BIONUMERICS Tutorial:

Performing a de novo assembly on the cloud calculation engine

1 Aim

In this tutorial, we will perform a de novo assembly on the cloud calculation engine.

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, "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 https://www.ncbi.nlm.nih.gov/sra.

3 Preparing the demo database

A de novo assembly on the cloud calculation engine can only be performed after installation of the *WGS tools plugin* in the BIONUMERICS database (***File*** > ***Install / remove plugins...*** ( )).

During installation of the plugin, make sure to select the options ***Use default Cloud Calculation Engine*** and ***Enable running jobs on Cloud Calculation Engine*** to unlock the full potential of the default cloud calculation Engine. Note that this installation procedure requires a password and a project name, linked to a certain amount of credits. Please contact Applied Maths to obtain more information.

The **WGS demo database** for *Staphylococcus aureus* already contains the installed *WGS tools plugin* (but without any credits). This demo database can be downloaded directly from the *BIONUMERICS Startup* window (see 3.1), or restored from the back-up file available on our website (see 3.2).

3.1 Option 1: Download demo database from the Startup Screen

1. To download the database directly from the *BIONUMERICS Startup* window, click the  button, located in the toolbar in the *BIONUMERICS Startup* window.

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

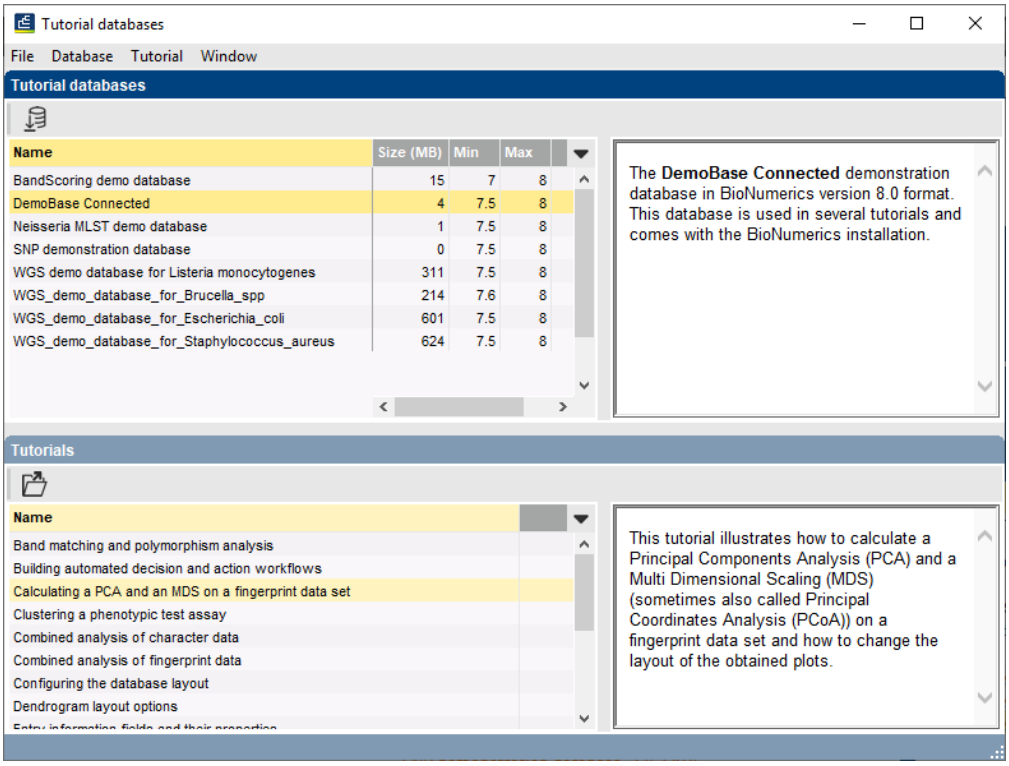

Figure 1: The *Tutorial databases* window, used to download the demonstration database.

2. Select the **WGS_demo_database_for_Staphylococcus_aureus** 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_Staphylococcus_aureus** appears in the *BIONUMERICS Startup* window.

5. Double-click the **WGS_demo_database_for_Staphylococcus_aureus** in the *BIONUMERICS Startup* window to open the database.

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

A BIONUMERICS back-up file of the whole genome demo database for Staphylococcus aureus is also available on our website. This backup can be restored to a functional database in BIONUMERICS.

6. Download the file `wgMLST_SAUR.bnbk` file from https://www.applied-maths.com/download/sample-data, under 'WGS_demo_database_for_Staphylococcus_aureus'.

In contrast to other browsers, some versions of Internet Explorer rename the `wgMLST_SAUR.bnbk` database backup file into `wgMLST_SAUR.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 *BIONUMERICS 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. "Whole genome Staphylococcus aureus demobase".
10. Click <***OK***> to start restoring the database from the backup file (see Figure 2).

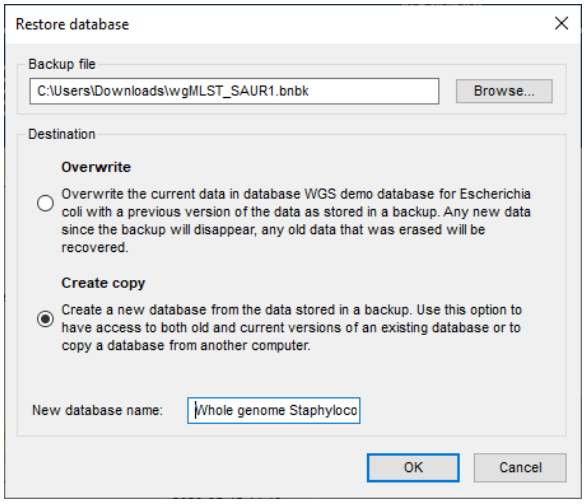

Figure 2: Restoring the whole genome demonstration database from the BioNumerics backup file `wgMLST_SAUR.bnbk`.

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

The *Main* window is displayed (see Figure 3).

4 Importing sequence read sets

1. Open the **WGS_demo_database_for_Staphylococcus_aureus** database or your own database with the *WGS tools* plugin installed.
2. Select ***File*** > ***Import...*** (, **Ctrl+I**) to open the *Import* dialog box.
3. Make sure the ***Import sequence read set data as links*** option is selected in the Import tree and press <***Import***>.

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***

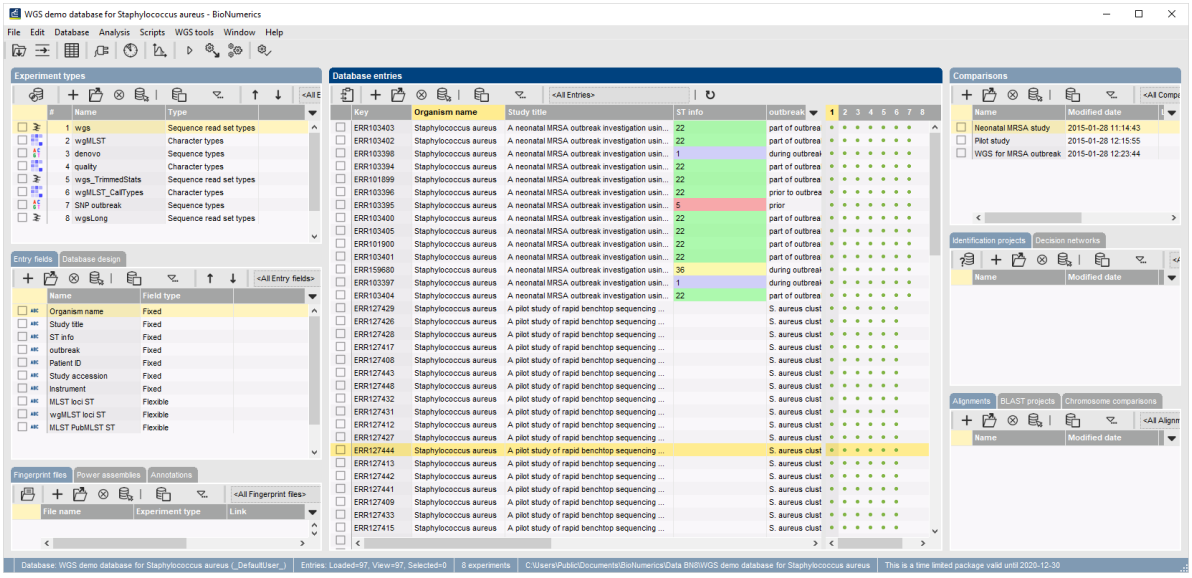

Figure 3: The *Staphylococcus aureus* demonstration database: the *Main* window.

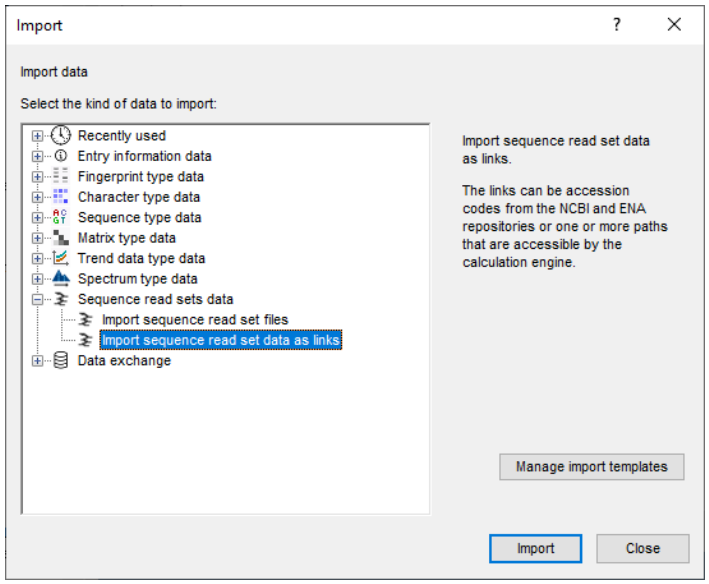

Figure 4: Import sequence read set data as links.

(see Figure 5). Depending on the choice of import, different parameters may be queried in the next steps.

In this tutorial, the import of FASTQ files from a local file server is covered. For more information about the other options, please consult the *WGS tools plugin* manual or the tutorial "Importing links to online repositories".

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

5. Press <***Browse***>, navigate to the correct location, select both `ERR1143520_1.fastq.gz` and `ERR1143520_2.fastq.gz` and 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

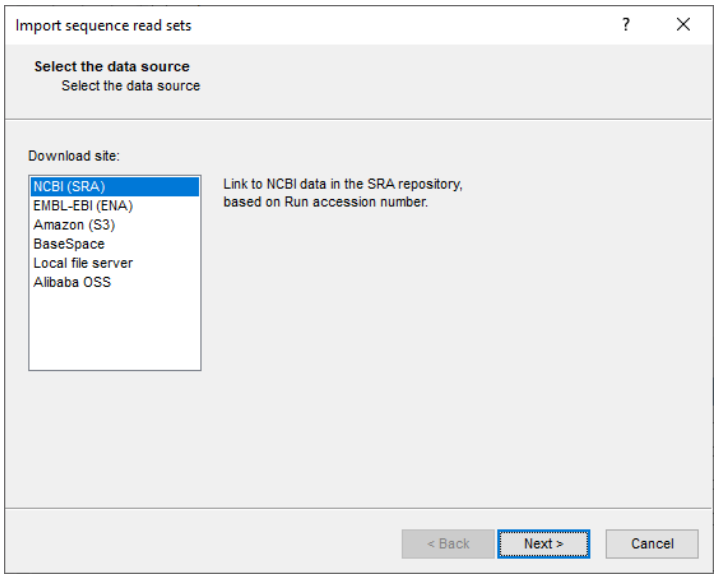

Figure 5: Data sources.

or _2 suffix.

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

7. 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.

8. Select the ***wgs*** experiment and press <***Next***>.

9. 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 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.

10. 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 entry **ERR1143520** is created/updated and has one green dot next to it in the column of the sequence read set experiment type **wgs**.

11. Click on the green colored dot of the imported entry corresponding to the experiment type **wgs**.

The data links are displayed in the *Sequence read set experiment* window (see Figure 7).

If the option ***Calculate sequence read set statistics*** was checked in the last step, the statistics

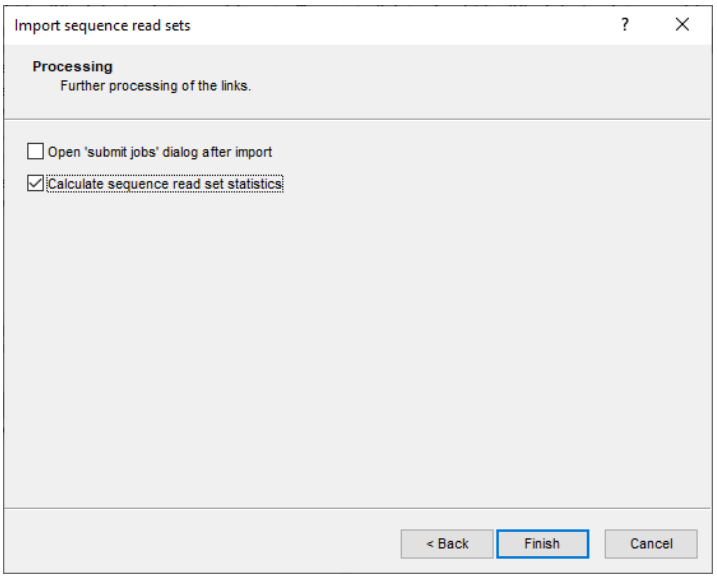

Figure 6: Processing of the links.

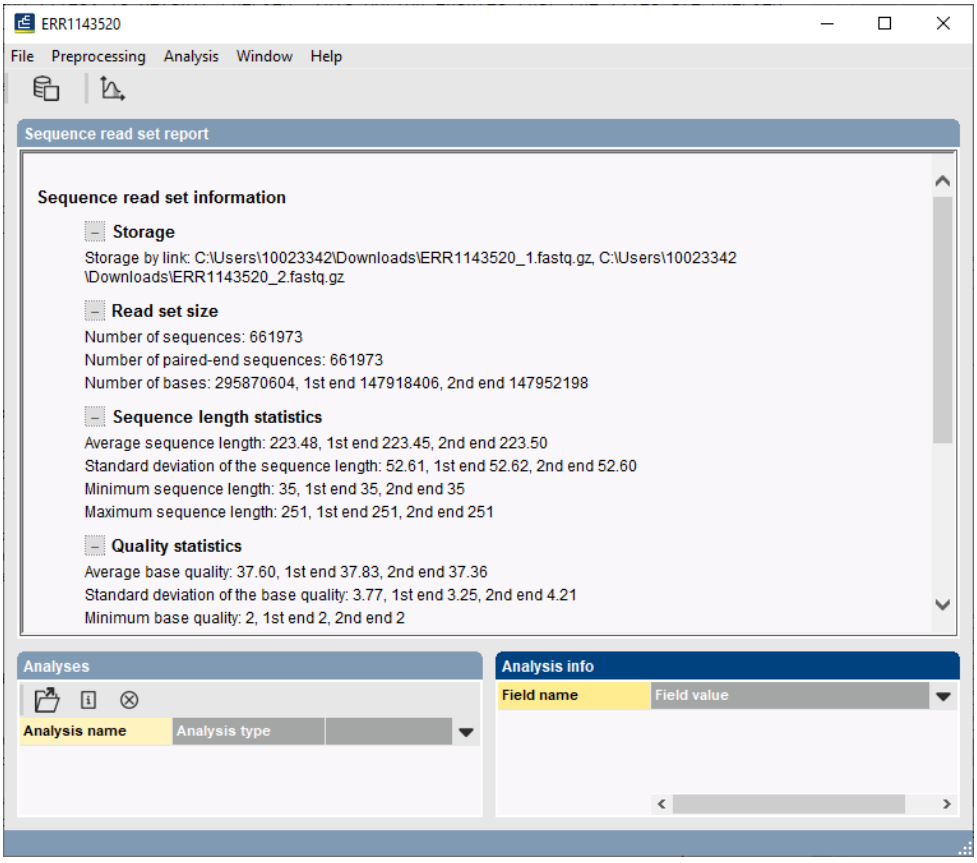

Figure 7: Data link to local file server.

are displayed below).

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

5 Performing a de novo assembly in the cloud

Launching the 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 **ERR1143520** is selected.

2. Select ***WGS tools*** > ***Submit jobs...*** ( ) to call the *Submit jobs* dialog box (see Figure 8).

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

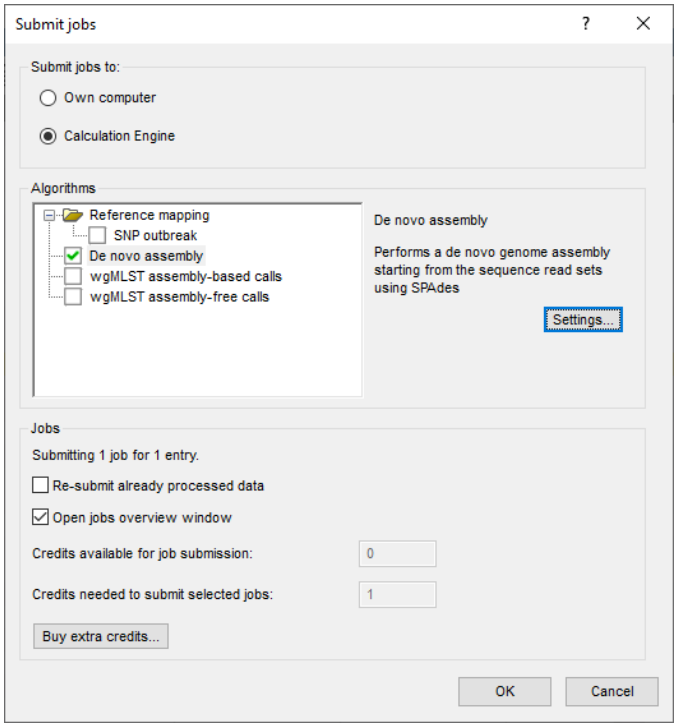

Figure 8: Submit de novo assembly job on the cloud calculation engine.

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 9).

6. Close the *Perform de novo assembly* dialog box.

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: 1 credit is counted for 1 de novo assembly.

7. Press <***OK***> to launch the job on the cloud calculation engine.

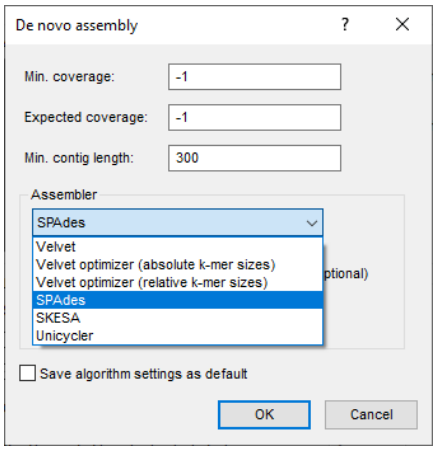

Figure 9: Cloud calculation engine: available de novo assemblers.

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.

When sufficient credits are available for the submission of the job(s) to the external calculation engine, and when links are present to `*.fastq` or `*.fastq.gz` files stored on a local hard drive or a local file server a message will pop up asking to upload the files to an Amazon S3 temporary storage (called the **CE Store**), which the calculation engine can access (see Figure 10). Press <***OK***> to start the **CE Store Uploader** (see Figure 11).

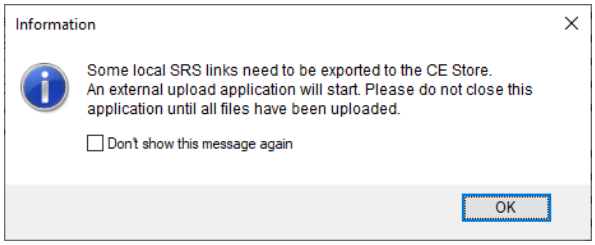

Figure 10: Upload to CE store.

8. By default, the *Job overview* window will open after submission of the job(s). The same dialog can be called at any time with ***File*** > ***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 novo assembly on the cloud calculation engine is around **20-30 min**.

9. To refresh the overview, press ***View*** > ***Refresh*** (, **F5**).

10. 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

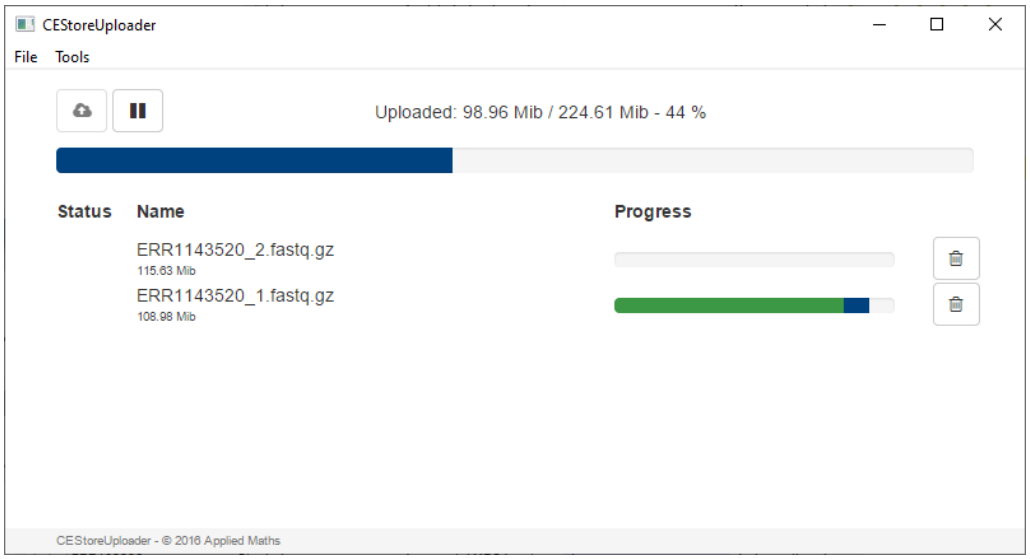

Figure 11: CE Store Uploader.

with coverage information are stored in the sequence experiment type **denovo**.