BioNumerics Tutorial:

Performing a de novo assembly locally

1 Aim

In this tutorial, we will perform a de novo assembly without using the external calculation engine.

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”.

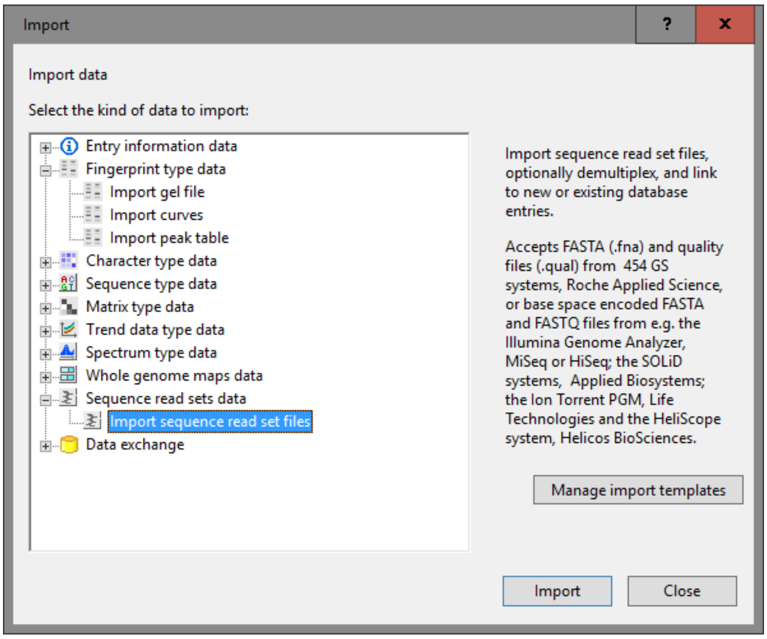
3. Under ***Sequence read sets data***, select the option ***Import sequence read set files*** (see Figure 1).

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.



Import sequence read set files, optionally demultiplex, and link to new or existing database entries.

Accepts FASTA (.fna) and quality files (.qual) from 454 GS systems, Roche Applied Science, or base space encoded FASTA and FASTQ files from e.g. the Illumina Genome Analyzer, MiSeq or HiSeq; the SOLiD systems, Applied Biosystems; the Ion Torrent PGM, Life Technologies and the HeliScope system, Helicos BioSciences.

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.

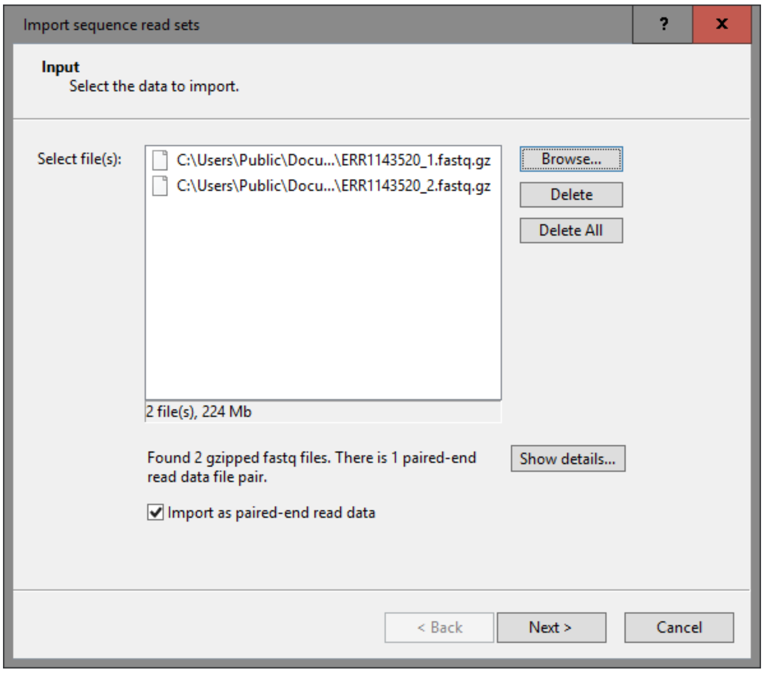
8. Select the ***Example import*** template and press the <***Preview***> button to check the outcome of the parsing. Close the preview.

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***>.

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).

© 2018 Applied Maths NV. All rights reserved.



Import sequence read sets

Input

Select the data to import.

Select file(s):

C:\Users\Public\Docu...\ERR1143520_1.fastq.gz

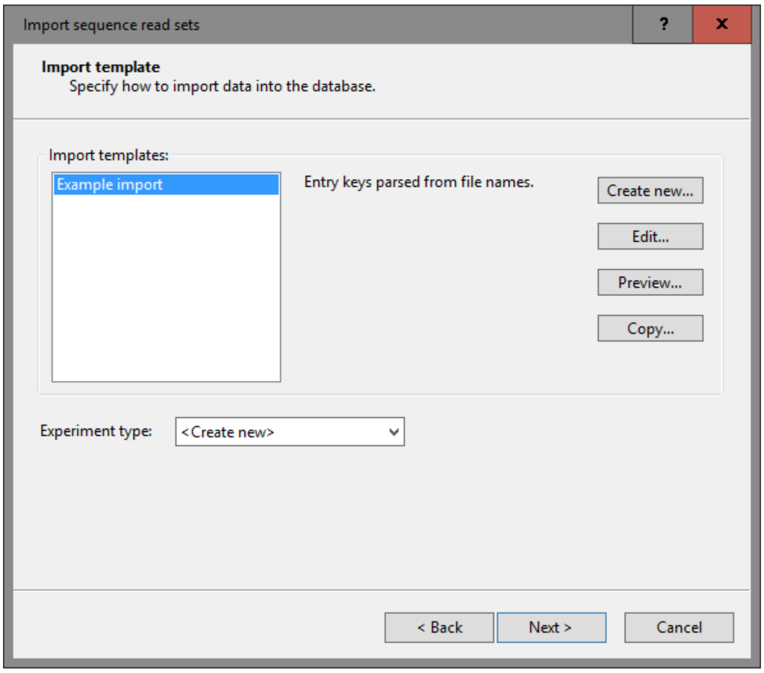
C:\Users\Public\Docu...\ERR1143520_2.fastq.gz

2 file(s), 224 Mb

Found 2 gzipped fastq files. There is 1 paired-end read data file pair.

☑ Import as paired-end read data

Figure 2: Select the FASTQ files.



Import sequence read sets

Import template

Specify how to import data into the database.

Import templates:

Example import

Entry keys parsed from file names.

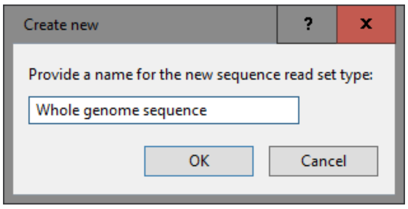
Experiment type: <Create new>

Figure 3: Select import template and experiment.

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**.

© 2018 Applied Maths NV. All rights reserved.



Provide a name for the new sequence read set type:

Whole genome sequence

Figure 4: New experiment type.

3.2 Quality assessement 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.

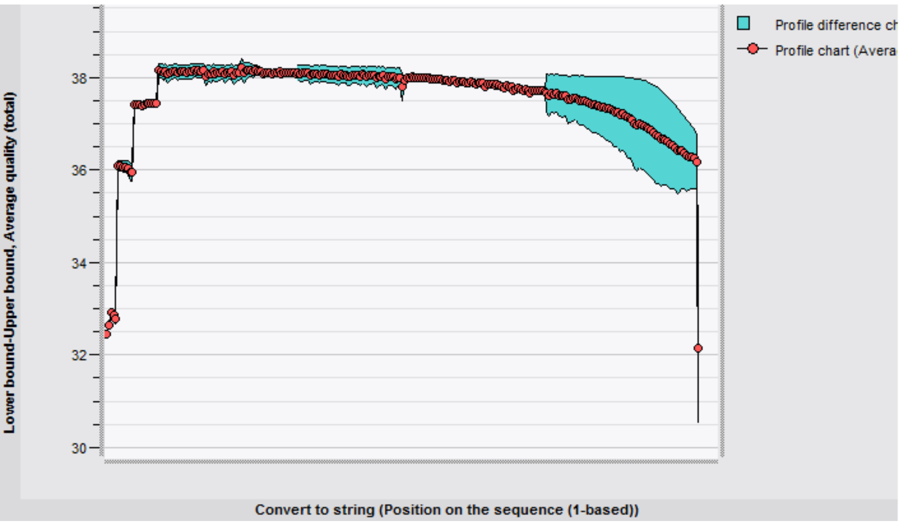


Figure 5: The chart displaying the sequence read set quality distribution (average).

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

© 2018 Applied Maths NV. All rights reserved.

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.

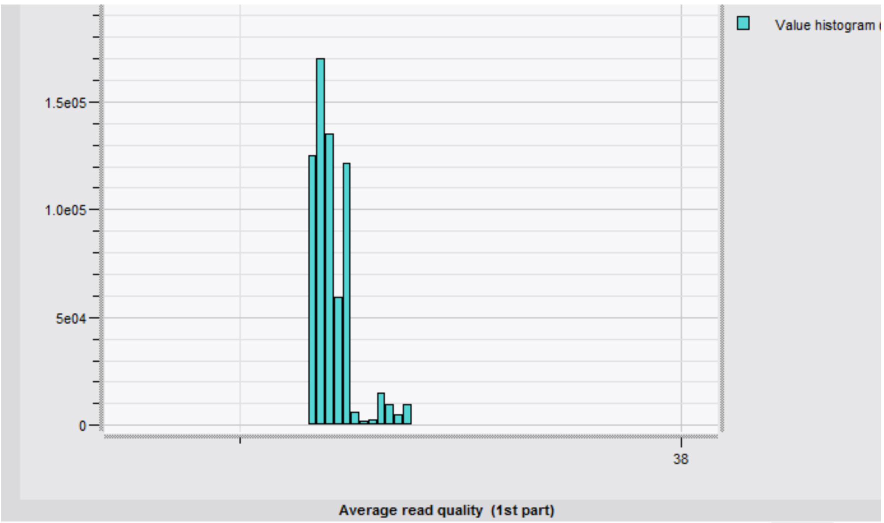


Figure 6: The 'Sequence read set average read quality distribution' plot.

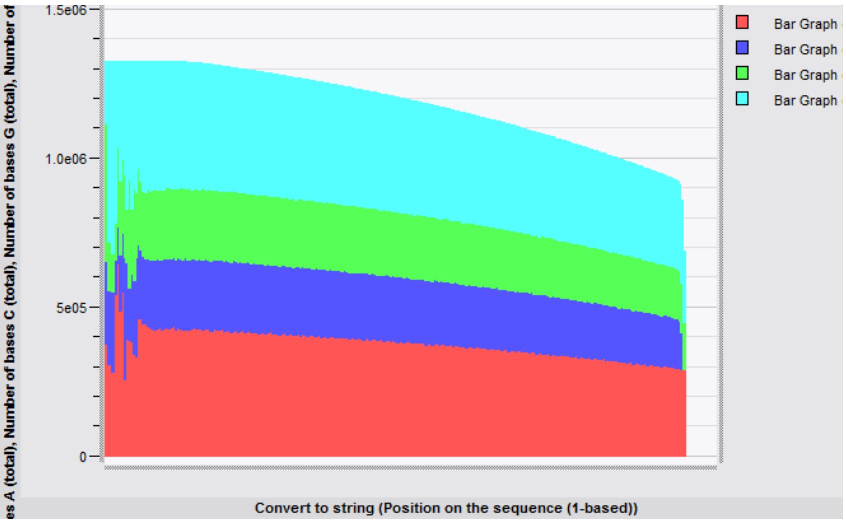


Figure 7: The 'Sequence read set base distribution' plot.

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...*** (✚).

© 2018 Applied Maths NV. All rights reserved.

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***.

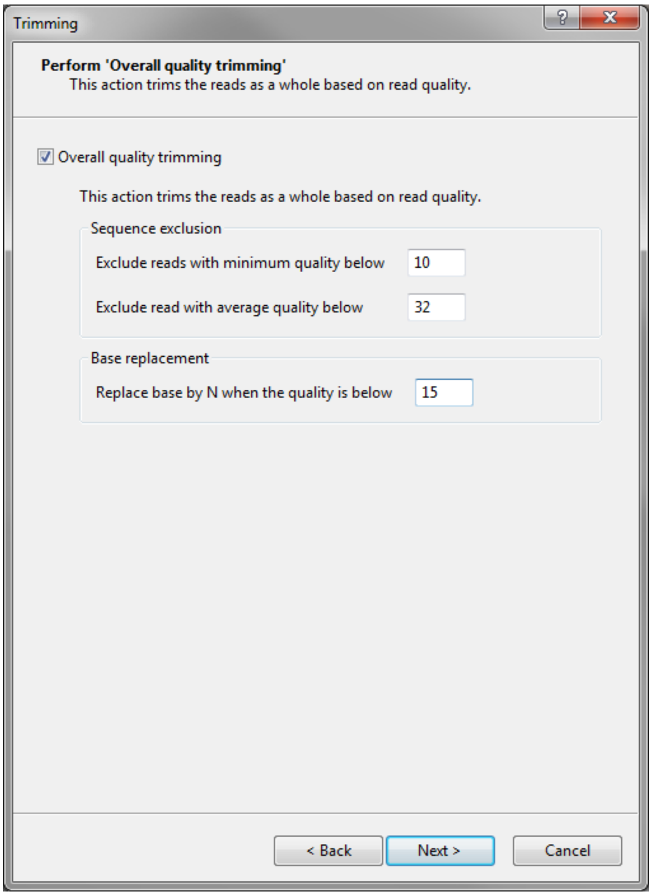
22. In the *Trimming* dialog box, leave **Whole genome sequence** as the ***Input sequence read set experiment type***, and press <***Next***>.

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***>.



Trimming

Perform 'Overall quality trimming'
This action trims the reads as a whole based on read quality.

☑ Overall quality trimming

This action trims the reads as a whole based on read quality.

Sequence exclusion

Exclude reads with minimum quality below: 10

Exclude read with average quality below: 32

Base replacement

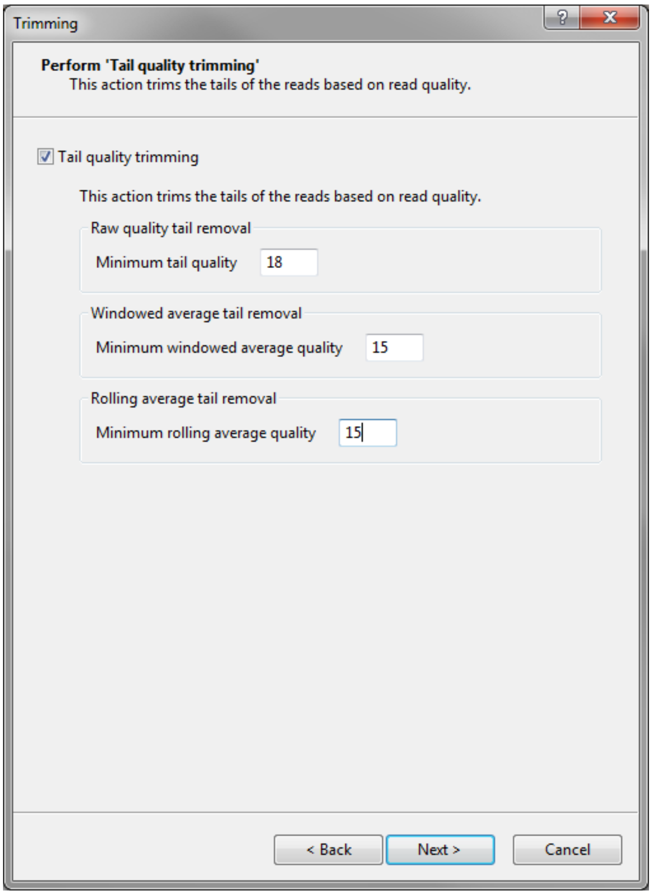
Replace base by N when the quality is below: 15

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***>.

© 2018 Applied Maths NV. All rights reserved.



Perform 'Tail quality trimming'
This action trims the tails of the reads based on read quality.

☑ Tail quality trimming

This action trims the tails of the reads based on read quality.

Raw quality tail removal
Minimum tail quality 18

Windowed average tail removal
Minimum windowed average quality 15

Rolling average tail removal
Minimum rolling average quality 15

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).

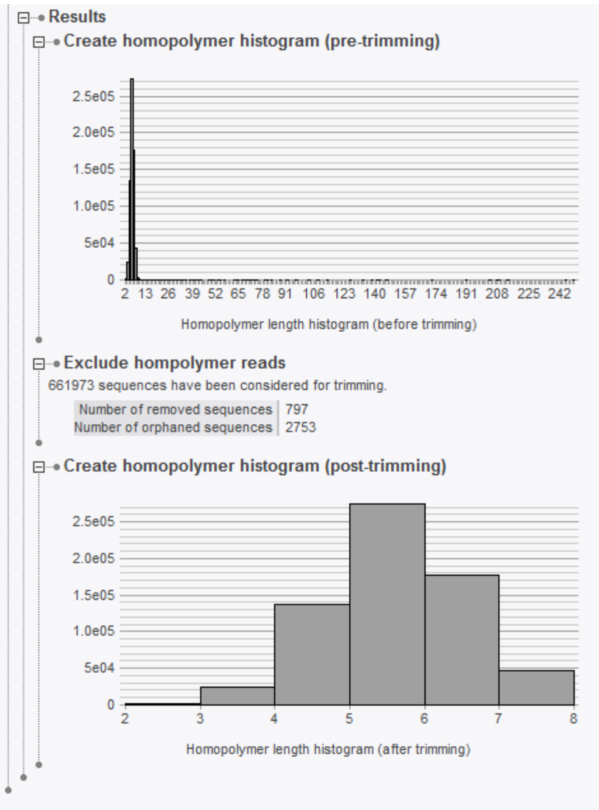
A number of parameters can be changed via a drag-and-drop procedure on the charts:

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.

© 2018 Applied Maths NV. All rights reserved.



Results

Create homopolymer histogram (pre-trimming)

Exclude hompolymer reads

661973 sequences have been considered for trimming.

Number of removed sequences	797
Number of orphaned sequences	2753

Create homopolymer histogram (post-trimming)

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).

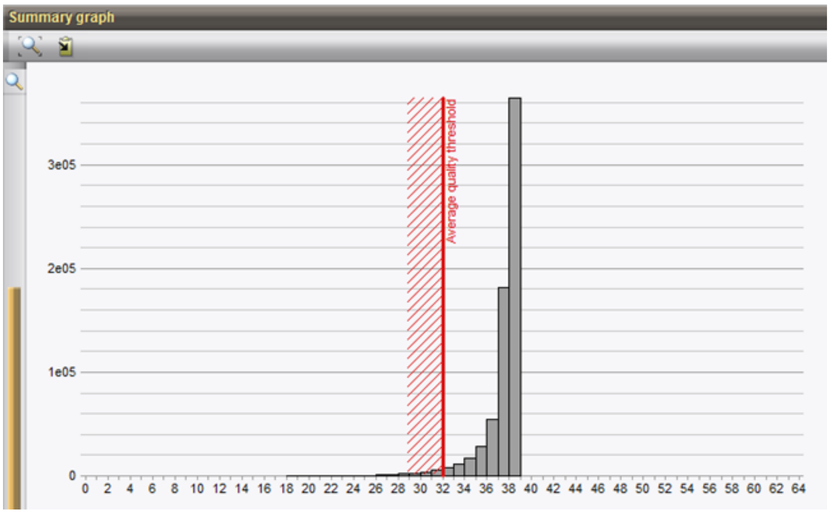


Figure 11: The ’Average read quality histogram (before trimming)’ plot.

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

© 2018 Applied Maths NV. All rights reserved.

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 Performing a de novo assembly locally

4.1 Starting a de novo assembly

If no suitable sequence experiment type is present in the database, a new one needs to be created:

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

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

3. Enter e.g. "Denovo" as ***Sequence type name*** and press <***Next***>.

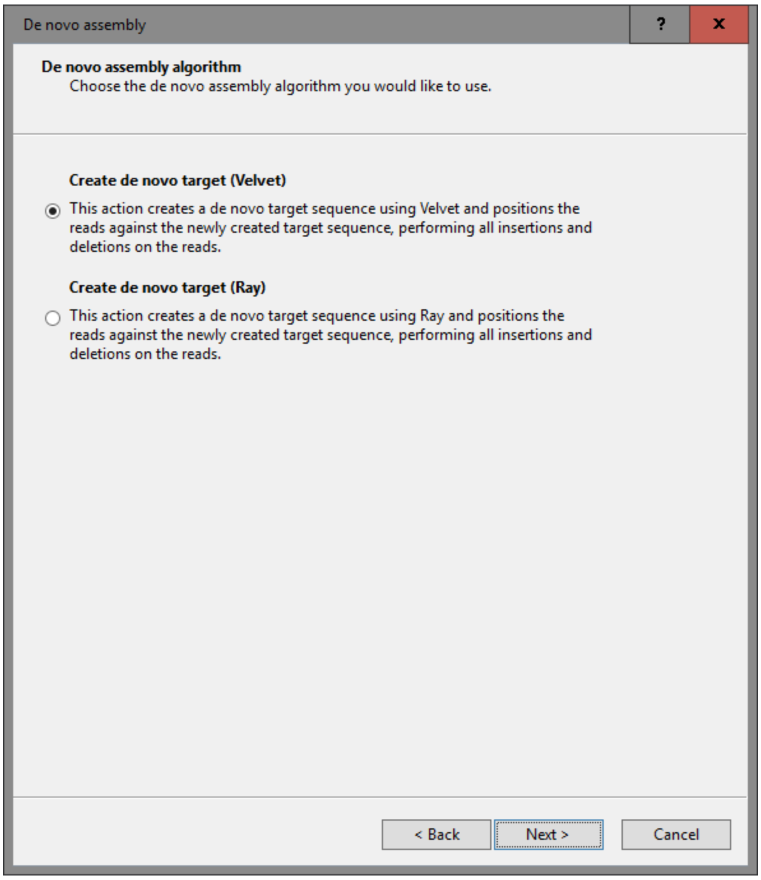
4. Leave ***Nucleic acid sequences*** selected and press <***Finish***> to create the new sequence type.

5. If not already selected, select entry **ERR1143520** in the database and use ***Analysis*** > ***Sequence read set types*** > ***De novo assembly***.

The wizard that appears allows the user to tweak several parameters for de novo genome assembly.

6. Select "wgs_trimmed" for ***Input sequence read set experiment type*** and press <***Next***>.

7. Check ***Create de novo target (Velvet)*** and press <***Next***> (see Figure 12).



De novo assembly algorithm
Choose the de novo assembly algorithm you would like to use.

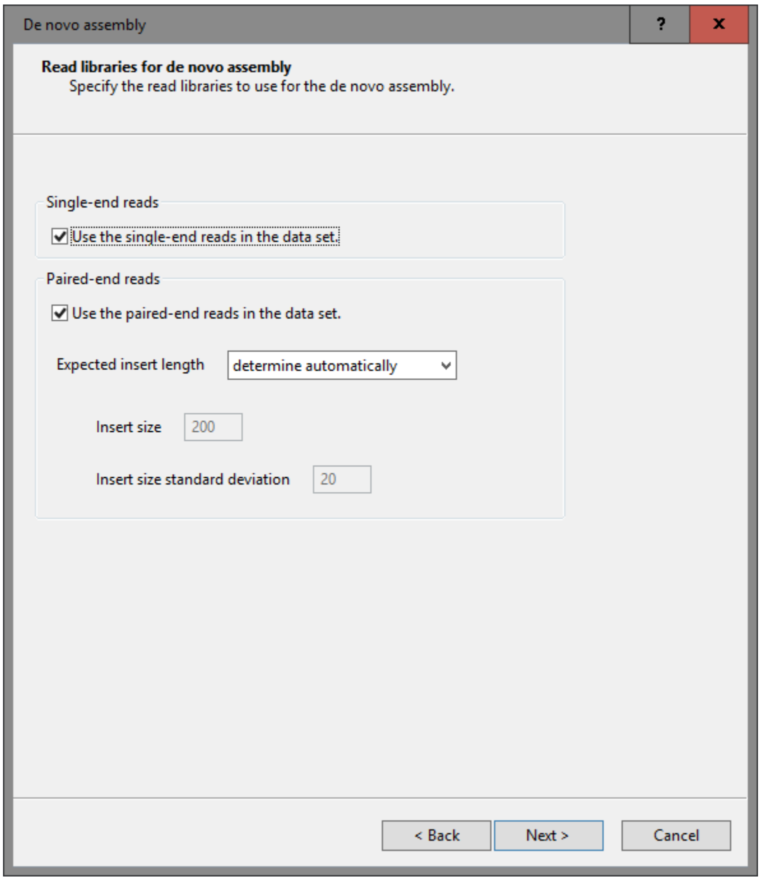
Create de novo target (Velvet)
● This action creates a de novo target sequence using Velvet and positions the reads against the newly created target sequence, performing all insertions and deletions on the reads.

Create de novo target (Ray)
○ This action creates a de novo target sequence using Ray and positions the reads against the newly created target sequence, performing all insertions and deletions on the reads.

Figure 12: De novo assembly algorithm.

© 2018 Applied Maths NV. All rights reserved.

8. Check both ***Use the single-end reads in the data set*** and ***Use the paired-end reads in the data set*** (see Figure 13). Leave the other settings to their defaults and press <***Next***>.



Read libraries for de novo assembly

Specify the read libraries to use for the de novo assembly.

Single-end reads

☑ Use the single-end reads in the data set.

Paired-end reads

☑ Use the paired-end reads in the data set.

Expected insert length: determine automatically

Insert size: 200

Insert size standard deviation: 20

Figure 13: Select read libraries.

The k-mer length is a very important parameter for the Velvet algorithm. An optimal size is around 65% of the read length.

9. Leave the default settings for ***Expected coverage*** and ***Coverage cutoff*** (determine automatically), enter a ***k-mer length*** of “131” and press <***Next***> (see Figure 14).

10. Check both ***Allow gaps in the reads*** and ***Allow gaps in the reference***. Leave the other settings to their defaults and press <***Next***>.

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

12. Check ***Enforce paired-end read constraints***, enter “2000” for ***Expected inter-read distance***, enter “2500” for ***Maximum distortion of inter-read distance*** and press <***Next***> (see Figure 15).

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

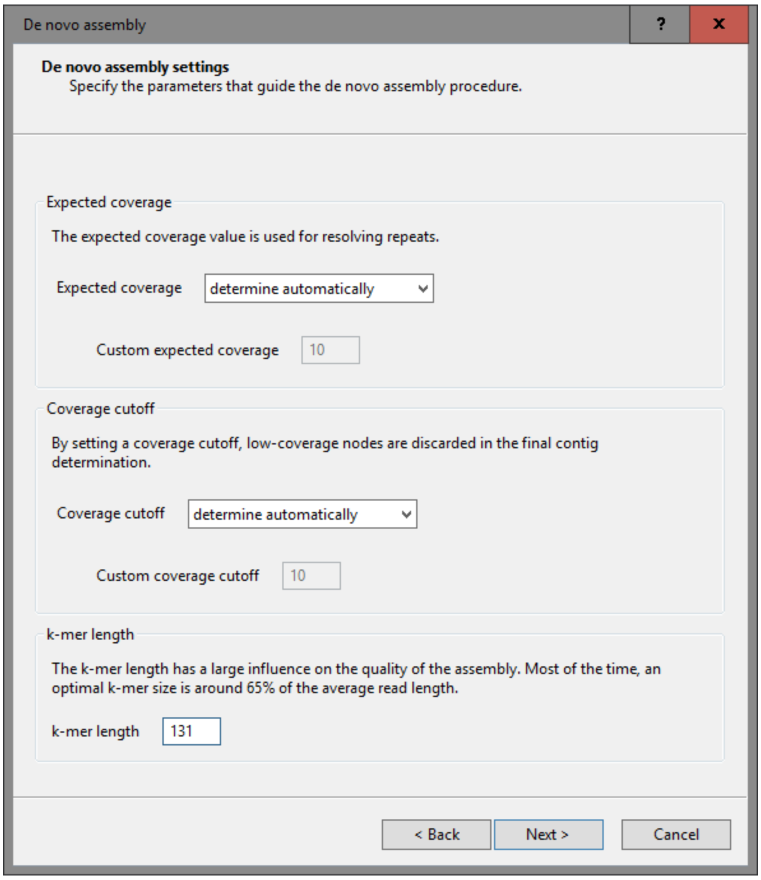
14. For the ***Output sequence experiment type***, specify “Denovo” and press <***Finish***>.

Now, the question ”Run analysis in dedicated window?” pops up.

15. Choose either <***Yes***> or <***No***> to start the de novo assembly.

The calculations should take less than ten minutes on an average desktop computer.

© 2018 Applied Maths NV. All rights reserved.



De novo assembly settings

Specify the parameters that guide the de novo assembly procedure.

Expected coverage

The expected coverage value is used for resolving repeats.

Expected coverage: determine automatically

Custom expected coverage: 10

Coverage cutoff

By setting a coverage cutoff, low-coverage nodes are discarded in the final contig determination.

Coverage cutoff: determine automatically

Custom coverage cutoff: 10

k-mer length

The k-mer length has a large influence on the quality of the assembly. Most of the time, an optimal k-mer size is around 65% of the average read length.

k-mer length: 131

Figure 14: Additional de novo settings.

4.2 Examining a de novo assembly

When the calculations are completed, we will check the result of the de novo assembly via the *Report* panel in the *Power assembly* window. 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:

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

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

18. In the *Project pipeline* panel of the *Power assembly* window, click on the action **Create de novo target (Velvet)**. If the *Report* panel is not shown, click on the corresponding tab to display the report that corresponds to this action.

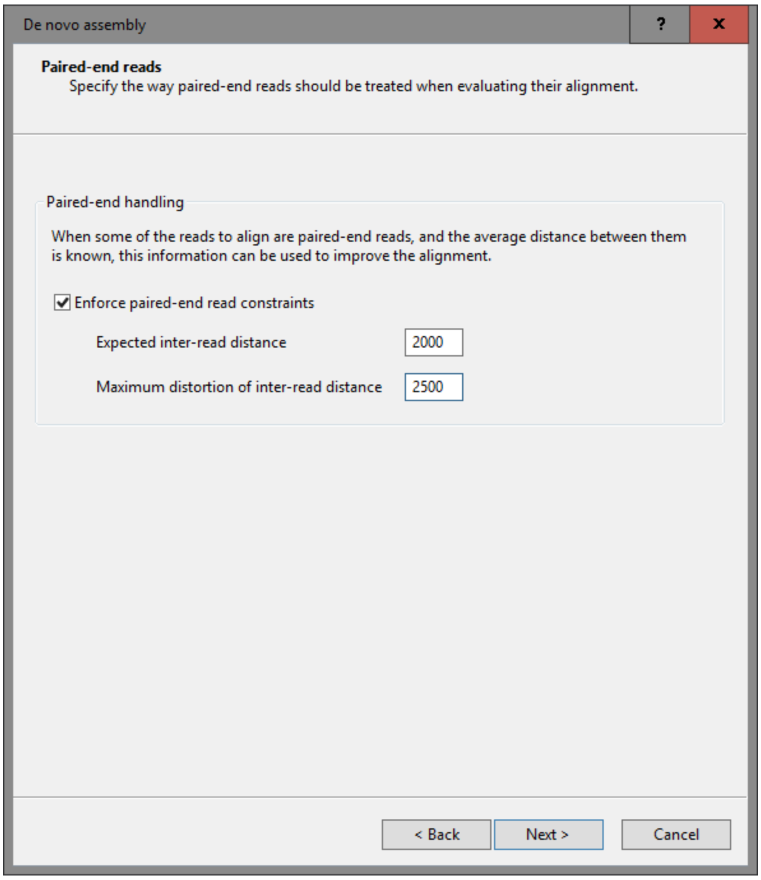
From the **Results** section in the report, it can be seen that 93% of the sequence reads were used and the large majority of bases in the consensus sequences were called unambiguously. However, a large number of contigs were produced, meaning that this particular de novo assembly was actually not optimal.

We can inspect the sequence assembly by launching the assembly map to the *Assembly* panel:

19. In the *Project pipeline* panel, highlight the **Create de novo target (Velvet)** 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 16).

© 2018 Applied Maths NV. All rights reserved.



Paired-end reads
Specify the way paired-end reads should be treated when evaluating their alignment.

Paired-end handling

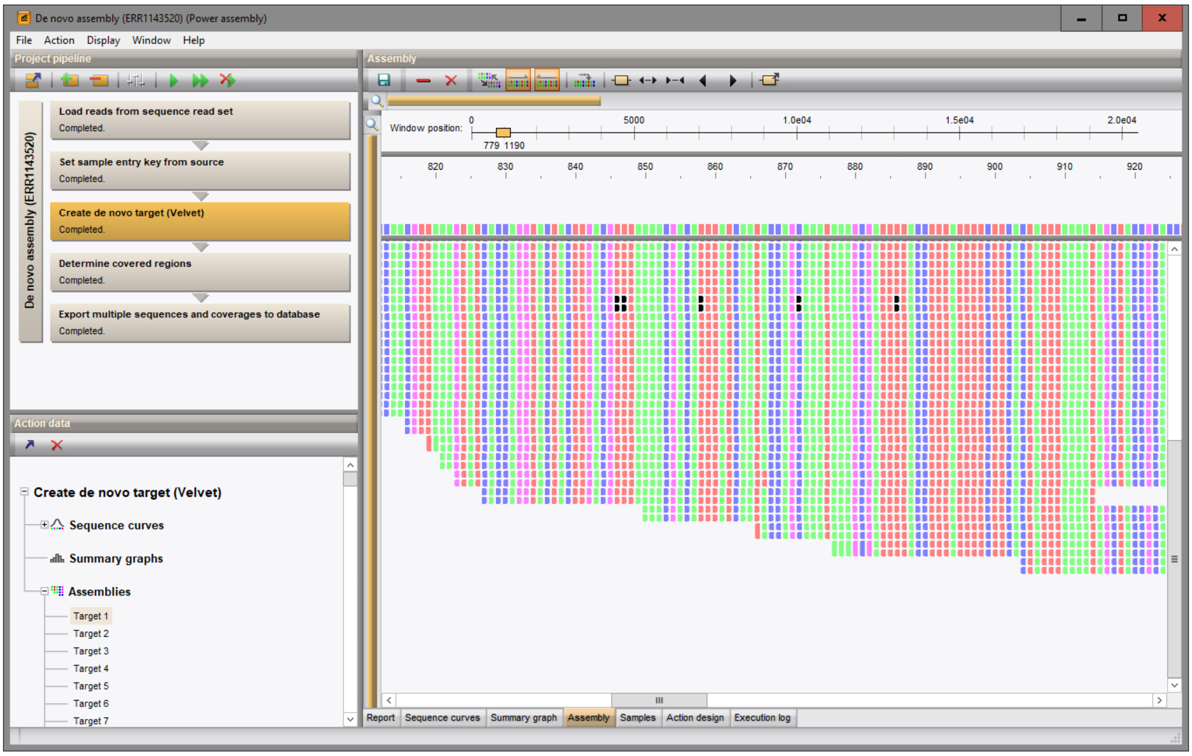
When some of the reads to align are paired-end reads, and the average distance between them is known, this information can be used to improve the alignment.

☑ Enforce paired-end read constraints

Expected inter-read distance: 2000

Maximum distortion of inter-read distance: 2500

Figure 15: Paired-end reads.



Load reads from sequence read set
Completed.

Set sample entry key from source
Completed.

Create de novo target (Velvet)
Completed.

Determine covered regions
Completed.

Export multiple sequences and coverages to database
Completed.

Create de novo target (Velvet)
- Sequence curves
- Summary graphs
- Assemblies
  - Target 1
  - Target 2
  - Target 3
  - Target 4
  - Target 5
  - Target 6
  - Target 7

Figure 16: Part of Assembly map.

© 2018 Applied Maths NV. All rights reserved.

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 de novo 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.

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 de novo 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 de novo 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.

© 2018 Applied Maths NV. All rights reserved.

© 2018 Applied Maths NV. All rights reserved.

Bibliography