BIONUMERICS Tutorial:

wgSNP with mapping performed locally on your own computer

1 Introduction

A Single Nucleotide Polymorphism (SNP) is a variation in a single nucleotide, which occurs at a specific position of the genome. When performed on whole genome sequences (WGS), this analysis is referred as **whole genome SNP (wgSNP) analysis**. When performed in BIONUMERICS a typical workflow for wgSNP analysis consists of following steps:

1. Choose a reference sequence
2. Map sequence reads against the reference sequence (locally or on the cloud calculation engine)
3. Perform wgSNP analysis and filter out relevant SNPs
4. wgSNP clustering

In this tutorial we will focus on the first two steps of the workflow, with local mapping of the reads against the reference sequence. The last two steps are covered in the tutorial ”wgSNP: analysis and clustering”.

2 Preparing the database

Mapping of reads against the reference sequence can only be performed locally on your own computer after installation of the *WGS tools plugin* in the BIONUMERICS database (***File*** > ***Install / remove plugins...*** (🔌)).

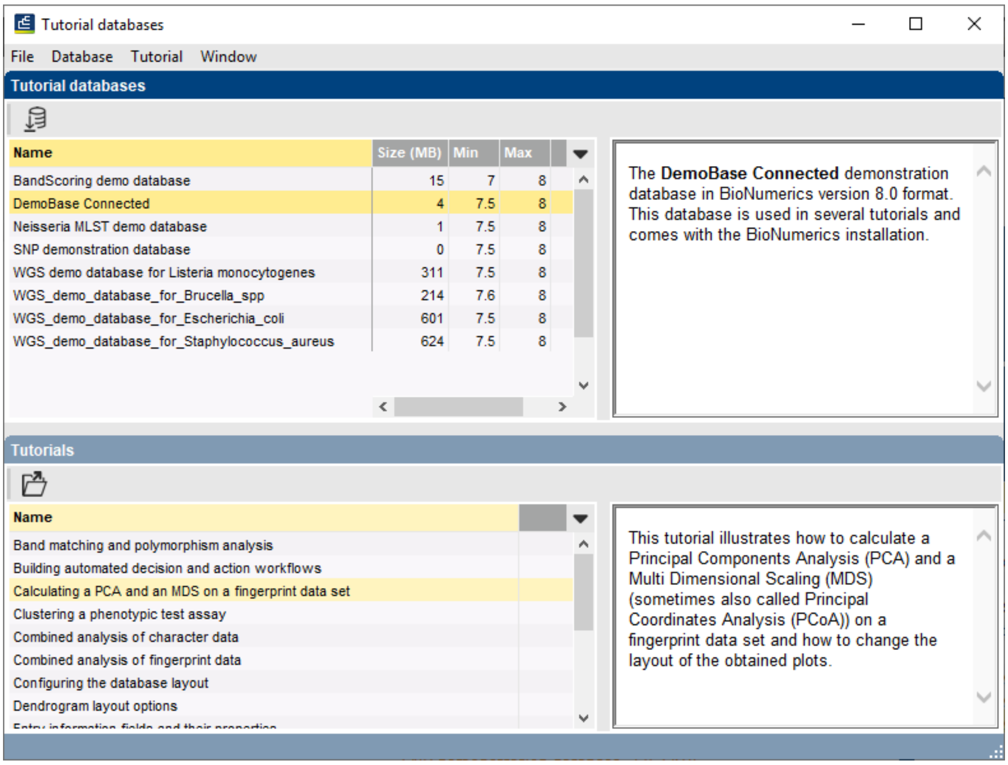
During installation of the *WGS tools plugin* plugin, make sure to select the options ***Use default Cloud Calculation Engine*** and ***Local calculations only***.

The **WGS demo database** for *Staphylococcus aureus* already contains the installed *WGS tools plugin*. This demo database can be downloaded directly from the *BIONUMERICS Startup* window (see 2.1), or restored from the back-up file available on our website (see 2.2)

2.1 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).



Name	Size (MB)	Min	Max
BandScoring demo database	15	7	8
DemoBase Connected	4	7.5	8
Neisseria MLST demo database	1	7.5	8
SNP demonstration database	0	7.5	8
WGS demo database for Listeria monocytogenes	311	7.5	8
WGS_demo_database_for_Brucella_spp	214	7.6	8
WGS_demo_database_for_Escherichia_coli	601	7.5	8
WGS_demo_database_for_Staphylococcus_aureus	624	7.5	8

The **DemoBase Connected** demonstration database in BioNumerics version 8.0 format. This database is used in several tutorials and comes with the BioNumerics installation.

Name
Band matching and polymorphism analysis
Building automated decision and action workflows
Calculating a PCA and an MDS on a fingerprint data set
Clustering a phenotypic test assay
Combined analysis of character data
Combined analysis of fingerprint data
Configuring the database layout
Dendrogram layout options
Entry information fields and their properties

This tutorial illustrates how to calculate a Principal Components Analysis (PCA) and a Multi Dimensional Scaling (MDS) (sometimes also called Principal Coordinates Analysis (PCoA)) on a fingerprint data set and how to change the layout of the obtained plots.

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.

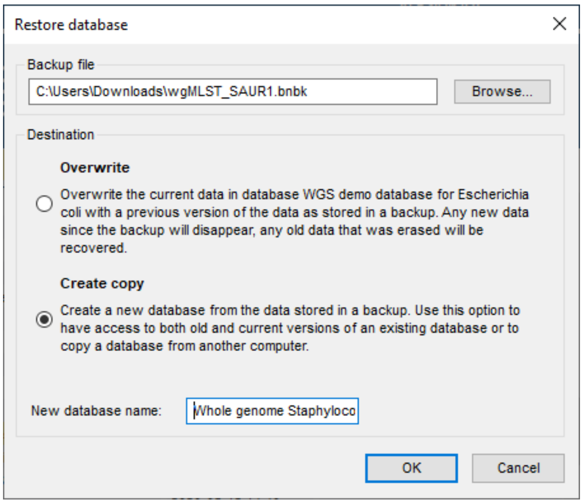
2.2 Option 2: Restore demo database from back-up file

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).



Restore database ×

Backup file

C:\Users\Downloads\wgMLST_SAUR1.bnbk

Browse...

Destination

Overwrite

○ Overwrite the current data in database WGS demo database for Escherichia coli with a previous version of the data as stored in a backup. Any new data since the backup will disappear, any old data that was erased will be recovered.

Create copy

● Create a new database from the data stored in a backup. Use this option to have access to both old and current versions of an existing database or to copy a database from another computer.

New database name: Whole genome Staphyloco

OK Cancel

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

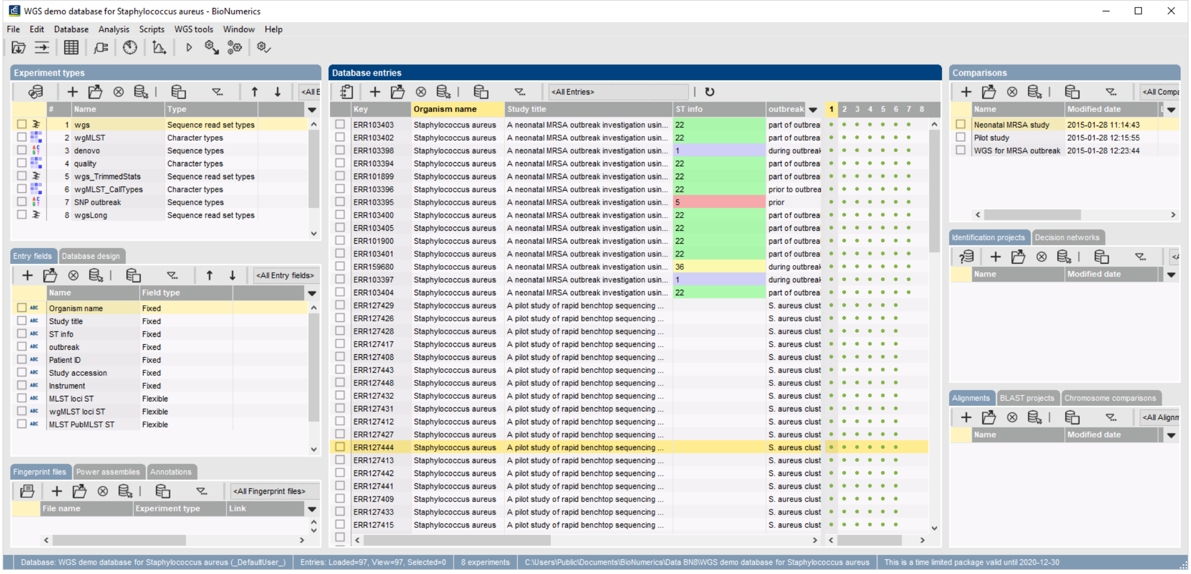
11. Once the process is complete, click <***Yes***> to open the database.

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

3 Creating a reference mapped sequence type

First, we will create a sequence type to store the reference mapped sequences in:

1. Click on the *Experiment types* panel to activate it and select ***Edit*** > ***Create new object...*** (+).
2. From the *Create a new experiment type* dialog box that pops up, select ***Sequence type*** and press <***OK***>.
3. In the *New sequence type* wizard, enter a ***Sequence type name*** (e.g. “My wgSNP”).
4. Press <***Next***> (see Figure 4).



	#	Name	Type
□	1	wgs	Sequence read set types
□	2	wgMLST	Character types
□	3	denovo	Sequence types
□	4	quality	Character types
□	5	wgs_TrimmedStats	Sequence read set types
□	6	wgMLST_CallTypes	Character types
□	7	SNP outbreak	Sequence types
□	8	wgsLong	Sequence read set types

	Name	Field type
□	Organism name	Fixed
□	Study title	Fixed
□	ST info	Fixed
□	outbreak	Fixed
□	Patient ID	Fixed
□	Study accession	Fixed
□	Instrument	Fixed
□	MLST loci ST	Flexible
□	wgMLST loci ST	Flexible
□	MLST PubMLST ST	Flexible

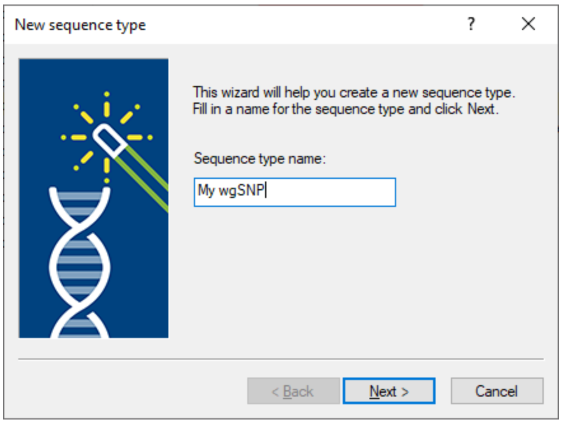
File name	Experiment type	Link

	Key	Organism name	Study title	ST info	outbreak
□	ERR103403	Staphylococcus aureus	A neonatal MRSA outbreak investigation usin...	22	part of outbrea
□	ERR103402	Staphylococcus aureus	A neonatal MRSA outbreak investigation usin...	22	part of outbrea
□	ERR103398	Staphylococcus aureus	A neonatal MRSA outbreak investigation usin...	1	during outbreak
□	ERR103394	Staphylococcus aureus	A neonatal MRSA outbreak investigation usin...	22	part of outbrea
□	ERR101899	Staphylococcus aureus	A neonatal MRSA outbreak investigation usin...	22	part of outbrea
□	ERR103396	Staphylococcus aureus	A neonatal MRSA outbreak investigation usin...	22	prior to outbrea
□	ERR103395	Staphylococcus aureus	A neonatal MRSA outbreak investigation usin...	5	prior
□	ERR103400	Staphylococcus aureus	A neonatal MRSA outbreak investigation usin...	22	part of outbrea
□	ERR103405	Staphylococcus aureus	A neonatal MRSA outbreak investigation usin...	22	part of outbrea
□	ERR101900	Staphylococcus aureus	A neonatal MRSA outbreak investigation usin...	22	part of outbrea
□	ERR103401	Staphylococcus aureus	A neonatal MRSA outbreak investigation usin...	22	part of outbrea
□	ERR159680	Staphylococcus aureus	A neonatal MRSA outbreak investigation usin...	36	during outbreak
□	ERR103397	Staphylococcus aureus	A neonatal MRSA outbreak investigation usin...	1	during outbreak
□	ERR103404	Staphylococcus aureus	A neonatal MRSA outbreak investigation usin...	22	part of outbrea
□	ERR127429	Staphylococcus aureus	A pilot study of rapid benchtop sequencing ...		S. aureus clust
□	ERR127426	Staphylococcus aureus	A pilot study of rapid benchtop sequencing ...		S. aureus clust
□	ERR127428	Staphylococcus aureus	A pilot study of rapid benchtop sequencing ...		S. aureus clust
□	ERR127417	Staphylococcus aureus	A pilot study of rapid benchtop sequencing ...		S. aureus clust
□	ERR127408	Staphylococcus aureus	A pilot study of rapid benchtop sequencing ...		S. aureus clust
□	ERR127443	Staphylococcus aureus	A pilot study of rapid benchtop sequencing ...		S. aureus clust
□	ERR127448	Staphylococcus aureus	A pilot study of rapid benchtop sequencing ...		S. aureus clust
□	ERR127432	Staphylococcus aureus	A pilot study of rapid benchtop sequencing ...		S. aureus clust
□	ERR127431	Staphylococcus aureus	A pilot study of rapid benchtop sequencing ...		S. aureus clust
□	ERR127412	Staphylococcus aureus	A pilot study of rapid benchtop sequencing ...		S. aureus clust
□	ERR127427	Staphylococcus aureus	A pilot study of rapid benchtop sequencing ...		S. aureus clust
□	ERR127444	Staphylococcus aureus	A pilot study of rapid benchtop sequencing ...		S. aureus clust
□	ERR127413	Staphylococcus aureus	A pilot study of rapid benchtop sequencing ...		S. aureus clust
□	ERR127442	Staphylococcus aureus	A pilot study of rapid benchtop sequencing ...		S. aureus clust
□	ERR127441	Staphylococcus aureus	A pilot study of rapid benchtop sequencing ...		S. aureus clust
□	ERR127409	Staphylococcus aureus	A pilot study of rapid benchtop sequencing ...		S. aureus clust
□	ERR127433	Staphylococcus aureus	A pilot study of rapid benchtop sequencing ...		S. aureus clust
□	ERR127415	Staphylococcus aureus	A pilot study of rapid benchtop sequencing ...		S. aureus clust

	Name	Modified date
□	Neonatal MRSA study	2015-01-28 11:14:43
□	Pilot study	2015-01-28 12:15:55
□	WGS for MRSA outbreak	2015-01-28 12:23:44

Database: WGS demo database for Staphylococcus aureus (_DefaultUser_) | Entries: Loaded=97, View=97, Selected=0 | 8 experiments | C:\Users\Public\Documents\BioNumerics\Data BN8\WGS demo database for Staphylococcus aureus | This is a time limited package valid until 2020-12-30

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

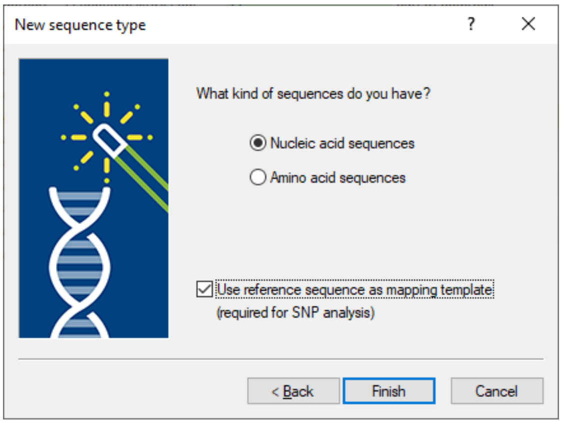


This wizard will help you create a new sequence type.
Fill in a name for the sequence type and click Next.

Sequence type name: My wgSNP

Figure 4: New sequence type experiment.

5. Leave the default ***Nucleic acid sequences*** option checked and check the option ***Use reference sequence as mapping template*** (see Figure 5).



What kind of sequences do you have?

● Nucleic acid sequences

○ Amino acid sequences

☑ Use reference sequence as mapping template
(required for SNP analysis)

Figure 5: Use as mapping template.

6. Press <***Finish***> to create the reference mapped sequence type.

The reference mapped sequence type is added to the *Experiment types* panel.

4 Assigning a reference sequence

The first sequence that is imported in the reference mapped sequence type will automatically be assigned as the reference. The reference sequence might be a closed, fully annotated genome sequence (e.g. downloaded from an online repository such as NCBI), but could as well be a de novo assembled sequence, consisting of multiple contigs (i.e. a draft genome).

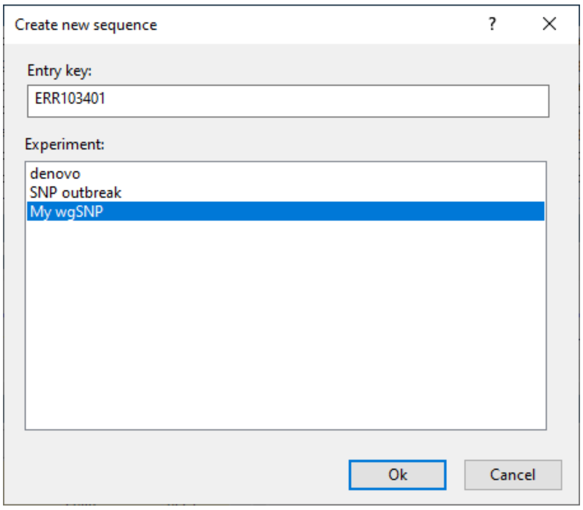
Tutorials are available on our website explaining how to import genome sequences (***File*** > ***Import...*** (📂, **Ctrl+I**), ”Sequence data”) and how to import and assemble sequence reads (***File*** > ***Import...*** (📂, **Ctrl+I**), ”Sequence read sets”).

In this tutorial, we will copy the sequence from the ***denovo*** experiment type of entry ***ERR103401*** to our new sequence type experiment so it serves as a reference sequence:

1. Click on the green dot in the *Experiment presence* panel that corresponds to the ***denovo*** experiment for entry ***ERR103401***. In default configuration, the ***denovo*** experiment corresponds to the third column in the *Experiment presence* panel.

2. In the *Sequence editor* window that opens, select ***File*** > ***Save as...*** (**Ctrl+Shift+S**).

3. Highlight ***My wgSNP*** in the list and press <***OK***> (see Figure 6).



Entry key: ERR103401

Experiment:
denovo
SNP outbreak
My wgSNP

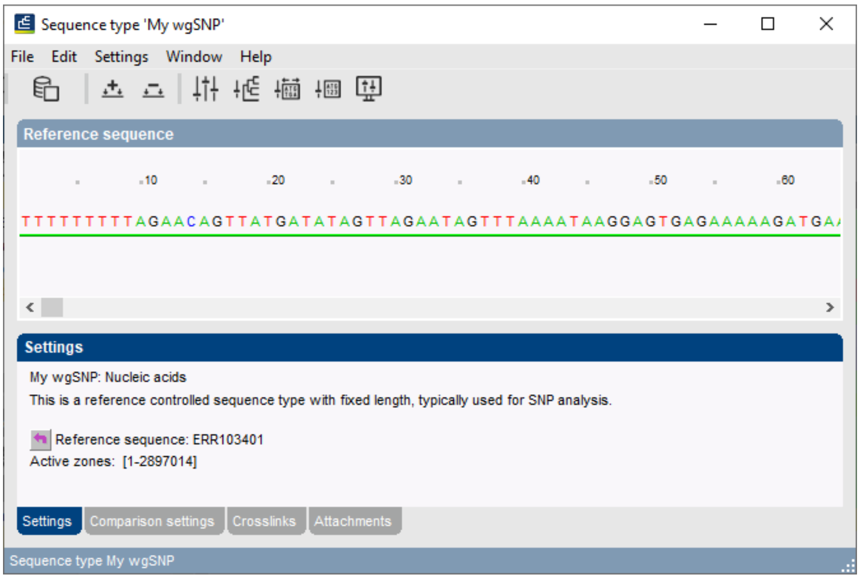
Figure 6: Save sequence to a new sequence type experiment.

4. Close the *Sequence editor* window.

We can check if this sequence is indeed used as reference:

5. In the *Experiment types* panel, double-click on ***My wgSNP*** to open its *Sequence type* window: the reference sequence is displayed here (see Figure 7).

6. Close the *Sequence type* window.



TTTTTTTTTAGAACAGTTATGATATAGTTAGAATAGTTTAAAATAAGGAGTGAGAAAAAGATGA

My wgSNP: Nucleic acids

This is a reference controlled sequence type with fixed length, typically used for SNP analysis.

Reference sequence: ERR103401

Active zones: [1-2897014]

Figure 7: The new sequence type experiment and the reference sequence.

5 Importing sequence read sets

Entries that we want to analyze using wgSNP need a sequence experiment that is obtained via mapping to the reference sequence, as this ensures collinearity of the sequences. For this, we need to start from the corresponding sequence read sets.

The preferred way of importing sequence read sets in BIONUMERICS is with the ***Import sequence read set files as links*** import routine (***File*** > ***Import...*** (📂, **Ctrl+I**)). Importing sequence read sets as links is only possible when the *WGS tools plugin* is installed in the BIONUMERICS database (***File*** > ***Install / remove plugins...*** (🔌)).

In our demonstration database, the *WGS tools plugin* is already installed and all read sets are imported as links:

1. Click on the green colored dot for one of the entries in the first column in the *Experiment presence* panel. Column 1 corresponds to the first experiment type listed in the *Experiment types* panel, which is **wgs** in the default configuration.

The data link is displayed in the *Sequence read set experiment* window (see Figure 8).

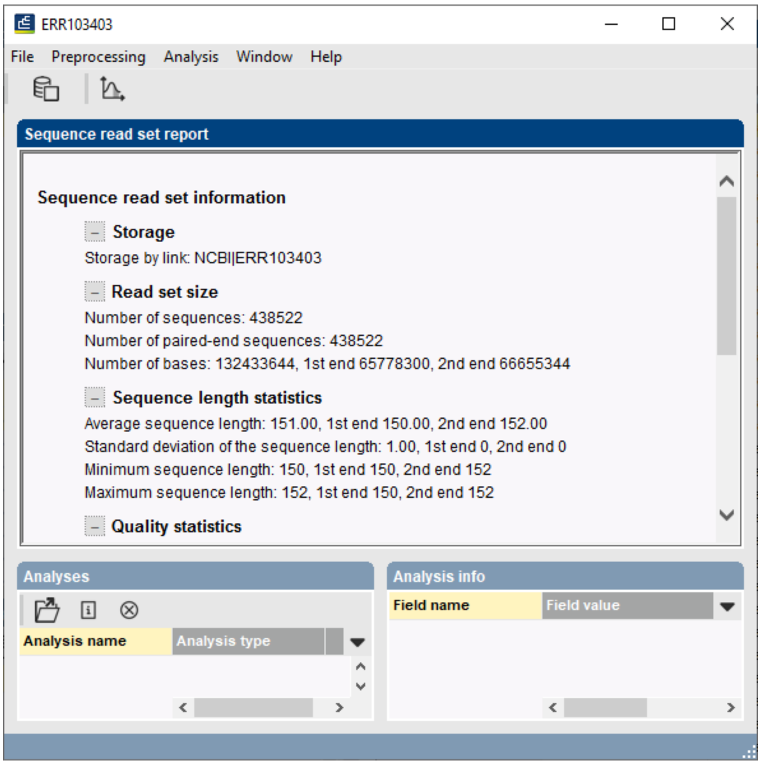
2. Close the *Sequence read set experiment* window.

Sequence read sets can also be stored inside the database with the ***Import sequence read set files*** import routine in the Import tree (***File*** > ***Import...*** (📂, **Ctrl+I**)). One disadvantage of this option is the storage size of the sequence read sets.

6 Mapping sequence reads against the reference sequence

1. Select a few entries in the *Database entries* panel to include in the SNP analysis, using the ballot boxes next to the entries. Selected entries are marked by a checked ballot box: ☑.

2. Select ***WGS tools*** > ***Submit jobs...*** (▷) to display the *Submit jobs* dialog box (see Figure 9).



Sequence read set report

Sequence read set information

Storage

Storage by link: NCBI|ERR103403

Read set size

Number of sequences: 438522
Number of paired-end sequences: 438522
Number of bases: 132433644, 1st end 65778300, 2nd end 66655344

Sequence length statistics

Average sequence length: 151.00, 1st end 150.00, 2nd end 152.00
Standard deviation of the sequence length: 1.00, 1st end 0, 2nd end 0
Minimum sequence length: 150, 1st end 150, 2nd end 152
Maximum sequence length: 152, 1st end 150, 2nd end 152

Quality statistics

Figure 8: Data link.

The ***Submit jobs to*** option deals with the location where the jobs will be sent to for execution: either to your ***Own computer*** or to the cloud ***Calculation Engine***.

3. Since this tutorial covers wgSNP with mapping performed on the local computer, make sure the ***Own computer*** option is selected.

4. Select the ***My wgSNP*** option as ***Reference mapping*** option (see Figure 9).

5. With the ***My wgSNP*** option highlighted, press the <***Settings***> button.

The only reference mapper available is ***SNAP***.

6. Close the *Reference mapping settings* dialog box.

Since no credits are required to run jobs on your own computer, this section in the *Submit jobs* dialog box can be ignored.

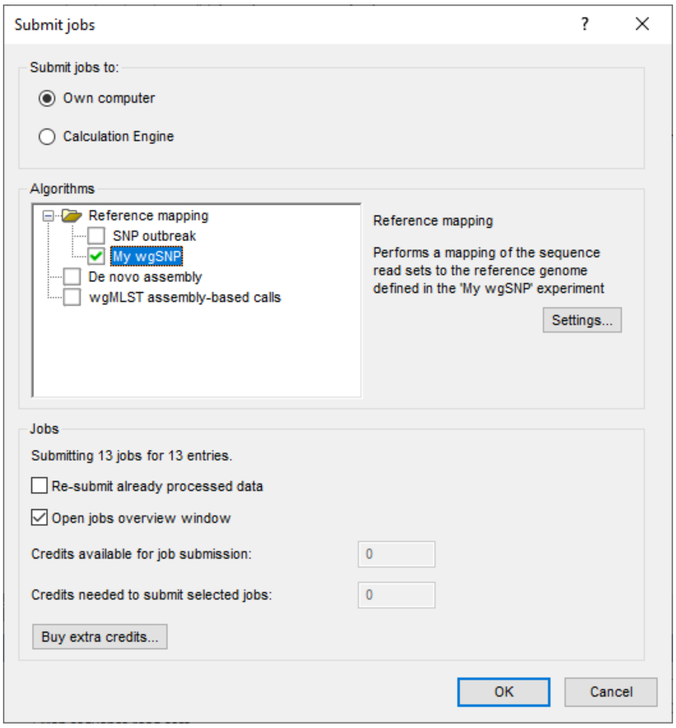
7. Press <***OK***> to launch the jobs.

When a reference mapping is already present for the submitted entries (and the ***Resubmit already processed data*** option was unchecked in the *Submit jobs* dialog box), an information message will appear, saying that no jobs are submitted.

The *Job overview* window will open, in which the status of the jobs can be followed (see Figure 11).

There are two options available in the *Job overview* window to import the job results in your BIONUMERICS database:

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).



Submit jobs

Submit jobs to:

● Own computer

○ Calculation Engine

Algorithms

- Reference mapping
  - □ SNP outbreak
  - ☑ My wgSNP
- □ De novo assembly
- □ wgMLST assembly-based calls

Reference mapping

Performs a mapping of the sequence read sets to the reference genome defined in the 'My wgSNP' experiment

Jobs

Submitting 13 jobs for 13 entries.

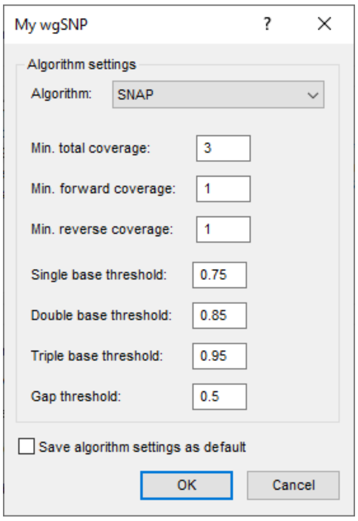
□ Re-submit already processed data

☑ Open jobs overview window

Credits available for job submission: 0

Credits needed to submit selected jobs: 0

Figure 9: The *Submit jobs* dialog box.



My wgSNP

Algorithm settings

Algorithm: SNAP

Min. total coverage: 3

Min. forward coverage: 1

Min. reverse coverage: 1

Single base threshold: 0.75

Double base threshold: 0.85

Triple base threshold: 0.95

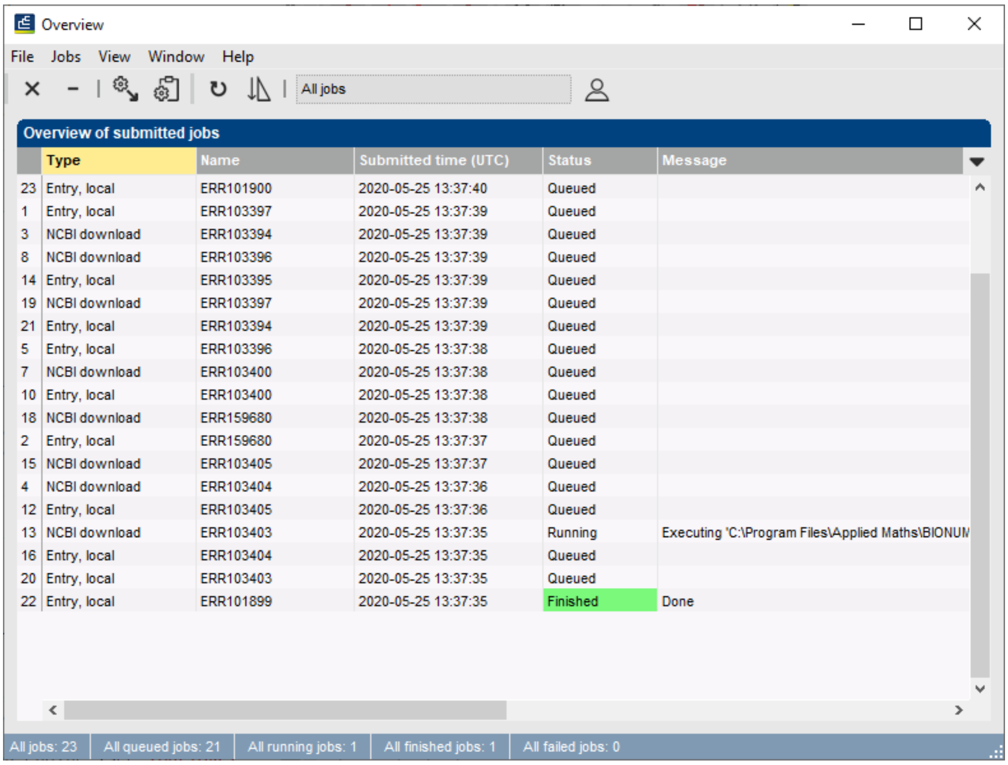
Gap threshold: 0.5

□ Save algorithm settings as default

Figure 10: Reference mapper.

9. Close the *Job overview* window.

The results of the mapping, i.e. the consensus sequences for the samples in the same frame as the reference sequence, are stored in the mapped sequence experiment type (here: **My wgSNP**).



Overview of submitted jobs

	Type	Name	Submitted time (UTC)	Status	Message
23	Entry, local	ERR101900	2020-05-25 13:37:40	Queued	
1	Entry, local	ERR103397	2020-05-25 13:37:39	Queued	
3	NCBI download	ERR103394	2020-05-25 13:37:39	Queued	
8	NCBI download	ERR103396	2020-05-25 13:37:39	Queued	
14	Entry, local	ERR103395	2020-05-25 13:37:39	Queued	
19	NCBI download	ERR103397	2020-05-25 13:37:39	Queued	
21	Entry, local	ERR103394	2020-05-25 13:37:39	Queued	
5	Entry, local	ERR103396	2020-05-25 13:37:38	Queued	
7	NCBI download	ERR103400	2020-05-25 13:37:38	Queued	
10	Entry, local	ERR103400	2020-05-25 13:37:38	Queued	
18	NCBI download	ERR159680	2020-05-25 13:37:38	Queued	
2	Entry, local	ERR159680	2020-05-25 13:37:37	Queued	
15	NCBI download	ERR103405	2020-05-25 13:37:37	Queued	
4	NCBI download	ERR103404	2020-05-25 13:37:36	Queued	
12	Entry, local	ERR103405	2020-05-25 13:37:36	Queued	
13	NCBI download	ERR103403	2020-05-25 13:37:35	Running	Executing 'C:\Program Files\Applied Maths\BIONUM
16	Entry, local	ERR103404	2020-05-25 13:37:35	Queued	
20	Entry, local	ERR103403	2020-05-25 13:37:35	Queued	
22	Entry, local	ERR101899	2020-05-25 13:37:35	Finished	Done

All jobs: 23 | All queued jobs: 21 | All running jobs: 1 | All finished jobs: 1 | All failed jobs: 0

Figure 11: Overview of local jobs.