BIONUMERICS Tutorial:

Fast character-based identification

1 Introduction

The BIONUMERICS software offers a tool for screening an entry against the database, based upon a character type experiment. This identification tool benefits from a bulk-fetching mechanism, which makes it many times faster for identification against large databases.

This tutorial describes how to perform such character-based identification of entries.

2 Preparing the database

2.1 Introduction to the demonstration database

We provide a **WGS demo database** for *Listeria monocytogenes* containing sequence read set data links for 51 samples, calculated de novo assemblies and wgMLST results (allele calls and quality information).

✎ The *de novo* assembly, wgMLST workflow and results will not be discussed in this tutorial.

The **WGS demo database** for *Listeria monocytogenes* can be downloaded directly from the *BIONUMERICS Startup* window (see 2.2), or restored from the back-up file available on our website (see 2.3).

2.2 Option 1: Download the demo database from the Startup screen

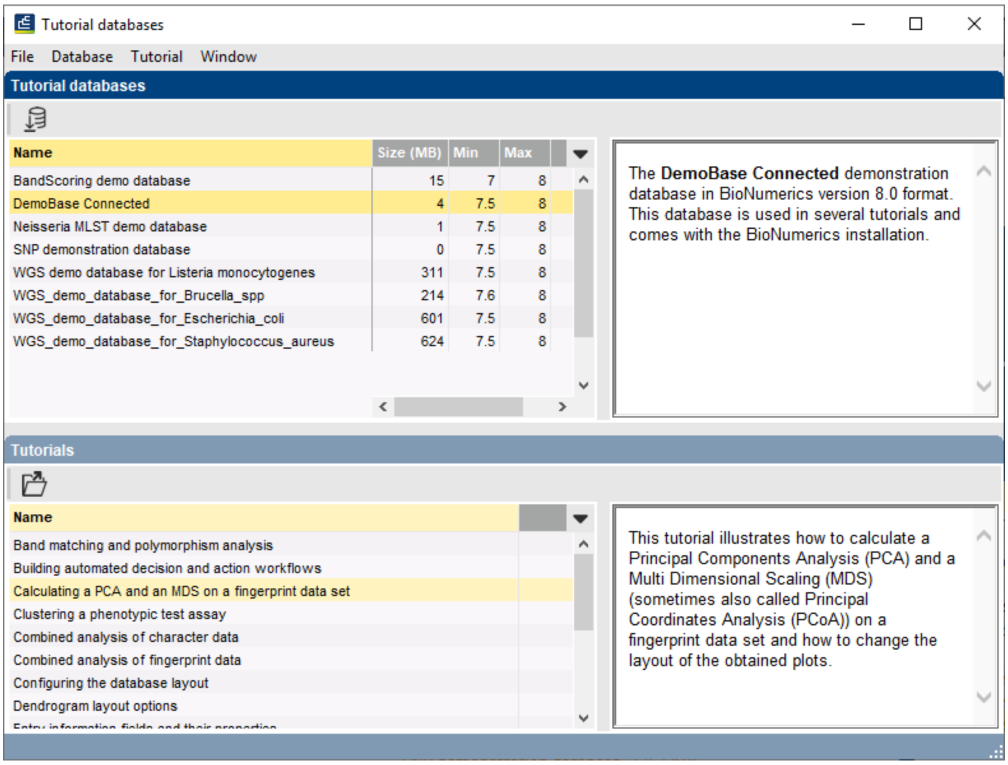
1. Click the ↓ button, located in the toolbar in the *BIONUMERICS Startup* window.

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

2. Select the **WGS_demo_database_for_Listeria_monocytogenes** 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***.



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.

The **WGS_demo_database_for_Listeria_monocytogenes** appears in the *BIONUMERICS Startup* window.

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

2.3 Option 2: Restore the demo database from back-up file

A BIONUMERICS back-up file of the **WGS demo database** for *Listeria monocytogenes* is also available on our website. This backup can be restored to a functional database in BIONUMERICS.

6. Download the file `WGS_LM01.bnbk` file from https://www.bionumerics.com/download/sample-data, under 'WGS_demo_database_for_Listeria_monocytogenes'.

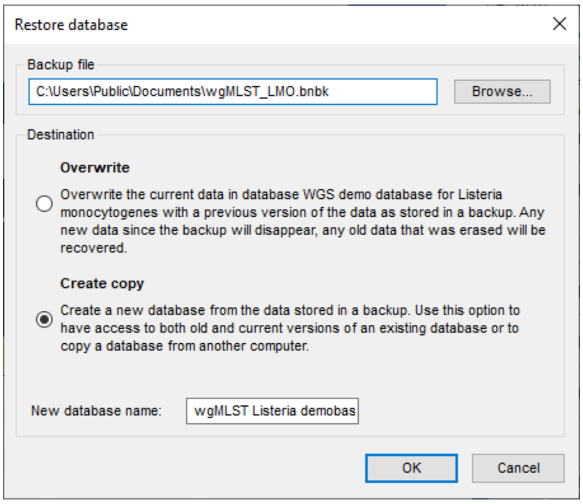
In contrast to other browsers, some versions of Internet Explorer rename the `WGS_LM01.bnbk` database backup file into `WGS_LM01.zip`. If this happens, you should manually remove the `.zip` file extension and replace with `.bnbk`. A warning will appear ("If you change a file name extension, the file 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 and make sure the name does not contain any spaces to ensure the successful installation of the *Listeria functional genotyping plugin*. Specify for example: “WGS_Listeria_demobase”.

10. Click <***OK***> to start restoring the database from the backup file (see Figure 2).



Restore database

Backup file

C:\Users\Public\Documents\wgMLST_LMO.bnbk

Destination

Overwrite

○ Overwrite the current data in database WGS demo database for Listeria monocytogenes 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

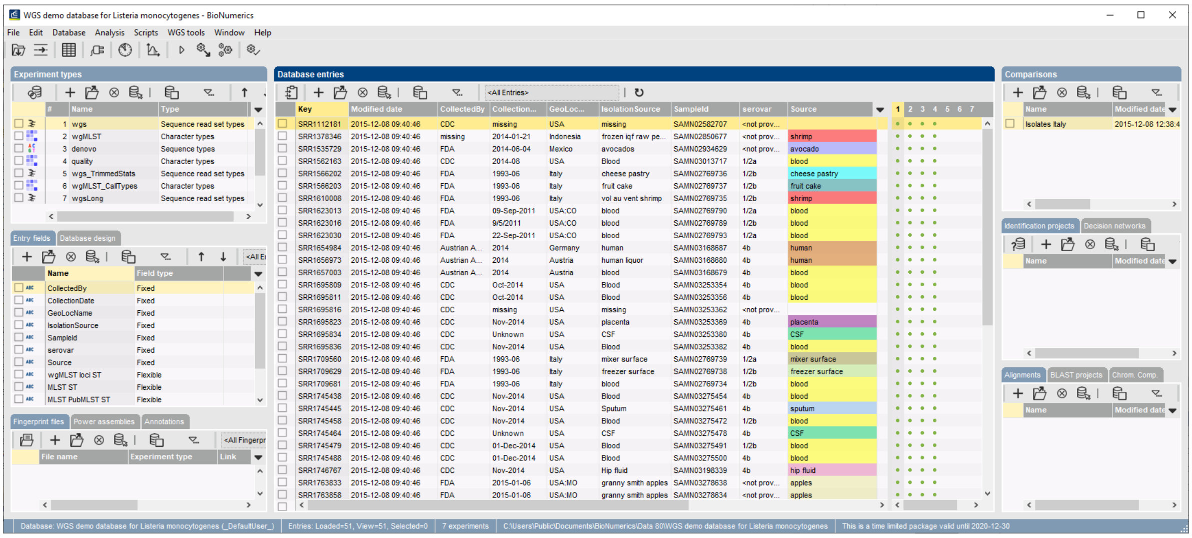
● Create a new database from the data stored in a backup. Use this option to have access to both old and current versions of an existing database or to copy a database from another computer.

New database name: wgMLST Listeria demobas

Figure 2: Restoring the **WGS demonstration database** from the backup file `WGS_LMO1.bnbk`.

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

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



WGS demo database for Listeria monocytogenes - BioNumerics

Experiment types

#	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	wgsLong	Sequence read set types

Entry fields

Name	Field type
CollectedBy	Fixed
CollectionDate	Fixed
GeoLocName	Fixed
IsolationSource	Fixed
SampleId	Fixed
serovar	Fixed
Source	Fixed
wgMLST loci ST	Flexible
MLST ST	Flexible
MLST PubMLST ST	Flexible

Fingerprint files

File name	Experiment type	Link

Database entries

Key	Modified date	CollectedBy	Collection...	GeoLoc...	IsolationSource	SampleId	serovar	Source	1	2	3	4	5	6	7
SRR1112181	2015-12-08 09:40:46	CDC	missing	USA	missing	SAMN02582707	<not prov...		●	●	●	●			
SRR1378346	2015-12-08 09:40:46	missing	2014-01-21	Indonesia	frozen iqf raw pe...	SAMN02850677	<not prov...	shrimp	●	●	●	●			
SRR1535729	2015-12-08 09:40:46	FDA	2014-06-04	Mexico	avocados	SAMN02934629	<not prov...	avocado	●	●	●	●			
SRR1562163	2015-12-08 09:40:46	CDC	2014-08	USA	Blood	SAMN03013717	1/2a	blood	●	●	●	●			
SRR1566202	2015-12-08 09:40:46	FDA	1993-06	Italy	cheese pastry	SAMN02769736	1/2b	cheese pastry	●	●	●	●			
SRR1566203	2015-12-08 09:40:46	FDA	1993-06	Italy	fruit cake	SAMN02769737	1/2b	fruit cake	●	●	●	●			
SRR1610008	2015-12-08 09:40:46	FDA	1993-06	Italy	vol au vent shrimp	SAMN02769735	1/2b	shrimp	●	●	●	●			
SRR1623013	2015-12-08 09:40:46	FDA	09-Sep-2011	USA:CO	blood	SAMN02769790	1/2a	blood	●	●	●	●			
SRR1623016	2015-12-08 09:40:46	FDA	9/5/2011	USA:CO	blood	SAMN02769789	1/2b	blood	●	●	●	●			
SRR1623030	2015-12-08 09:40:46	FDA	22-Sep-2011	USA:CO	blood	SAMN02769793	1/2a	blood	●	●	●	●			
SRR1654984	2015-12-08 09:40:46	Austrian A...	2014	Germany	human	SAMN03168687	4b	human	●	●	●	●			
SRR1656973	2015-12-08 09:40:46	Austrian A...	2014	Austria	human liquor	SAMN03168680	4b	human	●	●	●	●			
SRR1657003	2015-12-08 09:40:46	Austrian A...	2014	Austria	blood	SAMN03168679	4b	blood	●	●	●	●			
SRR1695809	2015-12-08 09:40:46	CDC	Oct-2014	USA	Blood	SAMN03253354	4b	blood	●	●	●	●			
SRR1695811	2015-12-08 09:40:46	CDC	Oct-2014	USA	Blood	SAMN03253356	4b	blood	●	●	●	●			
SRR1695816	2015-12-08 09:40:46	CDC	missing	USA	missing	SAMN03253362	<not prov...		●	●	●	●			
SRR1695823	2015-12-08 09:40:46	CDC	Nov-2014	USA	placenta	SAMN03253369	4b	placenta	●	●	●	●			
SRR1695834	2015-12-08 09:40:46	CDC	Unknown	USA	CSF	SAMN03253380	4b	CSF	●	●	●	●			
SRR1695836	2015-12-08 09:40:46	CDC	Nov-2014	USA	Blood	SAMN03253382	4b	blood	●	●	●	●			
SRR1709560	2015-12-08 09:40:46	FDA	1993-06	Italy	mixer surface	SAMN02769739	1/2a	mixer surface	●	●	●	●			
SRR1709629	2015-12-08 09:40:46	FDA	1993-06	Italy	freezer surface	SAMN02769738	1/2b	freezer surface	●	●	●	●			
SRR1709681	2015-12-08 09:40:46	FDA	1993-06	Italy	blood	SAMN02769734	1/2b	blood	●	●	●	●			
SRR1745438	2015-12-08 09:40:46	CDC	Nov-2014	USA	Blood	SAMN03275454	4b	blood	●	●	●	●			
SRR1745445	2015-12-08 09:40:46	CDC	Nov-2014	USA	Sputum	SAMN03275461	4b	sputum	●	●	●	●			
SRR1745458	2015-12-08 09:40:46	CDC	Nov-2014	USA	Blood	SAMN03275472	1/2b	blood	●	●	●	●			
SRR1745464	2015-12-08 09:40:46	CDC	Unknown	USA	CSF	SAMN03275478	4b	CSF	●	●	●	●			
SRR1745479	2015-12-08 09:40:46	CDC	01-Dec-2014	USA	Blood	SAMN03275491	1/2b	blood	●	●	●	●			
SRR1745488	2015-12-08 09:40:46	CDC	01-Dec-2014	USA	Blood	SAMN03275500	4b	blood	●	●	●	●			
SRR1746767	2015-12-08 09:40:46	CDC	Nov-2014	USA	Hip fluid	SAMN03198339	4b	hip fluid	●	●	●	●			
SRR1763833	2015-12-08 09:40:46	FDA	2015-01-06	USA:MO	granny smith apples	SAMN03278638	<not prov...	apples	●	●	●	●			
SRR1763858	2015-12-08 09:40:46	FDA	2015-01-06	USA:MO	granny smith apples	SAMN03278634	<not prov...	apples	●	●	●	●			

Comparisons

Name	Modified date
Isolates Italy	2015-12-08 12:38:4

Identification projects

Name	Modified date

Alignments

Name	Modified date

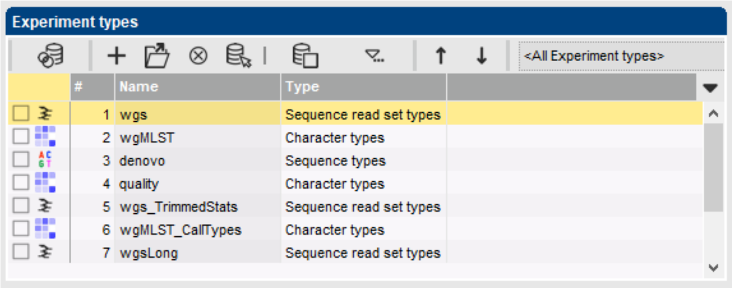
Database: WGS demo database for Listeria monocytogenes (_DefaultUser_) | Entries: Loaded=51, View=51, Selected=0 | 7 experiments | C:\Users\Public\Documents\BioNumerics\Data 80\WGS demo database for Listeria monocytogenes | This is a time limited package valid until 2020-12-30

Figure 3: The *Listeria monocytogenes* demonstration database: the *Main* window.

3 About the demonstration database

The WGS demo database contains links to sequence read set data on NCBI’s sequence read archive (SRA) for 51 publicly available sequencing runs. Additional information, stored in entry info fields (CollectionDate, CollectedBy, serovar, etc.) was collected from the corresponding publications and added to the demo database.

Seven experiments are present in the demo database and are listed in the *Experiment types* panel (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	wgsLong	Sequence read set types

Figure 4: The *Experiment types* panel in the *Main* window.

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 *Sequence read set experiment* window, the link to the sequence read set data on NCBI (SRA) with a summary of the characteristics of the sequence read set is displayed: *Read set size*, *Sequence length statistics*, *Quality statistics*, *Base statistics* (see Figure 5).

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

3. Click on the green colored dot for one of the entries in the third column in the *Experiment presence* panel. Column 3 corresponds to the third experiment type listed in the *Experiment types* panel, which is **denovo**.

The *Sequence editor* window opens, containing the results from the de novo assembly algorithm, i.e. concatenated de novo contig sequences (see Figure 6).

4. Close the *Sequence editor* window.

The sequence read set experiment type **wgs_TrimmedStats** contains some data statistics about the reads retained after trimming, used for the de novo assembly.

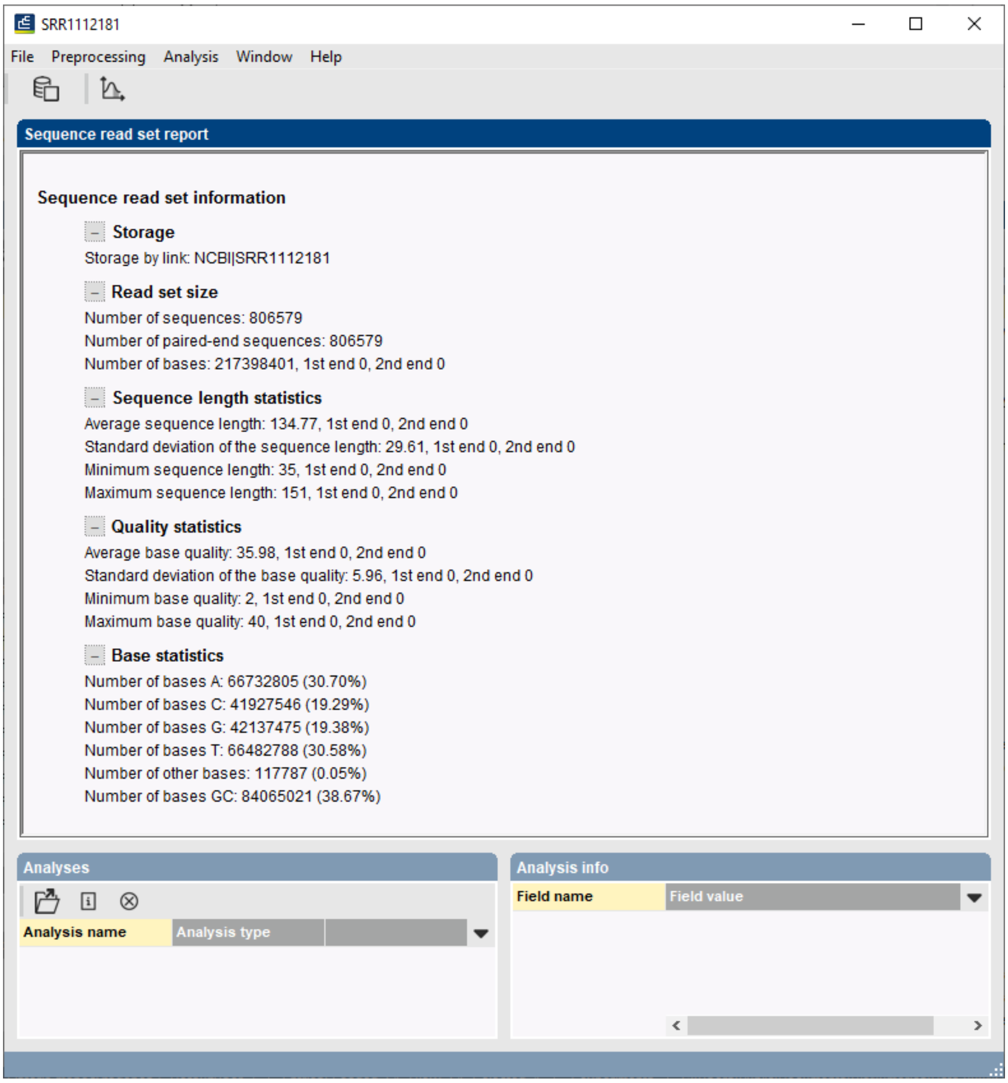
The sequence read set experiment type **wgsLong** contains the links to long read sequence read data (typically PacBio or MinION datasets). In this demo database, no links are defined for this experiment.

The other three experiments contain data related to the wgMLST analysis performed on the samples:

- Character experiment type **wgMLST** contains the allele calls for detected loci in each sample, where the consensus from assembly-based and assembly-free calling resulted in a single allele ID.

- Character experiment type **quality** contains quality statistics for the raw data, the de novo assembly and the different allele identification algorithms.

- Character experiment type **wgMLST_CallTypes**: contains details on the call types.



Sequence read set information

Storage

Storage by link: NCBI|SRR1112181

Read set size

Number of sequences: 806579
Number of paired-end sequences: 806579
Number of bases: 217398401, 1st end 0, 2nd end 0

Sequence length statistics

Average sequence length: 134.77, 1st end 0, 2nd end 0
Standard deviation of the sequence length: 29.61, 1st end 0, 2nd end 0
Minimum sequence length: 35, 1st end 0, 2nd end 0
Maximum sequence length: 151, 1st end 0, 2nd end 0

Quality statistics

Average base quality: 35.98, 1st end 0, 2nd end 0
Standard deviation of the base quality: 5.96, 1st end 0, 2nd end 0
Minimum base quality: 2, 1st end 0, 2nd end 0
Maximum base quality: 40, 1st end 0, 2nd end 0

Base statistics

Number of bases A: 66732805 (30.70%)
Number of bases C: 41927546 (19.29%)
Number of bases G: 42137475 (19.38%)
Number of bases T: 66482788 (30.58%)
Number of other bases: 117787 (0.05%)
Number of bases GC: 84065021 (38.67%)

Figure 5: The sequence read set experiment card for an entry.

4 Fast identification based on characters

4.1 General identification

We will use the allele calls of the wgMLST profile stored in the **wgMLST** character experiment of entry 'SRR1695834' to screen for similar wgMLST profiles in the database.

1. In the *Main* window, click on ***Edit*** > ***Find object in list...*** (, **Ctrl+Shift+F**).

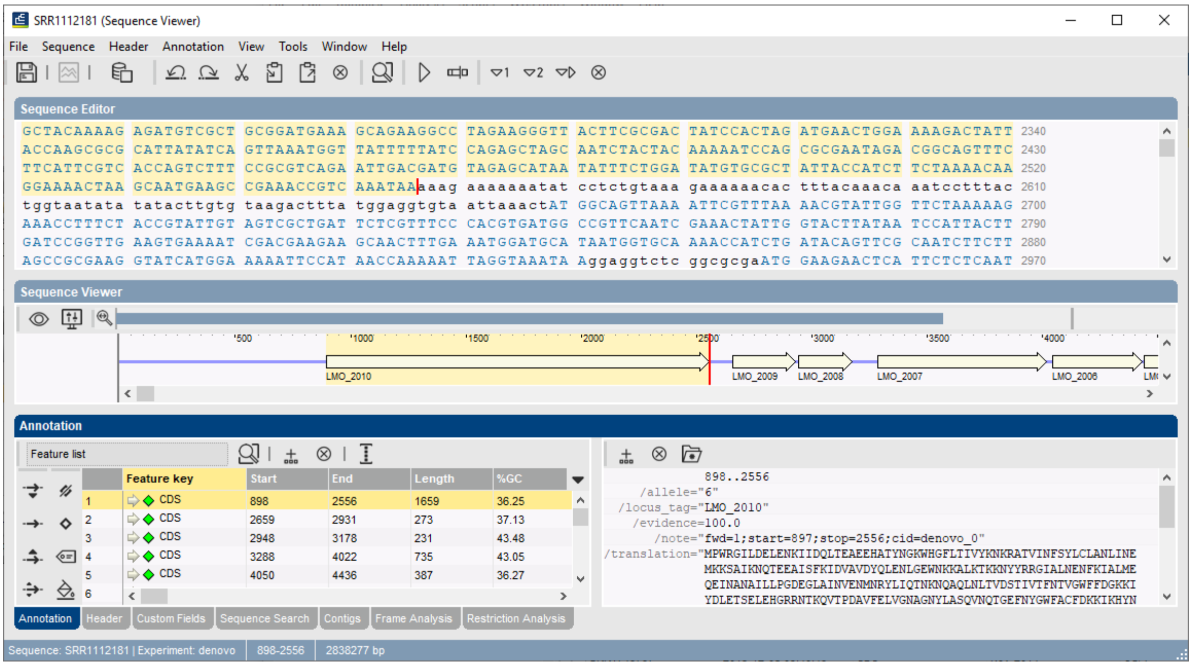
2. Type 'SRR1695834' (without the quotes) in the text field of the *Find* dialog box and click <***Select all***>.

The entry 'SRR1695834' is now selected in the database.

3. Click on the green dot corresponding with the **wgMLST** character experiment for the selected entry 'SRR1695834'.

The character experiment card, containing the different allele calls for all the wgMLST loci will open.

4. Close the experiment card by clicking on the triangle in the upper left corner of the card.



SRR1112181 (Sequence Viewer)
File Sequence Header Annotation View Tools Window Help
Sequence Editor

```
GCTACAAAAG AGATGTCGCT GCGGATGAAA GCAGAAGGCC TAGAAGGGTT ACTTCGCGAC TATCCACTAG ATGAACTGGA AAAGACTATT 2340
ACCAAGCGCG CATTATATCA GTTAAATGGT TATTTTTATC CAGAGCTAGC AATCTACTAC AAAAATCCAG CGCGAATAGA CGGCAGTTTC 2430
TTCATTCGTC ACCAGTCTTT CCGCGTCAGA ATTGACGATG TAGAGCATAA TATTTCTGGA TATGTGCGCT ATTACCATCT TCTAAAACAA 2520
GGAAAACTAA GCAATGAAGC CGAAACCGTC AAATAAaaag aaaaaaatat cctctgtaaa gaaaaaacac tttacaaaca aatcctttac 2610
tggtaatata tatacttgtg taagacttta tggaggtgta attaaactAT GGCAGTTAAA ATTCGTTTAA AACGTATTGG TTCTAAAAAG 2700
AAACCTTTCT ACCGTATTGT AGTCGCTGAT TCTCGTTTCC CACGTGATGG CCGTTCAATC GAAACTATTG GTACTTATAA TCCATTACTT 2790
GATCCGGTTG AAGTGAAAAT CGACGAAGAA GCAACTTTGA AATGGATGCA TAATGGTGCA AAACCATCTG ATACAGTTCG CAATCTTCTT 2880
AGCCGCGAAG GTATCATGGA AAAATTCCAT AACCAAAAAT TAGGTAAATA Aggaggtctc ggcgcgaATG GAAGAACTCA TTCTCTCAAT 2970
```

	Feature key	Start	End	Length	%GC
1	CDS	898	2556	1659	36.25
2	CDS	2659	2931	273	37.13
3	CDS	2948	3178	231	43.48
4	CDS	3288	4022	735	43.05
5	CDS	4050	4436	387	36.27
6					

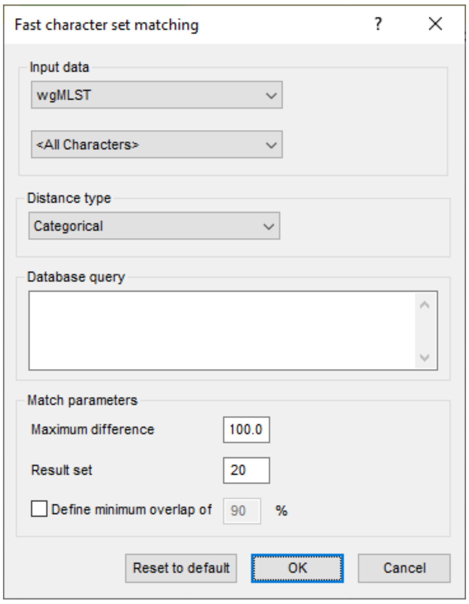
```
             898..2556
     /allele="6"
  /locus_tag="LMO_2010"
    /evidence=100.0
       /note="fwd=1;start=897;stop=2556;cid=denovo_0"
/translation="MPWRGILDELENKIIDQLTEAEEHATYNGKWHGFLTIVYKNKRATVINFSYLCLANLINE
             MKKSAIKNQTEEAISFKIDVAVDYQLENLGEWNKKALKTKKNYYRRGIALNENFKIALME
             QEINANAILLPGDEGLAINVENMNRYLIQTNKNQAQLNLTVDSTIVTFNTVGWFFDGKKI
             YDLETSELEHGRRNTKQVTPDAVFELVGNAGNYLASQVNQTGEFNYGWFACFDKKIKHYN
```

Sequence: SRR1112181 | Experiment: denovo | 898-2556 | 2838277 bp

Figure 6: The *Sequence editor* window.

5. Choose ***Analysis*** > ***Fast matching*** > ***Fast character matching...*** in the *Main* window.

The *Fast character set matching* dialog box will open (see Figure 7).



Input data: wgMLST

<All Characters>

Distance type: Categorical

Database query:

Match parameters

Maximum difference: 100.0

Result set: 20

□ Define minimum overlap of 90 %

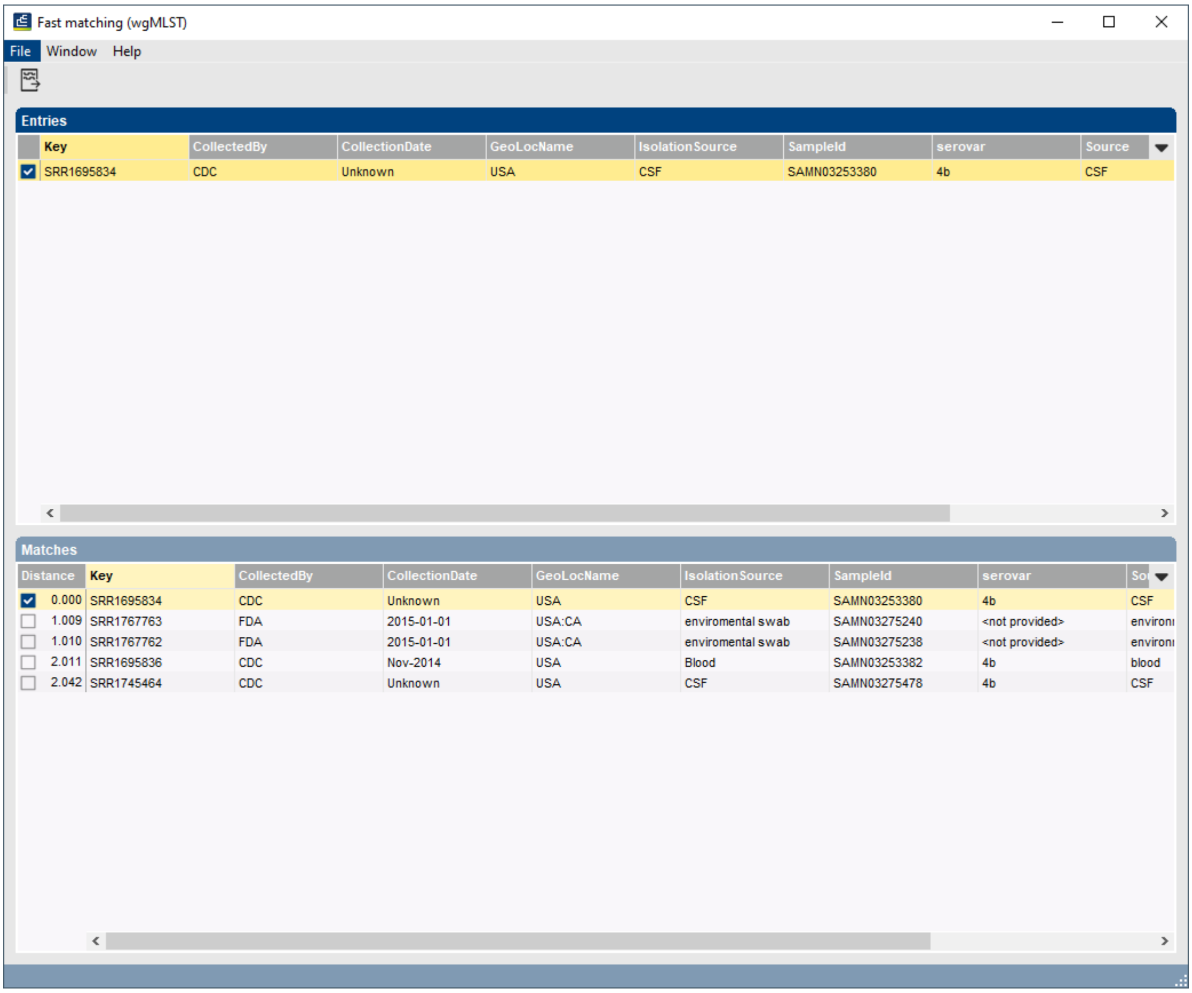
Figure 7: The *Fast character set matching* dialog box.

6. Change the **Distance type** to **Categorical**.

7. Leave the other parameters at their default values and click <***OK***>.

More information on the other parameters can be found in the reference manual.

The *Fast matching* window will open showing the results of the screening (see Figure 8). In the *Entries* panel you will see the entries and their corresponding information fields that were used to screen the database. In this case, we only used one entry (entry ’SRR1695834’) for the screening. If you would use multiple entries, all these entries will be shown here.



Entries

	Key	CollectedBy	CollectionDate	GeoLocName	IsolationSource	SampleId	serovar	Source
☑	SRR1695834	CDC	Unknown	USA	CSF	SAMN03253380	4b	CSF

Matches

Distance	Key	CollectedBy	CollectionDate	GeoLocName	IsolationSource	SampleId	serovar	So
☑ 0.000	SRR1695834	CDC	Unknown	USA	CSF	SAMN03253380	4b	CSF
□ 1.009	SRR1767763	FDA	2015-01-01	USA:CA	enviromental swab	SAMN03275240	<not provided>	environ
□ 1.010	SRR1767762	FDA	2015-01-01	USA:CA	enviromental swab	SAMN03275238	<not provided>	environ
□ 2.011	SRR1695836	CDC	Nov-2014	USA	Blood	SAMN03253382	4b	blood
□ 2.042	SRR1745464	CDC	Unknown	USA	CSF	SAMN03275478	4b	CSF

Figure 8: The *Fast character set matching* dialog box.

8. Click on the entry ’SRR1695834’ to show its matches.

In the *Matches* panel of the *Fast matching* window the matches for the selected entry are shown. In this case 5 matches should be present, one of which is a 100% match with entry ’SRR1695834’ which we used for the screening (see Figure 8). The column **Distance** indicates the difference in distance between your entry and the match. In this particular case, this number can be interpreted as the number of allelic differences between the wgMLST profiles.

By looking at the values in the **CollectedBy** column we can see that matches with strains collected by both the CDC and FDA were detected (see Figure 8). In the next part of this tutorial you will learn to limit the search for matches to a certain subset of entries that have a certain (combination of) information field value(s) (see 4.2).

The results of the *Fast matching* window can also be exported to a comma separated text file.

9. Choose ***File*** > ***Export...*** (🖫) to export the results to text format.

10. Close the *Fast matching* window.

4.2 Limiting identification to a certain subset of entries in the database

This part of the tutorial will describe how you can limit the search for matches to a certain subset of entries that have a certain (combination of) information field value(s) by using a SQL statement. We will continue working on the previous example and try to obtain only matches of strains which were collected by the FDA.

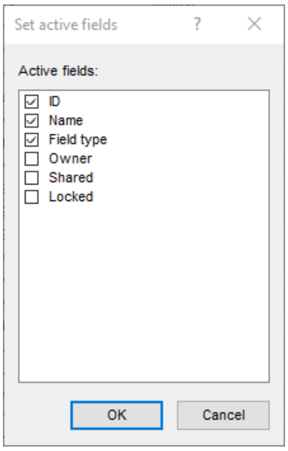
The SQL query will work only if you use the information field ID instead of the field name.

The information field IDs can be obtained from the *Entry fields* panel in the *Main* window window.

11. Click on the downward arrow to the right of the table headers in the *Entry fields* panel.

12. From the menu, select ***'Set active fields...'*** or press **Ctrl+F5**.

The *Set active fields* dialog box dialog will open (see Figure 9).



Set active fields

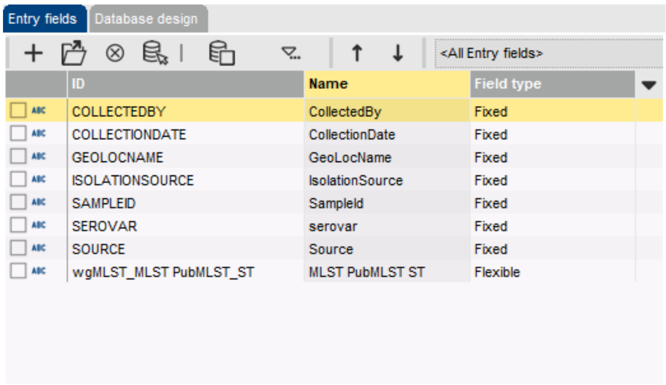
Active fields:

☑ ID
☑ Name
☑ Field type
☐ Owner
☐ Shared
☐ Locked

Figure 9: The *Set active fields* dialog box.

13. Check the box in front of **ID** (see Figure 9) and press <***OK***> to close the *Set active fields* dialog box.

The information field ID will now be visible in the *Entry fields* panel (see Figure 10). The ID of the **CollectedBy** information field is **COLLECTEDBY**.



	ID	Name	Field type
☐	COLLECTEDBY	CollectedBy	Fixed
☐	COLLECTIONDATE	CollectionDate	Fixed
☐	GEOLOCNAME	GeoLocName	Fixed
☐	ISOLATIONSOURCE	IsolationSource	Fixed
☐	SAMPLEID	SampleId	Fixed
☐	SEROVAR	serovar	Fixed
☐	SOURCE	Source	Fixed
☐	wgMLST_MLST PubMLST_ST	MLST PubMLST ST	Flexible

Figure 10: The 'ID' field is now active in the *Entry fields* panel.

14. Make sure entry 'SRR1695834' is still selected in the *Main* window.

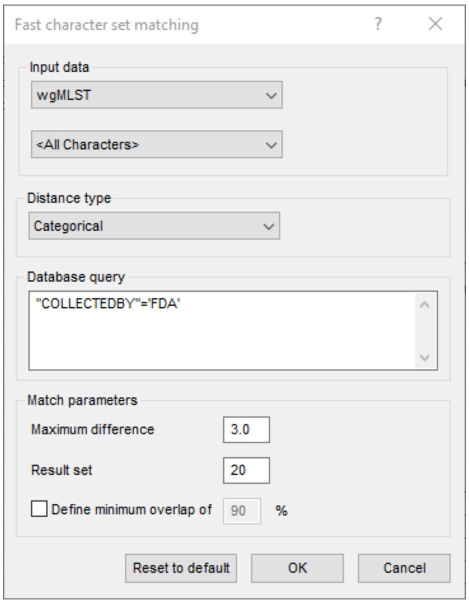
15. Choose ***Analysis*** > ***Fast matching*** > ***Fast character matching...*** in the *Main* window.

The *Fast character set matching* dialog box will open (see Figure 7).

16. Change the **Distance type** to **Categorical**.

17. In the **Database query** field, type the SQL statement ”COLLECTEDBY”=’FDA’ (see Figure 11).

Two SQL statements can be combined by ’AND’, when both statements should be fulfilled, or ’OR’, when either statement should be fulfilled.



Fast character set matching

Input data

wgMLST

<All Characters>

Distance type

Categorical

Database query

"COLLECTEDBY"='FDA'

Match parameters

Maximum difference 3.0

Result set 20

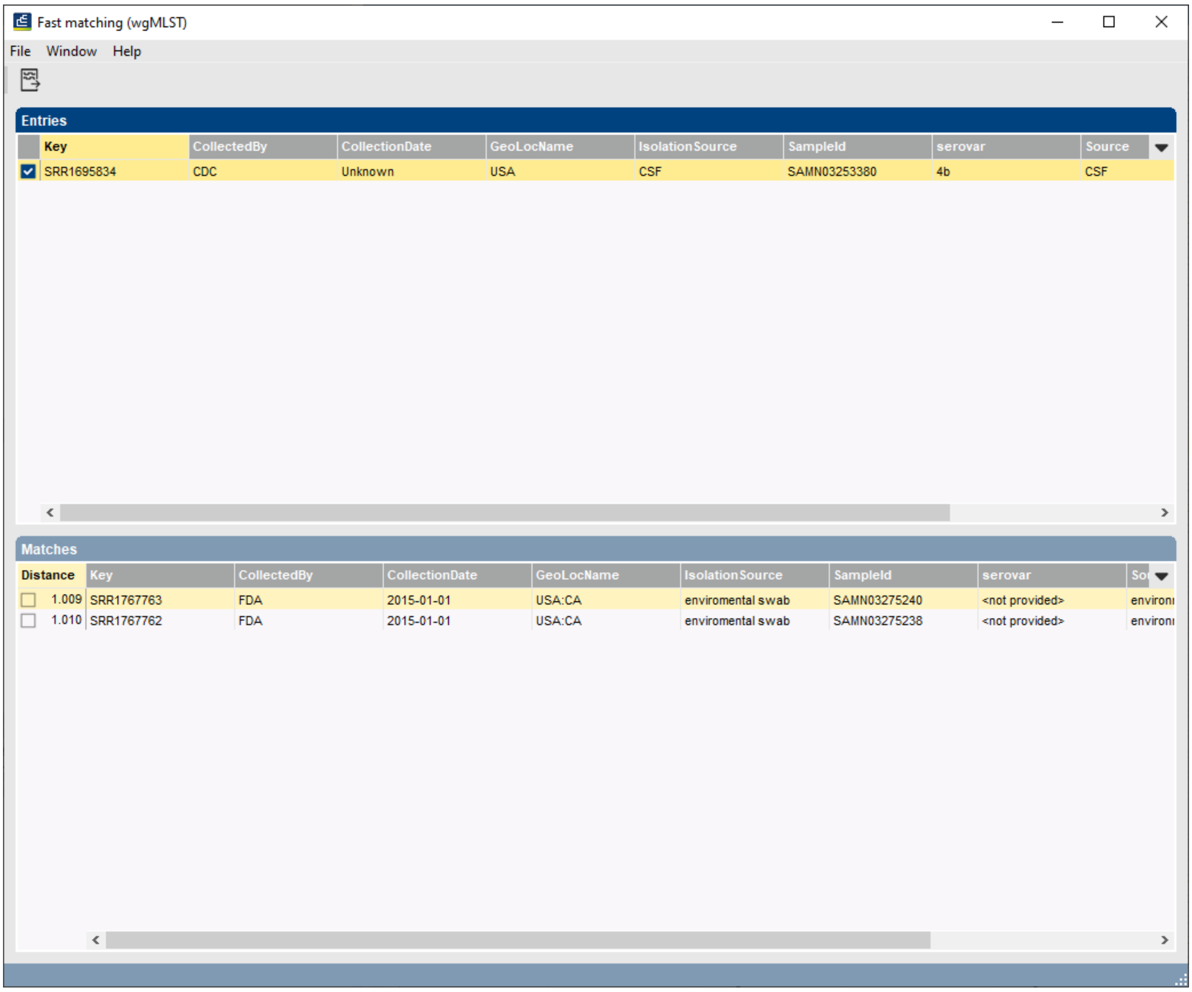
□ Define minimum overlap of 90 %

Figure 11: The SQL statement in the *Fast character set matching* dialog box.

18. Press <***OK***> to execute the fast matching.

The *Fast matching* window will open showing the results of the screening (see Figure 12). You will notice that now only 2 matches were found, both of strains which were collected by the FDA.

19. Close the *Fast matching* window.



	Key	CollectedBy	CollectionDate	GeoLocName	IsolationSource	SampleId	serovar	Source
☑	SRR1695834	CDC	Unknown	USA	CSF	SAMN03253380	4b	CSF

Distance		Key	CollectedBy	CollectionDate	GeoLocName	IsolationSource	SampleId	serovar	So
□	1.009	SRR1767763	FDA	2015-01-01	USA:CA	enviromental swab	SAMN03275240	<not provided>	environ
□	1.010	SRR1767762	FDA	2015-01-01	USA:CA	enviromental swab	SAMN03275238	<not provided>	environ

Figure 12: The expected results in the *Fast matching* window.