APPLIED MATHS
A BIOMÉRIEUX COMPANY

BioNumerics Tutorial:

Importing links to online repositories

1 Aim

In this tutorial the steps to import links to following online repositories are described:

- ***NCBI (SRA)***: link to data from the Sequence Read Archive (SRA) repository, based on the NCBI run accession number (see 3).
- ***EMBL-EBI (ENA)***: link to data from the ENA repository, based on the EMBL-EBI run accession number (see 3).
- ***Amazon (S3)***: link to data uploaded to a client-specific data bucket hosted at the Amazon S3 storage repository (see 4).
- ***BaseSpace***: link to data uploaded to a data folder hosted on Illumina BaseSpace (see 5).

2 Preparing the database

Importing links to sequence read sets available on NCBI, EMBL-EBI, Amazon or BaseSpace is only possible when the *WGS tools plugin* is installed in the BioNumerics database (***File*** > ***Install / remove plugins...*** ()). 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 acquire your credentials.

3 Import links to NCBI or EMBL-EBI

1. Create a new database (see tutorial ”Creating a new database”) or open an existing database.
2. Select ***File*** > ***Import...*** (, **Ctrl+I**) to call the Import tree.
3. Make sure the ***Import sequence read set data as links*** option is selected in the Import tree. This option is only available after installation of the *WGS tools plugin* (see Figure 1).
4. Press <***Import***>.

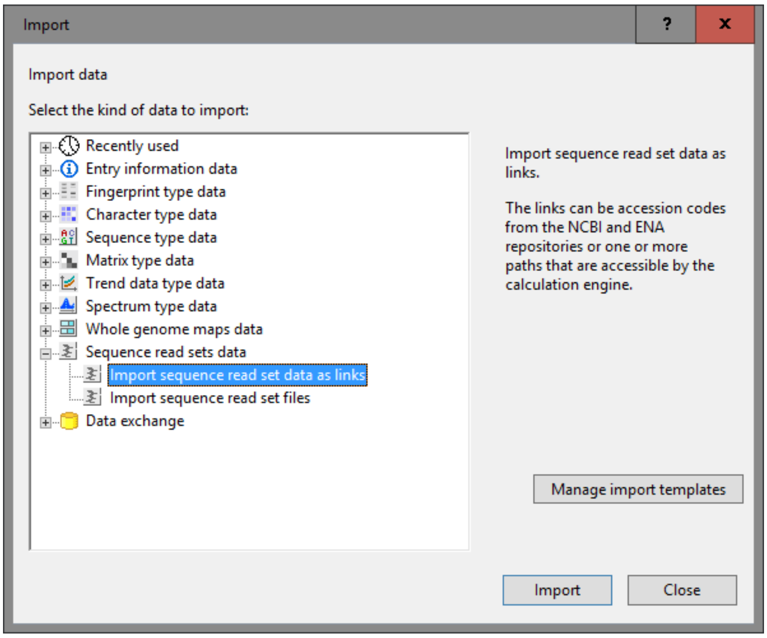
Links to multiple data sources are available, including online and offline data repositories (see Figure 2).

5. Select ***NCBI (SRA)*** or ***EMBL-EBI (ENA)*** and press <***Next***> to go to the next step.

The only required information when importing data from NCBI or EMBL-EBI, are the run ***Accession code(s)*** for the read data. When fetching multiple runs in the same import routine, the different accession codes need to be separated by the same separation character in the ***Accession code(s)*** input box.

With the ***Pick up accession codes from field*** option, accession codes stored in an entry information field in the database can be added to the ***Accession code(s)*** panel by selecting the entry field from the list and pressing the <***Fetch***> button.

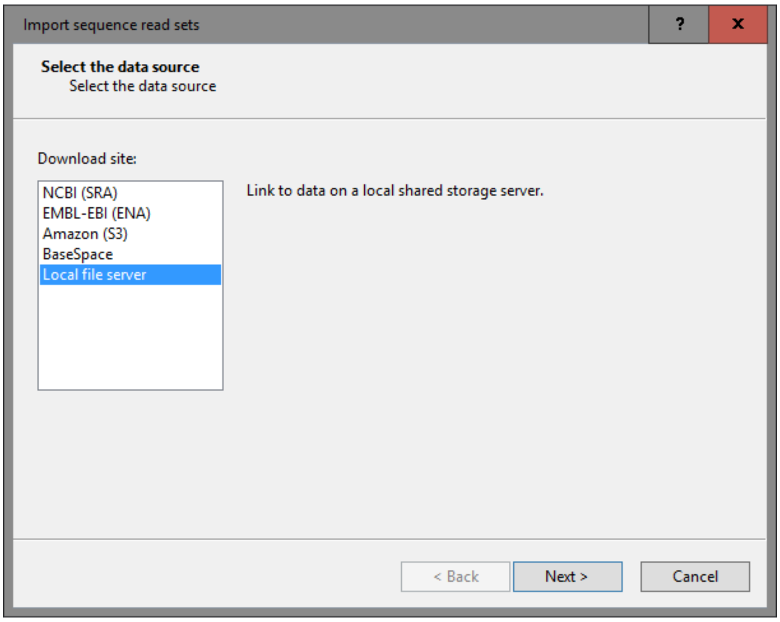
6. Specify the accession number(s) (see Figure 3 for an example) and press <***Next***>.



Import sequence read set data as links.

The links can be accession codes from the NCBI and ENA repositories or one or more paths that are accessible by the calculation engine.

Figure 1: Import sequence read set data as links.



Link to data on a local shared storage server.

Figure 2: Data sources.

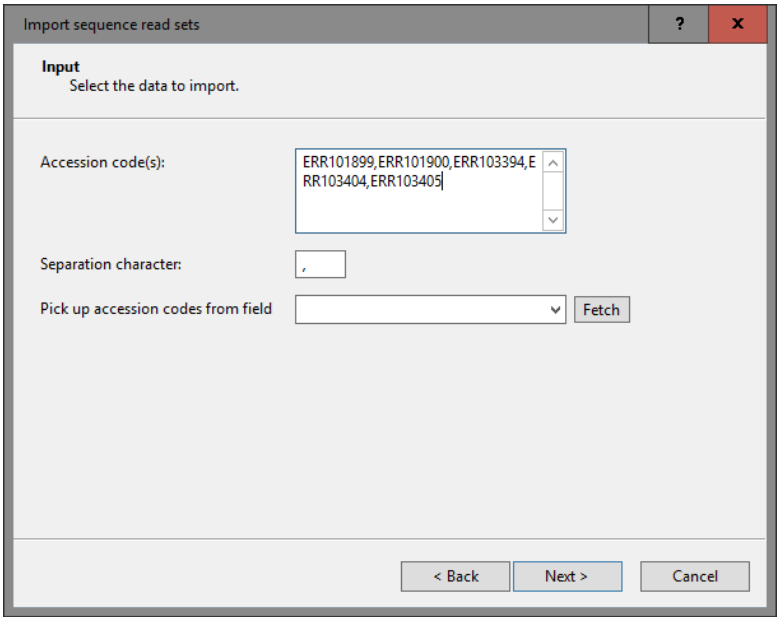
Now you need to define to which field you would like to link the accession number (e.g. to the ***Key*** field or to any other non-default entry field).

7. Double-click the only row (= accession number) in the grid, select the ***Key*** field from the tree or a new or existing entry field under ***Entry info field*** and press <***OK***>.

The grid is updated.

8. Optionally, you can do a preview of what you are about to import. Press <***Preview...***> to open the preview. Close the preview again.

9. Click <***Next***> and <***Finish***> to finish the creation of the import template.



Import sequence read sets

Input

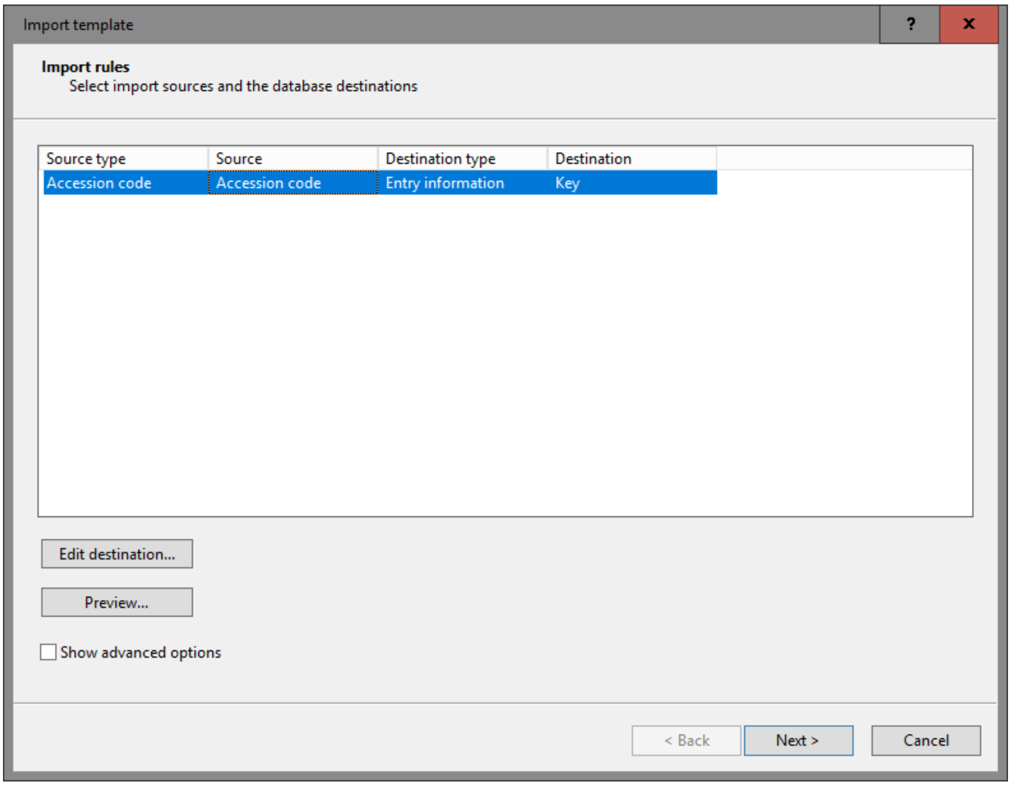
Select the data to import.

Accession code(s): ERR101899,ERR101900,ERR103394,ERR103404,ERR103405

Separation character: ,

Pick up accession codes from field

Figure 3: Accession number(s).



Import template

Import rules

Select import sources and the database destinations

Source type	Source	Destination type	Destination
Accession code	Accession code	Entry information	Key

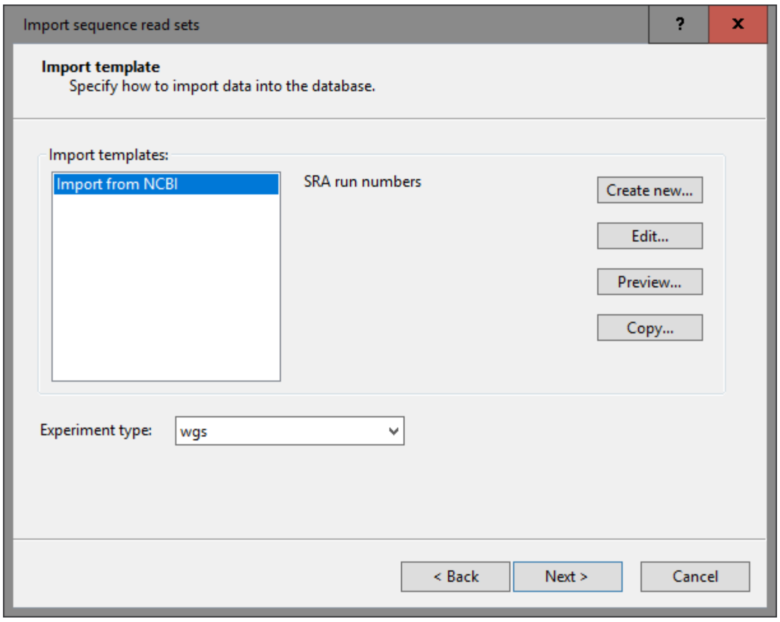
□ Show advanced options

Figure 4: Import rule.

10. Enter a meaningful name (and optionally a description) for the created import template e.g. "Import from NCBI", and click <***OK***>.

11. Choose the newly created import template from the list and click <***Next***>.

12. Select the created import template and a new or existing experiment from the drop-down list and press <***Next***>.



Import sequence read sets

Import template

Specify how to import data into the database.

Import templates:

Import from NCBI — SRA run numbers

Experiment type: wgs

Figure 5: Import template.

13. When an experiment name is prompted for, specify a sequence type name, e.g. “wgs”. Click <***OK***> and confirm the creation of the experiment.

14. Press <***Next***> to start the import of the sequence read set links.

In the last step, calculation jobs on the external calculation engine can be launched on the imported data links (***Open submit jobs dialog after import***). Jobs include de novo assembly, assembly-based and assembly-free calling and reference mapping for wgSNP. The same dialog can be called from the *Main* window at any time with ***WGS tools*** > ***Submit jobs...*** ( ).

15. Uncheck ***Open submit jobs dialog after import*** and press <***Finish***> to start the import of the data links.

Once the import is completed, the entries are created/updated and have one green dot next to it in the column of the selected sequence read set experiment type (e.g. **wgs**).

16. Click on a green colored dot corresponding to the sequence read set experiment type.

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

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

4 Import links to Amazon

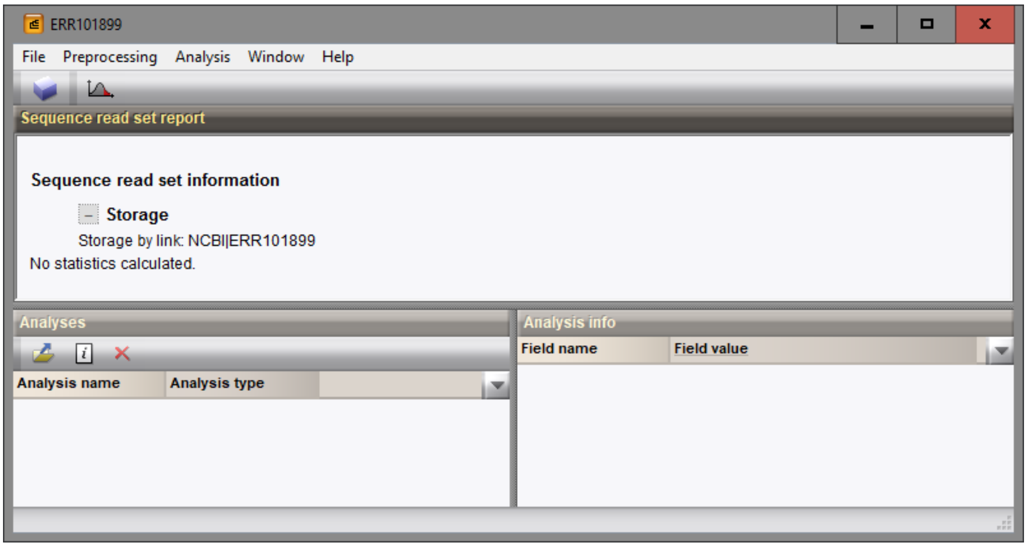
1. Create a new database (see tutorial ”Creating a new database”) or open an existing database.

When using the Amazon import routine, make sure the read set files you wish to import in the same import routine are grouped in the same folder of your Amazon bucket.

2. Select ***File*** > ***Import...*** (, **Ctrl+I**) to call the Import tree.

3. Make sure the ***Import sequence read set data as links*** option is selected in the Import tree. This option is only available after installation of the *WGS tools plugin* (see Figure 1).

4. Press <***Import***>.



Sequence read set information

Storage

Storage by link: NCBI|ERR101899

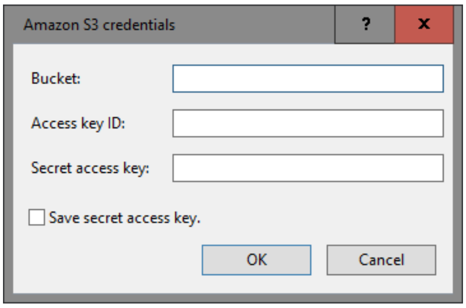
No statistics calculated.

Figure 6: Data link to NCBI.

Links to multiple data sources are available, including online and offline data repositories (see Figure 2).

5. Select ***Amazon (S3)*** and press <***Next***> to go to the next step.

6. The first time you use this import routine, you need to specify your Amazon S3 credentials: ***Bucket name***, ***Access key ID*** and the ***Secret access key***.



Bucket:

Access key ID:

Secret access key:

□ Save secret access key.

Figure 7: Amazon (S3) credentials.

7. Check the option ***Save secret access key*** to save the credentials to the database and press <***OK***>.

All detected folders and files in the bucket are listed in the next dialog.

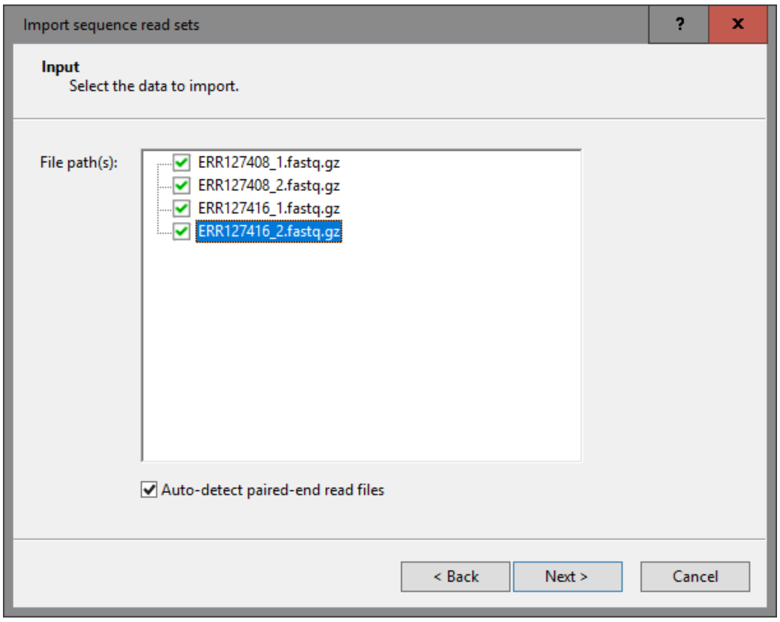
8. Check the files you wish to import and leave the option ***Auto-detect paired-end read files*** checked. Press <***Next***>.

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 run accession numbers from the file names and store this information in the BioNumerics ***Key*** field.

9. 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.

10. Make sure <***Create new***> is selected from the ***Experiment type*** list or select an existing experiment and press <***Next***>.



Import sequence read sets

Input
Select the data to import.

File path(s):
- ☑ ERR127408_1.fastq.gz
- ☑ ERR127408_2.fastq.gz
- ☑ ERR127416_1.fastq.gz
- ☑ ERR127416_2.fastq.gz

☑ Auto-detect paired-end read files

Figure 8: Files detected in the Amazon S3 bucket.

11. When an experiment name is prompted for, specify a sequence type name, e.g. “wgs”. Click <***OK***> and confirm the creation of the experiment.

12. Press <***Next***> to start the import of the sequence read set links.

In the last step, calculation jobs on the external calculation engine can be launched on the imported data links (***Open submit jobs dialog after import***). Jobs include de novo assembly, assembly-based and assembly-free calling and reference mapping for wgSNP. The same dialog can be called from the *Main* window at any time with ***WGS tools*** > ***Submit jobs...*** ( ).

13. Uncheck ***Open submit jobs dialog after import*** and press <***Finish***> to start the import of the data links.

Once the import is completed, the entries are created/updated and have one green dot next to it in the column of the selected sequence read set experiment type (e.g. **wgs**).

14. Click on a green colored dot corresponding to the sequence read set experiment type.

The data link is displayed in the *Sequence read set experiment* window (see Figure 9).

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

5 Import links to BaseSpace

1. Create a new database (see tutorial ”Creating a new database”) or open an existing database.

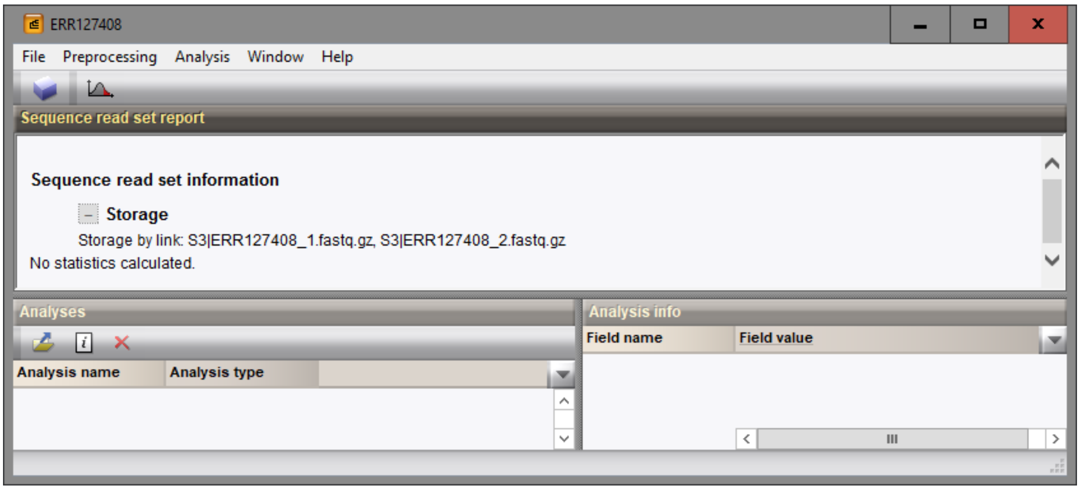
2. Select ***File*** > ***Import...*** (, **Ctrl+I**) to call the Import tree.

3. Make sure the ***Import sequence read set data as links*** option is selected in the Import tree. This option is only available after installation of the *WGS tools plugin* (see Figure 1).

4. Press <***Import***>.

Links to multiple data sources are available, including online and offline data repositories (see Figure 2).

5. Select ***BaseSpace*** and press <***Next***> to go to the next step.



Sequence read set information

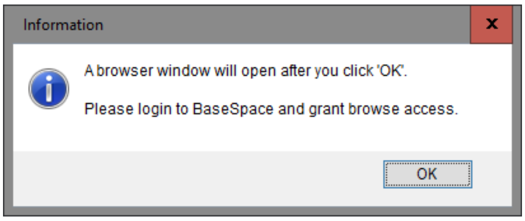
Storage

Storage by link: S3|ERR127408_1.fastq.gz, S3|ERR127408_2.fastq.gz

No statistics calculated.

Figure 9: Data link to Amazon S3 bucket.

6. The first time you use this import routine, you need to login to your BaseSpace account and grant browse access (see Figure 10).

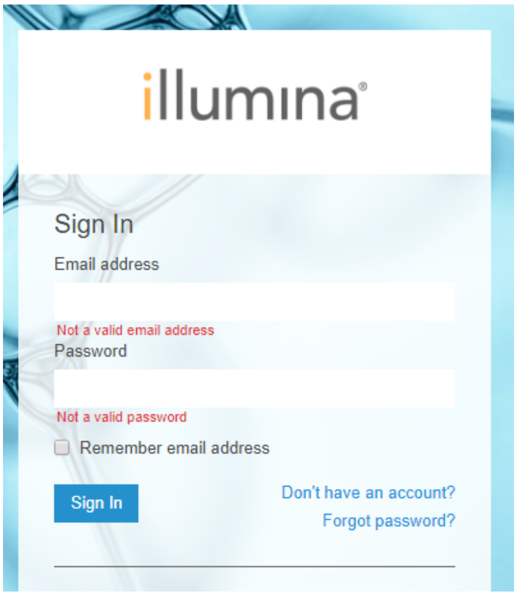


A browser window will open after you click 'OK'.

Please login to BaseSpace and grant browse access.

Figure 10: Information window.

7. Provide your ***Email address*** and ***Password*** (see Figure 11). After authorization, close the browser window.



Not a valid email address

Not a valid password

□ Remember email address

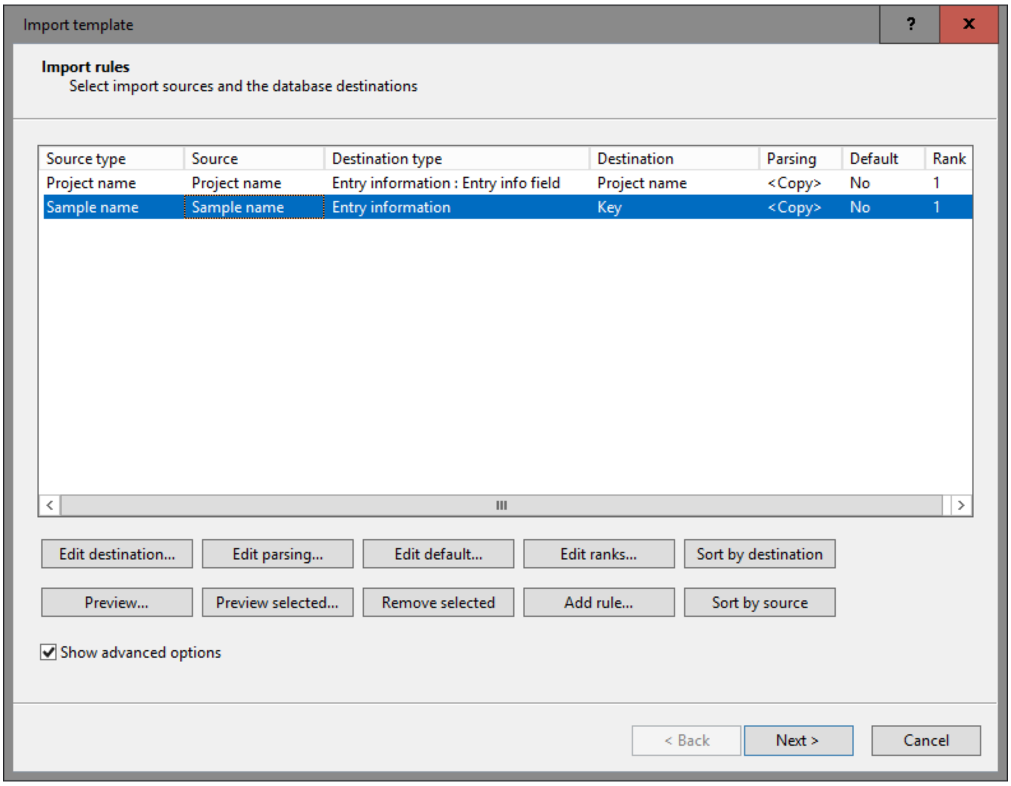
Figure 11: Sign in.

8. In the next step in the BioNumerics wizard, select your BaseSpace ***Project*** and select the ***Sample(s)*** you wish to import. Multiple samples can be selected with the **Ctrl**- and **Shift**-keys.

9. Press <*Next*> to go to the next step.

Now you need to define how the data should be stored in the database. A database destination can be specified for the ***Project name*** and ***Sample name***.

10. You might for example want to link the ***Project name*** to an entry field and the ***Sample name*** to the ***Key*** field (see Figure 12 for these rules). Linking is done by double-clicking the row in the grid and selecting the correct destination from the tree.



Source type	Source	Destination type	Destination	Parsing	Default	Rank
Project name	Project name	Entry information : Entry info field	Project name	<Copy>	No	1
Sample name	Sample name	Entry information	Key	<Copy>	No	1

☑ Show advanced options

Figure 12: Import rules: an example.

11. Optionally, you can do a preview of what you are about to import. Press <***Preview...***> to open the preview. Close the preview again.

12. Click <*Next*> and <***Finish***> to finish the creation of the import template.

13. Enter a meaningful name (and optionally a description) for the created import template e.g. “Import from BaseSpace”, and click <***OK***>.

The new template is automatically selected.

14. Make sure <***Create new***> is selected from the ***Experiment type*** list or select an existing experiment and press <***Next***>.

15. When an experiment name is prompted for, specify a sequence type name, e.g. “wgs”. Click <***OK***> and confirm the creation of the experiment.

16. Press <***Next***> to start the import of the sequence read set links.

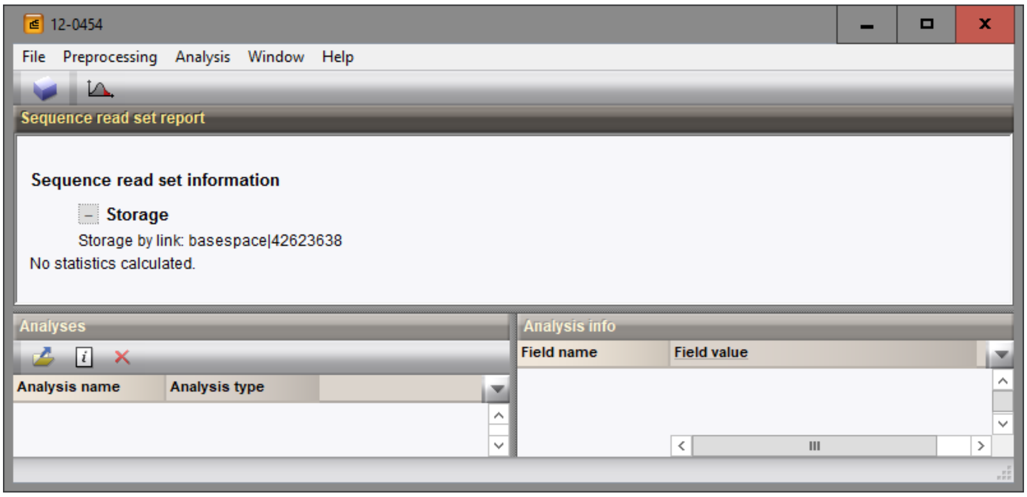
In the last step, calculation jobs on the external calculation engine can be launched on the imported data links (***Open submit jobs dialog after import***). Jobs include de novo assembly, assembly-based and assembly-free calling and reference mapping for wgSNP. The same dialog can be called from the *Main* window at any time with ***WGS tools*** > ***Submit jobs...*** ().

17. Uncheck ***Open submit jobs dialog after import*** and press <***Finish***> to start the import of the data links.

Once the import is completed, the entries are created/updated and have one green dot next to it in the column of the selected sequence read set experiment type (e.g. **wgs**).

18. Click on a green colored dot corresponding to the sequence read set experiment type.

The data link is displayed in the *Sequence read set experiment* window (see Figure 13).



Sequence read set information

Storage

Storage by link: basespace|42623638

No statistics calculated.

Figure 13: Data link to BaseSpace.

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

6 Analysis tools

Calculation jobs on the external calculation engine include de novo assembly, assembly-based and assembly-free calling (wgMLST) and reference mapping (wgSNP). More information about posting jobs on the external calculation engine can be found following tutorials:

- ”Performing a de novo assembly on the external calculation engine”
- ”wgMLST typing in BioNumerics: routine workflow”
- ”Performing whole genome SNP analysis with mapping performed on the external calculation engine”