BIONUMERICS Tutorial:

Minhashing based cluster analysis of sequences

1 Aim

Similarity and distance-based trees (e.g. UPGMA and Neighbor joining) and phylogenetic trees (e.g. Maximum likelihood and Maximum parsimony trees) can be calculated in the *Comparison* window in BIONUMERICS, reflecting the relationships between the analyzed sequences. This tutorial covers clustering of sequences using a minhashing based calculated similarity.

MinHash techniques allow the comparison of large datasets of genomic sequences which is currently infeasible with alignment based approaches. During minhashing the k-mer content of each sequence is first determined and each k-mer is then passed through a hash function to obtain hashes. Retaining the lowest hashes enables the sampling of a random set of k-mers, which is called a ”sketch” or ”MinHash signature”. Using only these signatures the similarity of the original sequences can be compared in a rapid and accurate way.

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).

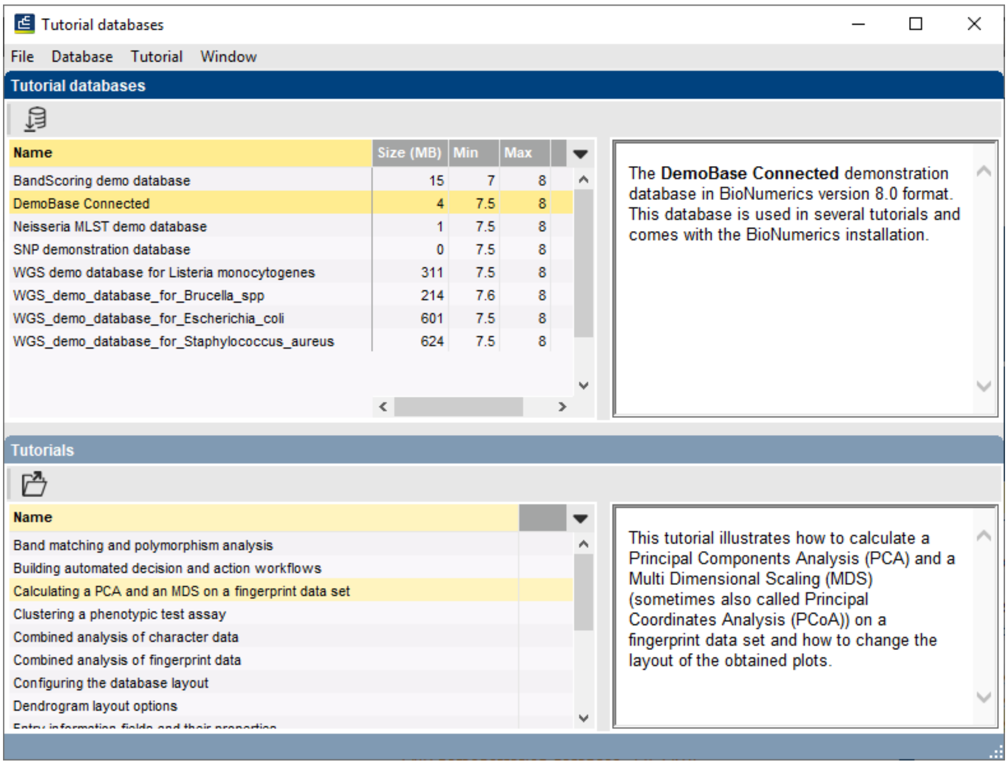
The wgMLST workflow and results will not be discussed in this tutorial.

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).



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_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***.

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 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'.

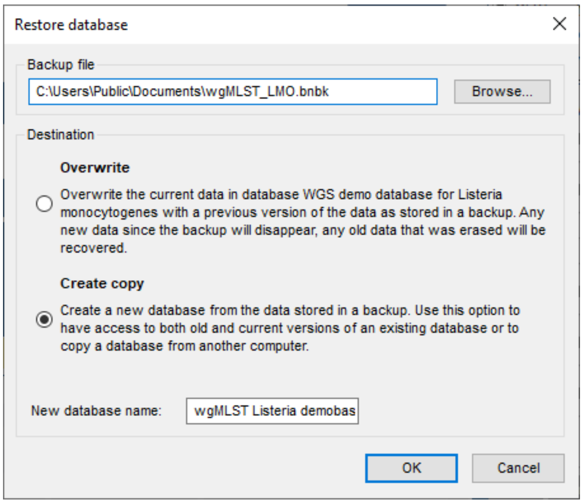
In contrast to other browsers, some versions of Internet Explorer rename the `WGS_LM01.bnbk` database backup file into `WGS_LM01.zip`. If this happens, you should manually remove the `.zip` file extension and replace with `.bnbk`. A warning will appear ("If you change a file name extension, the file 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

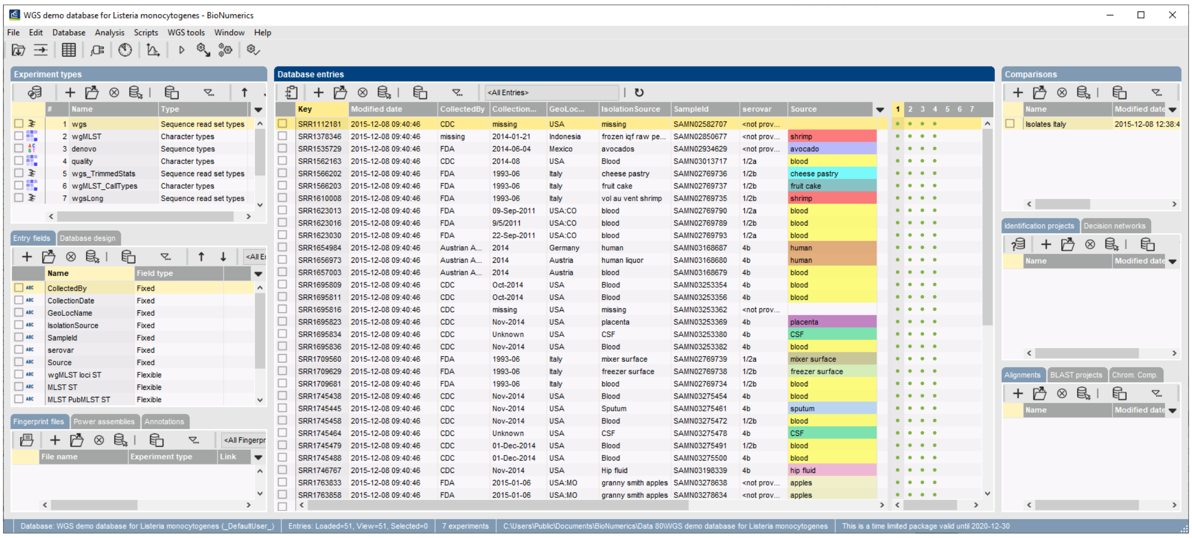
● Create a new database from the data stored in a backup. Use this option to have access to both old and current versions of an existing database or to copy a database from another computer.

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

	#	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

	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

File name	Experiment type	Link

	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	●	●	●	●			

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

Name	Modified date

Name	Modified date

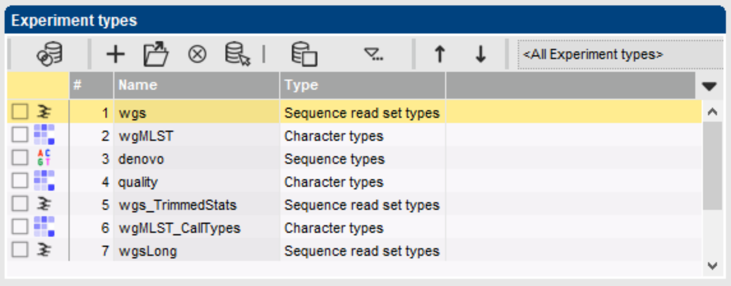
Database: WGS demo database for Listeria monocytogenes (_DefaultUser_) | Entries: Loaded=51, View=51, Selected=0 | 7 experiments | C:\Users\Public\Documents\BioNumerics\Data 80\WGS demo database for Listeria monocytogenes | This is a time limited package valid until 2020-12-30

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).

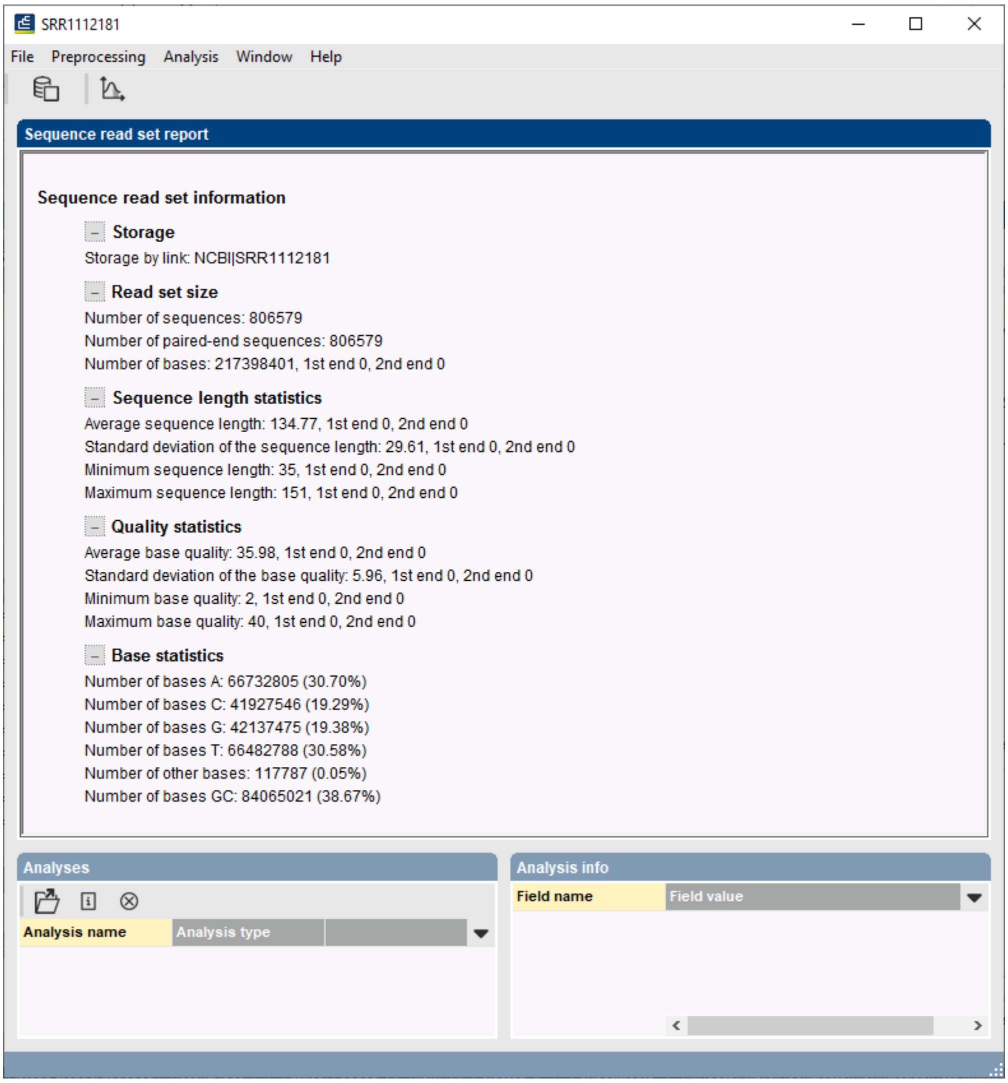
4. Close the *Sequence editor* window.

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 Performing a minhashing based clustering of sequences

A minhashing based clustering is performed in the *Comparison* window.

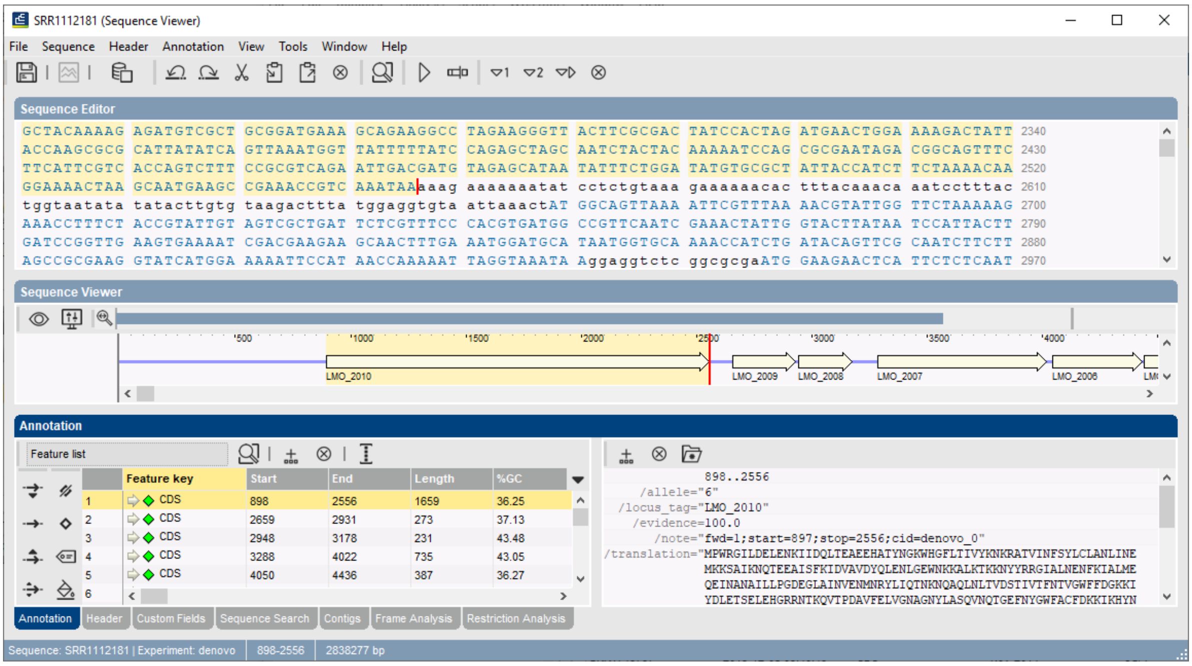
1. In the *Database entries* panel of the *Main* window, select all entries that have an associated **denovo** experiment. To select all entries at once, use the **CTRL+A** shortcut combination.

2. Highlight the *Comparisons* panel in the *Main* window and select ***Edit*** > ***Create new object...*** (+) to create a new comparison for the selected entries.

3. Click on the experiment name **denovo** in the *Experiments* panel to highlight it.

Clicking on the 👁 next to the experiment name **denovo** in the *Experiments* panel will display the sequences in the *Experiment data* panel. However, when analysing a lot of sequences at the same time, it could take quite a while before all sequences are displayed.

There is no need to align the sequences first. The similarities between all the sequences are calculated from the sequences' k-mers by the minhashing algorithm.



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
```

Figure 6: The *Sequence editor* window.

4. Select ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...*** or press and select ***Calculate cluster analysis***.

5. Choose ***Minhashing based*** with the default settings and press <***Next***>.

The **K-mer size** allows to specify a particular k-mer size for the minHashing algorithm to use. It is advised to use k = 31 or k = 51. A k-mer size with k = 31 is more sensitive at the genus level and k = 51 is more stringent, leading to fewer false positive matches. The amount of hashes that will be calculated can either be specified directly or can depended on the specified sourmash scaling factor, depending on whether the **Use sourmash scaling factor** option is checked or not.

More information on the different minHashing options can be found in the reference manual.

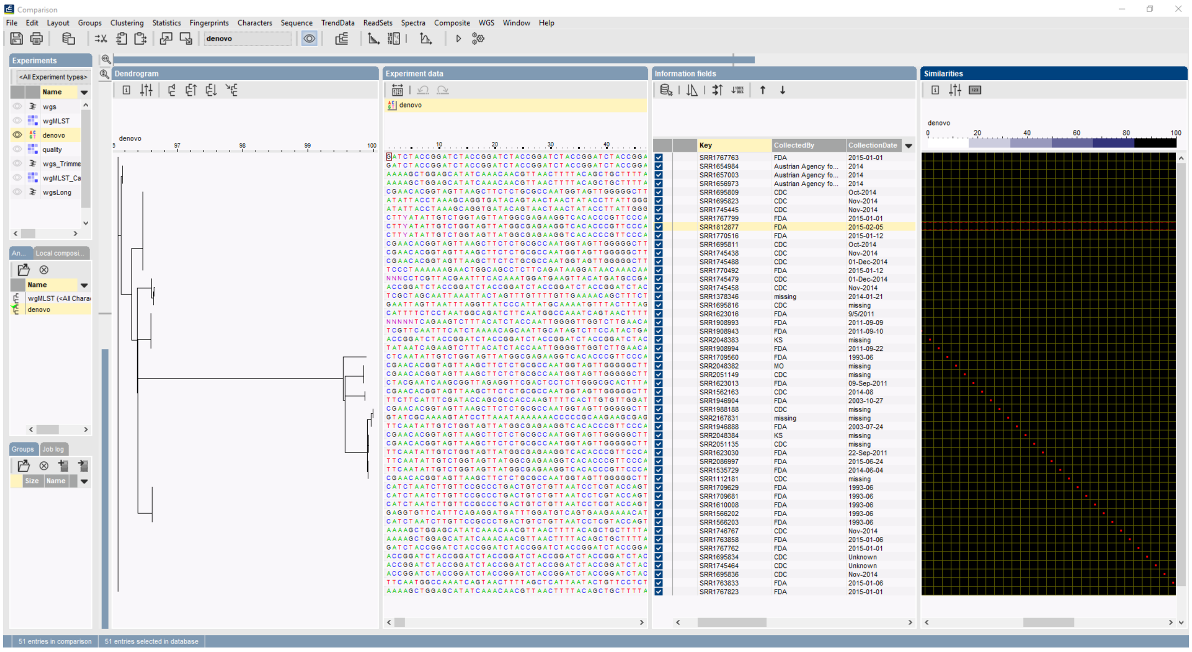
6. In the second step of the wizard, select ***Neighbor joining*** as dendrogram type and click <***Finish***>.

After the minHashing has been performed, the dendrogram and similarity matrix will be shown in the *Comparison* window (see Figure 7).

7. Click on the in the *Similarities* panel to show the similarity values between the sequences.

8. Click on the in the *Dendrogram* panel to show the settings that have been used for calculation of the minHashing based similarities and the dendrogram.

9. Save the comparison with ***File*** > ***Save*** (, **Ctrl+S**).



		Key	CollectedBy	CollectionDate
☑		SRR1767763	FDA	2015-01-01
☑		SRR1654984	Austrian Agency fo...	2014
☑		SRR1657003	Austrian Agency fo...	2014
☑		SRR1656973	Austrian Agency fo...	2014
☑		SRR1695809	CDC	Oct-2014
☑		SRR1695823	CDC	Nov-2014
☑		SRR1745445	CDC	Nov-2014
☑		SRR1767799	FDA	2015-01-01
☑		SRR1812877	FDA	2015-02-05
☑		SRR1770516	FDA	2015-01-12
☑		SRR1695811	CDC	Oct-2014
☑		SRR1745438	CDC	Nov-2014
☑		SRR1745488	CDC	01-Dec-2014
☑		SRR1770492	FDA	2015-01-12
☑		SRR1745479	CDC	01-Dec-2014
☑		SRR1745458	CDC	Nov-2014
☑		SRR1378346	missing	2014-01-21
☑		SRR1695816	CDC	missing
☑		SRR1623016	FDA	9/5/2011
☑		SRR1908993	FDA	2011-09-09
☑		SRR1908943	FDA	2011-09-10
☑		SRR2048383	KS	missing
☑		SRR1908994	FDA	2011-09-22
☑		SRR1709560	FDA	1993-06
☑		SRR2048382	MO	missing
☑		SRR2051149	CDC	missing
☑		SRR1623013	FDA	09-Sep-2011
☑		SRR1562163	CDC	2014-08
☑		SRR1946904	FDA	2003-10-27
☑		SRR1988188	CDC	missing
☑		SRR2167831	missing	missing
☑		SRR1946888	FDA	2003-07-24
☑		SRR2048384	KS	missing
☑		SRR2051135	CDC	missing
☑		SRR1623030	FDA	22-Sep-2011
☑		SRR2086997	FDA	2015-06-24
☑		SRR1535729	FDA	2014-06-04
☑		SRR1112181	CDC	missing
☑		SRR1709629	FDA	1993-06
☑		SRR1709681	FDA	1993-06
☑		SRR1610008	FDA	1993-06
☑		SRR1566202	FDA	1993-06
☑		SRR1566203	FDA	1993-06
☑		SRR1746767	CDC	Nov-2014
☑		SRR1763858	FDA	2015-01-06
☑		SRR1767762	FDA	2015-01-01
☑		SRR1695834	CDC	Unknown
☑		SRR1745464	CDC	Unknown
☑		SRR1695836	CDC	Nov-2014
☑		SRR1763833	FDA	2015-01-06
☑		SRR1767823	FDA	2015-01-01

Figure 7: Neighbor joining based on minhashing.