

# 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).
- *Alibaba*: link to data uploaded to a client-specific data bucket hosted at the Alibaba OSS storage repository (see 6).

# 2 Preparing the database

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

Importing links to sequence read sets available on NCBI, EMBL-EBI, Amazon, BaseSpace or Alibaba is only possible when the *WGS tools plugin* is installed in the BIONUMERICS database:

- 2. Call the *Plugins* dialog box from the *Main* window with *File* > *Install* / *remove plugins...* (,,,).
- 3. Select the *WGS tools plugin* from the list in the *Applications tab* and press the *<Activate>* button.
- 4. Confirm the installation of the plugin.

The *Calculation engine URL* wizard page queries for the Uniform Resource Locator (URL) that uniquely identifies the calculation engine instance to connect to.

With the **Use default Cloud Calculation Engine** option clients will use the Applied Maths cloud instance (https://wgmlst.applied-maths.com), which is hosted on Amazon servers in the US. This option should also be selected if you do not intend to run jobs on the calculation engine, but instead run all calculations on your own computer.

5. Make sure the *Use default Cloud Calculation Engine* option is selected and press <*Next*>.

In the next step of the WGS tools installation wizard, two options are available:

- Choose *Local calculations only* if you do not intend to run jobs on the calculation engine and instead wish to run all calculations on your own computer.
- Choose *Enable running jobs on Cloud Calculation Engine* to unlock the full potential of the default Cloud Calculation Engine.

In this tutorial we will describe the install steps to run all calculations locally. Please consult the *WGS tools plugin* manual for more information about the *Enable running jobs on Cloud Calculation Engine* option.

 Make sure Local calculations only is checked, and select your organism from the Organism drop-down list (see Figure 1 for an example). If your organism is not listed, select the No organism option from the list.

:ttings		?	×
Calculation Er Organism	ngine and project		
Local calc	ulations only		
Run jobs o	n your own computer, the Calculation Engine connection is only us	ed to download search	data.
	ming jobs on cloud calculation Engine		
You will ne	ed your own project credentials and credits.		
You will ne Organism:	ed your own project credentials and credits.		
You will ne Organism: Project name:	ed your own project credentials and credits.     Staphylococcus aureus  >    LOCAL_saur  >		
You will ne Organism: Project name: Password:	ed your own project credentials and credits.     Staphylococcus aureus  ~    LOCAL_saur		
You will ne Organism: Project name: Password: Request a new	ed your own project credentials and credits.     Staphylococcus aureus  ~    LOCAL_saur		
You will ne Organism: Project name: Password: Request a new	ed your own project credentials and credits.     Staphylococcus aureus  ~    LOCAL_saur		

Figure 1: Calculation and organism settings.

7. Press < *Next*> to proceed with the installation.

If an organism was selected in the previous step, BIONUMERICS will download organism-specific settings and search data. A confirmation message pops up when the download is completed.

- 8. Press < OK > twice to finalize the installation of the plugin.
- 9. Press < *Exit* > to close the *Plugins* dialog box.
- 10. Close and reopen the database to activate the features of the WGS tools plugin.

After installation of the WGS tools plugin, sequence reads sets can now be imported as links.

# 3 Import links to NCBI or EMBL-EBI

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

Import	? ×
Import data Select the kind of data to import:	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.
1	Import Close

Figure 2: Import sequence read set data as links.

#### 3. Press <*Import*>.

Links to multiple data sources are available, including online and offline data repositories (see Figure 3).

Import sequence read sets			?	×
Select the data source Select the data source				
Download site: NCBI (SRA) EMBL-EBI (ENA) Amazon (S3) BaseSpace Local file server Alibaba OSS	Link to NCBI data in the SRA repository, based on Run accession number.			
	< Back	Next >	Cano	cel

Figure 3: Data sources.

4. 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.

3

5. Specify the accession number(s) (see Figure 4 for an example) and press < *Next*>.

Import sequence read sets			?	×
Input Select the data to import.				
Accession code(s):	ERR101899,ERR101900,ERR10339 4,ERR103404,ERR103405	^ ~		
Separation character:				
Pick up accession codes from field		∨ Fetch		
	< Back	Next >	Can	cel

Figure 4: Accession number(s).

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).

6. 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 (see Figure 5).

	-					
Source type	Source	Destination type	Destination			
Accession code	Accession code	Entry information	Кеу			
Fdit destination	7					
Edit destination						
Edit destination	]					
Edit destination						
Edit destination Preview	]					
Edit destination Preview						
Edit destination Preview						
Edit destination Preview Show advanced	options					
Edit destination Preview Show advanced	options					
Edit destination Preview Show advanced	options					
Edit destination Preview Show advanced	options					
Edit destination Preview ] Show advanced	options			Post	Manut	

Figure 5: Import rule.

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

- 8. Click *<Next>* and *<Finish>* to finish the creation of the import template.
- Enter a meaningful name (and optionally a description) for the created import template e.g. "Import from NCBI", and click < OK >.
- 10. Choose the newly created import template from the list and click < Next>.
- Select the created import template and a new or existing experiment from the drop-down list and press < *Next*> (see Figure 6).

mport sequence rea	ad sets			?	Х
Import template Specify how t	o import data into	the database.			
- Import templates:					
Import from NCBI	Import from NCBI			Create new	
				Edit	
				Preview	
				Сору	
Experiment type:	wgs	~			
			< Back	Next > Car	ncel

Figure 6: Import template.

- 12. When an experiment name is prompted for, specify a sequence type name, e.g. "wgs". Click < OK > and confirm the creation of the experiment.
- 13. Press < *Next* > to start the import of the sequence read set links.

In the last step, calculation jobs can be launched on the imported data links (*Open submit jobs dialog after import*). The same dialog can be called from the *Main* window at any time with *WGS tools* > *Submit jobs...* ( $\triangleright$ ).

14. 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**).

15. 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 7).

16. Close the Sequence read set experiment window.

### 4 Import links to Amazon

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.

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



Figure 7: Data link to NCBI.

- 2. 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 2).
- 3. Press <*Import*>.

Links to multiple data sources are available, including online and offline data repositories (see Figure 3).

- 4. Select *Amazon (S3)* and press < *Next*> to go to the next step.
- 5. 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* (see Figure 8).

Amazon S3 credentia	?	×		
Bucket:				
Access key ID:				
Secret access key:				
Save secret acces	ey.			
		ОК	Can	cel

Figure 8: Amazon (S3) credentials.

6. 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.

- 7. Check the files you wish to import and leave the option *Auto-detect paired-end read files* checked (see Figure 9).
- 8. Press <*Next*>.

Now you need to define how the data should be stored in the database. The default template **Ex-ample import** can be applied to most file names. This template will only retain the run accession



Figure 9: Files detected in the Amazon S3 bucket.

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

 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*>.
- 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 can be launched on the imported data links (*Open submit jobs dialog after import*). The same dialog can be called from the *Main* window at any time with *WGS tools* > *Submit jobs...* ( $\triangleright$ ).

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 10).

15. Close the Sequence read set experiment window.

# 5 Import links to BaseSpace

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



Figure 10: Data link to Amazon S3 bucket.

- 2. 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 2).
- 3. Press <*Import*>.

Links to multiple data sources are available, including online and offline data repositories (see Figure 3).

- 4. Select *BaseSpace* and press *<Next>* to go to the next step.
- 5. The first time you use this import routine, you need to login to your BaseSpace account and grant browse access (see Figure 11).





- 6. Provide your *Email address* and *Password* (see Figure 12). After authorization, close the browser window.
- 7. 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.
- 8. 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*.

9. You might for example want to link the *Project name* to an entry field and the *Sample name* to the *Key* field (see Figure 13 for these rules). Linking is done by double-clicking the row in the grid and selecting the correct destination from the tree.



Figure 12: Sign in.

ort template				?	×
mport rules Select import so	ources and the database	destinations			
Source type	Source	Destination type	Destination		
Project name	Project name	Entry information : Entry info field	Project name		
Sample name	Sample name	Entry information	Key		
Edit destination					
Edit destination Preview					
Edit destination Preview ] Show advanced	options				

Figure 13: Import rules: an example.

- 10. Optionally, you can do a preview of what you are about to import. Press < *Preview...* > to open the preview. Close the preview again.
- 11. Click < Next> and < Finish> to finish the creation of the import template.
- 12. 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.

- Make sure < *Create new*> is selected from the *Experiment type* list or select an existing experiment and press < *Next*>.
- 14. When an experiment name is prompted for, specify a sequence type name, e.g. "wgs". Click  $\langle OK \rangle$  and confirm the creation of the experiment.
- 15. Press *<Next>* to start the import of the sequence read set links.

In the last step, calculation jobs can be launched on the imported data links (*Open submit jobs dialog after import*). The same dialog can be called from the *Main* window at any time with *WGS tools* > *Submit jobs...* ( $\triangleright$ ).

 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**).

17. 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 14).

<b>E</b> 12-0171								-	×
File Preprocessing	Analysis	Window	Help						
6 h,									
Sequence read set	report								
Sequence read	set infor	mation							
- Storag	je								
Storage by	link: base	space 426	23635						
No statistics calcu	lated.								
Analyses					Analysis info				
[ <sup>7</sup> ] i ⊗					Field name	Fie	eld valu	е	•
Analysis name	Analysis	s type		-					
						<			>

Figure 14: Data link to BaseSpace.

18. Close the Sequence read set experiment window.

## 6 Import links to Alibaba OSS

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

- 2. 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 2).
- 3. Press <*Import*>.

Links to multiple data sources are available, including online and offline data repositories (see Figure 3).

- 4. Select *Alibaba OSS* and press < *Next* > to go to the next step.
- 5. The first time you use this import routine, you need to specify your Alibaba OSS credentials: *Bucket name*, *Access key ID*, *Secret access key* and *OSS Region* (see Figure 15).

Alibaba OSS credenti		?	×	
Bucket:				
Access key ID:				
Secret access key:				
OSS Region:				
Save secret acces	s key.			
		ОК	Cano	el

Figure 15: Alibaba OSS credentials.

6. 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 (see Figure 16).

Import sequenc	e read sets	?	×
Input Select the	e data to import.		
File path(s):	382634_SRR8792696_Ecoli_1.fastq.gz 382634_SRR8792696_Ecoli_2.fastq.gz		
	, ☑ Auto-detect paired-end read files		
	< Back Next >	Cano	el

Figure 16: Files detected in the Alibaba OSS bucket.

- 7. Check the files you wish to import and leave the option *Auto-detect paired-end read files* checked.
- 8. 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 stored the (parsed) file

names 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*>.
- 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 can be launched on the imported data links (*Open submit jobs dialog after import*). The same dialog can be called from the *Main* window at any time with *WGS tools* > *Submit jobs...* ( $\triangleright$ ).

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 17).

🖆 382634_SRR879269	96_Ecoli				-	×
File Preprocessing	Analysis Window	Help				
Ē 14,						
Sequence read set	report					
Sequence read	set information					^
- Storag	<b>je</b> link: 088l382634_8	RR8792696	Ecoli 1 fasto oz			
OSS 38263 No statistics calcu	84_SRR8792696_E0	coli_2.fastq.gz				~
Analyses			Analysis info			
[ <sup>77</sup> ] i ⊗			Field name	Field val	lue	-
Analysis name	Analysis type	-				
				<		>

Figure 17: Data link to Alibaba OSS bucket.

15. Close the Sequence read set experiment window.

## 7 Analysis tools

Analysis tools are covered in following tutorials:

- "Performing a de novo assembly on the local calculation engine"
- "Performing a de novo assembly on the external calculation engine"
- "Performing whole genome SNP analysis with mapping performed on the local calculation engine"
- "Performing whole genome SNP analysis with mapping performed on the cloud calculation engine"
- "wgMLST typing: routine workflow starting from sequence read sets"