



BIONUMERIC Tutorial:

Performing a long reads only de novo assembly on the cloud calculation engine

1 Aim

In this tutorial, we will perform a *long reads only* de novo assembly on the Cloud Calculation Engine.



A long reads only de novo assembly can only be performed on the Cloud Calculation Engine and not on your own computer due to the large amount of required resources.

2 Preparing the demo database

2.1 Option 1: Download demo database from the Startup Screen

1. Click the  button, located in the toolbar in the *BIONUMERIC Startup* window.

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

2. Select **WGS_demo_database_for_Escherichia_coli** from the list and select **Database** > **Download** (.
3. Confirm the installation of the database and press <OK> after successful installation of the database.
4. Close the *Tutorial databases* window with **File** > **Exit**.

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

5. Double-click the **WGS_demo_database_for_Escherichia_coli** in the *BIONUMERIC Startup* window to open the database.

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

A BIONUMERIC 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 BIONUMERIC.

6. Download the file `WGS_EC.bnbk` file from <https://www.bionumerics.com/download/sample-data>, under 'WGS_demo_database_for_Escherichia_coli'.

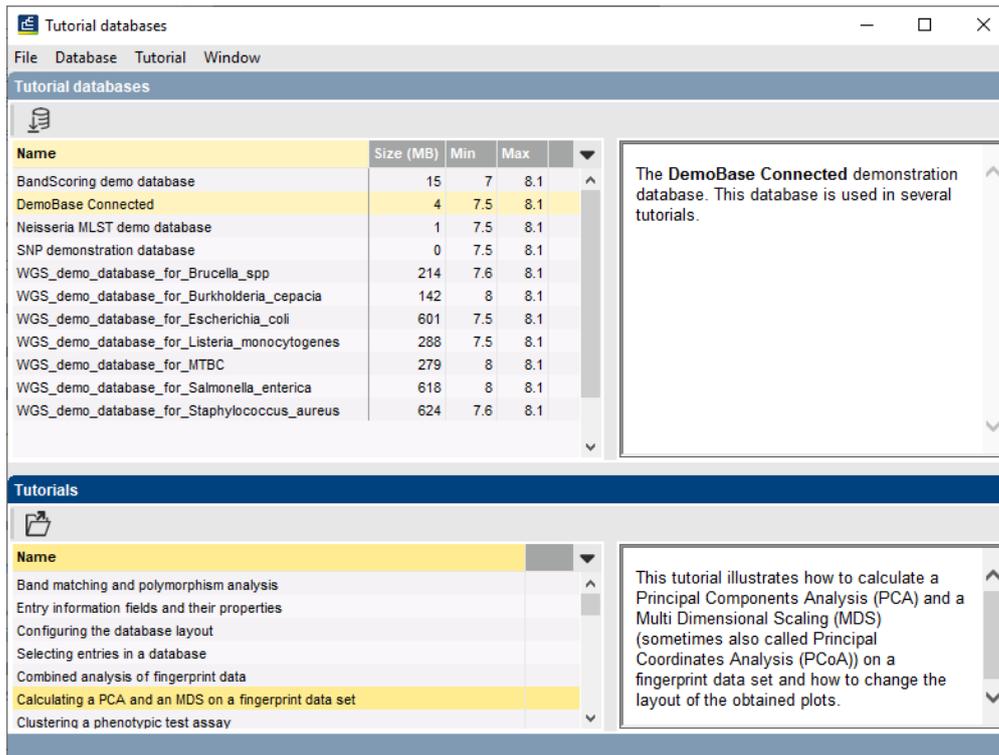


Figure 1: The *Tutorial databases* window, used to download the demonstration database.



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.

7. In the *BIONUMERICIS Startup* window, press the  button. From the menu that appears, select **Restore database....**
8. Browse for the downloaded file and select **Create copy**. Note that, if **Overwrite** is selected, an existing database will be overwritten.
9. Specify a new name for this demonstration database, e.g. "WGS_Escherichia_coli_demobase".
10. Click **<OK>** to start restoring the database from the backup file.
11. Once the process is complete, click **<Yes>** to open the database.

The *Main* window is displayed (see Figure 2).

3 Importing sequence read sets

You will be importing sequence read set data of *Escherichia coli* strain 397404 from an Oxford Nanopore minION sequencing run performed by Public Health England. First, you will create a

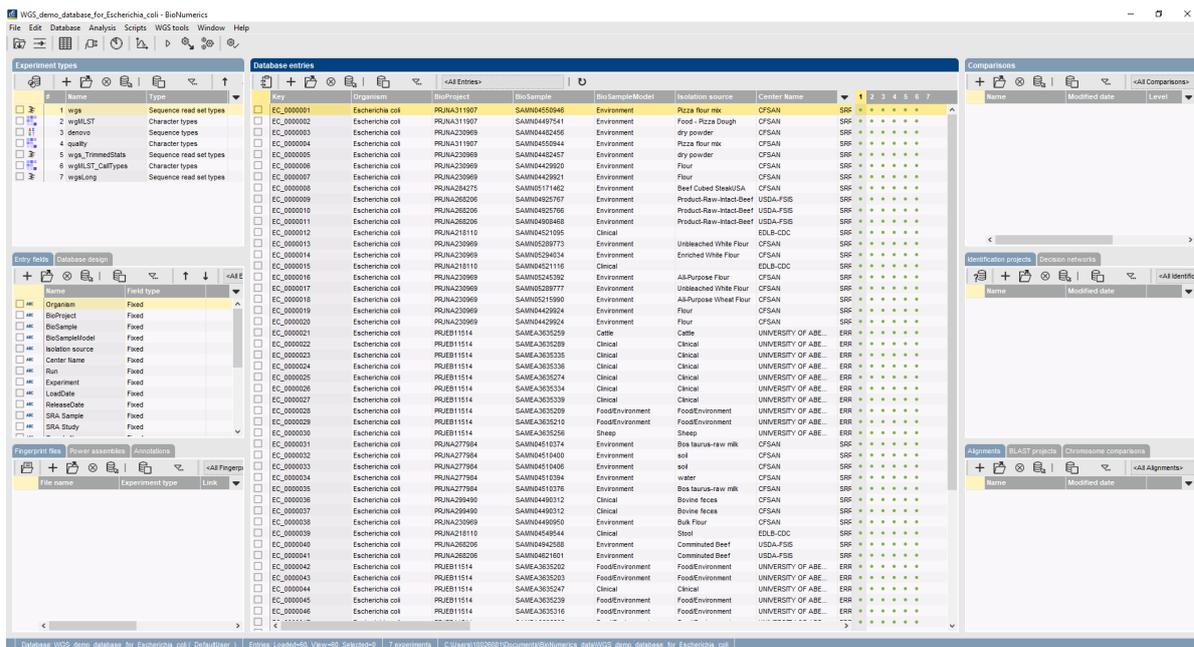


Figure 2: The *Escherichia coli* demonstration database: the *Main* window.

new entry in the database to which you will link the long read sequencing data.

1. Open the **WGS_demo_database_for_Escherichia_coli** database in which the *WGS tools* plugin installed.
2. Create a new entry in the database by clicking **+** in the *Database entries* panel of the *Main* window.
3. Type “397404” in the **Database key** text field and click **<OK>**.

A new entry with key **397404** will be created in the database. For this entry you will need to specify the NCBI SRA accession code of the long reads sequencing run. This accession code will later be used to link the data from the NCBI SRA repository to the entry.

4. Type “SRR6001344” in the entry field **Run** for entry **397404**.

You will now import the sequence read set data and link it to the newly created entry.

5. Make sure entry **397404** is selected by clicking the ballot box in front of the entry’s **Key**.
6. Select **File > Import...** (, **Ctrl+I**) to open the *Import data* wizard.
7. Highlight the **Import sequence read set data as links** option and press **<Next>**.

Links to multiple data sources are available, including online and offline data repositories such as: **NCBI (SRA)**, **EMBL-EBI (ENA)**, **Amazon (S3)**, **BaseSpace**, **Alibaba OSS** or **Local file server** (see Figure 3). Depending on the choice of import, different parameters may be queried in the next steps.

In this tutorial, the import of FASTQ files as links from NCBI is covered. For more information about the other options, please consult the *WGS tools plugin* manual.

8. Select **Import data link** under **NCBI** as download site and press **<Next>** (see Figure 3).
9. In the **Pick up accession codes from field** drop-down menu select the entry field **Run** and press **<Fetch>**.

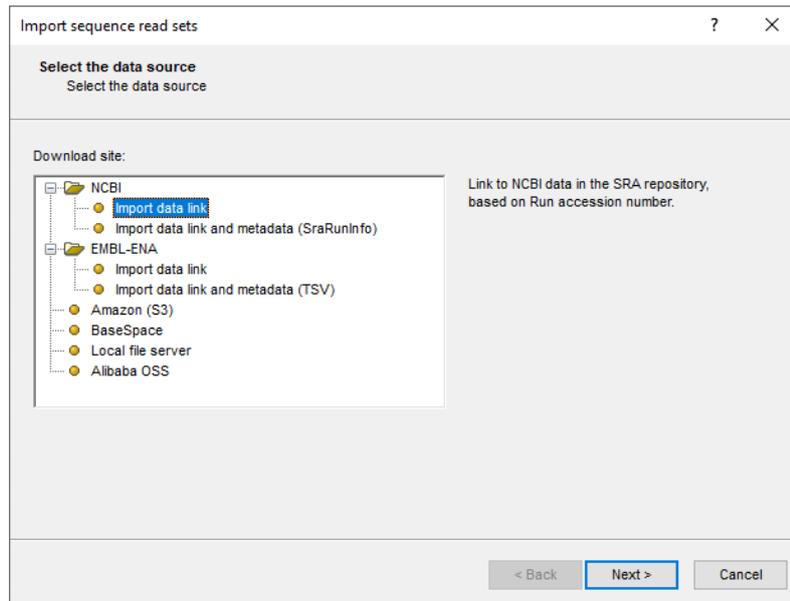


Figure 3: Data sources.

The accession number from the entry field **Run** will now be added to the **Accession code(s)** field.

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

Now you need to define how the data should be stored in the database. The dialog box allows you to set import rules. For each import source (in this case we have only the accession code), a database destination can be specified.

You will link the accession code to the **Run** field.

11. Double-click the first row in the grid, and click the small plus ('+') sign in front of **Entry info field** and select **Run** from the list (see Figure 4).

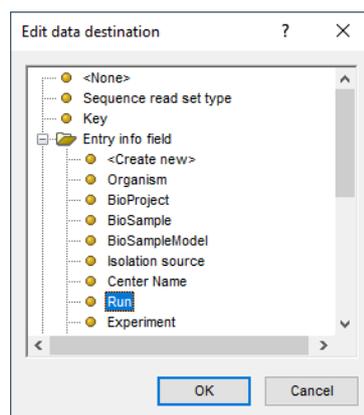


Figure 4: The **import rules** grid after assigning the file name as **Key**.

12. Press **<OK>**.

The import rules are updated in the grid (see Figure 5).

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

14. Press the **<Close>** button to close the preview and press **<Next>** to proceed to the *Import links* dialog box.

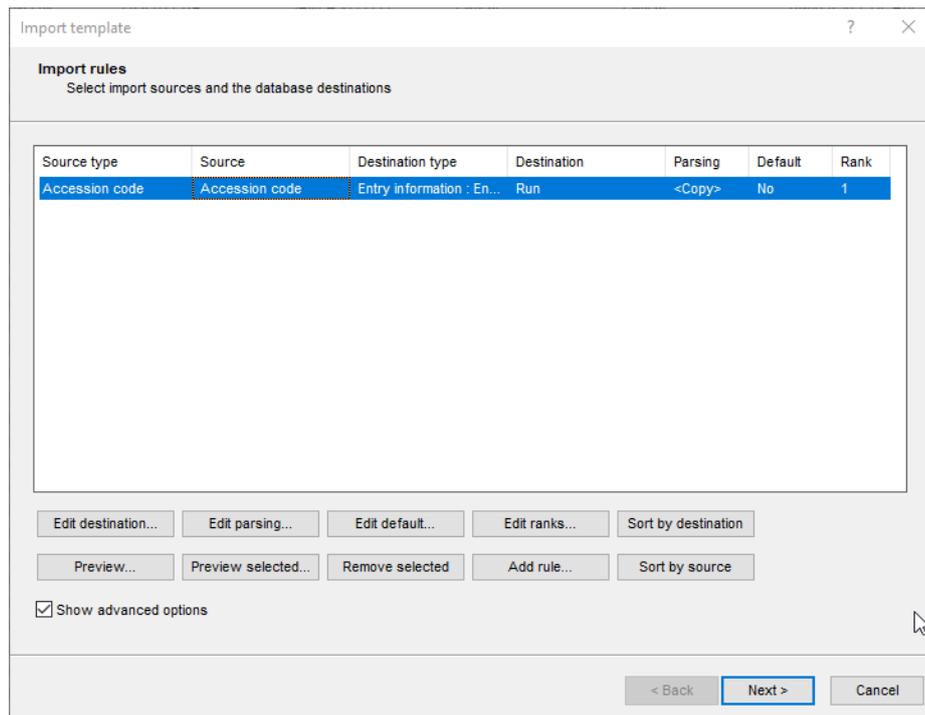


Figure 5: The **import rules** grid after assigning the file name as **Key**.

15. Make sure **Run** is checked as link and press <**Finish**>.

The import template needs to be saved to be able to use it again later on.

16. Enter a **Name** for the import template (e.g. "import reads") and optionally a **Description**. Next, press <**OK**>.

17. In the *Import template* dialog box, highlight the newly created template.

18. Make sure the **wgsLong** experiment is selected and click <**Next**>.

19. Press <**Next**> once more.

In the last step, calculation jobs (e.g. de novo assembly) can be launched on the imported data links (**Open submit jobs dialog after import**). Note that the same dialog can be called from the *Main* window at any time with **WGS tools** > **Submit jobs...** (▶).

20. Press <**Finish**> to start the import of the data links.

Once the import is completed, the entry **397404** is created/updated and has one green dot next to it in the column of the sequence read set experiment type **wgsLong**.

21. Click on the green colored dot of the imported entry corresponding to the experiment type **wgsLong**.

The data links are displayed in the *Sequence read set experiment* window (see Figure 6).

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

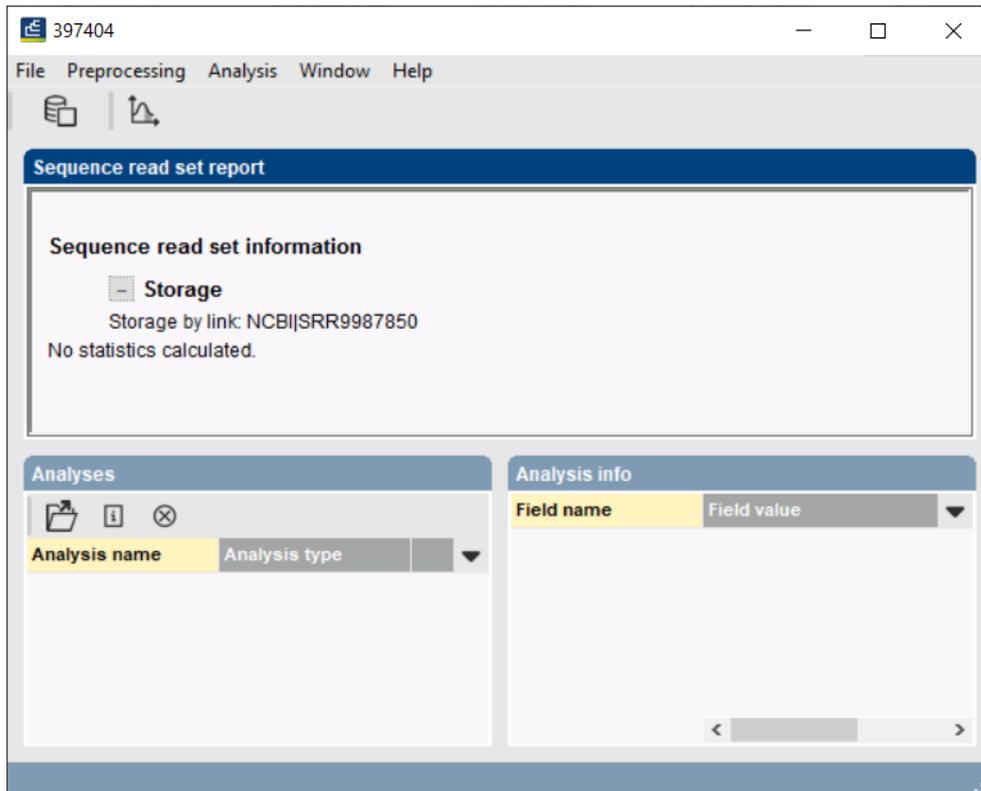


Figure 6: The data links are displayed in the *Sequence read set experiment* window.

4 Performing a long reads only de novo assembly in the cloud

Launching the long reads only de novo assembly job on the Cloud Calculation Engine is a very easy process.

1. In the *Main* window, select the entries that you want to analyze using the check-boxes next to the entries or with the **Ctrl**- or **Shift**-keys. In this example, make sure entry **397404** is selected.
2. Select **WGS tools** > **Submit jobs...** () to call the *Submit jobs* dialog box (see Figure 7).



Alternatively check the **Open submit jobs dialog after import** option in the *Processing* wizard page during import of the data.

3. To run the de novo assembly job on the Cloud Calculation Engine, check the **Calculation Engine** option.
4. If you are only interested in performing a de novo assembly based on the reads obtained after trimming (automated trimming step), check the **De novo assembly** option and uncheck all other options.
5. With the **De novo assembly** option highlighted, press the < **Settings** > button.

Following de novo assemblers are available on the Cloud Calculation Engine: **Velvet (Optimizer)**, **SPAdes** (default), **SKESA** and **Unicycler** (see Figure 8). A long reads only novo assembly can only be performed by **Unicycler**.

6. Select **Unicycler** from the **Assembler** drop-down menu (see Figure 9).
7. Check the box before **Only long reads** (see Figure 9).

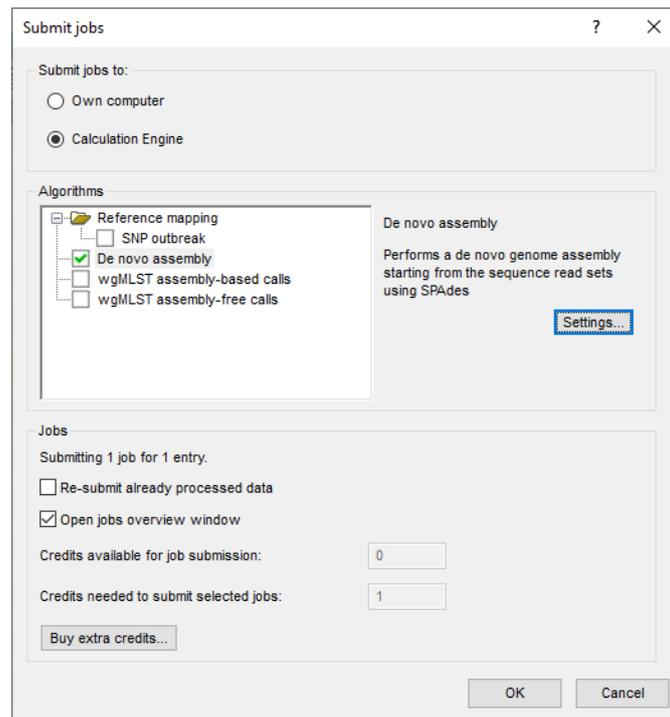


Figure 7: Submit de novo assembly job on the Cloud Calculation Engine.

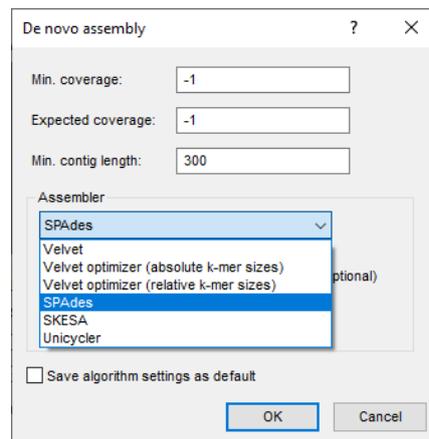


Figure 8: Cloud Calculation Engine: available de novo assemblers.

8. Close the *Perform de novo assembly* dialog box by pressing **<OK>**.

Jobs that already have been submitted and have been imported successfully, will not be re-launched for analysis, unless the check box in front of ***Re-submit already processed data*** in the **Jobs** part is checked.

Credit costs depend on the job that is submitted: 5 credits are counted for 1 long reads only de novo assembly.

9. Press **<OK>** to launch the job on the Cloud Calculation Engine.

When not sufficient credits are available for the submission of the job(s) to the external calculation engine, an error message pops up. Since no credits are assigned to the demo project, this error message will pop up when following this workflow in the demonstration database. Please consult Applied Maths for more information about the purchase of credits.

Figure 9: Settings for the Unicycler assembly algorithm.

When sufficient credits are available for the submission of the job(s) to the external Cloud Calculation Engine, the Cloud Calculation Engine will start the job by retrieving the sequencing read set data directly from the NCBI SRA repository.

10. By default, the *Job overview* window will open after submission of the job(s). The same dialog can be called at any time with **WGS tools > Jobs overview...** (⚙️).

The *Entry* key, the *Submitted time*, the job *Status*, a *Description* of the job and its *Progress* and much more can be monitored. In the *Message* field, the run comments are displayed in real time.

On average, the calculation time for a long reads only de novo assembly on the Cloud Calculation Engine is around **60 min**.

11. To refresh the overview, press **View > Refresh** (🔄, F5).
12. Finished jobs can be imported with a manual action (**Jobs > Get results** (🔧)) or through an automatic update: select **File > Settings**, check both options and specify an interval (e.g. 10 min).

Once the results are imported, the corresponding jobs and their underlying data sets are automatically deleted from the calculation engine and as such, from the *Job overview* window.

The results from the de novo assembly algorithm, i.e. concatenated de novo contig sequences with coverage information are stored in the sequence experiment type **denovo**.

13. Click on the green colored dot for the entry **397404** 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 10).

The de novo assembly can be used for bacterial typing using whole genome Multi Locus Sequence Typing (wgMLST). It can also be used as reference genome in a whole genome Single Nucleotide Polymorphism (wgSNP) analysis, or it can be compared to other whole genome assemblies using minHashing. For all these analyses, please refer to the corresponding tutorials on our website <https://www.bionumerics.com/tutorials>.

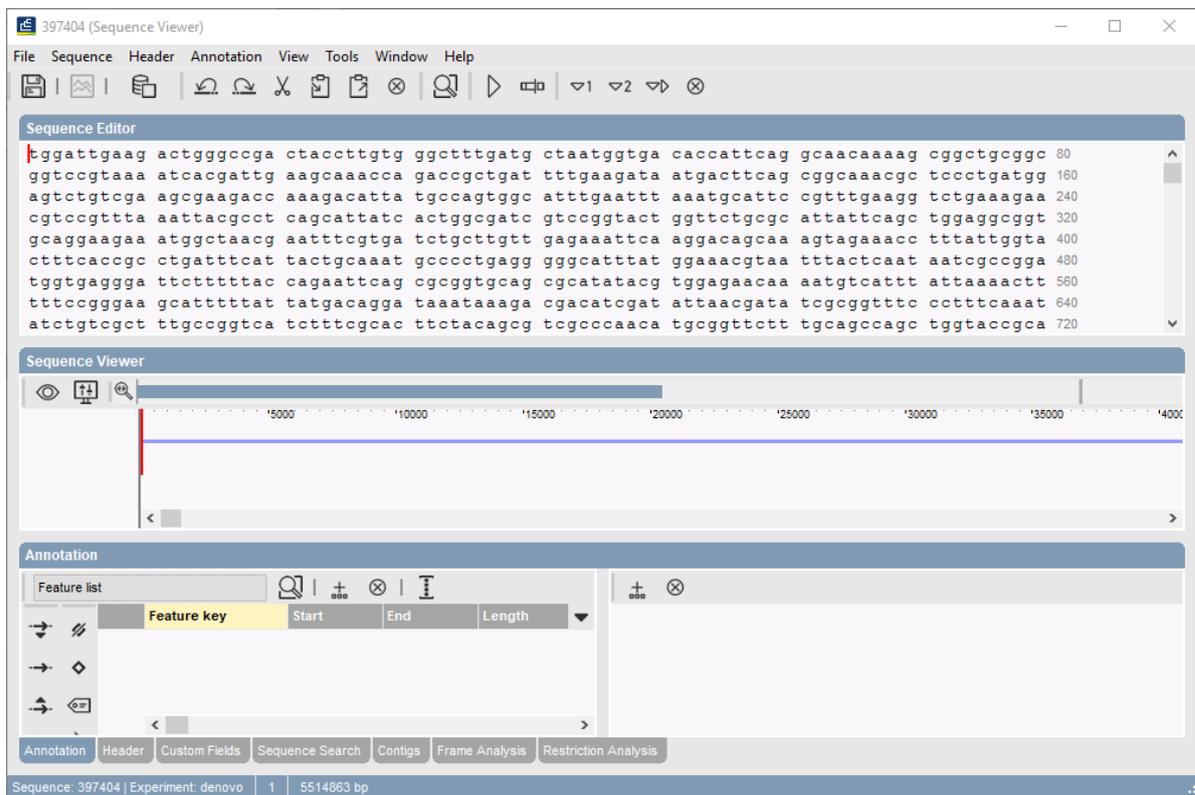


Figure 10: The *Sequence editor* window containing the concatenated de novo contig sequences.