BioNumerics Tutorial:

Performing a de novo assembly on the external calculation engine

1 Aim

In this tutorial, we will perform a de novo assembly on 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 Preparing the demo database

A de novo assembly on the external calculation engine can only be performed after installation of the *WGS tools plugin* in the BioNumerics database. The installation of the *WGS tools 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*.

A *Staphylococcus aureus* demo database is available on our website, already containing the installed *WGS tools plugin* (but without any credits).

1. To access this database, click the ***Download example databases*** link, located in the lower right corner of the *BioNumerics Startup* window.

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

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 <***Yes***> 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.

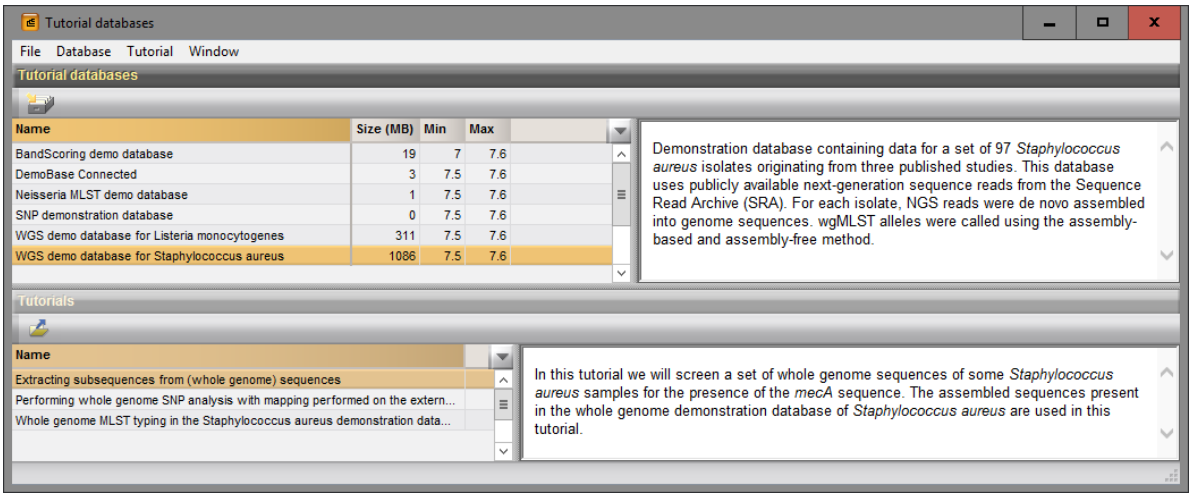

Figure 1: The *Tutorial databases* window, used to download the demonstration database.

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***>.

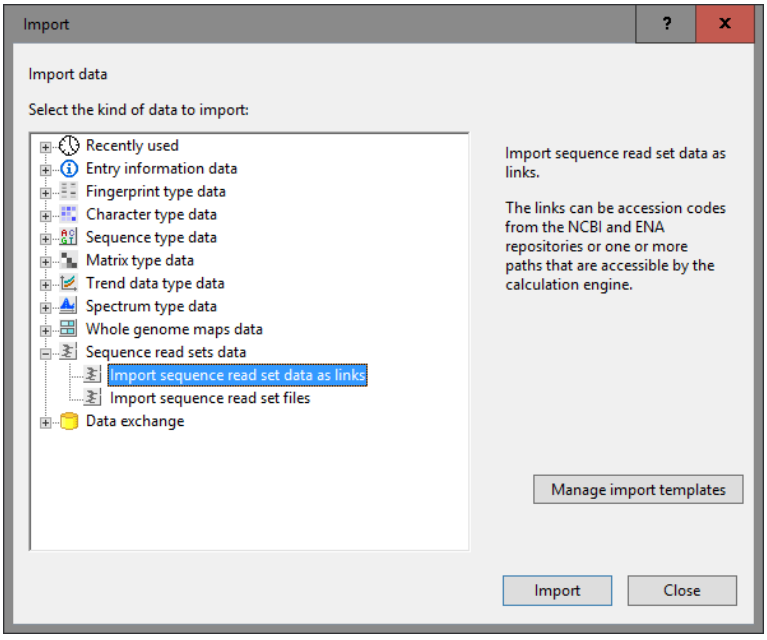

Figure 2: Import sequence read set data as links.

Links to multiple data sources are available, including online and offline data repositories such as: ***NCBI (SRA)***, ***EMBL-EBI (ENA)***, ***Amazon (S3)***, ***BaseSpace*** 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 from a local file server is covered. For more information about

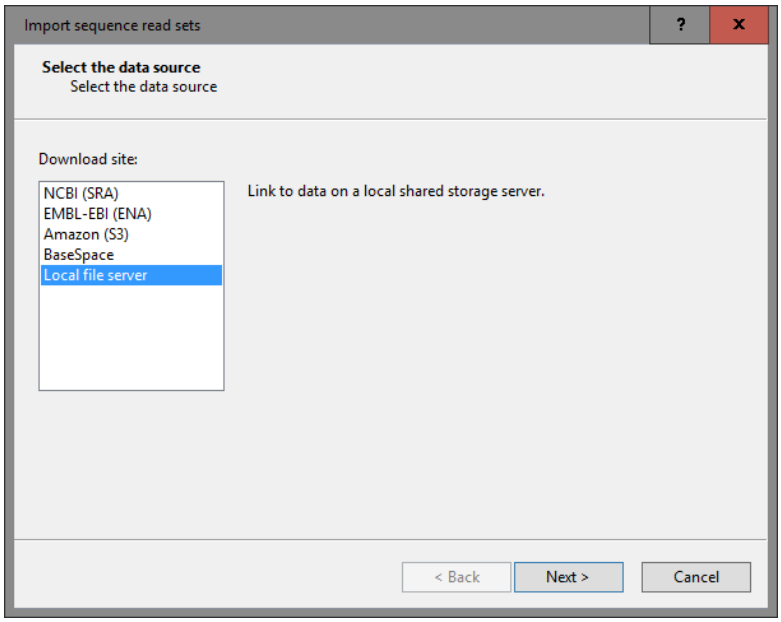

Figure 3: Data sources.

the other options, please consult the *WGS tools plugin* manual.

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` while holding the **Ctrl**-key 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.

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 on the external calculation engine (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.

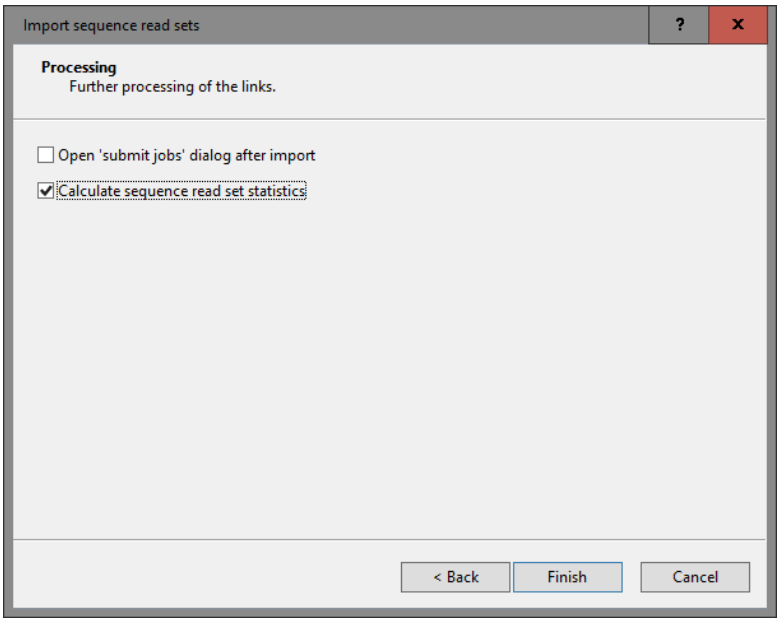

Figure 4: Processing of the links.

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.

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

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

In the *Submit jobs* dialog box you can define which algorithms can be run on the samples.

3. 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.

4. Press the <***Details***> button next to the ***De novo assembly*** option. Two algorithms are available: ***Velvet Optimizer*** and ***SPAdes***.

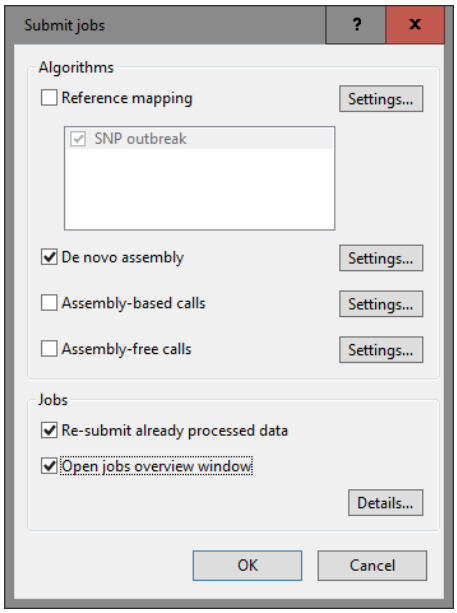

Figure 5: Submit de novo assembly job.

Jobs that already have been submitted and have been imported successfully, will not be relaunched 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. For more information on the credits press the <***Details***> button.

5. Press <***OK***> to launch the job on the 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.

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 6). Press <***OK***> to start the **CE Store Uploader** (see Figure 7).

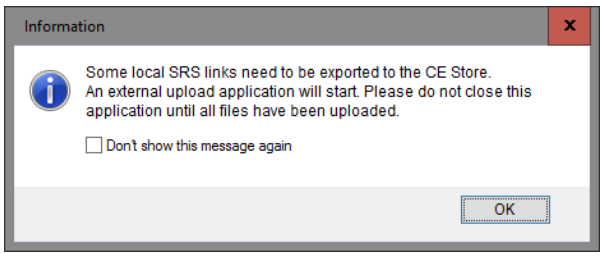

Figure 6: Upload to CE store.

6. By default, the *Calculation engine 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 novo assembly is around **20-30 min**.

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

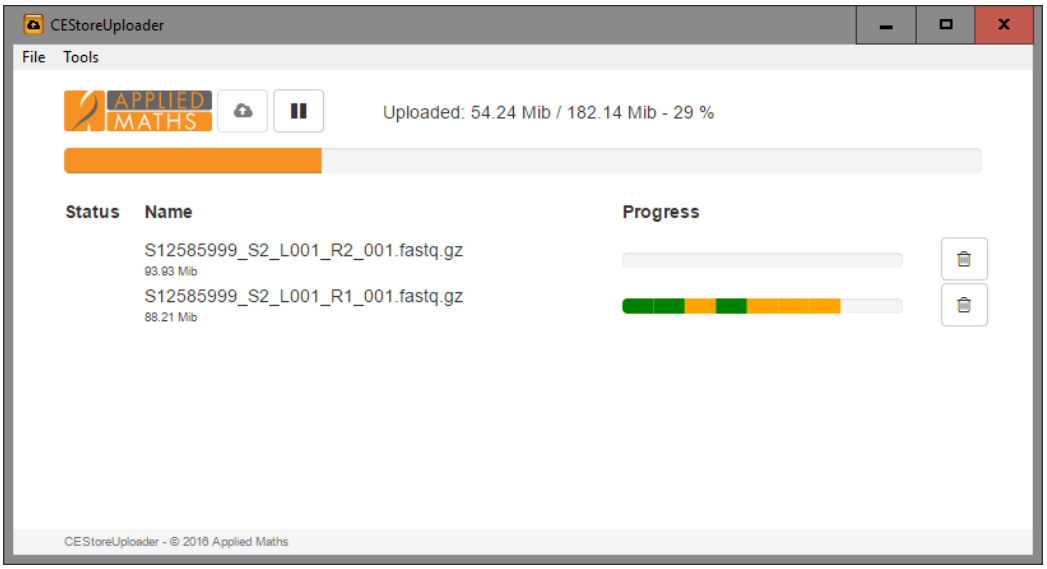

Figure 7: CE Store Uploader.

8. 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 *Calculation engine 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**.