

BIONUMERICS Tutorial:

wgMLST typing in the Staphylococcus aureus demonstration database

1 Introduction

This guide is designed for users to explore the wgMLST functionality present in BIONUMERICS without having to post calculation jobs on their own computer or on the external calculation engine. The whole genome demonstration database used in this tutorial contains the results obtained from the full wgMLST analysis in BIONUMERICS on publicly available sequence read sets of *Staphylococcus aureus* from three studies, as they were published on NCBI's sequence read set archive.

Although this guide provides the necessary information to start working with the wgMLST functionality present in BIONUMERICS, it is recommended to read the following documentation available for download on the tutorial page on our website:

- Tutorial "wgMLST typing: routine workflow starting from sequence read sets"
- Tutorial "wgMLST typing: routine workflow starting from imported genomes"
- Tutorial "wgMLST typing: detailed exploration of results"
- WGS tools plugin manual

2 Preparing the database

The **WGS demo database** for *Staphylococcus aureus* can be downloaded directly from the *BION-UMERICS 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).

2. Select the WGS_demo_database_for_Staphylococcus_aureus from the list and select Database > Download ().

🖆 Tutorial databases					- 0
File Database Tutorial Window					
Tutorial databases					
Name	Size (MB)	Min	Max	-	
BandScoring demo database	15	7	8	~	The DemoBase Connected demonstration
DemoBase Connected	4	7.5	8		database in BioNumerics version 8.0 format.
Neisseria MLST demo database	1	7.5	8		This database is used in several tutorials and comes with the BioNumerics installation.
SNP demonstration database	0	7.5	8		comes with the biologinencs installation.
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		
				~	
	<			>	
Tutoriala					
Tutorials					
C ²					
Name				-	
Band matching and polymorphism analysis				^	This tutorial illustrates how to calculate a
Building automated decision and action workflows					Principal Components Analysis (PCA) and a
Calculating a PCA and an MDS on a fingerprint data set					Multi Dimensional Scaling (MDS) (sometimes also called Principal
Clustering a phenotypic test assay					Coordinates Analysis (PCoA)) on a
					fingerprint data set and how to change the
Combined analysis of character data					layout of the obtained plots.
Combined analysis of character data Combined analysis of fingerprint data					layout of the obtained plots.
Combined analysis of character data Combined analysis of fingerprint data Configuring the database layout Dendrogram layout options					layout of the obtained plots.

Figure 1: The *Tutorial databases* window, used to download the demonstration database.

- 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 BIONU-MERICS.

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).

0.1	Users\Downloads\wgMLST_SAUR1.bnbk Browse	
6.1	Disersible withoads/wgmEST_SAURT.bltok	_
Dest	ination	
	Overwrite	
0	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	
۲	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.	
Nev	v database name: Whole genome Staphyloco	

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).

Edit	Database Analysis	Scripts WGS tool	s Window	Help											
			, °o 0												
		N. D.	, ⊨ ∞⊜ κ	~											
perime	ent types					Datab	ase entries						Comparisons		
କ୍ଷ	+ 🖻 🛛	3. fh	2. 1	1	<aile< th=""><th>ŧ</th><th>+ 🖻</th><th>⊗ €. €</th><th>\ ∇</th><th><all entries=""></all></th><th>U</th><th></th><th>+ 13 ⊗ €.</th><th>6 ~</th><th><ai c<="" th=""></ai></th></aile<>	ŧ	+ 🖻	⊗ €. €	\ ∇	<all entries=""></all>	U		+ 13 ⊗ €.	6 ~	<ai c<="" th=""></ai>
400	Name	Туре		•		_	l L	Organism name		v title	ST info	outbreak 🔽 1 2 3 4 5 6 7 8	Name	Modified date	
										,					
¥	1 wgs		ead set types		^		RR103403	Staphylococcus aure		onatal MRSA outbreak investigation usin		part of outbrea			
-	2 WgMLST	Character ty			- 11		RR103402	Staphylococcus aure		onatal MRSA outbreak investigation usin		partoroutoroa	Pilot study	2015-01-28 12:15:5	
	3 denovo	Sequence ty					RR103398	Staphylococcus aure		onatal MRSA outbreak investigation usin		during outbreak	WGS for MRSA outbre	ak 2015-01-28 12:23:4	44
¥	4 quality	Character ty			- 11		ERR103394	Staphylococcus aure		onatal MRSA outbreak investigation usin		part of outoreal			
	5 wgs_TrimmedStat		ead set types				RR101899	Staphylococcus aure		onatal MRSA outbreak investigation usin		part of outbrea • • • • • • •			
45	6 wgMLST_CalTyp				- 11		RR103396	Staphylococcus aure		onatal MRSA outbreak investigation usin		prior to outbrea • • • • • • •			
	7 SNP outbreak	Sequence ty					RR103395	Staphylococcus aure		onatal MRSA outbreak investigation usin		pror			
≩	8 wgsLong	Sequence re	ead set types				RR103400	Staphylococcus aure		onatal MRSA outbreak investigation usin		partoroutorea	<		
					~		RR103405	Staphylococcus aure		onatal MRSA outbreak investigation usin		partoroadorea	Identification projects Deci	sion networks	
_							RR101900	Staphylococcus aure		onatal MRSA outbreak investigation usin		part of outbrea • • • • • • •			
fields	Database design						RR103401	Staphylococcus aure		onatal MRSA outbreak investigation usin		part of outbreal • • • • • • •	29 + 12 ⊗		∽
· P	9 ⊗ €. 6	1 2 1	E I D	All Entry fi	oblex		RR159680	Staphylococcus aure		onatal MRSA outbreak investigation usin		during outbreat • • • • • • •	Name	Modified date	
-			• E	- or cital y in			RR103397	Staphylococcus aure		onatal MRSA outbreak investigation usin		during outbreat • • • • • • •			_
	Name	Field type			-		RR103404	Staphylococcus aure		onatal MRSA outbreak investigation usin	22	part of outbreal • • • • • • •			
10C (Organism name	Fixed			^		RR127429	Staphylococcus aure		it study of rapid benchtop sequencing .		S. aureus clust • • • • • •			
ABC 2	Study title	Fixed					RR127426	Staphylococcus aure		t study of rapid benchtop sequencing .		S. aureus clust • • • • • •			
	ST info	Fixed					RR127428	Staphylococcus aure		it study of rapid benchtop sequencing .		S. aureus clust • • • • • •			
ac a	outbreak	Fixed					RR127417	Staphylococcus aure		t study of rapid benchtop sequencing .		S. aureus clust • • • • • •			
ac	Patient ID	Fixed					RR127408	Staphylococcus aure		t study of rapid benchtop sequencing .		S. aureus clust • • • • • •			
ac :	Study accession	Fixed					RR127443	Staphylococcus auro		t study of rapid benchtop sequencing .		S. aureus clust • • • • • •			
ĸ	nstrument	Fixed					RR127448	Staphylococcus aure	us Apik	t study of rapid benchtop sequencing .		S. aureus clust • • • • • •			
ac	MLST loci ST	Flexible					RR127432	Staphylococcus aure	us Apik	t study of rapid benchtop sequencing .		S. aureus clust • • • • • •	Alignments BLAST project		
к ,	wgMLST loci ST	Flexible					RR127431	Staphylococcus aure	us A pik	t study of rapid benchtop sequencing .		S. aureus clust • • • • • •	+ 🖻 🛛 🔍 🛛	6 ~	<
8C	MLST PubMLST ST	Flexible					RR127412	Staphylococcus aure	us Apik	t study of rapid benchtop sequencing .		S. aureus clust • • • • • •			<a1< td=""></a1<>
							RR127427	Staphylococcus aure	us Apik	t study of rapid benchtop sequencing .		S. aureus clust • • • • • •	Name		
					~		RR127444	Staphylococcus aure	us Apik	t study of rapid benchtop sequencing .		S. aureus clust • • • • • •			
_							RR127413	Staphylococcus aure	us Apik	t study of rapid benchtop sequencing .		S. aureus clust • • • • • •			
	fies Power assemble	s Annotations					RR127442	Staphylococcus aure	us Apik	t study of rapid benchtop sequencing .		S. aureus clust • • • • • •			
3	+ 13 8 8.		All Fin	oerprint file			RR127441	Staphylococcus aure	us Apik	t study of rapid benchtop sequencing .		S. aureus clust • • • • • •			
	· 🗗			Acchange und			RR127409	Staphylococcus aure	us A pik	t study of rapid benchtop sequencing .		S. aureus clust • • • • • •			
Fil	e name	Experiment type	Link		•		RR127433	Staphylococcus aure	us A pik	t study of rapid benchtop sequencing .		S. aureus clust • • • • • •			
					0		RR127415	Staphylococcus aure	us A pik	t study of rapid benchtop sequencing .		S. aureus clust • • • • • •			
<		_			>		<					> < >			

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

3 About the demonstration database

The demobase contains links to sequence read set data on NCBI's sequence read archive (SRA) for 97 publicly available sequencing runs of three *Staphylococcus aureus* whole genome sequencing studies ([1] [2] [3]) (see Figure 3). Sequence read set experiment type **wgs** contains the link to the sequence read set data on NCBI (SRA) with some raw data statistics.

The full wgMLST analysis (de novo assembly, assembly-based calls and assembly-free calls) was performed on this set of samples using default settings and the *S. aureus* wgMLST scheme on the Applied Maths Cloud Calculation Engine.

1. Select *WGS tools* > *Settings...* to access the settings of the plugin.

The calculation engine project is linked to the *Staphylococcus aureus* allele database. No credits are assigned to this project so no jobs can be submitted to the external calculation engine, however since the option *Enable running jobs on my own computer* is checked in the *Calculation engine* tab, it is possible to run jobs on your own computer (see Figure 5).

alculation engine	Organism	Experiment typ	es	waMLST		
Remote calculatio						
Url	-		\sim	On premises	_	
		ed Maths cloud	0	On premises	5	
	https://w	gmlst.applied-ma	ths.c	om		
Project name:	demobase	e-saur]	
Password:	•••••]	
Time out (sec):	30]	
Own computer						
Enable runnin	g jobs on m	y own computer				
Interval (min):	1]				

Figure 4: The Calculation engine tab of the Calculation engine settings dialog box.

Click on the *wgMLST tab* (see Figure 5) and press the <*Auto submission criteria*> button (see Figure 6).

By default, the **Use nomenclature acceptance criteria** option will be checked, meaning that the automatic submission settings are defined by the curator of the allele database.

3. Click <*Cancel*> in both dialog boxes.

Experiment types linked to wgMLST are present in the database for each of the entries and are displayed in the *Experiment types* panel:

- 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.
- Sequence experiment type **denovo** contains the results from the de novo assembly algorithm, i.e. concatenated de novo contig sequences.

Settings				?	×
Calculation engine C	Organism	Experiment types	wgMLST		
New allele submiss	ion				
Lab ID: LOCAL	_saur				
Submit new alle	les autom	atically			
Auto submissi	ion criteria				
Auto submissi	ion chich				
Multiple allele calls of	character	values			
Store lowest c	common al	lele ID			
O Store as abser	nt value				
Allele mapping expe					
wgMLST_MLS	T PubMLS	ST			
		_			

Figure 5: The *wgMLST tab* of the *Calculation engine settings* dialog box.

Auto submission criteria		?	\times							
Use nomenclature acceptance criteria										
 Set custom submission crit 	eria									
Require start/stop codons (for CDS	S loci on	ly)							
Allow internal stop codons	Allow internal stop codons (for CDS loci only)									
Minimum homology	85.0	%								
Maximum number of gaps	99									
ОК		Cance	el							

Figure 6: The Auto submission criteria dialog box.

- Character experiment type **quality** contains quality statistics for the raw data, the de novo assembly and the different allele identification algorithms.
- Sequence read set experiment type wgs_TrimmedStats: contains some data statistics about the reads retained after trimming.
- Character experiment type wgMLST_CallTypes: contains details on the call types.



No data is available for the sequence read set type **wgsLong** in the demo database. This sequence read set is used to store links to long read sequence read data (e.g. PacBio or MinION datasets).

A reference mapping has been calculated for all entries from the Neonatal MRSA study and the resulting sequences are stored in the **SNP outbreak** sequence type. These sequences are used in the wgSNP tutorials to illustrate the wgSNP functionality present in BIONUMERICS.

Additional information (in entry info fields Organism name, Instrument, Study accession, etc.) was collected from the corresponding publications and added to the demonstration database. Additionally, a number of comparisons were created that include all the samples together or grouped per study.

By clicking on one of the green dots next to an entry in the database, the corresponding results can be viewed, either in a separate window or in an experiment card for the character data types:

4. 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.

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 7).

	\times
File Preprocessing Analysis Window Help	
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	
Average base quality: 34.62, 1st end 34.92, 2nd end 34.33 Standard deviation of the base guality: 7.45, 1st end 6.72, 2nd end 8.10	
Minimum base quality: 2, 1st end 2, 2nd end 2	
Maximum base quality: 41, 1st end 41, 2nd end 41 Q20: 126071395 (95.20%), 1st end 63308955 (96.25%), 2nd end 62762440 (94.16%)	
Q25: 124374102 (93.91%), 1st end 62477601 (94.98%), 2nd end 62762440 (94.16%)	
Q30: 117580110 (88.78%), 1st end 59225531 (90.04%), 2nd end 58354579 (87.55%)	
- Base statistics	
Number of bases A: 44358027 (33.49%) Number of bases C: 21735791 (16.41%)	
Number of bases G: 21590930 (16.30%)	
Number of bases T: 44748888 (33.79%) Number of other bases: 8 (< 0.01%)	
Number of bases GC: 43326721 (32.72%)	Ý
Analyses Analysis info	
Field name Field value	-
Analysis name Analysis type	^
Ĉ <	> ~

Figure 7: The sequence read set experiment card for an entry.

- 5. Close the Sequence read set experiment window.
- 6. Click on the green colored dot for one of the entries in the second column in the *Experiment presence* panel. Column 2 corresponds to the second experiment type listed in the *Experiment types* panel, which is **wgMLST** in the default configuration.

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 (see Figure 8).

7. Close the character experiment card by clicking on the triangle in the top left corner.

Character	Value	Mapping	
SAUR_26	6	ŝ <+>	^
SAUR_27	8	3 <+>	
SAUR_28	7	/ <+>	
SAUR_29	8	3 <+>	
SAUR_30	1	<+>	
SAUR_31	1	<+>	
SAUR_32	8	3 <+>	
SAUR_33	10) <+>	
SAUR_34	11	<+>	
SAUR_35	7	/ <+>	
SAUR_36	11	<+>	
SAUR_37	9) <+>	
SAUR_38	7	/ <+>	
SAUR_39	5	i <+>	v

Figure 8: The character experiment card for an entry.

8. 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** in the default configuration.

The *Sequence editor* window opens, containing the results from the de novo assembly algorithm, i.e. concatenated de novo contig sequences (see Figure 9).

	2402 /6-	quence Viewer)													×
															^
-		Header Annotation			1										
	\approx	6 <u>A</u> A	X 🖺 [3 ⊗ Q		▽1 ▽2 ▽	$\sim \otimes$								
Sequen	ce Edito	i de la companya de la													
			t tottoa	ttta acac	atotto a	toccatto	ggaaa	rata	c atatoca	.gc aatcgacgt	a cotaaacat	t traatcora	t ttotttaaa	c 32230	
										jat ctatagatt					
										Jog gcaagagog					17
										tt cattgtcag					
										tt gcccagctt					
gaat	ccttt	a ccttccaag	t caactg	tatt cacg	ccaata t	gaatgacta	tgtcca	acac	c ttcttct	jaa cgaatacco	a tagcatgtt	t agtaggtac	a atcatagat	a 32780	
ttaa	accat	t gaacggtgd	g attact	ttgg atto	ttcatg a	gctttgatt	gccaa	goot	t cacctac	at tttctctct	a aaaattgaa	t ctcttactt	c tttaagtaa	a 32890	
atac	gacga	c cagegatto	g tgettt	cact aaaa	toggtg t	tgtcacttt	tgett	gagg	t gtataac	tg tatcttggd	t attaatatt	g tcagtattc	g caccctctt	c 33000	
aaca	ttago	a toottaatt	a atcgct	gtaa ttgc	tgatac a	cttcatcga	(cattt	ccgt	t taatttt	tg actaactgo	g catcattt	c tgctacaga	t gtcacatca	a 33110	
ctat	ttgtt	t cacttcatt	t gctgta	gaag gaat	tgtatg t	tgcatatga	attgt:	aatg	c ccttcgt	itc atacgttac	a tcaacaatg	t tttcgacac	c acccacage	t 33220	
Somuon	ce View	or												-	
© l	<u>₽</u> @														
		10000	15000	20000	25000	30000	35000	4000	0 '45000	50000	'55000 '6000	65000	*70000	75000	
			ka k	JK-0K-	ג¢ג¢ג		<mark>``</mark> -\¢ikix	Ĭ	SAUR 3559 SAU						-С
						01			. short_ssss and		anon anan an	. ano anoana	n ar ananon	. 340	
		<													
Annotat			01.	0 J T											
Feature	e list		Q] #	⊗ 1					÷ ⊗ [
	e.	Feature key	Start					•	(-11-1-	complement (31290	32454)				
÷ >	32	🖕 🔶 CDS	28168	29058	891	33.60	6	^	/allele /locus_tag						
·	o 33	🖕 🔷 CDS	29068	30414	1347	32.47	7			=95.77821429					
	34	🖕 🔷 CDS	31290	32454	1165	37.03	1		/note	"fwd=0;start=0;st					
	= 35	🖨 🔷 CDS	33410	34114	705	33.10	8		/translation	"MTKLPQNFMWGGALAA					
	36	🖕 🔶 CDS	34340	36631	2292	35.05	8			EDIALFKENGLKCLRT					
⇒ ≤	37	A one	00070	07070	700	00.00	-	, [*]		YGGFRNREVVDYFVHF AHHELLASALAVRLGK					
Annotatio	n Head	er Custom Fields	Sequence Search	Contigs Fra	me Analysis R	estriction Analysis									
	5004000		0.000 000						_						
		03 Experiment: denov													

Figure 9: The Sequence editor window.

- 9. Close the Sequence editor window.
- 10. Click on the green colored dot in column 4 to open the **quality** character card (default configuration) for an entry in the database.

The **quality** character card contains quality statistics for the raw data, the de novo assembly and the different allele identification algorithms (see Figure 10).

11. Close the character experiment card by clicking on the triangle in the top left corner.

Character	Value	Mapping	
AvgQuality	35	<+>	^
AvgReadCoverage	47	<+>	
N50	31810	<+>	
NrContigs	198	<+>	
NrNonACGT	538	<+>	
Length	2862093	<+>	
KeywordCov	30	<+>	
NrAFMultiple	70	<+>	
NrAFPerfect	2584	<+>	
NrAFPresent	2727	<+>	
NrBAFMultiple	0	<->	
NrBAFPerfect	2247	<+>	
NrToBeSubmitted	144	<+>	
NrAlreadySubmitted	1932	<+>	~

Figure 10: The character experiment card for an entry.

4 Subschemes

1. In the *Main* window double-click the character experiment type **wgMLST** in the *Experiment types* panel to call the *Character type* window (see Figure 11).

aracters	1 1010						
	ı∕ ⊠× ↑ ↓ .	<all characters=""></all>	ย				
Character	Enabled	Min. Max.	Color scale	LocusName	LocusTag	Gene	•
SAUR_1	✓	0	5000	SAUR00001	SACOL0001	dnaA	
SAUR_2	×	0	5000	SAUR00002	SACOL1613		
SAUR_3	✓	0	5000	SAUR00003	SACOL2111	tdk	
SAUR_4	×	0	5000	SAUR00004	SACOL0005	gyrB	
SAUR_5	×	0	5000	SAUR00005	SACOL0180		
SAUR_6	✓	0	5000	SAUR00006	SACOL1055	sspC	
SAUR_7	×	0	5000	SAUR00007	SACOL2327	hutG	
SAUR_8	×	0	5000	SAUR00008	SACOL1248	rnc	
SAUR_9	×	0	5000	SAUR00009	SACOL2467		
SAUR_10	×	0	5000	SAUR00010	SACOL0429		
SAUR 11	✓	0	5000	SAUR00011	SACOL0589	rpoC	
<							>
aracters Mapping							_
mparison settings							_
gMLST settings							
	/alues, open data set (390	(characters)					
gineor, numericari		4 characters)					
*******	son						
- Compari							
	nilarity coefficient						
– Sir	milarity coefficient prical (values)						

Figure 11: The Character type window.

Within a character experiment type, a character view can be defined that specifies a particular subset of characters.

2. Click on the drop-down bar in the toolbar (see Figure 12).

In this database, four views have been defined at the curator level and are synchronized upon installation: the default view **All loci**, the **MLST PubMLST** view for the traditional seven house-keeping loci, the **Core loci** view and the **wgMLST loci** view containing all loci except the ones

Character type 'wgMLST	•				_	
	Fields Mapping Wind	law. Hala				
-		iow help				
6 바냰 🗔	1					
haracters						
+ 🛛 🗟 🗠	⊠× ↑ ↓ ↓\	<all characters=""> U</all>				
Character	Enabled M	in. 🗸 (All Characters)	LocusName	LocusTag	Gene	
SAUR_1	✓	<selected characters=""></selected>	SAUR00001	SACOL0001	dnaA	
SAUR_2 SAUR_3	✓		SAUR00002	SACOL1613		
] SAUR_3	✓	<manage defined="" user="" views=""></manage>	SAUR00003	SACOL2111	tdk	
] SAUR_4	✓		SAUR00004	SACOL0005	gyrB	
] SAUR_5	✓	All loci	SAUR00005	SACOL0180		
] SAUR_6	✓	Core loci	SAUR00006	SACOL1055	sspC	
SAUR_7	✓		SAUR00007	SACOL2327	hutG	
SAUR_8	✓	wgMLST loci	SAUR00008	SACOL1248	rnc	
] SAUR_9	✓	MLST PubMLST	SAUR00009	SACOL2467		
SAUR_10	✓	0 5000	SAUR00010	SACOL0429		
SAUR_11	✓	0 5000	SAUR00011	SACOL0589	rpoC	
<						>
aracters Mapping						
omparison settings						
vgMLST settings						
	es, open data set (3904 c	haractere)				
 Comparison 	l					
- Simila	arity coefficient					
Categorica						
	as distance: No					
omparison settings Cross	inks Attachments					

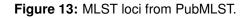
present in the MLST PubMLST view.

Figure 12: Views defined at the curator side.

3. Select the MLST PubMLST view from the list.

After selecting a character view, the window is updated (see Figure 13), and the number of characters in view is displayed in the status bar at the bottom of the window.

Character type 'wgMLS	Т					_	
le Settings Characters	Fields Mapping Window	/ Help					
Ê 排⊮:							
	910 910						
Characters							
+ 🛛 🗟 🗠	⊠×∣↑↓∣↓∖	MLST PubMLST	ບ				
Character	Enabled Min.	Max.	Color scale	LocusName	LocusTag	Gene	
SAUR_3898	✓	0	5000	SAUR03898	MLST:arcC		1
SAUR_3899	×	0	5000	SAUR03899	MLST:aroe		
SAUR_3900	×	0	5000	SAUR03900	MLST:glpf		
SAUR_3901	×	0	5000	SAUR03901	MLST:gmk		
SAUR_3902	×	0	5000	SAUR03902	MLST:pta		_
SAUR_3903	×	0	5000	SAUR03903	MLST:tpi		
SAUR_3904	×	0	5000	SAUR03904	MLST:yqil		
<							>
Characters Mapping							
mapping							
Comparison settings							
							_
wgMLST settings							
	ues, open data set (3904 cha						
-		acters)					
 Compariso 	n						
– Simi	larity coefficient						`
	slinks Attachments						
Somparison settings							
naracter type wgMLST C	haracters in view=7						



4. To view all characters again, select < All loci > again from the drop-down list.

Besides these curator views, the user can create as many additional local character views as needed and use them as subscheme e.g. for clustering or when inspecting the allele calls for a subset of loci. Creating a character view can be done in two ways:

- The first method is based on a character *selection*.
- The second method is based on a *dynamic query* using the character information fields.
 - 5. Select a few characters by selecting the characters directly in the *Character type* window (Ctrl+click or Shift+click).

The selection is synchronized with the database: any selection of characters made in the *Character type* window is reflected in other windows, e.g. the *Comparison* window, and vice versa.

- 6. Click on the drop-down bar in the toolbar and choose *Manage user defined views*, alternatively select *Characters* > *Character Views* > *Manage user defined views...* (<all Characters>).
- 7. Press <*Add...*>, specify a name, e.g. **MySubsetExample**, make sure **Subset based** is selected, and press <*OK*> and <*Exit*>.

The new view is added to the database and is automatically selected in the *Character type* window. The new view is available for use e.g. in the *Character type* window, *wgMLST quality assessment* window or *Comparison* window.

8. To view all characters again, select <**All loci**> again from the drop-down list.

As a second example we will create a query-based view of all loci encoding a ribosomal protein. Because all those loci have a gene name starting with "rpl" (ribosomal proteins of the large subunit) or "rps" (ribosomal proteins of the small subunit), this subset can be easily defined with a query-based view.

- 9. Click on the drop-down bar in the toolbar and choose *Manage user defined views*, alternatively select *Characters* > *Character Views* > *Manage user defined views...* (<all Characters>).
- Select < *Add...*>, specify a name, e.g. "ribosomal proteins", make sure *Query based* is selected and click < *OK*>.
- 11. Select the *Gene* field, change the *Equals* condition to *Contains* and type "rpl" in the white box.
- 12. Press < Add new> in the Statements panel and edit it to Gene Contains "rps".
- 13. Press < *Remove all unused*>.
- 14. Finally, select both remaining rules (use Ctrl+click) and press < OR> in the Group by panel.

The query should now look like in Figure 14.

15. Press *<OK>* to validate the query and *<Yes>* to confirm and press *<Exit>*.

The new query-based view is created with the 46 characters that fulfill the specified criteria (see Figure 15). The new view is available for use e.g. in the *Character type* window, *wgMLST quality assessment* window or *Comparison* window.

- 16. To view all characters again, select <**All loci**> again from the drop-down list.
- 17. Close the Character type window.

Query vie	w editor				? ×
Chara	cter field values				Statements
	Gene	<u>Contains</u>	rpl ~		Add new
OR	Gene	<u>Contains</u>	rps v		Remove selected
					Remove all unused
					Include all visible fields
					Group by AND OR NOT
					Reset query
					Share this view
<				>	OK Cancel

Figure 14: Query based view.

<u>د</u> C	haracter type 'wgMLST						_	X
ile	Settings Characters	Fields Mapping \	Vindow Help					
£	그 부부 카운 귀했							
_								
Cha	racters							
+	• 🛛 🗟 🖉	⊠× ↑ ↓	↓ ribosomal proteins	ບ				
	Character	Enabled	Min. Max.	Color scale	LocusName	LocusTag	Gene	•
	SAUR_12	×	0	5000	SAUR00012	SACOL0545	rplY	^
	SAUR_39	×	0	5000	SAUR00039	SACOL2239	rpIC	
	SAUR_61	×	0	5000	SAUR00061	SACOL0015	rpll	
	SAUR_275	×	0	5000	SAUR00275	SACOL1254	rpsP	
	SAUR_280	×	0	5000	SAUR00280	SACOL0583	rplK	
	SAUR_290	×	0	5000	SAUR00290	SACOL0584	rplA	
	SAUR_298	×	0	5000	SAUR00298	SACOL0585	rpIJ	~
	<							>
Char	acters Mapping							
Com	nparison settings							
we	MLST settings							
	MLST: numerical value	as onen data set (30	04 characters)					
wy								
	 Comparison 							
	– Simila	rity coefficient						~
Comp	oarison settings Crossi	inks Attachments						
narac	ter type wgMLST Cha	aracters in view=46						

Figure 15: New query based view.

5 Obtaining MLST profiles and sequence types

Using the *WGS tools plugin*, MLST profiles with public allele numbers can be obtained, i.e. using the same allele numbering as PubMLST. Additionally, the plugin allows the retrieval of public sequence types.

First, we need to activate the corresponding allele mapping experiment in the wgMLST settings:

- 1. Select *WGS tools* > *Settings...* to open the *Calculation engine settings* dialog box.
- 2. Click on the *wgMLST* tab to bring the wgMLST settings into focus.
- 3. Under Allele mapping experiments, check wgMLST_MLST PubMLST and press < OK >.

A character experiment type called **wgMLST_MLST PubMLST** is created in the database in case it did not exist yet. Now, MLST profiles with exactly the same allele IDs as used on PubMLST can be obtained for all entries with a **wgMLST** experiment:

- 4. In the *Experiment types* panel, highlight the **wgMLST** experiment type and select *Database* > *Entries* > *Select entries with experiment* to make the entry selection.
- 5. Select *WGS tools* > *Get alleles mapping*.

The allele numbers from the **wgMLST** experiments are translated into public nomenclature. The public allele numbers are then retrieved and stored in the **wgMLST_MLST PubMLST** experiments. Optionally, this can be verified in the *Comparison* window:

- 6. Highlight the *Comparisons* panel and select *Edit* > *Create new object...* (+) to open a comparison with the selected entries.
- 7. In the *Experiments* panel, click on the icon next to wgMLST_MLST PubMLST to visualize the MLST profiles in the *Experiment data* panel. Select *Characters* > *Show values* (iii) to display the values (see Figure 16).

Comparison									-		
Edit Layout Groups C	lustering Statistics Fingerpri	ints Characters	Sequer	nce TrendData ReadSets Spectra Composite	Window Help						
⊜ :∞ ⊳ €	6 ≭X £1 C≩ 6	wgt	ILST MLS	ST Pub 🔘 🖻 🗽 🛐	· ·						
	€U →^		ILSI_MLS		<u>-</u> ,						
eriments											
		٢		Experiment data	Information fields	-				Simil	1
II Experiment types>	ା ଓ										
		-		122 123 ANC ANC Charact	₿. \ . \$	1,882 ↓ ↓					
wgMLST	<all characters=""></all>	^		wgMLST_MLST PubMLST							
	<default></default>										
				0 w 1, z 1							
quality	<all characters=""></all>			ancC ghr yqL yqL	Group	Key	Organism name	Study title ST info			
≵ wgs_TrimmedStats	<default></default>			7 6 1 5 8 8 6		ERR103403	Staphylococcus au	A neonatal MRSA ou 22			
wgMLST_CallTypes	<all characters=""></all>			7 6 1 5 8 8 6		ERR103402	Staphylococcus au	A neonatal MRSA ou 22			
SNP outbreak	<default></default>			1 1 1 1 1 1 1		ERR103398	Staphylococcus au	A neonatal MRSA ou 1			
≩ wasLong	<default></default>			7 6 1 5 8 8 6		ERR103394	Staphylococcus au	A neonatal MRSA ou 22			
				6 1 5 8 8 6	v	ERR101899	Staphylococcus au	A neonatal MRSA ou 22			
wgMLST_MLST PubMLS	ST <all characters=""></all>	× -	r	7 6 1 5 8 8 6	V	ERR103396	Staphylococcus au	A neonatal MRSA ou 22			
ses Local composite datas	ets .			1 4 1 4 12 1 10	✓	ERR103395	Staphylococcus au	A neonatal MRSA ou 5			
		_		7 6 1 5 8 8 6	✓	ERR103400	Staphylococcus au	A neonatal MRSA ou 22			
$\rightarrow \otimes$				7 6 1 5 8 8 6	✓	ERR103405	Staphylococcus au	A neonatal MRSA ou 22			
		-		7 1 5 8 8 6	✓	ERR101900	Staphylococcus au	A neonatal MRSA ou 22			
				7 6 1 5 8 8 6	✓	ERR103401	Staphylococcus au	A neonatal MRSA ou 22			
Job log				2 2 2 2 3 3 2	Z	ERR159680	Staphylococcus au	A neonatal MRSA ou 36			
		_		1 1 1 1 1 1 1		ERR103397	Staphylococcus au	A neonatal MRSA ou 1			
› ⊗ 1 1 1 ₽				7 6 1 5 8 8 6		ERR103404	Staphylococcus au	A neonatal MRSA ou 22			
		-				ERR127429 ERR127426	Staphylococcus au	A pilot study of rapi			
				1 1 1 1 22 1 1	✓ ✓	ERR127428	Staphylococcus au Staphylococcus au	A pilot study of rapi A pilot study of rapi			
				22 1 14 23 12 4 31		ERR127426	Staphylococcus au	A plot study of rapi			
				22 1 14 23 12 4 31		ERR127417	Staphylococcus au	A pilot study of rapi			
						ERR127400	Staphylococcus au	A pilot study of rapi			
				13 13 1 1 12 11 13	 			· · · · · · · · · · · · · · · · · · ·	>	<	÷.

Figure 16: The Comparison window.

8. Close the *Comparison* window.

Next, sequence types can be assigned for the selected entries, based on the **MLST PubMLST** subscheme.

9. In the *Main* window, select *WGS tools* > *Assign wgMLST sequence types...*.

This opens the *Assign sequence types* dialog box, where available typing schemes can be checked to be included in the assignment of the sequence types (see Figure 17).

10. Leave the subscheme **MLST PubMLST** checked and press < *OK* > to assign a sequence typing based on the 7 loci used for traditional MLST analysis.

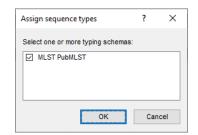


Figure 17: The Assign sequence types dialog box, with a single typing scheme listed.

Per entry and typing scheme, a list of allele identifications is sent to the allele database and sequence type information is returned. The sequence types are then saved to a dedicated entry information field.

In our example database, a sequence type is added in the field *MLST PubMLST ST* for the selected entries (see Figure 18).

Edit	Database Analysi	s Scripts WGS to	ols Windo	ow Help																
) 		O L D	Q, %	0,																
		01041.	N 0		_															
perim	ient types				Data	base entries														Comparisons
A	+ 🖻 🛛		₽	↑ ↓	f°	1 + 🖻	⊗ ₿ ჩ	2	<all entri<="" td=""><td>es></td><td></td><td>U</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>+ 🖻 🛛 🗎</td></all>	es>		U								+ 🖻 🛛 🗎
	# Name	Type				Key	Organism name	Study title		ST info	outbreak	Patient ID	Study acce	Instrument	MLST PubML	1 2	3 4	5 6 7 8	9	Name
3	1 was		read set typ	es ^	~	ERR103403	Staphylococcus aureus	A neonatal	MRSA	22	part of outbreak	MRSA 6C	ERP001256	Ilumina MiSeq	publicST22					Neonatal MRSA s
	2 woMLST	Character			~	ERR103402	Staphylococcus aureus			22	part of outbreak		ERP001256	Ilumina MiSeq	publicST22					Plot study
45	3 denovo	Sequence			~	ERR103398	Staphylococcus aureus			1	during outbreak		ERP001256	Ilumina MiSeg	publicST1					WGS for MRSA o
÷.	4 quality	Character			~	ERR103394	Staphylococcus aureus			22	part of outbreak		ERP001256	Ilumina MiSeg	publicST22					
3	5 wgs_TrimmedS		read set typ	es	~	ERR101899	Staphylococcus aureus			22	part of outbreak		ERP001256	Ilumina MiSeg						
	6 wgMLST_Call				~	ERR103396	Staphylococcus aureus			22	prior to outbreak		ERP001256	Iumina MiSeq	publicST22					
15	7 SNP outbreak	Sequence			~	ERR103395	Staphylococcus aureus			5	prior	MRSA 14C prior	ERP001256	Ilumina MiSeq	publicST5					
3	8 wasLong		read set typ	es	~	ERR103400	Staphylococcus aureus			22	part of outbreak		ERP001256	Ilumina MiSeq	publicST22					<
82	9 WOMLST MLST	T PubMLST Character	types		~	ERR103405	Staphylococcus aureus	A neonatal	MRSA	22	part of outbreak	MRSA 8C	ERP001256	Ilumina MiSeq	publicST22				11	identification projects
-				~	~	ERR101900	Staphylococcus aureus	A neonatal	MRSA	22	part of outbreak	MRSA 11C	ERP001256	Ilumina MiSeg					115	
ry field	S Database design				~	ERR103401	Staphylococcus aureus		MRSA	22		MRSA 1B Index	ERP001256	Ilumina MiSeg	publicST22					2월 🕂 🖻
					~	ERR159680	Staphylococcus aureus		MRSA	36	during outbreak	MRSA 18B	ERP001256	Ilumina MiSeq	publicST36					Name
+ [" ⊗ €.	6 .	↑ ↓	<ai entry="" fie<="" td=""><td>~</td><td>ERR103397</td><td>Staphylococcus aureus</td><td>A neonatal</td><td>MRSA</td><td>1</td><td>during outbreak</td><td>MRSA 16B</td><td>ERP001256</td><td>Ilumina MiSeq</td><td>publicST1</td><td></td><td></td><td></td><td></td><td>Marrie</td></ai>	~	ERR103397	Staphylococcus aureus	A neonatal	MRSA	1	during outbreak	MRSA 16B	ERP001256	Ilumina MiSeq	publicST1					Marrie
	Name	Field type		-	~	ERR103404	Staphylococcus aureus		MRSA	22	part of outbreak	MRSA 7C	ERP001256	Ilumina MiSeq	publicST22					
ABC	Organism name	Fixed			~	ERR127429	Staphylococcus aureus	A pilot stud	v of rap		S, aureus clu	s	ERP001413	Ilumina MiSeq	publicST772					
ARC	Study title	Fixed			~	ERR127426	Staphylococcus aureus	A pilot stud	y of rap		S. aureus clu	U	ERP001413	Ilumina MiSeg	publicST772					
ABC	ST info	Fixed			~	ERR127428	Staphylococcus aureus	A pilot stud	y of rap		S. aureus clu	т	ERP001413	Ilumina MiSeq	publicST772			• •		
ABC	outbreak	Exed			~	ERR127417	Staphylococcus aureus	A pilot stud	y of rap		S. aureus clu	H nasal swab	ERP001413	Ilumina MiSeq	publicST88		• •	• •		
ABC	Patient ID	Fixed			~	ERR127408	Staphylococcus aureus	A pilot stud	y of rap		S. aureus clu	с	ERP001413	Ilumina MiSeq	publicST88		• •	• •		
AIC	Study accession	Fixed			~	ERR127443	Staphylococcus aureus	A pilot stud	y of rap		S. aureus clu	L	ERP001413	Ilumina MiSeq	publicST6		• •	• •		<
ARC	Instrument	Fixed			~	ERR127448	Staphylococcus aureus	A pilot stud	y of rap		S. aureus clu	0	ERP001413	Ilumina MiSeq	publicST15		• •	• •	1	Alignments BLST CH
ARC	MLST loci ST	Flexible			~	ERR127432	Staphylococcus aureus	A pilot stud	y of rap		S. aureus clu	٥	ERP001413	Ilumina MiSeq	publicST772		• •	• •		
ARC	woMLST loci ST	Flexible			~	ERR127431	Staphylococcus aureus	A pilot stud	y of rap		S. aureus clu	w	ERP001413	Ilumina MiSeq	publicST772		• •	• •		+ 🖻 🛛 🗎
	MLST PubMLST ST	Flexible			~	ERR127412	Staphylococcus aureus	A pilot stud	y of rap		S. aureus clu	A	ERP001413	Illumina MiSeq	publicST88		• •	• •		Name
				*	~	ERR127427	Staphylococcus aureus	A pilot stud	y of rap		S. aureus clu	v	ERP001413	Illumina MiSeq	publicST772		• •	• •		
gerprin		bles Anostations			~	ERR127444	Staphylococcus aureus	A pliot stud	y of rap		S. aureus clu	N	ERP001413	Ilumina MiSeq	publicST15		• •	• •		
					~	ERR127413	Staphylococcus aureus	A pilot stud	y of rap		S. aureus clu	G	ERP001413	Ilumina MiSeq	publicST88		• •	• •		
3	+ 🖻 🛛 🗎		. <ai< td=""><td>Fingerprint files</td><td>~</td><td>ERR127442</td><td>Staphylococcus aureus</td><td>A pilot stud</td><td>y of rap</td><td></td><td>S. aureus clu</td><td>к</td><td>ERP001413</td><td>Ilumina MiSeq</td><td>publicST5</td><td></td><td>• •</td><td>• •</td><td></td><td></td></ai<>	Fingerprint files	~	ERR127442	Staphylococcus aureus	A pilot stud	y of rap		S. aureus clu	к	ERP001413	Ilumina MiSeq	publicST5		• •	• •		
_	ile name	Experiment type	Link		~	ERR127441	Staphylococcus aureus	A pilot stud	y of rap		S. aureus clu	1	ERP001413	Ilumina MiSeq	publicST30		• •	• •		
		caper and the	LINK		~	ERR127409	Staphylococcus aureus	A pilot stud	y of rap		S. aureus clu	E	ERP001413	Illumina MiSeq	publicST88		• •	• •		
				2	~	ERR127433	Staphylococcus aureus	A pilot stud	y of rap		S. aureus clu	R	ERP001413	Ilumina MiSeq	publicST772	• •	• •		~	
	(~	<									>	<				<

Figure 18: MLST PubMLST ST numbers.

In case an entry has an incomplete profile for the **MLST PubMLST** subscheme, no sequence type can be assigned and an error message will be generated for that entry.

6 Import of sample-specific allele sequences into the database

Once the wgMLST allele results have been imported in the database, it is possible to import the actual allele sequences for a specific wgMLST locus or a combination of loci, as defined in a subscheme, using *WGS tools* > *Store wgMLST locus sequences...*.

As an example, we will describe how to retrieve the allele sequences for the seven MLST loci into the database, using sequence type names that can be recognized by the *MLST online plugin*. First, a character info field should be created and the exact locus names as defined in the MLST scheme should be entered for those seven loci.

1. Open the **wgMLST** *Character type* window by double-clicking the character experiment type in the *Experiment types* panel (top right of *Main* window).

- 2. In the character views drop down menu, select *MLST PubMLST*.
- 3. Fill in the names of the seven MLST loci as they are defined in the *S. aureus* MLST scheme on http://saureus.mlst.net/, in the *Gene* field: "arcc", "aroe", "glpf", "gmk", "pta", "tpi", and "yqil" (see Figure 19). A field becomes editable by clicking it after it was selected (click twice slowly).

e C	haracter type 'wgMLST'						_	
	Settings Characters		indow Help					
£								
_	racters							
+		⊠× ↑ ↓ .	MLST PubMLST	<u>ی</u>				
-		Enabled	Min. Max.	Color scale	LocusName	LocusTag	Gene	
	SAUR_3898	∠	0	5000	SAUR03898	MLST:arcC	arcc	
	SAUR 3899		0	5000	SAUR03899	MLST:aroe	aroe	
	SAUR_3900	×	0	5000	SAUR03900	MLST:glpf	glpf	
	SAUR_3901	×	0	5000	SAUR03901	MLST:gmk	gmk	
	SAUR_3902	×	0	5000	SAUR03902	MLST:pta	pta	
	SAUR_3903	×	0	5000	SAUR03903	MLST:tpi	tpi	
	SAUR_3904	✓	0	5000	SAUR03904	MLST:yqil	yqil	
	<							>
Char	acters Mapping							
Com	nparison settings							
	ipaneen eetange							
								/
	gMLST settings							
wg	MLST: numerical value	es, open data set (390	4 characters)					
	- Comparison							
	– Simila	rity coefficient						`
Comp	parison settings Crossli	nks Attachments						
arac	ter type wgMLST Cha	racters in view=7					_	

Figure 19: The *Character type* window for **wgMLST**, with locus names for the 7 MLST loci, as known on PubMLST.net, filled in in the 'Gene' character information field.

4. Close the Character type window.

Now the allele sequences can be imported into sequence type experiments that have the correct name for analysis by the *MLST online plugin*.

- 5. Make sure the *Database entries* panel is the active panel and select *Edit* > *Select all* (Ctrl+A) to select all entries at once.
- 6. Select *WGS tools* > *Store wgMLST locus sequences...* (see Figure 20). Specify "MLST PubMLST" as the *Subschema* and select "Gene" for the *Sequence experiment type*.
- 7. Click <*OK*> to start importing the allele sequences and <*Yes*> to confirm the creation of new experiment types.

The database now contains the allele sequences for the 7 MLST loci, stored in 7 sequence experiment types that can be accessed by the *MLST online plugin*. This can be illustrated as follows:

- 8. Install the *MLST online plugin*, via *File* > *Install / remove plugins...* (,): Select *MLST online*, press <*Activate*> and confirm.
- 9. Choose Select organism from on-line list and select Staphylococcus aureus from the list. Leave all the other settings at default: press < Next> several times, then < Finish> and confirm with < OK> twice. Close the Plugins dialog box.
- 10. In the *Main* window, make sure the *Database entries* panel is the active panel and select *Edit* > *Select all* (Ctrl+A) to select all entries at once and choose *MLST* > *Identify alleles and profiles*.

Subschema	
For each locus in the subsche sequence experiment type wi	
Core loci (1861 loci)	_
wgMLST loci (3897 loci)	
MLST PubMLST (7 loci)	
ribosomal proteins (46 loci)	
Sequence experiment type	
Select the identifier you want	e
sequence experiment type na	

Figure 20: The Store sequences dialog box.

The character type **MLST** now contains the allele numbers for the 7 loci as they are known in the public PubMLST scheme, the public sequence types are written to the entry field **MLST ST** (see Figure 21). For two entries, one of the loci was not called, so no sequence was stored in the database and no sequence type could be assigned.

WGS d	emo database for Staphyloc	occus aureus - BioNumeric	s										- 0
Edit	Database Analysis MLS												
	I D D 1	A D Q 00 1	©_										
-													
xperim	ent types		Data	ibase entries									Comparisons
67	+ 🖒 🛛 🗟	6 2 1	1 1	1 + 13 ⊗ 6.	6	2	All Entries>		U				+ 13 8 8 1 6 2 4
Ų,	Name	Type		Organism name	Study title	ST info	outbreak	Patient ID		MLST PubMLS	MLST ST	1 2345678	Name Modified date
₹			× 🗸			22	part of outbreak				22		
2	1 wgs 2 wgMLST	Sequence read set types Character types	- î 💆	Staphylococcus aureus Staphylococcus aureus		22	part of outbreak			publicST22 publicST22	22		Pilot study 2015-01-28 12:15:55
	3 denovo	Sequence types		Staphylococcus aureus			during outbreak			publicST1	1		WGS for MRSA outbreak 2015-01-28 12:23:44
5	4 quality	Character types		Staphylococcus aureus		22	part of outbreak			publicST22	22		U WGS for MRSA Outpreak 2013-01-2012:23.44
3	4 quality 5 wos TrimmedStats	Sequence read set types		Staphylococcus aureus Staphylococcus aureus		22	part of outbreak			public3122			
	6 woMLST CallTypes	Character types		Staphylococcus aureus	A neonatal	22	prior to outbreak			publicST22	22		
	7 SNP outbreak	Sequence types		Staphylococcus aureus	A neonatal	6	prior to outpreak	MRSA_16C MRSA_14C_prior		publicST5	5		
] ≩	8 wgsLong	Sequence read set types		Staphylococcus aureus	A neonatal	5 22	part of outbreak			publicST22	22		<
	9 wgMLST MLST PubML				A neonatal	22	part of outbreak			publicST22	22		
	10 arcc	Sequence types		Staphylococcus aureus		22	part of outbreak			publicoree			Identification projects Decision networks
	11 aroe	Sequence types		Staphylococcus aureus	A neonatal	22	part of outbreak	-		publicST22	22		2 + 13 ⊗ €. 6 √.
	12 glof	Sequence types		Staphylococcus aureus	A neonatal	36	during outbreak			publicST36	36		
	13 gmk	Sequence types		Staphylococcus aureus	A neonatal	1	during outbreak	MRSA 16B		publicST1	1		Name Modified date
	14 pta	Sequence types		Staphylococcus aureus	A neonatal	22	part of outbreak			publicST22	22		
	14 pla	Seducince (ypea	- × 5	Staphylococcus aureus		66	S. aureus clu			publicST772	772		
ntry field	Database design			Staphylococcus aureus	A pilot study .		S. aureus clu			publicST772	772		
_					A pilot study		S. aureus clu	т		publicST772	772		
+ 6	9 ⊗ & 6	∽↑↓	<all entry<="" td=""><td>Staphylococcus aureus</td><td></td><td></td><td>S. aureus clu</td><td>Hinseelewsh</td><td></td><td>publicST88</td><td>88</td><td></td><td></td></all>	Staphylococcus aureus			S. aureus clu	Hinseelewsh		publicST88	88		
	Name Fie	ld type	-	Staphylococcus aureus	A pilot study		S aureus clu	C		publicST88	88		
AK	Organism name Fixe			Staphylococcus aureus	A pilot study		S. aureus clu	1		publicST6	6		
	Study title Fixe			Staphylococcus aureus			S, aureus clu	0		publicST15	15		
	ST info Fixe			Staphylococcus aureus	A pilot study		S. aureus clu	0		publicST772	772		Alignments BLAST projects Chromosome comparisons
	outbreak Ebe			Staphylococcus aureus			S. aureus clu	W		publicST772	772		+ 13 8 8 6 7 4
	Patient ID Else			Staphylococcus aureus			S. aureus clu	A		publicST88	88		Name Modified date
	Study accession Fixe			Staphylococcus aureus	A pilot study		S. aureus clu	v		publicST772	772		Modified date
				Staphylococcus aureus			S. aureus clu	N		publicST15	15		
ngerprin	files Power assemblies A	nnotations	Z	Staphylococcus aureus			S. aureus clu	G		publicST88	88		
. E	1 12 0 8 1			Staphylococcus aureus	A pilot study		S. aureus clu	к		publicST5	5		
巴	+ 🖻 🛛 🗞 I	All Fin	igerprint :	Staphylococcus aureus	A pilot study		S. aureus clu	1		publicST30	30		
	le name Expe	riment type Link		Staphylococcus aureus	A pilot study		S. aureus clu	E		publicST88	88		
			^ Z	Staphylococcus aureus	A pilot study		S. aureus clu	R		publicST772	772		
			> × 🗹	e						1		> < >	

Figure 21: The Main window.

- 11. Click on the green colored dot for one of the entries in the **MLST** column in the *Experiment presence* panel.
- 12. Close the character experiment card by clicking on the triangle in the top left corner.

Please consult the *MLST online plugin* manual for detailed instructions.

7 Follow-up analysis

A cluster analysis on the **wgMLST** character experiment (or a subscheme thereof) is created in the *Comparison* window or the *Advanced cluster analysis* window. We will detail here how a dendrogram and minimum spanning tree (MST) can be created from the *Comparison* window and the *Advanced cluster analysis* window, using data from [2].

aroE 6 <+> glpF 1 <+> gmk 5 <+> pta 8 <+> tpi 8 <+>	Character	Value Mapping 🗨
glpF 1 <+> gmk 5 <+> pta 8 <+> tpi 8 <+>	arcC	7 <+>
gmk 5 <+> pta 8 <+> tpi 8 <+>	aroE	6 <+>
pta 8 <+> tpi 8 <+>	glpF	1 <+>
tpi 8 <+>	gmk	5 <+>
	pta	8 <+>
yqiL 6 <+>	tpi	8 <+>
	yqiL	6 <+>

Figure 22: The MLST character experiment card.

7.1 Comparison window

In the WGS demonstration database, three comparisons are already created, corresponding to the three studies (see Figure 23).

News Newslow different data Lawy Newslow of exteriors
Name Modified date Level Number of entries
Neonatal MRSA study 2020-06-02 18:36:39 14
Pilot study 2020-06-03 10:25:10 26
WGS for MRSA outbreak 2020-06-03 10:25:13 57

Figure 23: The *Comparisons* panel with the three comparisons.

Creating a new comparison is easily achieved by selecting the entries you would like to include in the *Main* window and clicking on the + icon in the *Comparisons* panel. Here we will work with the selection of entries present in the saved **WGS for MRSA outbreak** comparison:

- 1. Open comparison **WGS for MRSA outbreak** by double-clicking it in the *Comparisons* panel in the *Main* window.
- 2. Select the **wgMLST** character experiment in the *Experiments* panel of the *Comparison* window.

A valuable addition in the analysis of wgMLST data is the use of character views, i.e. wgMLST subschemes consisting of a subset of loci for a specific research question. Default **All characters** are included in the analysis. Another character view can be selected from the drop-down list in the **Aspect** column (see Figure 24).

7.2 Similarity based clustering

The WGS for MRSA outbreak comparison contains saved cluster analyses, stored in the *Analyses* panel. The experiment and subscheme (between brackets) are indicated (e.g. wgMLST (Core loci)).

3. Switching between the analyses can be done by double-clicking them from the Analyses panel.

</th <th>All Ex</th> <th>periment types></th> <th>្រ ប</th>	All Ex	periment types>	្រ ប
		Name	Aspect
	₹	wgs	<default></default>
	85	wgMLST	Core loci
	AC 6 T	denovo	<all characters=""></all>
	15	quality	<selected characters=""></selected>
	₹	wgs_TrimmedStats	All loci
	85	wgMLST_CallTypes	Core loci
	A C 61	SNP outbreak	wgMLST loci
	₹	wgsLong	MLST PubMLST
	8	wgMLST_MLST PubMLST	ribosomal proteins

Figure 24: Character views in the wgMLST experiment type.

As an example we will perform a new cluster analysis, only based on the 7 traditional MLST loci.

- 4. Select the **MLST PubMLST** character view of the **wgMLST** character experiment in the *Experiments* panel.
- 5. In the *Experiments* panel click on the eye icon () that proceeds **wgMLST**. Select *Characters* > *Show values* () to display the values of the 7 MLST loci.
- 6. Select Clustering > Calculate > Cluster analysis (similarity matrix)..., select Categorical (values), make sure Calculate as distance is unchecked, press < Next>, choose UPGMA in the last step and press < Finish>.

The resulting dendrogram is displayed in the *Dendrogram* panel and the analysis is stored in the *Analyses* panel (see Figure 25). The subscheme that was used is indicated between brackets: **wgMLST (MLST PubMLST)**.

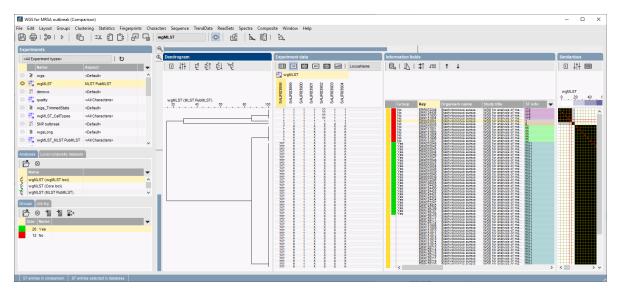


Figure 25: Dendrogram based on the MLST loci.

From the dendrogram it is clear that all the samples with ST 2371 cluster closely together. All these samples were isolated as part of an outbreak, either from patients or from one of the health care workers in the same facility. We will now calculate a dendrogram based on the core loci (alternatively double-click on the saved analysis **wgMLST (Core loci)** in the *Analyses* panel):

7. Select the **Core loci** character view of the **wgMLST** character experiment in the *Experiments* panel.

- 8. Select *Clustering* > *Calculate* > *Cluster analysis (similarity matrix)...* to start a cluster analysis.
- 9. Select the *Categorical (values)* similarity coefficient, press <*Next*>, and select the *UPGMA* clustering method. Press <*Finish*> to start the calculation of the dendrogram.

The resulting dendrogram is displayed in the *Dendrogram* panel. It is clear that the core loci provide a much higher resolution over the MLST set.

To study the relationships in ST 2371 cluster more closely, we can create a new comparison that includes only those entries. We can select only the entries from within the comparison by emptying the current selection and then clicking on the node that contains all the ST 2371 entries while holding the **Ctrl**-key. Alternatively, we can make a new selection in the *Main* window.

- 10. In the *Main* window, clear the current selection with *Database* > *Entries* > *Unselect all entries (all levels)* (F4), then use *Edit* > *Find object in list...* (B, Ctrl+Shift+F) to open the *Find* dialog box.
- 11. Type "2371" and press < *Select all*> to select the 45 entries.
- 12. Create a new comparison for the selected entries by clicking on the + icon in the *Comparisons* panel.

We can add some information to the MST we are about to create, by specifying comparison groups. In the database, samples isolated from a patient have the label "Yes" in the field **Outbreak**, whereas the samples isolated from a health care worker do not carry a label:

- 13. Right-click on the Outbreak column header in the Information fields panel and select Create groups from database field. In the Group creation preferences dialog box, leave the settings at their defaults and press < OK >.
- 14. Select the **wgMLST loci** aspect for **wgMLST** in the *Experiments* panel.
- 15. In the *Experiments* panel click on the eye icon () that proceeds **wgMLST**. Select *Characters* > *Show values* () to display the values of the wgMLST loci.
- 16. Select *Clustering* > *Calculate* > *Cluster analysis (similarity matrix)....*

A disadvantage of the *Categorical (values)* similarity coefficient is that the number of different loci cannot easily be deduced from the dendrogram or similarity matrix. The *Categorical (differences)* coefficient is more suitable for this purpose.

17. Select the *Categorical (differences)* coefficient from the list.

The *Categorical (differences)* coefficient treats each different value as a different state, and results in a distance matrix.

With the *Scaling factor* one can deal with the hard-coded maximum of 200 that can be calculated for a distance value. Values that make sense are 1, 10 and 100, allowing the correct visualization of maximally 200, 2000 and 20000 different character values, respectively, in a cluster analysis.

18. In this example, choose a *Scaling factor* of 1.

19. Press < *Next*>, choose *Complete Linkage* in the last step and press < *Finish*>.

The resulting dendrogram is displayed in the *Dendrogram* panel.

20. To view the number of allele differences on the branches, select *Clustering* > *Dendrogram display settings...* (11), and tick the option *Show node information*. Press < *OK* >.

To trace back the number of different loci from the branches or distance matrix, the displayed values needs to be multiplied with the *Scaling factor* used (in this example: 1).

21. The polymorphic loci for the set of samples in the selected scheme can be displayed with *Characters* > *Filter characters* > *Select polymorphic characters...*

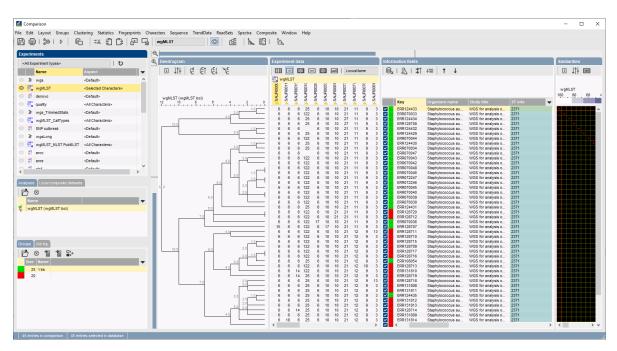


Figure 26: Complete linkage dendrogram.

22. Save the comparison with *File* > *Save as...*. Specify a name (e.g. **ST 2371**).

7.3 Minimum spanning tree

A minimum spanning tree is calculated in the *Advanced cluster analysis* window which is launched from the *Comparison* window.

- 23. Open the saved comparison **ST 2371** or create a new comparison containing all 45 entries in the database belonging to ST 2371.
- 24. Select the **wgMLST loci** character view of the **wgMLST** character experiment in the *Experiments* panel.
- 25. Select *Clustering* > *Calculate* > *Advanced cluster analysis...* in the *Comparison* window to launch the *Create network* wizard.

The predefined template *MST for categorical data* uses the categorical coefficient for the calculation of the similarity matrix, and will calculate a standard minimum spanning tree with single and double locus variance priority rules.

26. Specify an analysis name (for example wgMLST MST), make sure wgMLST (wgMLST loci) is selected, select MST for categorical data, and press < Next>.

A MST is now computed in the Advanced cluster analysis window.

27. To add more information to the MST, go to *Display > Display settings*. In the *Node labels and sizes panel* of the *Display settings* dialog box, check *Show node labels*. Choose *Patient ID* in the drop-down list and leave the other settings at their default values.

- 28. In the *Branch labels and sizes panel* of the *Display settings* dialog box, we can specify that we want to see the distances between the nodes (i.e. the number of allele differences): check **Show branch labels** and set **Number of digits** to "0".
- 29. Click < OK> to close the Display settings dialog box.

The MST is now displayed with node and branch labels.

- 30. Zooming can be done with the zoom slider on the left side of the image, and the size of the nodes can be adjusted with the zoom slider at the top. By holding the **Ctrl**-key and dragging a node with the mouse, the node can be repositioned in any direction.
- 31. Export the image via *File* > *Export image...* and save in the format of your choice.

The resulting MST gives a high resolution map of the outbreak. The colors allow to distinguish easily between patient samples (P) and MRSA colonies isolated from a health care worker (HC).

The branch labels indicate how many allele differences were found between each linked set of entries.

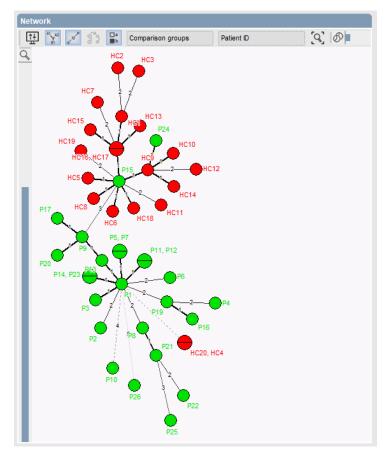


Figure 27: MST of the entries with ST2371.

By repeating the analysis steps using the character aspect *Core loci*, it can be demonstrated that wgMLST results in a higher-resolution MST then core genome MLST.

8 Core and pan genome analysis

The pan-genome of a bacterial species consists of a core and an accessory gene pool. As the wgMLST locus set is defined as pan-genomics scheme over all available organism genome se-

quences, the analysis can be limited to the pan-genomic and/or core genomic loci for the selected sample set in the comparison.

For a selected set of samples, the core set of loci can be defined as follows:

- 1. Select all entries in the *Main* window and click on the + icon in the *Comparisons* panel.
- In the *Experiments* panel of the *Comparison* window, highlight the wgMLST character experiment, make sure the "<All characters>" aspect is selected and select *Statistics* > *Core locus* analysis....

This opens the *Core locus analysis* dialog box where the *Number of repeats* and *Presence threshold* can be defined.

The determination of the number of core loci is based on sub-sampling the entries in the comparison. As such, the *Number of repeats* can be defined, i.e. the number of subsamples taken from the comparison set.

The **Presence threshold** indicates the minimum presence (expressed in %) for a locus to be called within the core. Entering 90%, will imply that only loci present in 90% of the entry selection will be identified as core loci. For a very strict analysis, one can put the presence threshold at 100%, limiting the core to only those loci which are present in all the entries under evaluation i.e. present in the comparison.

- 3. Set the *Presence threshold* to 100% and press < *OK* > to start the analysis. When the analysis has finished, the results open in the *Charts and statistics* window.
- 4. To create a Core genome analysis plot as shown in Figure 28, highlight Average number of loci, select Plot > Add new plot from selected properties... (+), choose Profile chart and press <Next> and <Finish>.
- 5. Repeat Instruction 4 for data sources *Minimum number of loci* and *Maximum number of loci*.

The result is shown in Figure 28.

6. The values used to create these curves can be viewed by making a selection of all data sources (click while holding the Ctrl-key) and selecting *Dataset* > *View selected properties* (Ⅲ).

For details on all the possibilities of the *Charts and statistics* window, please consult the BIONU-MERICS reference manual.

The core loci are now also selected in the **wgMLST** character experiment, in the form of a subsetbased character view.

7. Double-click on the **wgMLST** character experiment in the *Experiment types* panel of the *Main* window, create a selection based query, and specify a name that is different from the pre-defined **Core loci** subscheme, e.g. **Local core loci**.

From the same *Comparison* window, also a pan locus analysis can be done.

8. In the *Comparison* window select *Statistics* > *Pan locus analysis...*. As for the Core locus analysis, the *Number of repeats* and *Presence threshold* can be defined from the *Pan locus analysis* dialog box.

Similar to the determination of the number of core loci, the number of pan loci is also based on sub-sampling the entries in the comparison. As such, the *Number of repeats* can be defined, i.e. the number of subsamples taken from the comparison set.

The *Presence threshold* indicates the minimum presence (expressed in %) for a locus to be called within the pan loci. Entering 5%, will imply that only loci present in at least 5% of the selected entries will be identified as pan loci. For a very non-restrictive analysis, one can put the

Charts and statistics	- D X
File Dataset Plot Window Help	
Chart list	Chart area
⋧ + 図 ⊗ 井 ↑ ↓	· ◎ ዿ ନ 🖫 💷
Name Type Visible Profile chart (Maximum number of loci) Profile chart Yes Profile chart (Minimum number of loci) Profile chart Yes Profile chart (Average number of loci) Profile chart Yes	 Profie chart (Average number of loc) Profie chart (Maximum number of loc)
Ata source overview	age entruper of loci, Mi
Maximum number of loci (Value) 🗹	Legend Chart report Profile chart (Average number of loci) Image: Chart (Maximum number of loci) Profile chart (Maximum number of loci) Image: Chart (Maximum number of loci) Profile chart (Maximum number of loci) Image: Chart (Maximum number of loci) Profile chart (Maximum number of loci) Image: Chart (Maximum number of loci) Profile chart (Maximum number of loci) Image: Chart of loci Image: Chart (Maximum number of loci) Image: Chart of loci Image: Chart (Maximum number of loci) Image: Chart of loci Image: Chart (Maximum number of loci) Image: Chart of loci Image: Chart (Maximum number of loci) Image: Chart of loci Image: Chart (Maximum number of loci) Image: Chart of loci Image: Chart (Maximum number of loci) Image: Chart of loci Image: Chart (Maximum number of loci) Image: Chart of loci Image: Chart (Maximum number of loci) Image: Chart of loci Image: Chart (Maximum number of loci) Image: Chart of loci Image: Chart (Maximum number of loci) Image: Chart of loci Image: Chart (Maximum number of loci) Image: Chart of loci Image: Chart (Maximum number of loci) Image: Chart of loci Image: Chart (Maximum number of loci) Image: Chart of loci Image: Chart (Maximum number of loci) Image: Chart of loci

Figure 28: Core locus analysis for all samples in the wgMLST demonstration database (*Presence threshold* 100%).

presence threshold at 0%, defining the pan loci as all the loci which are present in at least one of the entries.

- 9. Set the *Presence threshold* to 5% and press <*OK*> to start the analysis. When the analysis has finished, the results open in the *Charts and statistics* window.
- 10. To create a Pan genome analysis plot as shown in Figure 29, perform exactly the same steps as in Instruction 4 and Instruction 5.

The Pan loci are now also selected in the **wgMLST** character experiment, in the form of a subsetbased character view.

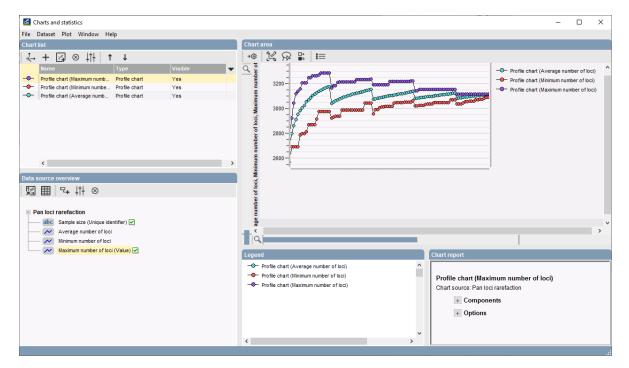


Figure 29: Pan locus analysis for all samples in the demonstration database (*Presence threshold*: 5%).

Bibliography

- [1] David W Eyre, Tanya Golubchik, N Claire Gordon, Rory Bowden, Paolo Piazza, Elizabeth M Batty, Camilla LC Ip, Daniel J Wilson, Xavier Didelot, Lily O'Connor, et al. A pilot study of rapid benchtop sequencing of staphylococcus aureus and clostridium difficile for outbreak detection and surveillance. *BMJ open*, 2(3):e001124, 2012.
- [2] Simon R Harris, Edward JP Cartwright, M Estée Török, Matthew TG Holden, Nicholas M Brown, Amanda L Ogilvy-Stuart, Matthew J Ellington, Michael A Quail, Stephen D Bentley, Julian Parkhill, et al. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant staphylococcus aureus: a descriptive study. *The Lancet infectious diseases*, 13(2):130–136, 2013.
- [3] Claudio U Köser, Matthew TG Holden, Matthew J Ellington, Edward JP Cartwright, Nicholas M Brown, Amanda L Ogilvy-Stuart, Li Yang Hsu, Claire Chewapreecha, Nicholas J Croucher, Simon R Harris, et al. Rapid whole-genome sequencing for investigation of a neonatal mrsa outbreak. *New England Journal of Medicine*, 366(24):2267–2275, 2012.