BioNumerics Tutorial:

Performing a resequencing assembly

1 Aim

In this tutorial, we will discuss the different options to obtain statistics about the sequence read set data and assess the quality, and we will perform a resequencing assembly.

2 Example data

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

The example data is stored as two gzipped fastq files in one paired end read data file pair coming from *Staphylococcus aureus*: ERR1143520_1.fastq.gz and ERR1143520_2.fastq.gz. This data was generated by Illumina MiSeq whole genome sequencing and downloaded from http://www.ncbi.nlm.nih.gov/sra.

3 Importing sequence read sets

3.1 Import wizard

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.

Two import routines are available for the import of fastq files:

- ***Import sequence read sets***: With this option, a multitude of different file types can be imported and stored inside the database.

- ***Import sequence read set data as links***: With this option, only the link to the samples is stored in BioNumerics, resulting in a lightweight database. This option is only available after installation of the *WGS tools plugin*. Installation of this plugin is only possible with a valid password and a project name, linked to a certain amount of credits. Please contact Applied Maths to obtain more information about the *WGS plugin*.

In this tutorial the first option is described. Please keep in mind that the storage by link is recommended, keeping the BioNumerics database lightweight and avoiding duplication of data. The storage by link workflow is illustrated in following tutorials: "Importing FASTQ files" and "Importing links to online repositories".

3. Under ***Sequence read sets data***, select the option ***Import sequence read set files*** (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.
- FASTQ files, with extensions .fq, .fastq or .txt.

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

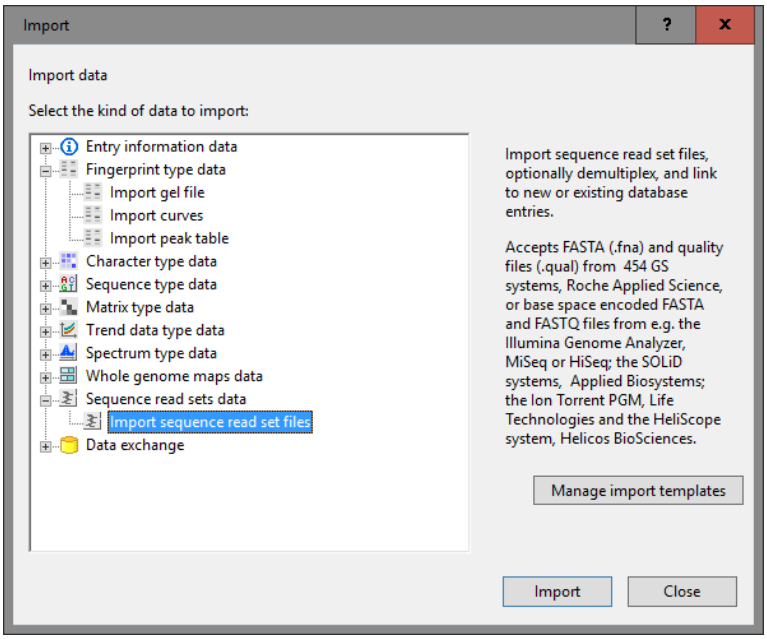

Figure 1: The Import tree.

5. A dialog box pops up, allowing you to browse for the sequence reads set files containing the data. Press <***Browse***>, navigate to the correct folder, select both `ERR1143520_1.fastq.gz` and `ERR1143520_2.fastq.gz` while holding the **Ctrl**-key and press <***Open***> to add the selected files to the import dialog (see Figure 2).

The *Import sequence read sets* wizard has detected that the two gzipped fastq files belong to one paired end read data file pair, because they have the same name apart from the _1 or _2 suffix.

6. Leave ***Import as paired-end read data*** checked and press <***Next***> to proceed.

7. It is possible to demultiplex the data during import, but no multiplexing was done in our current sample. Therefore, leave the option unchecked and press <***Next***>.

We now need to define how the data should be stored in the database. The accession number corresponds to the file name up until the underscore. We will use the NCBI run accession number as the entry key.

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 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. In the *Import template* dialog box, make sure the default import template is selected and select ***Create new*** from the experiment list (see Figure 3). Click <***Next***>.

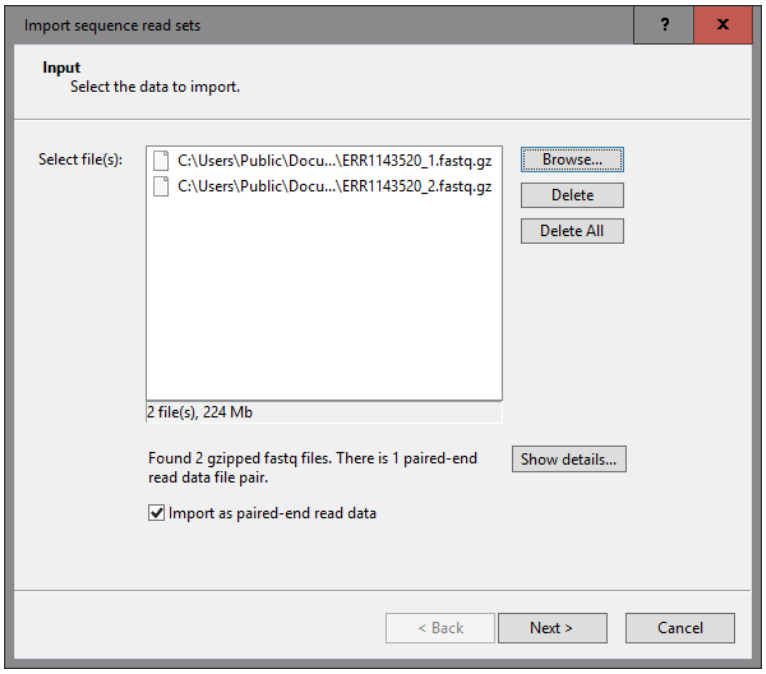

Figure 2: Select the FASTQ files.

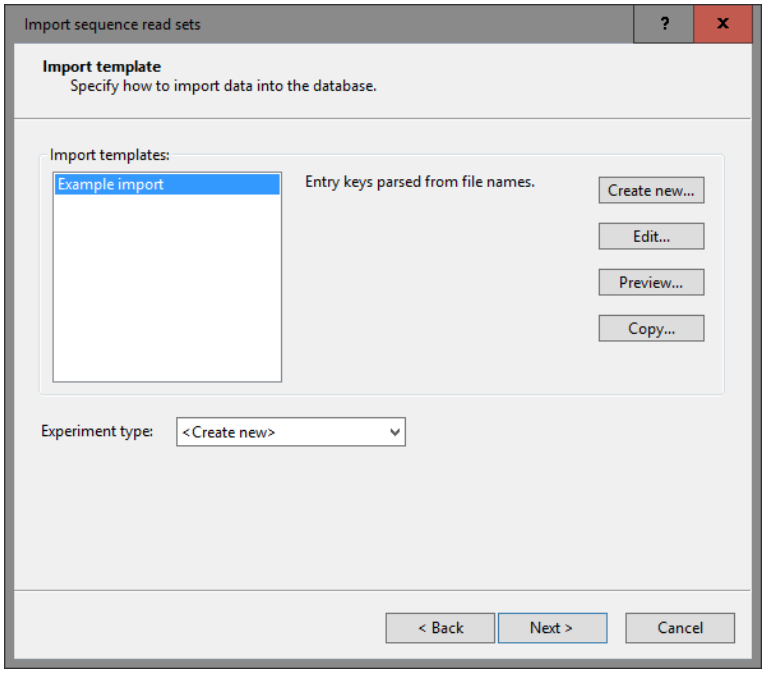

Figure 3: Select import template and experiment.

10. Specify a name for the new sequence type experiment (e.g. **Whole genome sequence**) and press <***OK***> and confirm the creation of the new experiment in the database (see Figure 4).

11. Click <***Finish***> to confirm the creation of 1 new entry and start the import.

Once the import is completed, entry **ERR1143520** is present in the BioNumerics database, and has one green dot next to it in the column of the sequence read set experiment type **Whole genome sequence**.

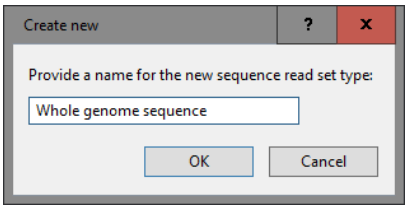

Figure 4: New experiment type.

3.2 Quality assessment of sequence read sets

A sequence read set experiment offers some high-level statistics on the number, length, quality, etc. of sequence reads.

12. Click on the colored dot of the imported sequence read set of entry **ERR1143520** to open the *Sequence read set experiment* window.

In this window, 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*.

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 ...

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

Selecting any of the chart templates and pressing <***OK***> will automatically create a dedicated chart upon the read information present in the sequence read set at hand.

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

This action will launch the *Charts and statistics* window, where the quality distribution is plotted (see Figure 5). From this figure, it can easily be seen that within the current data set, the average quality of the first 22 bases is lower than the rest of the reads. However, all average quality scores are 32 or higher, which is still acceptable.

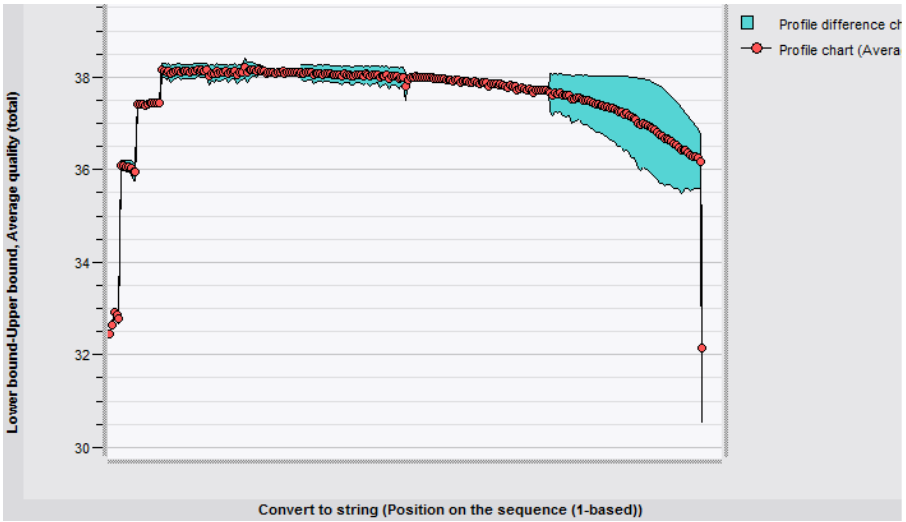

Figure 5: The chart displaying the sequence read set quality distribution (average).

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.

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

16. 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. See Figure 6 and Figure 7 for a few examples of predefined chart plots.

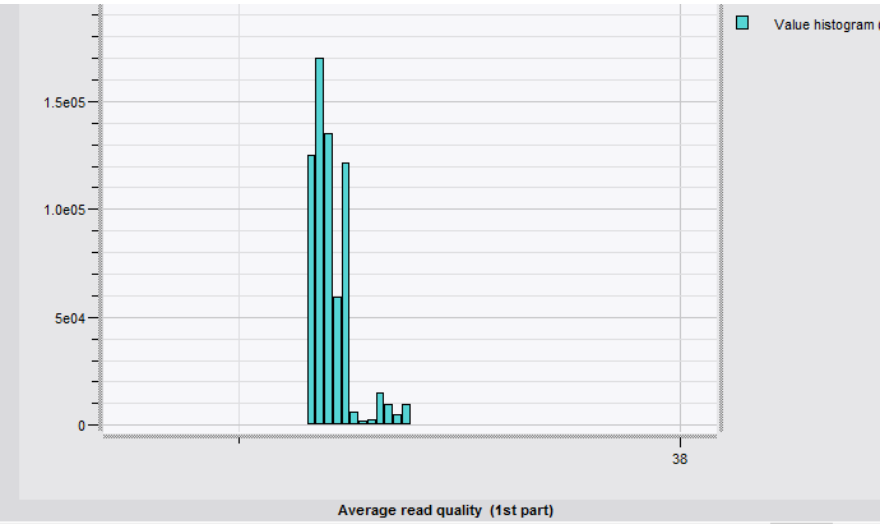

Figure 6: The 'Sequence read set average read quality distribution' plot.

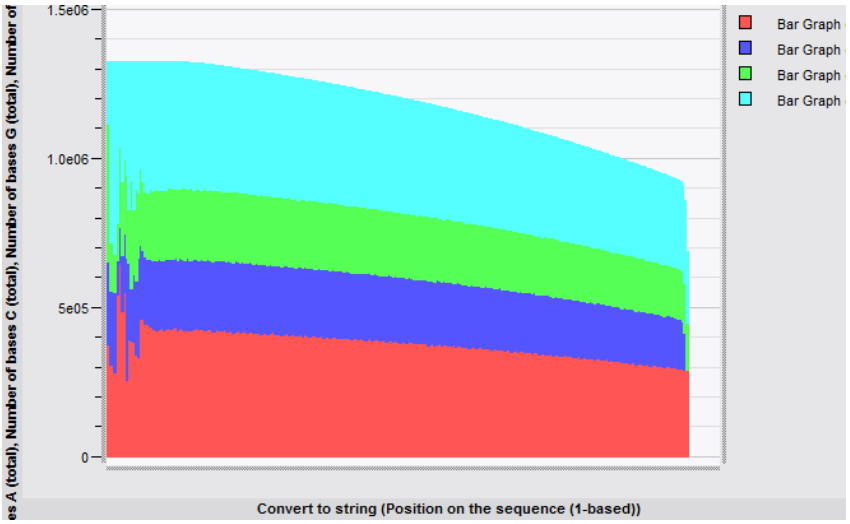

Figure 7: The 'Sequence read set base distribution' plot.

17. To export one of the generated charts, select ***File*** > ***Export...*** and choose the format of your choice.

3.3 Quality trimming of sequence read sets

A commonly used processing step after import of raw NGS data is quality trimming, to remove bases and reads of lower quality from the dataset. We will demonstrate this feature here on the newly imported entry and store the trimmed data in a new sequence read set experiment.

18. Click in the *Experiment types* panel to activate this panel and select ***Edit*** > ***Create new object...*** ().

19. From the *Create a new experiment type* dialog box that appears, highlight ***Sequence read set type*** and press <***OK***>.

20. Name it **wgs_trimmed** and click <***OK***>.

21. Select the entry with key **ERR1143520** and go to ***Analysis*** > ***Sequence read set types*** > ***Trimming***.

22. In the *Trimming* dialog box, leave **Whole genome sequence** as the ***Input sequence read set experiment type***, and press <***Next***>.

The first set of parameters defines the **Structural trimming** based on the sequence content, reads that do not qualify for these parameters will be removed as a whole.

23. Leave all values at their defaults and click <***Next***>.

The **Overall quality trimming** parameters define the quality threshold to remove reads as a whole based on read quality. The default values are based on fastq files with a different quality format and have to be modified for the example data.

24. Specify ***Exclude reads with minimum quality below*** as "10", ***Exclude reads with average quality below*** as "32", and ***Replace base by N when the quality is below*** as "15" (see Figure 8). Click <***Next***>.

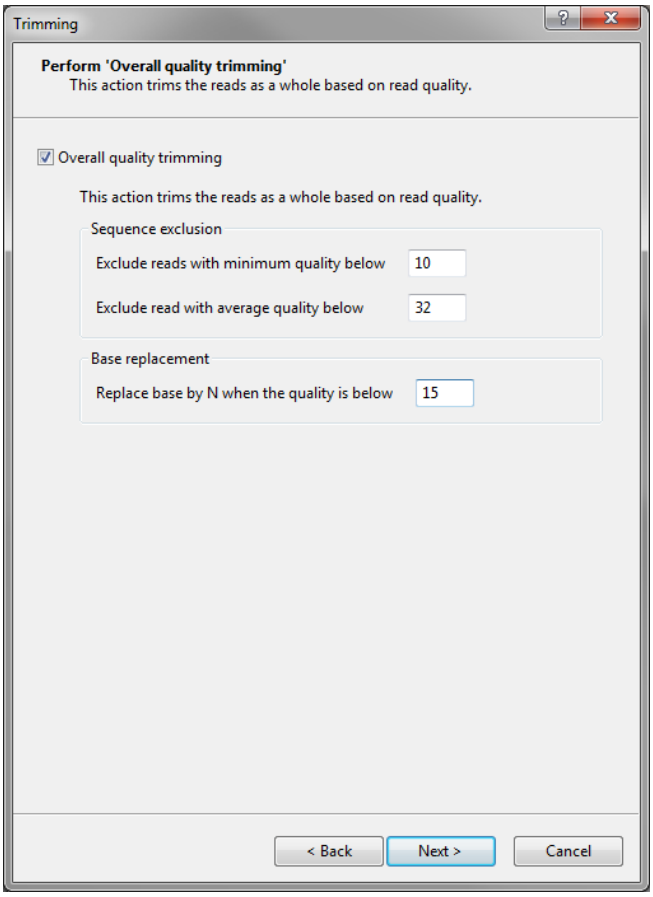

Figure 8: Overall quality trimming parameters.

The **Tail quality trimming** parameters define the removal of tails from the reads, based on the base quality.

25. Set ***Minimum tail quality*** to "18", ***Minimum windowed average quality*** to "15" and ***Minimum rolling average quality*** to "15" (see Figure 9). Click <***Next***>.

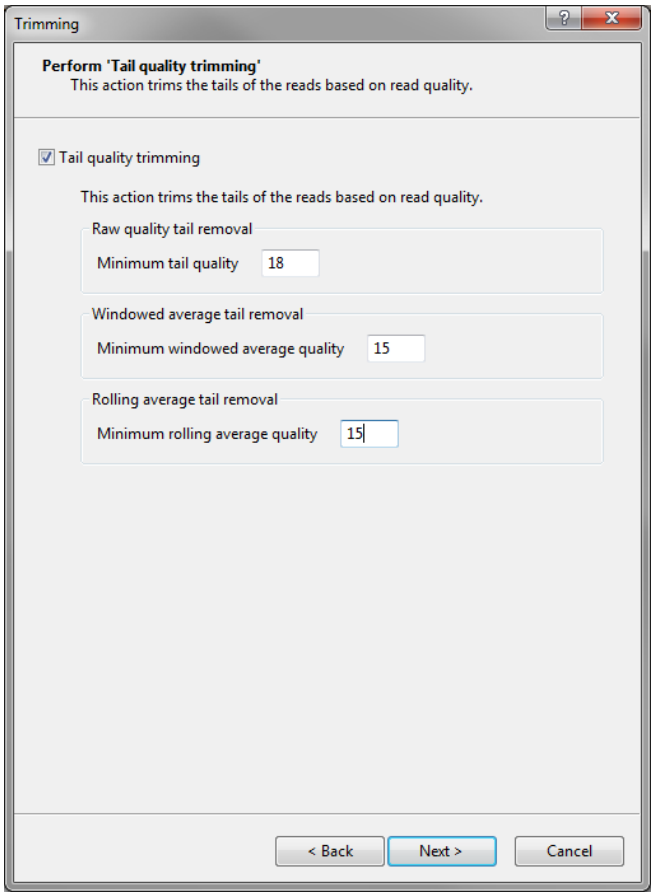

Figure 9: Tail quality trimming parameters.

Length trimming removes the reads that are too short and trims the reads that are too long.

26. Set ***Exclude reads shorter than*** to "70" bases and ***Restrict reads to at most*** to "251" bases. Click <***Next***>.

27. Change the output sequence read set experiment type to **wgs_trimmed**. Click <***Finish***>.

28. Confirm to run the analysis in a dedicated window by clicking <***Yes***>.

The *Power assembly* window opens with all the trimming actions running one by one, as shown in the *Project pipeline* panel. When the trimming is completed, we can have a look at the results:

29. For example, click on the action **Remove reads with long homopolymers** and make sure the *Report* panel is displayed.

From the homopolymer histogram pre-trimming in the results section, it is clear that few reads exist that have long homopolymers. This trimming action was ran on all 661,973 reads and 797 reads were completely removed by the action. 2,753 reads were orphaned, meaning that only one of the reads from a paired-end read pair was removed. In the homopolymer histogram created after trimming, the long tail is removed from the histogram (see Figure 10).

A number of parameters can be changed via a drag-and-drop procedure on the charts:

30. For example, highlight the action **Overall quality trimming**. In the *Action data* panel, right-click the graph ***Average read quality (before trimming)*** and select ***Show*** from the floating menu.

The graph is now displayed in the *Summary graph* panel.

31. Optionally, use ***Display*** > ***Sequence curves*** > ***Zoom to fit*** ( ) to automatically fit the graph to the window size.

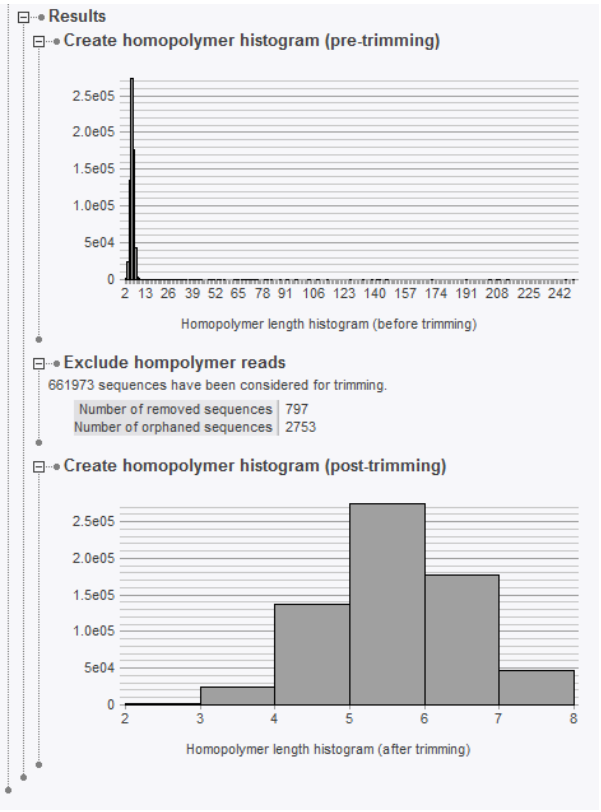

Figure 10: Results from homopolymer trimming

From this representation, the lower value threshold for the parameter is visualized by the red line and the shading. When navigating over the red line, the mouse cursor changes into a horizontal two-headed arrow, which allows to drag and drop the threshold at any value in the summary graph. From the moment a new parameter value has been changed graphically, the status of the action automatically changes to 'To be calculated'.

32. In the *Summary graph* panel, drag the red line to a position corresponding to e.g. "29". This way, a new value for the average read quality threshold is defined (see Figure 11).

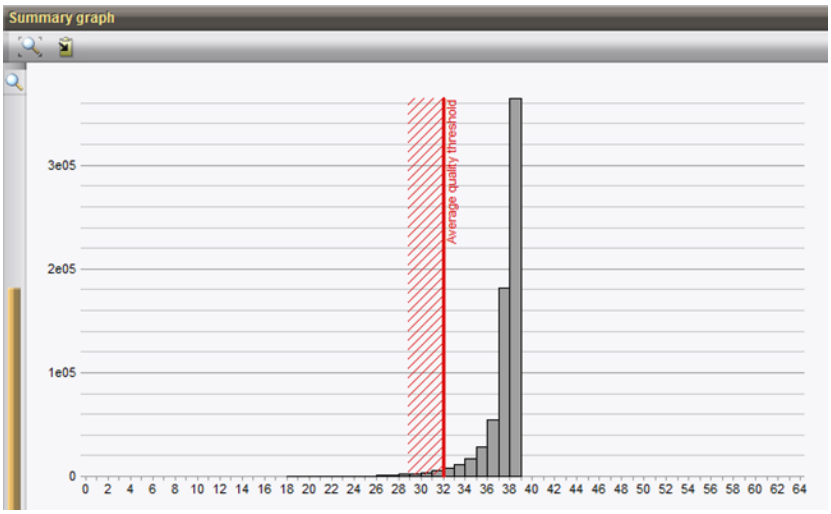

Figure 11: The 'Average read quality histogram (before trimming)' plot.

In the *Main* window, we can see that a new green dot is now available for experiment **wgs_trimmed** for

entry with key **ERR1143520**. By clicking this dot, we can have a look at the statistics for this experiment data and create charts, as described in the previous section.

4 Importing the reference sequence

In a resequencing assembly, NGS reads are assembled against a ”template”, i.e. the *reference sequence*. For this tutorial we will download the genome sequence of *Staphylococcus aureus* strain MRSA252 from EBI.

1. In the *Main* window, select ***File*** > ***Import...*** (, **Ctrl+I**) to open the *Import* dialog box.

2. Choose the option ***Download sequences from internet*** under the ***Sequence type data*** item in the tree and click <***Import***>.

3. Enter the accession code **BX571856** in the ***Accession codes*** input field and press <***Next***> (see Figure 12).

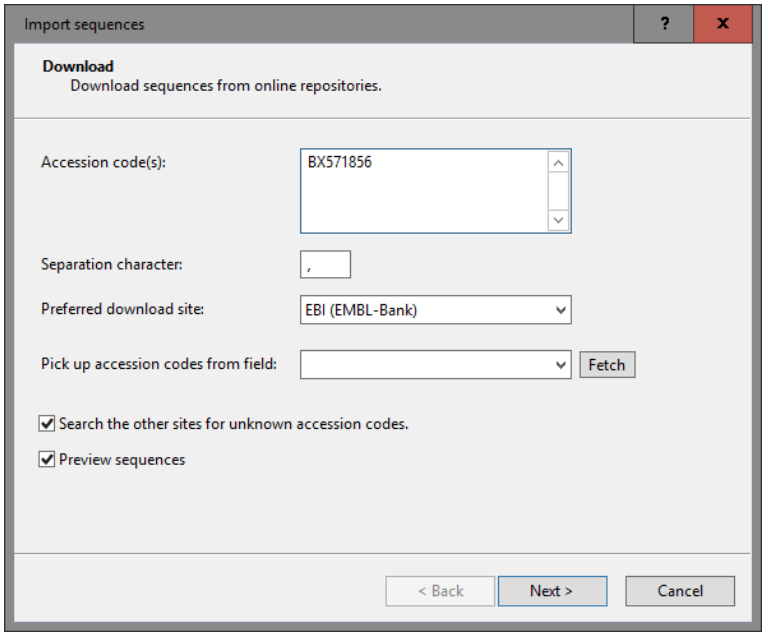

Figure 12: Download sequence from EBI.

The import routine fetches the sequence from EBI and shows detailed information in the next step (see Figure 13).

4. Press <***Next***>.

The next step of the import wizard lists the templates that are present to import sequence information in the database. As this is the first time we import a sequence from an online repository, we need to create a new import template by specifying ***Import rules***.

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

Each header tag (e.g. ID, AC, . . .) corresponds to a row in the grid panel.

6. Select ”AC - ACCESSION” in the list and click <***Edit destination***> or double-click on ”AC - ACCESSION”. Select ***Key***, and press <***OK***>.

The grid is updated (see Figure 14).

7. Click <***Next***> and press <***Finish***>.

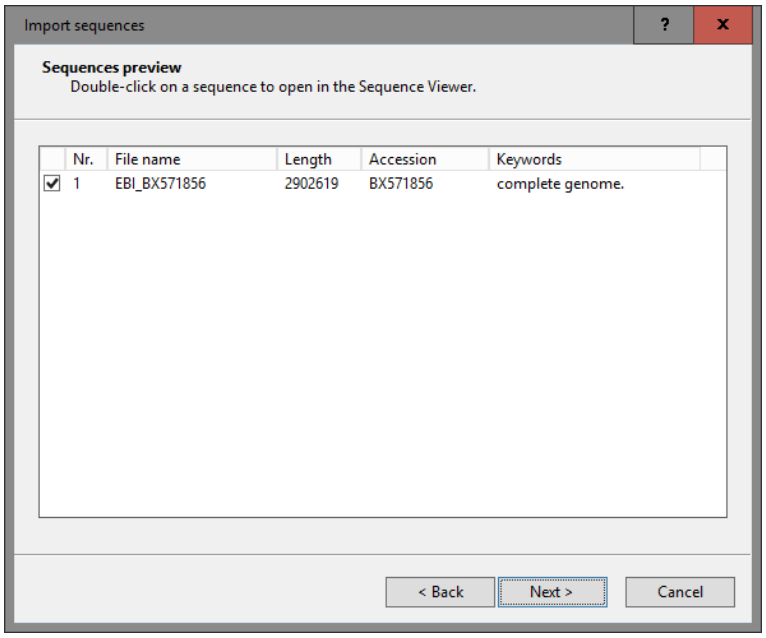

Figure 13: Fetched information.

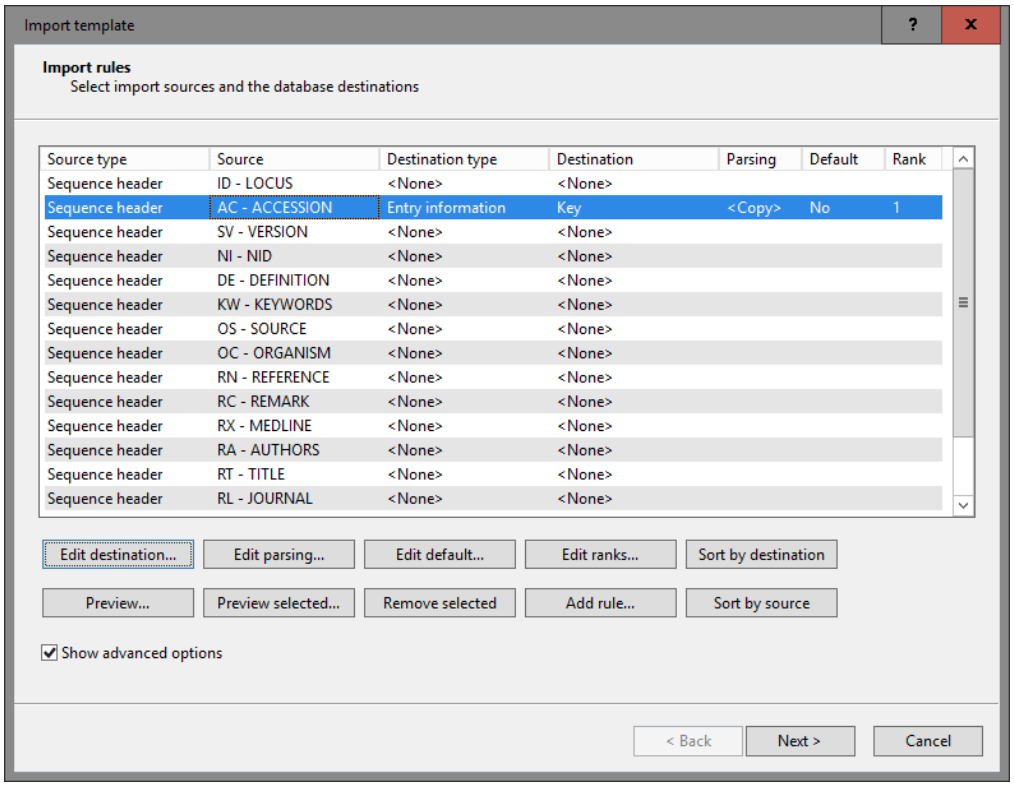

Figure 14: Import rule.

8. Specify a template name (e.g. “EBI”) and optionally enter a description. Press <***OK***>.

9. Highlight the newly created template and select ”Create new” as ***Experiment type*** (see Figure 15).

10. Press <***Next***>.

11. Specify a sequence type name (e.g. “WGS”) and press <***OK***> and confirm the action (see Figure 16).

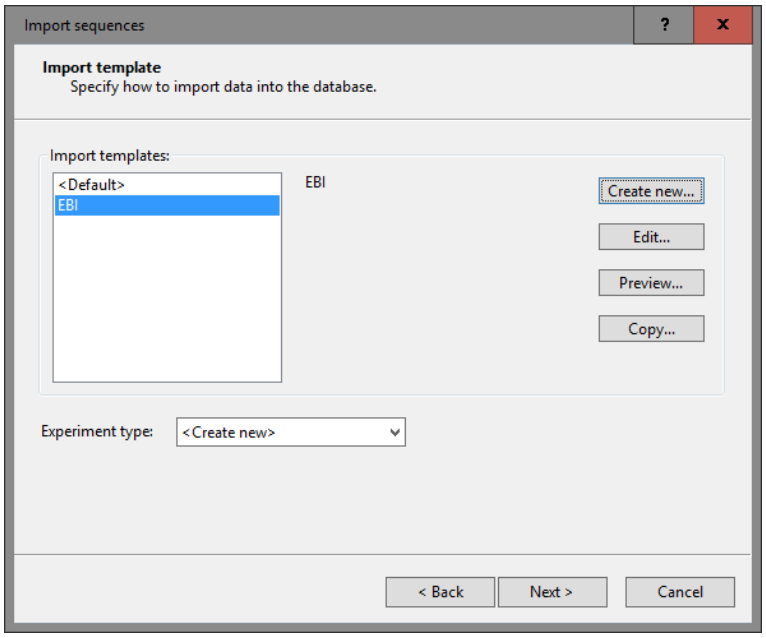

Figure 15: Import template and sequence type.

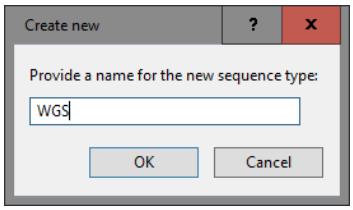

Figure 16: Create a new sequence type.

The *Database links* wizard page will indicate that 1 new entry will be created during import.

12. Press <***Finish***>.

The sequence is imported in the database and is automatically selected (see Figure 17).

13. Click on the green colored dot in the *Experiment presence* panel to open the *Sequence editor* window.

The sequence is displayed in the upper panel and a graphical representation of the sequence is displayed in the panel below. The *Annotation* panel holds the EMBL features, and the header information is stored in the *Header* panel (see Figure 18).

14. Close the *Sequence editor* window.

5 Performing a resequencing assembly

5.1 Starting a resequencing project

In a resequencing assembly, NGS reads are assembled against a ”template”, i.e. the *reference sequence*. For this tutorial we will use the genome sequence of *Staphylococcus aureus* strain MRSA252, which we have imported in our database in the previous chapter (see 4).

A resequencing analysis can be initiated from the *Sequence read set experiment* window or from the *Main*

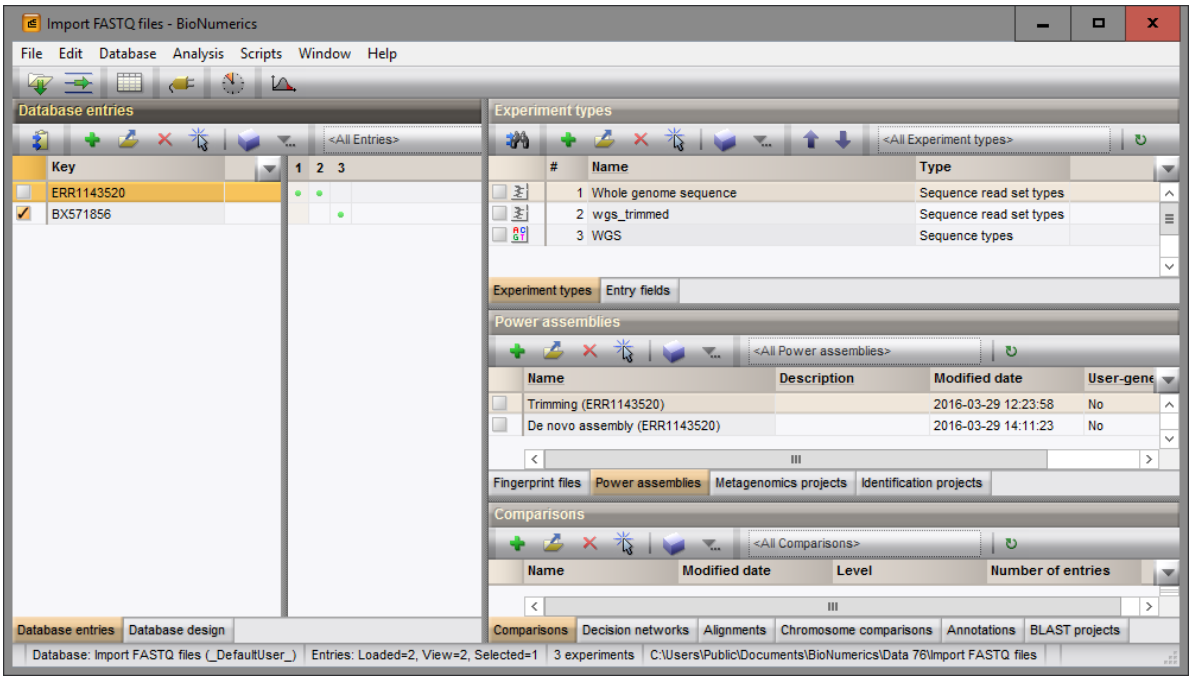

Figure 17: The *Main* window.

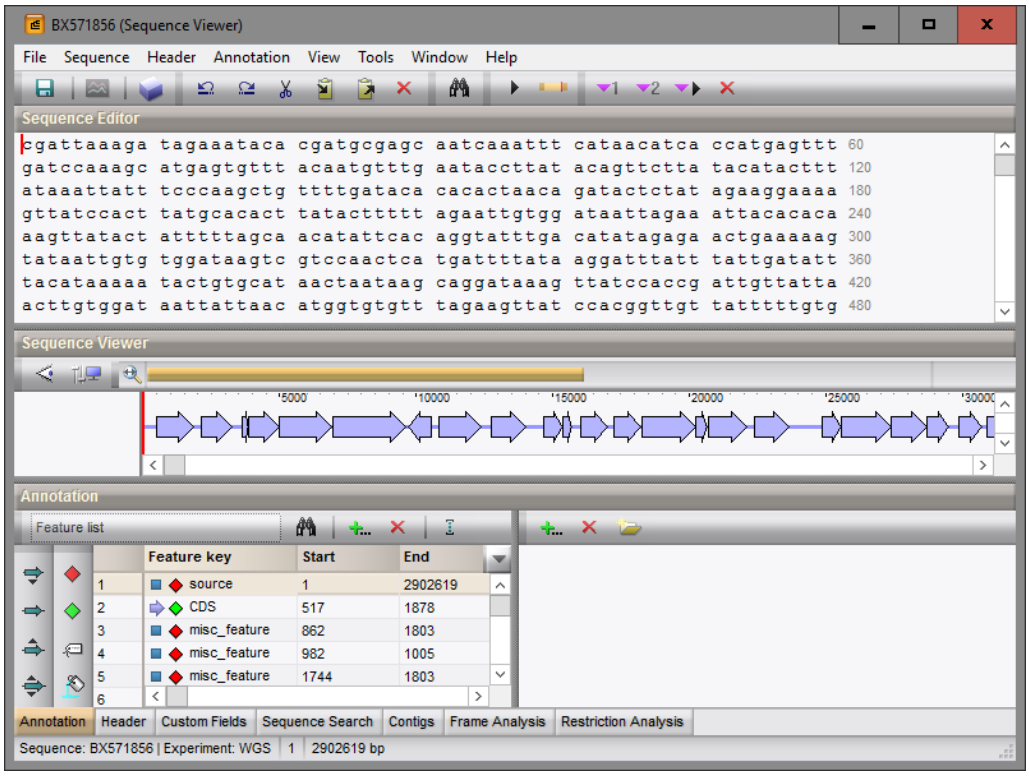

Figure 18: The *Sequence editor* window.

window, in which case a Power Assembly project is automatically created ”under the hood”. In this tutorial, we will start the analysis from the *Main* window:

1. Press the **F4**-key on your keyboard to unselect all entries in the database.

2. Check the check box next to entry **ERR1143520** in the *Database entries* panel and select ***Analysis*** > ***Sequence read set types*** > ***Resequencing assembly***.

In the wizard that appears, all relevant settings for resequencing assembly can be entered.

3. In the first wizard page, select "wgs_trimmed" for ***Input sequence read set experiment type*** and press <***Next***>.

The next page deals with the reference sequence(s):

4. Press the <***Add***> button to add one or more ***Input reference entries***.

This action displays a *Select entry* dialog box, which represents a convenient way to find one or more entries in the database.

5. Click on ***BX571856*** and press <***OK***> or hit the **Enter** key on the keyboard to select the reference entry.

6. The **WGS** sequence experiment is automatically selected (see Figure 19). Press <***Next***>.

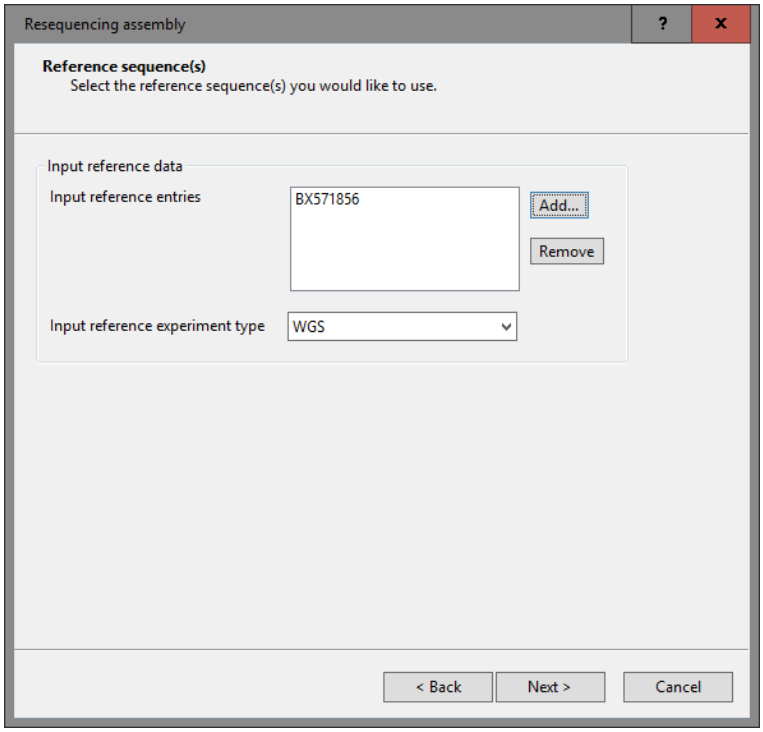

Figure 19: Select the reference sequence.

7. Check ***Create target (no inserts)*** and press <***Next***>.

8. Make sure ***Allow gaps in the reads*** and ***Allow gaps in the reference*** are selected, leave the other settings to their defaults and press <***Next***>.

9. In the next dialog, enter "2000" for ***Maximum penalty***, leave the other settings to their default values and press <***Next***>.

Since we do not know the exact insert size used, we will enter a relatively broad range for ***Enforce paired-end read constraints***:

10. Enter "2000" for ***Expected inter-read distance***, enter "2500" for ***Maximum distortion of inter-read distance*** and press <***Next***> (see Figure 20).

11. In the next two dialogs, leave all settings to their defaults and press <***Next***> twice.

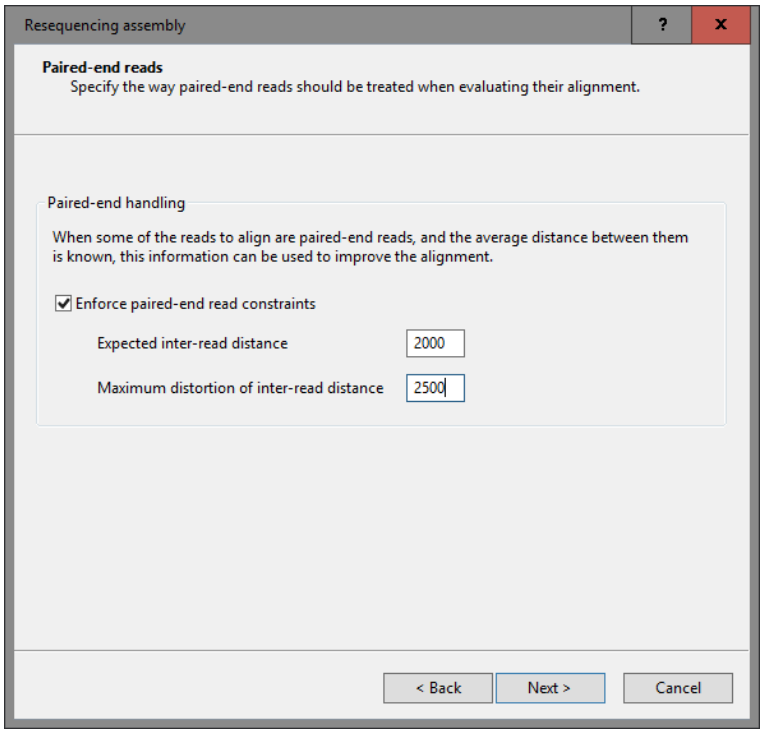

Figure 20: Paired-end reads.

12. For the ***Output sequence experiment type***, specify "WGS" and press <***Finish***>.

The question "Run analysis in dedicated window?" pops up. If you answer <***Yes***> to this question, the *Power assembly* window opens with the Power Assembly project that was created for this analysis and the project will execute automatically. If you answer <***No***>, the analysis will run in the background and no *Power assembly* window will open.

13. Choose either <***Yes***> or <***No***> to start the resequencing assembly.

Depending on your hardware specifications, the total calculation time will take a couple of minutes.

5.2 Examining the assembly

In this section, we will illustrate some of the information provided by the Power Assembly reports and graphs that were generated during the calculated resequencing assembly.

If you answered <***Yes***> to the question "Run analysis in dedicated window?", the project will already be open in the *Power assembly* window. If not, the project can be opened from the *Main* window:

14. Click on the tab of the *Power assemblies* panel to bring this panel into focus (center right of the *Main* window).

15. Highlight the last Power Assembly project in the list, called **Resequencing assembly (ERR1143520)**. Double-click on this project or select ***Edit*** > ***Open highlighted object...*** ( , **Enter**).

Next, we will examine the reports that were generated by the Power Assembly project:

16. In the *Project pipeline* panel of the *Power assembly* window, click on the action **Create target (no inserts)**. If the *Report* panel is not shown, click on the corresponding tab to display the report that corresponds to this action.

All settings used for this action are listed under **Parameters**.

Under **Results** and further **Map reads to reference**, it can be seen that around 89% of the reads were included in the assembly. From the section **Consensus base calling**, it is apparent that most of the bases were called unambiguously. This part of the report provides a first indication about the quality of the assembly.

Next, we will examine the coverage of the resequencing assembly:

17. In the *Action data* panel, highlight "Target 1" under "target coverage matrix" and select ***Action*** > ***Show...*** ( ). Click <***OK***>.

The coverage curve appears in the *Sequence curves* panel.

18. Select ***Display*** > ***Sequence curves*** > ***Zoom to fit*** ( ) to fit the coverage curve in the panel (see Figure 21).

The coverage in this assembly varies from 0 to 1250, with an average of about 88.

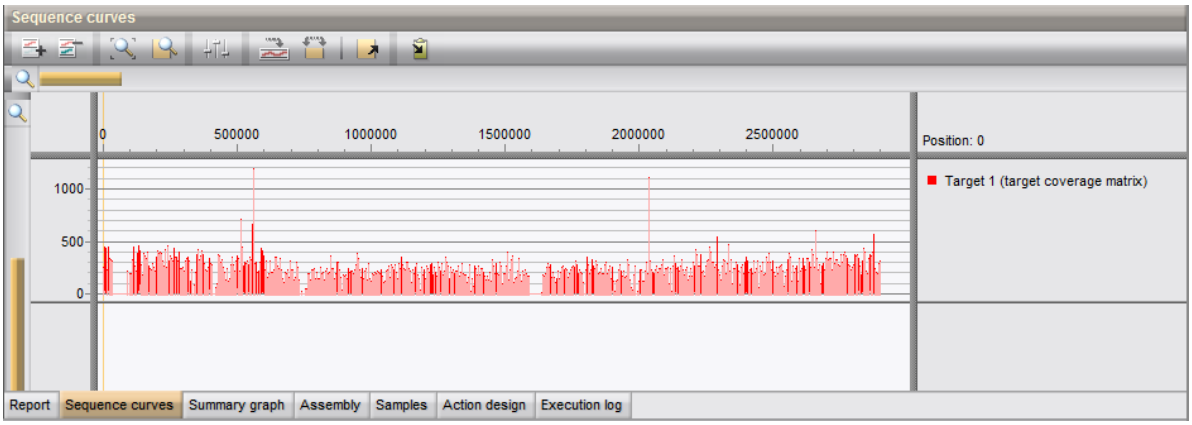

Figure 21: Sequence coverage.

The assembly, created by the **Create target (no inserts)** action, displays the mapping for each individual read onto the reference sequence. We can inspect the sequence assembly by launching the assembly map to the *Assembly* panel:

19. In the *Project pipeline* panel, highlight the **Create target (no inserts)** action by clicking it with the mouse.

20. In the *Action data* panel, under "Assemblies", highlight "Target 1" and select ***Action*** > ***Show...*** ( ). This will launch the selected assembly in the *Assembly* panel (see Figure 22).

By default, the zoom of the *Assembly* panel is set to nucleotide level. If desired, this can be changed by dragging the zoom slider on the left of the *Assembly* panel to zoom vertically (also **Ctrl+scroll**) and the zoom slider on top of the *Assembly* panel to zoom horizontally (also **Shift+scroll**).

Only the assembly of the first base pairs is shown by default, with the first nucleotide of the reference sequence as start position. However, the size and position of the *assembly viewport* (indicated with a yellow rectangle on the top ruler) can easily be altered.

21. Place the mouse pointer at the start or at the end of the orange rectangle and enlarge the assembly viewport by dragging the double arrow. Alternatively, one can also select ***Display*** > ***Assembly*** > ***Enlarge assembly viewport*** ( ) repeatedly to enlarge the assembly viewport stepwise.

22. To reposition the assembly viewport, move the mouse cursor over the yellow viewport and drag the four-headed arrow to another position. Alternatively, select ***Display*** > ***Assembly*** > ***Move assembly viewport to left*** ( ) or ***Display*** > ***Assembly*** > ***Move assembly viewport to right*** ( ).

Enlarging, shrinking or repositioning large viewports can take some time because the read information to be displayed in the panel needs to be loaded from the data set.

Figure 22: The *Assembly* panel, displaying a part of an assembly map.

Reads that are drawn at half of their normal size indicate paired-end reads from which the pairs overlap.

Furthermore, one has the option to highlight the nucleotide differences against the reference sequence or to display only the forward mapped, the reverse mapped or both forward and reverse mapped reads. For finding erroneously mapped reads, the first option is very useful, so we will enable this:

23. Select ***Display*** > ***Assembly*** > ***Show differences only*** ( ) to visualize only the nucleotide differences in the reads against the reference sequence. The nucleotide differences are highlighted in the corresponding nucleotide color.

Both forward and reverse mapped reads are by default displayed in the *Assembly* panel (buttons and are highlighted), but they can be displayed separately if desired:

24. Press the button in the toolbar to show only the reverse mapped reads.

25. Next, press the and buttons to show only the forward mapped reads.

5.3 Power Assembler templates

Any Power Assembler project can be customized and the modified pipeline can be stored as a template for later use. We can illustrate this on the resequencing pipeline:

First, we will make a modification to the pipeline.

26. In the *Project pipeline* panel, highlight the action **Determine covered regions** in the pipeline.

27. Select ***Action*** > ***Remove action*** ( ) to remove the action. Answer <***Yes***> twice to remove the associated curves and histograms as well.

28. Highlight the action **Create target (no inserts)** in the pipeline and click on the tab that corresponds to the *Action design* panel.

29. In the *Action design* panel, double-click the operator **Calculate coverage (no inserts)**.

30. Click on the *Operator parameters tab*, uncheck the option ***Allow extension of the reference sequence (begin and end)*** and press <***OK***>.

The current Power Assembler pipeline can now be saved as a template:

31. Select ***File*** > ***Store pipeline as template...***, enter e.g. “Resequencing in frame” as ***Name*** (see Figure 23) and press <***OK***>.

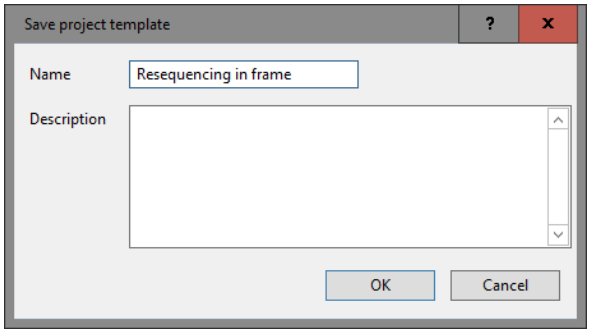

Figure 23: Save as new template.

32. Close the *Power assembly* window.

The modified pipeline actually does a resequencing that retains the frame of the reference sequence (same-length sequence is produced) and could be used to base a SNP analysis on.

Since BioNumerics version 7.6, custom templates can be run in batch on the sequence read sets from selected entries. This is achieved as follows:

33. In the *Main* window, select ***Analysis*** > ***Sequence read set types*** > ***Run custom template...***.

34. Highlight the template **Resequencing in frame** in the ”User-defined project templates” category and press the <***Settings***> button and proceed through the wizard to check the settings for this template.

Pressing <***OK***> will execute the Power Assembly template on the specified experiment type for the selected entries.