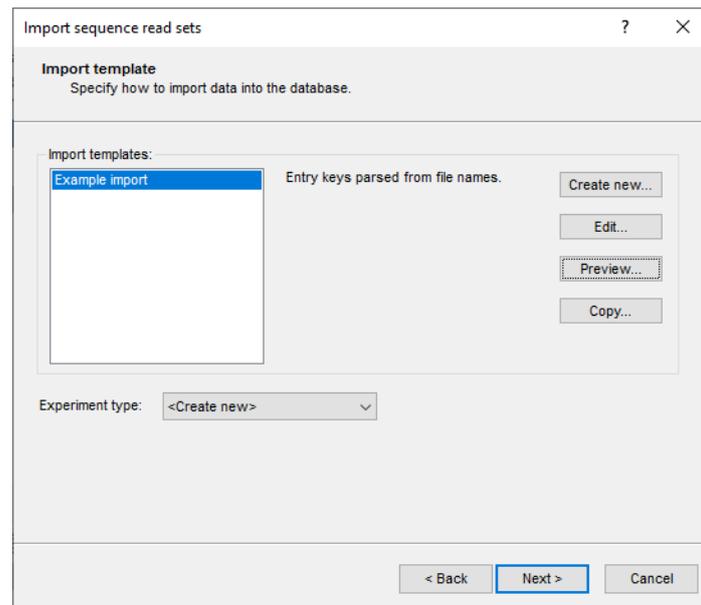
Figure 2: Import template.
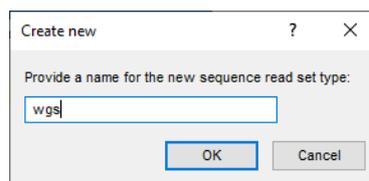
Figure 3: Create new experiment.

13. Close the sequence read set card.

14. Double-click the experiment **wgs** in the *Experiment types* panel and select **Settings** > **General settings...** (⏬) to call the *Sequence read set experiment type settings* dialog box.

By default, the option **Save in database** is unchecked, which implies that the imported sequence read set data are stored as separate files in the Source files location.

The default Source files location is `[DBPATH] \sourcefiles`. `[DBPATH]` hereby refers to the database folder in the BIONUMERICS home directory. However, the path can be any local directory or network path, for example on a server computer. More information on how to change the path can be found in the reference manual.

When the option **Save in database** is checked, the read sets are stored in the connected database. Please note that this may fill up your database very quickly as these data sets are typically beyond the reach of small database systems e.g. Access or SQL Server Express.

15. Close all windows.

4 Importing FASTQ file links in the BioNumerics database

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

Sequence reads can be imported as data links in BIONUMERICS using the **Import sequence**

read set data as links import routine in the Import tree. This is the recommended option, since it avoids duplication of the data and keeps the BIONUMERICS database lightweight.

Importing sequence read sets as links is only possible when the *WGS tools plugin* is installed in the BIONUMERICS database:

    2. Call the *Plugins* dialog box from the *Main* window with **File** > **Install / remove plugins...** (⏻).

    3. Select the *WGS tools plugin* from the list in the *Applications tab* and press the <**Activate**> button.

    4. Confirm the installation of the plugin.

The *Calculation engine URL* wizard page queries for the Uniform Resource Locator (URL) that uniquely identifies the calculation engine instance to connect to.

With the **Use default Cloud Calculation Engine** option clients will use the Applied Maths cloud instance (`https://wgmlst.applied-maths.com`), which is hosted on Amazon servers in the US. This option should also be selected if you do not intend to run jobs on the calculation engine, but instead run all calculations on your own computer.

    5. Make sure the **Use default Cloud Calculation Engine** option is selected and press <**Next**>.

In the next step of the *WGS tools installation* wizard, two options are available:

- Choose **Local calculations only** if you do not intend to run jobs on the calculation engine and instead wish to run all calculations on your own computer.

- Choose **Enable running jobs on Cloud Calculation Engine** to unlock the full potential of the default Cloud Calculation Engine.

In this tutorial we will run all calculations locally for our ***Staphylococcus aureus*** samples. Please consult the *WGS tools plugin* manual for more information about the **Enable running jobs on Cloud Calculation Engine** option.

    6. Make sure **Local calculations only** is checked, and select ***Staphylococcus aureus*** from the **Organism** drop-down list (see Figure 4).

Figure 4: Calculation and organism settings.

7. Press <***Next***> to proceed with the installation.

BIONUMERICS will now download organism-specific settings and search data. A confirmation message pops up when the download is completed.

8. Press <***OK***> twice to finalize the installation of the plugin.

9. Press <***Exit***> to close the *Plugins* dialog box.

10. Close and reopen the database to activate the features of the *WGS tools plugin*.

After installation of the *WGS tools plugin*, sequence reads sets can now be imported as links:

11. Select ***File*** > ***Import...*** (, **Ctrl+I**) to call the Import tree.

12. Make sure the ***Import sequence read set data as links*** option is selected in the Import tree and press <***Import***> (see Figure 5).
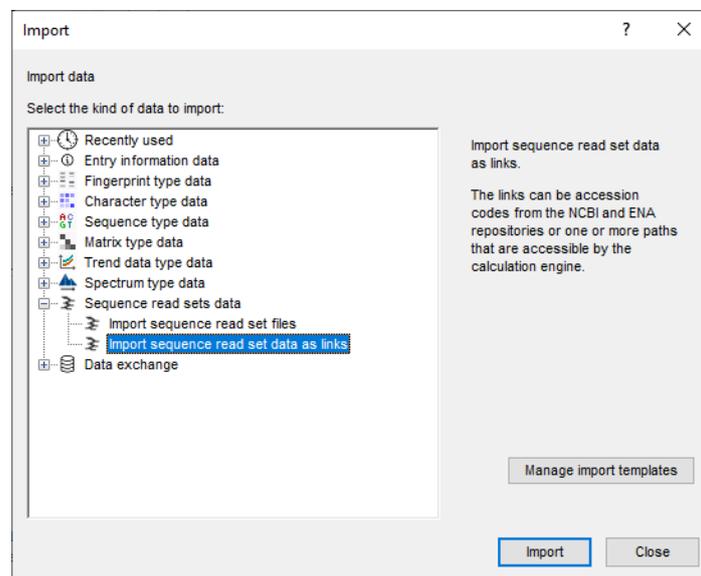
Figure 5: Import sequence read set data as links.

Links to multiple data sources are available, including online and offline data repositories. In this tutorial, the import of FASTQ files from a ***Local file server*** is covered.

> The import of links to ***NCBI (SRA)***, ***EMBL-EBI (ENA)***, ***Amazon (S3)***, ***BaseSpace*** and ***Alibaba OSS*** is covered in the tutorial "Importing links to online repositories".

13. Select the ***Local file server*** and press <***Next***>.

14. Press <***Browse***>, navigate to the correct location, select all 10 files in the `FastQ files` folder and press <***Open***> to add the selected files to the import dialog.

The option ***Auto-detect paired-end files*** is default checked. This option ensures that the files are checked for the presence of paired-end data. Files that contain paired-end data are recognized by the same file name except for paired-end specific characters: e.g. same name apart from the _1 or _2 suffix.

15. Select <***Next***> to go to the next step.

Now you need to define how the data should be stored in the database. The default template **Example import** can be applied to most file names. This template will only retain the SRA run
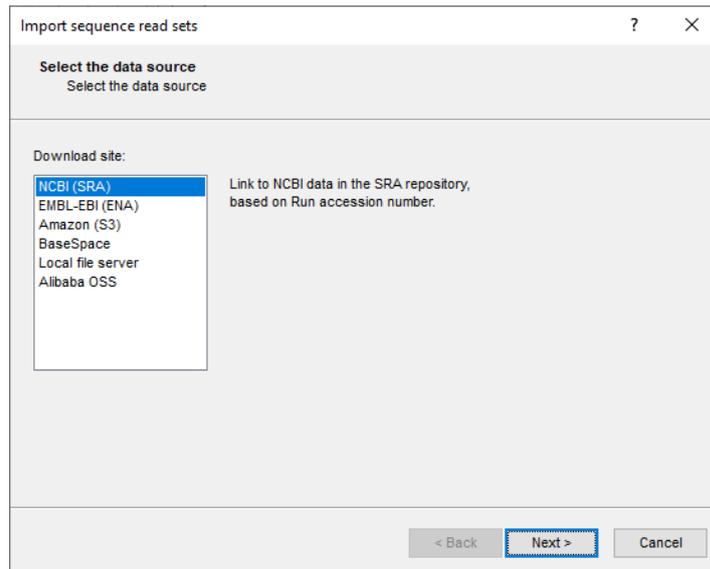
Figure 6: Data sources.

accession numbers from the file names and store this information in the BIONUMERICS **Key** field.

16. Select the **Example import** template and press the <**Preview**> button to check the outcome of the parsing. Close the preview.

> If the default template is not applicable to your files, press the <**Create new**> button to create your own template and rules.

17. Make sure <**Create new**> is selected from the **Experiment type** list or select an existing experiment (see Figure 7) and press <**Next**>.
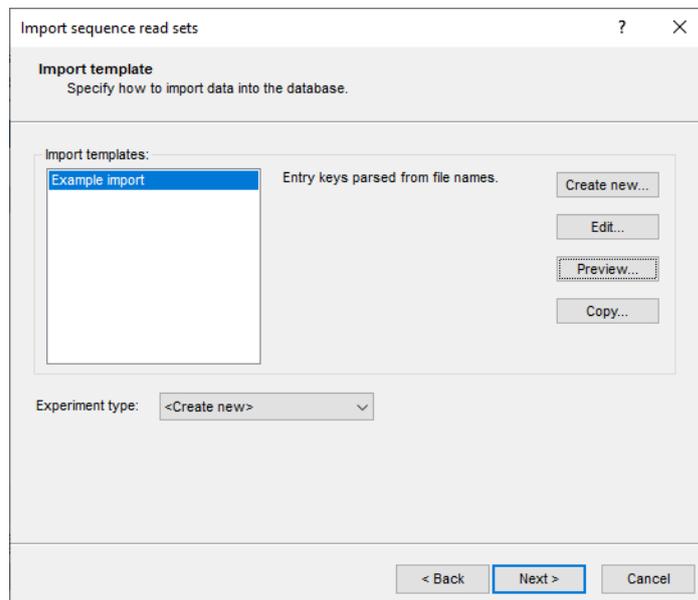
Figure 7: Import template.

18. Specify a sequence type name, e.g. "wgs". Click <**OK**> and confirm the creation of the experiment.

19. Leave the option **create 5 entries** checked and press <**Next**> to start the import of the sequence read set links.
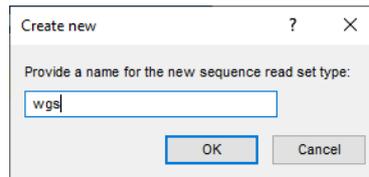
Figure 8: Create new experiment.

In the last step, calculation jobs can be launched on the imported data links (***Open submit jobs dialog after import***). The same dialog can be called from the *Main* window at any time with ***WGS tools*** > ***Submit jobs...*** (▷).

When the ***Local file server*** option was selected as data source, some basic statistics on the reads can be calculated upon import (***Calculate sequence read set statistics***). Based on the sequence read set statistics bad sequencing runs for which no jobs should be submitted can be filtered out.
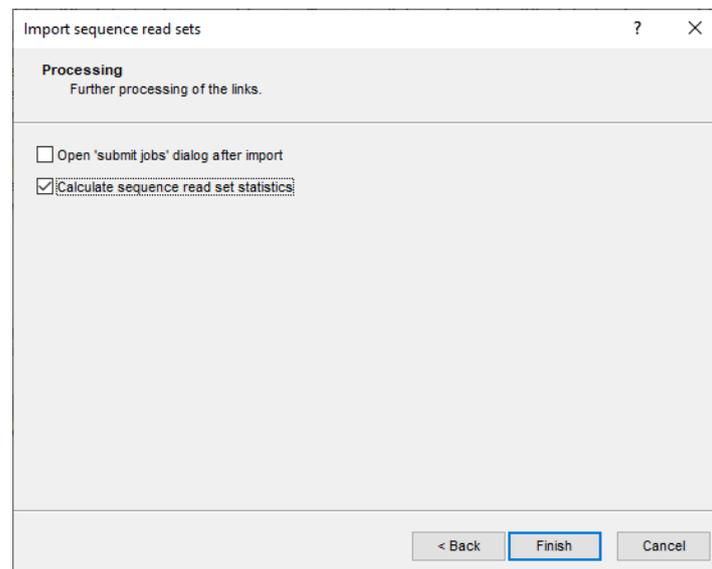
Figure 9: Processing of the links.

20. Make sure the ***Calculate sequence read set statistics*** option is selected, uncheck ***Open submit jobs dialog after import*** and press <***Finish***> to start the import of the data links.

Once the import is completed, the entries are created/updated and have one green dot next to it in the column of the sequence read set experiment type **wgs**.

21. Click on a green colored dot corresponding to the experiment type **wgs**.

The data link is displayed in the ***Storage*** section of the *Sequence read set experiment* window. *Sequence read set report* panel.

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

5 Importing metadata

Metadata on the sequence read sets can be imported from text, Excel, and other ODBC-compatible files using the BIONUMERICS import routines.

As an exercise, we will import data from the Excel file `Strain information.xlsx` into a BIONU-MERICS database.

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

2. Choose the option **Import fields (Excel file)** under the **Entry information data** in the tree (see Figure 10) and click <**Import**>.
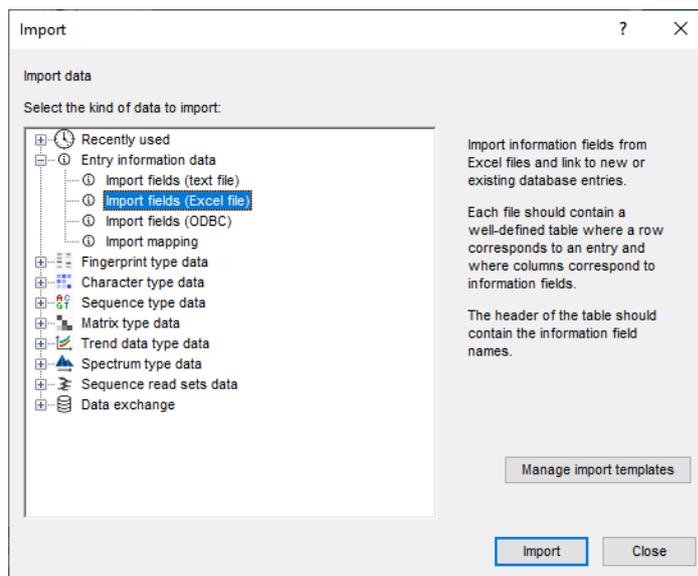
Figure 10: Import fields (Excel file).

3. Press <**Browse**>, navigate to the "Strain information.xlsx" file saved to your computer (see Figure 11), and press <**Next**>.
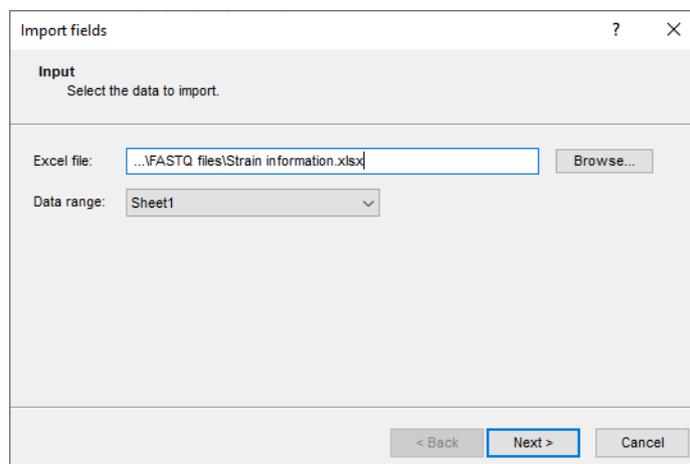
Figure 11: Browse for the Excel file.

The next dialog box allows you to set import rules. For each import source (i.e. Excel column), a database destination can be specified.

First we will link the ***Run number*** column in the Excel file to the ***Key*** field:

4. Double-click the first row in the grid, press ***Key*** from the list, and press <***OK***>.

5. Make a multiple selection for all other rows. Do this by selecting the second row (**Organism name**) and while holding the **Shift**-key, double-click on the last row (**Instrument**). Select <***Edit destination***>, select ***Entry info field*** as destination and click <***OK***>. Click <***OK***> once more and click <***Yes***> to confirm the creation of the new information fields.

The import rules are updated in the grid (see Figure 12).
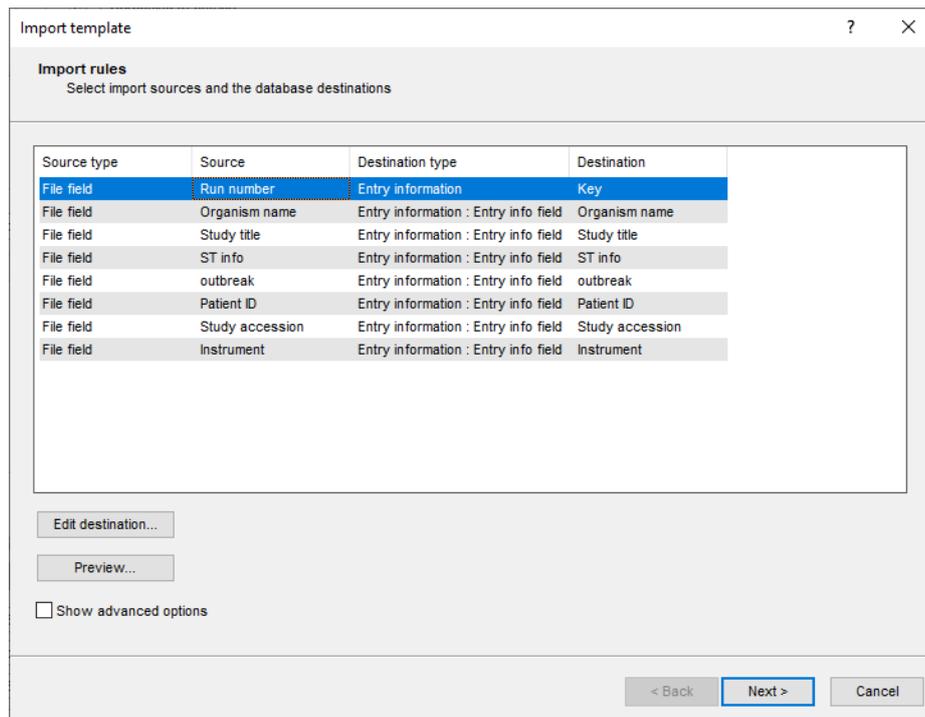
Figure 12: Import rules.

6. Optionally, you can do a preview of what you are about to import. Press <***Preview. . .***> to open the preview. Close the preview again.

7. Click <***Next***> and <***Finish***> to finish the creation of the import template for the database information fields.

8. Enter a meaningful name (and optionally a description) for the created import template e.g. "Import of run information", and click <***OK***>.

9. Then choose the newly created import template from the list and click <***Next***>.

10. The next dialog allows you to confirm the update of the 5 new entries in the database. Click <***Finish***>.

The information is added to the database (see Figure 14).

6   Quality assessment of sequence read sets

1. Click on the colored dot of an imported sequence read set to open the *Sequence read set experiment* window.
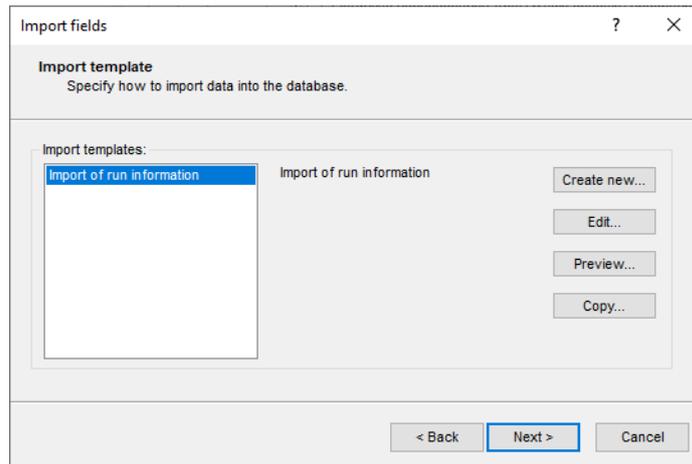
Figure 13: Import template.
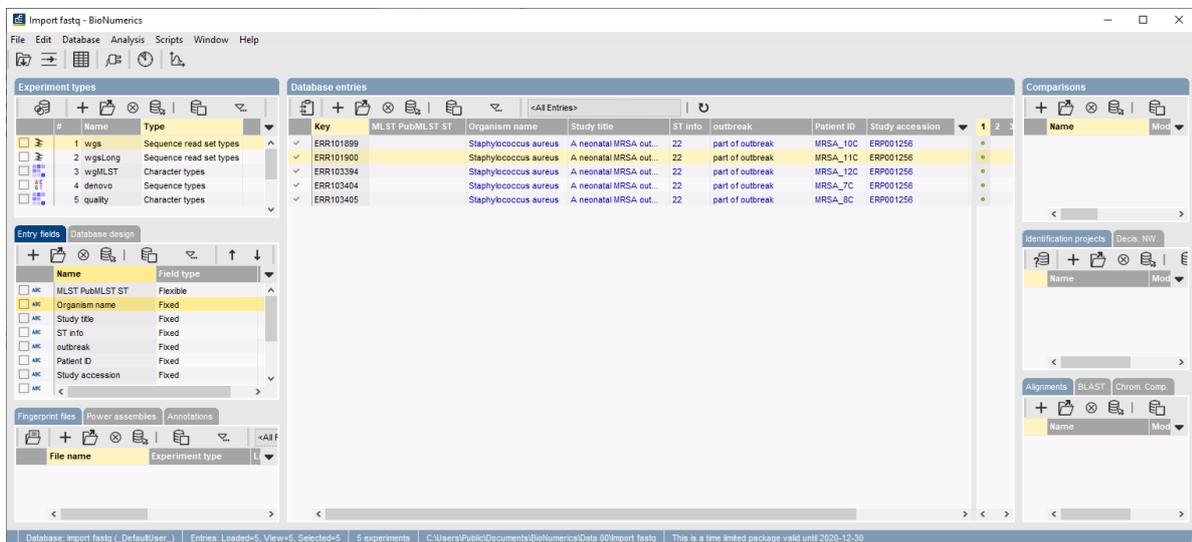
Figure 14: The *Main* window.

When the files are stored as links in the database, the paths are indicated in the *Sequence read set report* panel on top under the **Storage** section (see Figure 15).

When the files are stored in the BIONUMERICS database, no **Storage** section is present in the *Sequence read set report* panel (see Figure 16).

A summary of the characteristics of the sequence read set is displayed in the *Sequence read set report* panel, including information on *Read set size*, *Sequence length statistics*, *Quality statistics* and *Base statistics*. When files are stored as links, this information is only displayed if the option **Calculate sequence read set statistics** was checked in the last step.

On a more detailed level, it is very interesting to consult the predefined charts concerning the average read quality distribution, the base distribution, the read length distribution, read quality distribution by %GC . . .

2. Select **Analysis** > **Charts and statistics...** (, **F7**) to call the *Create chart* dialog box.

Selecting any of the charts and pressing <**OK**> will automatically create a dedicated chart upon
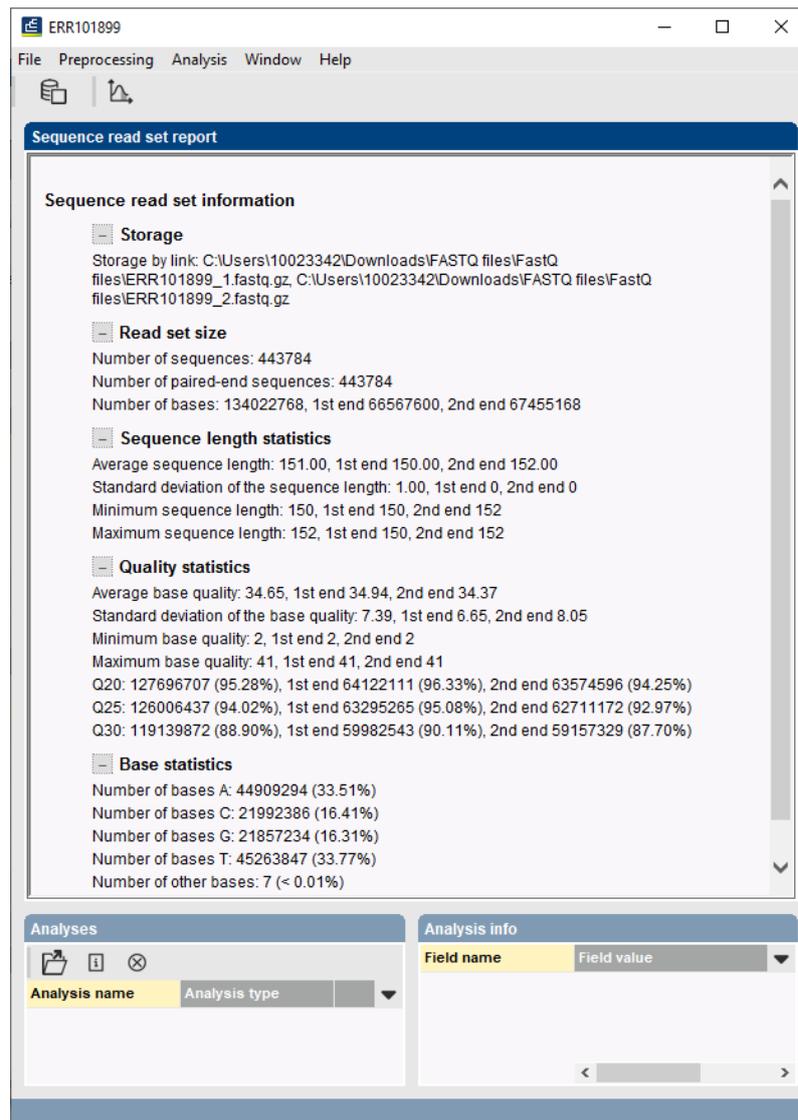
Figure 15: Sequence read set stored as link in the database.

the read information present in the sequence read set at hand.

3. Select the existing chart template **Sequence read set quality distribution (average)** and press <**OK**>.

This will launch the *Charts and statistics* window, where the quality distribution is plotted (see Figure 17).

The chart templates may provide insight in the sequence run quality and the possible presence of sequence artifacts in the run in a quick and easy way. From these preliminary insights, assessment can be made for the required preprocessing steps before starting the actual analysis.

4. Close the *Charts and statistics* window and return to the *Sequence read set experiment* window.

5. Select **Analysis** > **Charts and statistics...** (, **F7**) to call the *Create chart* dialog box again and select another chart template. Press <**OK**> to create the plot.
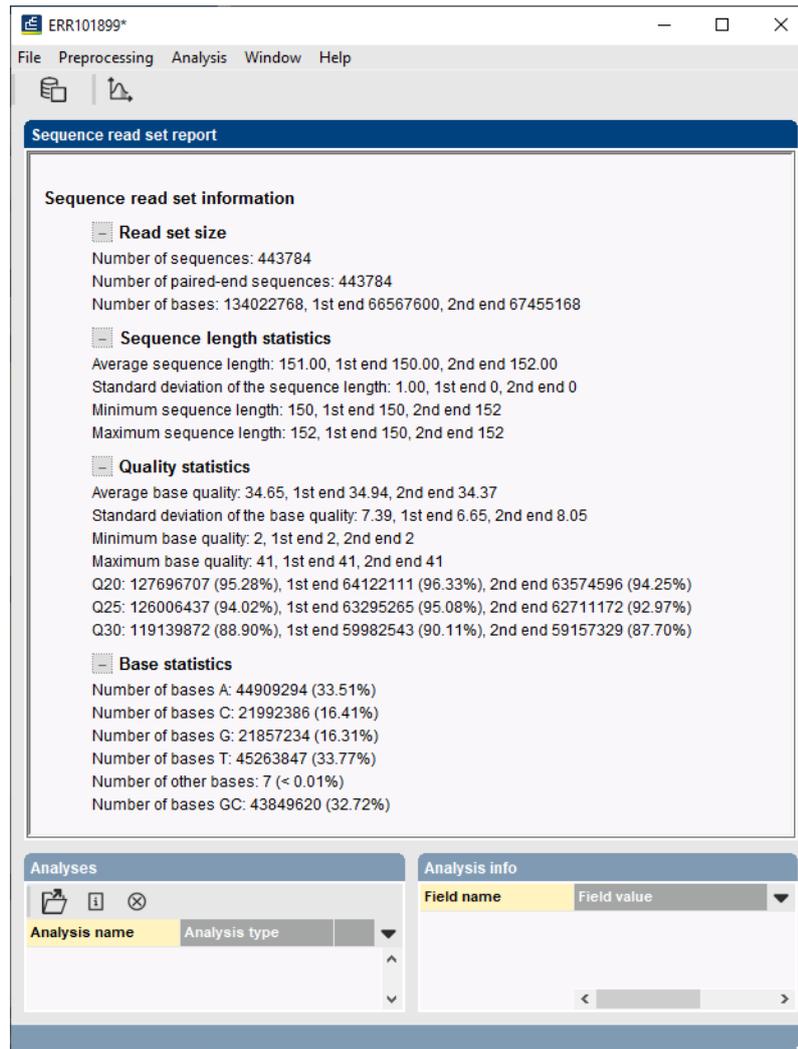
Figure 16: Sequence read set stored inside the database: no storage information is displayed.

7 Analysis tools

Analysis tools are covered in following tutorials:

- "Performing a de novo assembly on the local calculation engine"

- "Performing a de novo assembly on the cloud calculation engine"

- "Performing whole genome SNP analysis with mapping performed on the local calculation engine"

- "Performing whole genome SNP analysis with mapping performed on the cloud calculation engine"

- "wgMLST typing: routine workflow starting from sequence read sets"
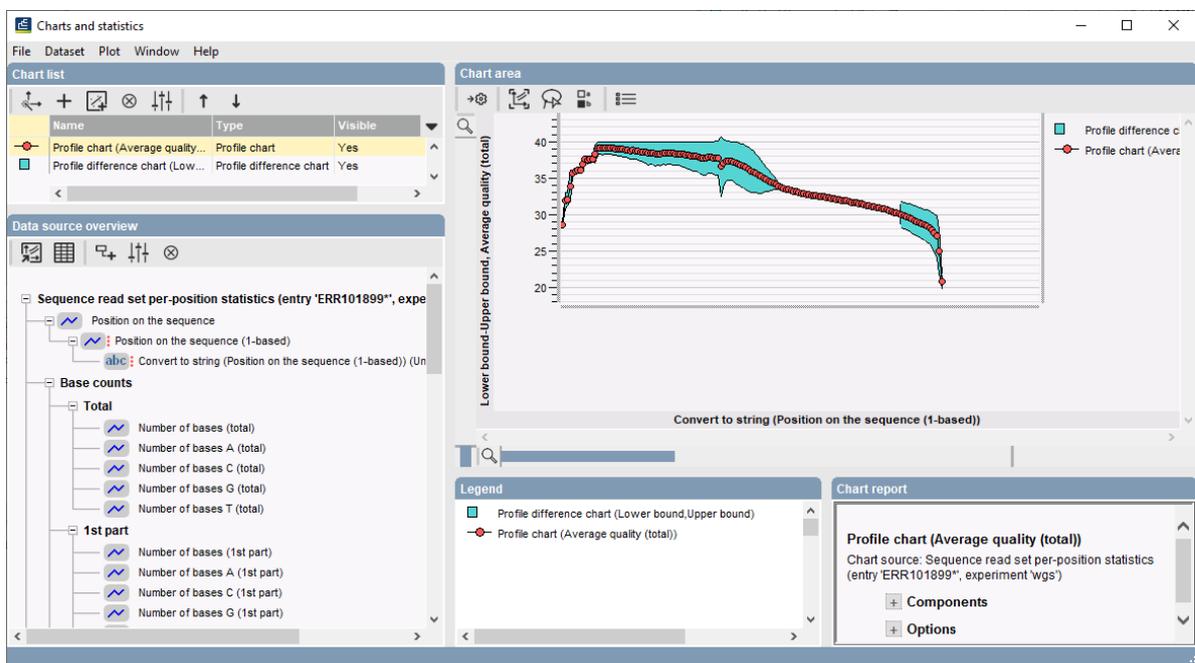
Figure 17: The chart displaying the sequence read set quality distribution (average) for an entry.