

BIONUMERICS Tutorial: Hash-based wgMLST typing

1 Introduction

Hash-based allele calling is a viable option to perform wgMLST without the need for a centrally maintained nomenclature and could therefore still be used after the Applied Maths Cloud Calculation Engine will be taken offline on December 31st, 2024, as part of the BIONUMERICS phase-out (see https://www.bionumerics.com/news/bionumerics-phasing-out).

In the hash-based allele calling algorithm implemented in the *WGS tools local plugin*, allele sequences that fulfill the nomenclature acceptance criteria are converted by a hash function into 17-digit integer numbers. Since hash functions are deterministic, a given allele sequence will always result into the same hash value. However, hash values as such cannot be used directly in the software, because the character experiment type in BIONUMERICS is not designed to reliably store and compare 17-digit integers. Hash values also look very different from the simple integer allele IDs that wgMLST users are familiar with.

To circumvent these issues, the *WGS tools local plugin* sets up a local allele nomenclature in which the hashes are converted into simple integer IDs. The latter are stored in the **wgMLST_Local** experiment type, which is automatically created and synchronized with the **wgMLST** experiment type. Each time a wgMLST assembly-based allele calling job result is retrieved, the **wgMLST_Local** experiment is automatically updated with the hash-based allele calls using the local nomenclature.



The **wgMLST_Local** experiment type only considers assembly-based wgMLST allele calls, assembly-free allele calling is not taken into account.

To allow exchange of wgMLST data between different labs and/or databases, local wgMLST profiles can be exported as hash values, and imported from hash values (see 6).

The process for setting up and running hash-based wgMLST allele calling will be illustrated in this tutorial using the *Listeria monocytogenes* demonstration database.

2 Preparing the database

2.1 Option 1: Download 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).

🖆 Tutorial databases						×
File Database Tutorial Window						
r Tutorial databases						
	_					
Name	Size (MB)	Min	Max	-		
BandScoring demo database	15	7	8.1	^	The DemoBase Connected demonstration	\sim
DemoBase Connected	4	7.5	8.1		database. This database is used in several	
Neisseria MLST demo database	1	7.5	8.1		tutoriais.	
SNP demonstration database	0	7.5	8.1			
WGS_demo_database_for_Brucella_spp	214	7.6	8.1			
WGS_demo_database_for_Burkholderia_cepacia	142	8	8.1			
WGS_demo_database_for_Escherichia_coli	601	7.5	8.1			
WGS_demo_database_for_Listeria_monocytogenes	288	7.5	8.1			
WGS_demo_database_for_MTBC	279	8	8.1			
WGS_demo_database_for_Salmonella_enterica	618	8	8.1			
WGS_demo_database_for_Staphylococcus_aureus	624	7.6	8.1			
						\sim
					Į.	
Tutorials						
Name				-		
Band matching and polymorphism analysis				^	This tutorial illustrates how to calculate a	\sim
Entry information fields and their properties					Principal Components Analysis (PCA) and a	
Configuring the database layout					Multi Dimensional Scaling (MDS)	
Selecting entries in a database					Coordinates Analysis (PCoA)) on a	
Combined analysis of fingerprint data					fingerprint data set and how to change the	
Calculating a PCA and an MDS on a fingerprint data set					layout of the obtained plots.	\sim
Clustering a phenotypic test assay				~		

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.

2.2 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_LMO1.bnbk database backup file into WGS_LMO1.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. "WGS Listeria demobase".
- 10. Click < OK > to start restoring the database from the backup file (see Figure 2).

C:1	\Users\Public\Docume	nts\wgMLST_LN	IO.bnbk		Browse
Des	tination				
	Overwrite				
0	Overwrite the curre monocytogenes wit new data since the recovered.	nt data in databa h a previous ver backup will disa	ase WGS dem sion of the da ppear, any old	o database for ta as stored in data that was	Listeria a backup. Any erased will be
	Create copy				
۲	Create a new datab have access to bott copy a database fro	ase from the da h old and curren om another comp	ta stored in a l t versions of a outer.	backup. Use th in existing data	is option to abase or to
Nev	w database name:	wgMLST Liste	ria demobas		

Figure 2: Restoring the WGS demonstration database from the backup file WGS_LM01.bnbk.

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

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

🔄 wg	6 demo database for Lister	ria monocytogen	es - BioNumerics													- п х
F11. F.4	b Database Assistants	C	and Mendaus II													
rile co	it Database Analysis	Scripts Wust	ools window H	ыb												
D E	± ⊞ ,c= ©) <u> [</u>]]	e, e, e,													
Experi	ment types			Da	tabase entries										Comparisons	
Æ			▽ †	114	ല – ല			<all entries<="" td=""><td></td><td>1.75</td><td></td><td></td><td></td><td></td><td>+ PÅ ⊗ B</td><td></td></all>		1.75					+ PÅ ⊗ B	
0				1 -	· · · ·		Collected	Collection	Castas		Parametetet		C			
	# Name	Type			ксу	Modified date	CollectedBy	Collection	GeoLoc	IsolationSource	sampleid	serovar	source	• 1 2 3 4 5 6 7	Name	Modified date
1 8	1 wgs	Sequenc	e read set types		SRR1112181	2015-12-08 09:40:46	CDC	missing	USA	missing	SAMN02582707	<not prov<="" td=""><td>a balance</td><td></td><td>Isolates taly</td><td>2015-12-08 12:38:4</td></not>	a balance		Isolates taly	2015-12-08 12:38:4
	2 WgMLST	Characte	rtypes		SRR1378346	2015-12-08 09:40:46	missing	2014-01-21	Indonesia	trozen igt raw pe	SAMN02850677	<not prov<="" td=""><td>shrimp</td><td></td><td></td><td></td></not>	shrimp			
	3 denovo	Sequenc	e types		SRR1535729	2015-12-08 09:40:46	PDA CDC	2014-06-04	Mexico	avocados	SAMN02934629	<not prov<="" td=""><td>avocado</td><td></td><td></td><td></td></not>	avocado			
	+ quality	Characte Seguene	r types		SPD1562163	2015-12-06 09:40:46	EDA	2014-08	Baby	chases pastor	SAMN03013/1/	1/28	obodo esetos			
	6 wg6_frimmedStat	is Sequenc	e reau set types		SRR 1500202	2010-12-00 09:40:40	EDA	1003-00	kaly habi	cricese pastry	CAMM02700737	1/20	tridese pastry			
	7 wool.con	es characte	r types		SRR1500203	2015-12-08 05:40:46	FDA	1993-06	Rahr	wel au west ebrime	SAMM02769737	1/20	nul cale			
	/ wyocony	Jequeiro	e reau set types		S004622042	2010-12-00 00:40:40	EDA	00 See 2011	USA CO	bleed	SAMM02703733	1/20	blad		<	>
	<		>		SRR1623013	2015-12-08 09:40:46	EDA	9/5/2011	USA-CO	blood	CAMM02769780	1/26	blood			
-					SRR1623016	2015-12-08 09:40:46	FDA	3/3/2011 22 Sep 2011	USA:CO	blood	SAMN02769769	1/20	blood		Identification projects Dec	ision networks
Entry fi	elds Database design				SRR1023030	2010-12-00 00:40:40	Austrian A	22-360-2011	Carrier	human	SAMM02103185	1/2.8	human		29 + P &	R B
+	P & R P	5 2	1 J <ai< td=""><td></td><td>CDD1666072</td><td>2015-12-08 09:40:46</td><td>Austrian A</td><td>2014</td><td>Austria</td><td>human louor</td><td>CAMM03100007</td><td>46</td><td>human</td><td></td><td></td><td></td></ai<>		CDD1666072	2015-12-08 09:40:46	Austrian A	2014	Austria	human louor	CAMM03100007	46	human			
		Tield trees			SRR1636973	2015-12-06 09:40:46	Austrian A	2014	Austria	human iquor	SAMN03100000	40	human		Name	Modified date
	Name	Pield type	_		SPD1695809	2015 12 08 09:40:46	CDC	Oct 2014	LISA A	Blood	SAMM03253354	46	blood			
AR	CollectedBy	Fixed			CDD1605011	2015-12-08 09:40:46	CDC	001-2014	UEA	Blood	SAMM03253354	40	blood			
AK	CollectionDate	Fixed			0004005040	2015-12-00 03:40:40	000	minning	USA	minoine	CAMM02253350	reat areas	0000			
AK	GeoLocName	Fixed			SRR1033010	2015-12-08-09-40-46	CDC	New 2014	USA	placenta	SAMM03253362	46	niscente			
AR	IsolationSource	Fixed			SRR1695623	2015-12-08 05:40:46	CDC	Nov-2014	UEA	ces	SAMN03253369	40	ces			
AIC	Sampleld	Fixed			CDD4005034	2010-12-00 03:40:40	000	New 2014	USA	Bland	CAMM02253300	40	blad			
AIC	serovar	Fixed			SRR1053030	2013-12-08 09:40:40	EDA	4002.00	Baby.	Dioda	SAMN03233302	40	bibbo		<	>
ARC	Source	Fixed			SRR1709500	2015-12-06 05:40:46	FDA	1993-06	Rohr	freezer ourface	SAMN02709739	1/28	freezer eurface			
AIC	wgMLST loci ST	Flexible			SRR1703023	2010-12-00 00:40:40	EDA	1003.00	Raly	hiered	SAMM02703730	1/26	hited		Alignments BLAST project	ts Chrom. Comp.
AR	MLST ST	Flexible			SRR1705001	2015-12-08 09:40:46	CDO.	1993-06	Laly UCA	Diood	SAMN02708734	1720	blood		+ 13 0 8	
ARC	MLST PubMLST ST	Flexible			SRR1745456	2015-12-08 09.40.46	000	NOV-2014	USA	Biolog	SAMINU3275454	40	biblio		1 [] 0 63	
-					SRR1745445	2015-12-00 09:40:40	CDC	Nov-2014	USA	Sputum	SAMN03275461	40	sputum		Name	Modified date
ringerp	nnt nes Power assemble	8 Annotations			SRR1745456	2015-12-08 09:40:46	CDC	Nov-2014	UDA	000	SAMN03275472	1720	0000			
同	+ 13 & &	161 1	<all fingers<="" td=""><td></td><td>SRR1745404</td><td>2015-12-06 09.40.46</td><td>000</td><td>Of Des Oot 4</td><td>USA</td><td>Cor</td><td>SAMINU3275476</td><td>40</td><td>Cor Maria</td><td></td><td></td><td></td></all>		SRR1745404	2015-12-06 09.40.46	000	Of Des Oot 4	USA	Cor	SAMINU3275476	40	Cor Maria			
	File name	Experiment type	Link		C004746499	2010-12-00 09:40:40	000	01-Dec-2014	UPA	Direct	CAMM03275491	1/20	blood			
	The Harrie	experiment type	CIIIX V		SRR1745488	2010-12-06 09:40:46	000	01-Dec-2014	UDA	Dioud	SAMN03275500	-0	blood			
			~		SHK1/46/6/	2015-12-06 09:40:46	EDA	NOV-2014	USA	np iuo	SAMNU3198339	40	nip nuo			
					SRR1763633	2010-12-00 09:40:40	FDA.	2010-01-00	UCA-NO	granny smith apples	CAMMP03270030	snot prov	appres		~	
	<		,		SHORT /63858	2010-12-06/09:40:46	rue.	2010-01-06	USA.MO	granny smith apples	5AMR03278634	snot prov	appres	> < >		>
																,
Datal																

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

3 About the demonstration database

The WGS Listeria demo database contains links to sequence read set data on NCBI's sequence read archive (SRA) for 51 publicly available sequencing runs. 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 *L. monocytogenes* wgMLST scheme on the Applied Maths Calculation Engine.

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

The calculation engine project is linked to the *Listeria monocytogenes* 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 4).

alculation engine	Organism	Experiment types	WgMLST			
Remote calculatio	n engine					
Urt: Applied Maths cloud On premises						
https://wgmlst.applied-maths.com						
Project name:	eval_lmo_	demo				
Password:	•••••	•••••				
Time out (sec):	120					
Own computer						
Interval (min):	1 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*> twice to close the *Calculation engine settings* dialog box.

Experiment types linked to wgMLST analysis are present in the database for each of the entries and are displayed in the *Experiment types* panel (see Figure 7):

- 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	Organism	Experiment types	wgMLST		
New allele submis	sion				
Lab ID: eval_	lmo_demo				
Submit new a	lleles autom	natically			
	sion criteria				
Multiple allele calls	character	values			
Store lowest	common al	llele ID			
O Store as abs	ent value				
Allele mapping ex	CT DubML				
wgMLST_ML	re Pasteur	51			
			OK	0	nool

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

Auto submission criteria	?	×					
Use nomenclature acceptance criteria							
O Set custom submission criteria							
Require start/stop codons (for CDS loci only)							
Allow internal stop codons (fo	or CDS loc	i only)					
Minimum homology 85	.0 %						
Maximum number of gaps 99	9						
ОК	Car	ncel					

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).

Additional information, stored in entry info fields (CollectionDate, CollectedBy, serovar, etc.) was collected from the corresponding publications and added to the demonstration database.

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.

Experim	Experiment types								
କ୍ଷ	-	+ 🗗 🛛 🗟	All Experiment types>						
		Name	Туре	-					
∎≩]≵ 1 wgs Se		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_CalTypes Character types		Character types						
⊒ ≩	7	wgsLong	Sequence read set types						
				~					

Figure 7: The Experiment types panel of the Main window.

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 8).



Figure 8: 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 9).

SRR1378346							
Character		Mapping	•				
LMO_1	1	<+>	^				
LMO_13	13	<+>					
LMO_19	2	<+>					
LMO_21	2	<+>					
LMO_46	2	<+>					
LMO_48	2	<+>					
LMO_104	2	<+>					
LMO_106	1	<+>					
LMO_111	7	<+>					
LMO_125	1	<+>	¥				
Press Insert to add chara	cter						

Figure 9: The character experiment card for an entry.

- 7. Close the character experiment card by clicking on the triangle in the top left corner.
- 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 10).



Figure 10: 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 11).

SRR1112181							
Character	Value	Mapping	•				
AvgQuality	36	<+>	^				
AvgReadCoverage	39	<+>					
N50	219659	<+>					
NrContigs	31	<+>					
NrNonACGT	128	<+>					
Length	2838247	<+>					
KeywordCov	52	<+>					
NrAFMultiple	8	<+>					
NrAFPerfect	2701	<+>					
NrAFPresent	2857	<+>	~				
Press Insert to add chara	acter						

Figure 11: The character experiment card for an entry.

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

4 Installation of the WGS tools local plugin

The WGS tools local plugin is designed to work with the latest BIONUMERICS version. Earlier versions are not tested an might have compatibility issues. Please check our software download page (https://www.bionumerics.com/download/software) to see if an update to your current version is available.

The WGS tools local plugin should be installed in a BIONUMERICS database in which the WGS tools plugin is already installed and connected to a Calculation Engine instance via a CE project and password.



Installation of the *WGS tools local plugin* should be done before December 31th, 2024 while the Applied Maths Cloud Calculation Engine is still up and running.

The *WGS tools local plugin* is made available as an online plugin, which can be installed in the relational database. This has as an advantage that no Windows administrator rights are required for installation. Furthermore, in a multi-user database setup, this procedure ensures that all database users work with the same plugin version.

Proceed as follows to install the WGS tools local plugin:

- 1. Select *File* > *Install / remove plugins...* () in the *Main* window to call the *Plugins and Scripts* dialog box.
- 2. Press the <*Manage database plugins*> button to open the *Manage database plugins* dialog box.

The *Manage database plugins* dialog box lists the plugins that are currently stored in the relational database. Likely, this list is initially empty.

- 3. Press the <*Add/Update*> button to open the *Add database plugins* dialog box (see Figure 12.
- 4. Check the check box in front of the WGS tools local plugin and click < OK >.

A message appears, indicating that the plugin will be loaded after the database is restarted.

5. Press *<Close>* in the *Add database plugins* dialog box and repeat the same action in the *Plugins and Scripts* dialog box.

		Add files with button [B	roweal		Browse Delete	
Add tiles with outdon [strowse] or drag and drop from explorer.						
_	Plugin Coo plugin		Installed version	Latest version		
	Geo plugin SARS-CoV-2			2.02 0.54		
~	wgMLST Local			1.00		



- 6. Close and restart the BIONUMERICS database.
- 7. Select *File* > *Install / remove plugins...* (,) in the *Main* window to call the *Plugins and Scripts* dialog box again.

The WGS tools local plugin is now displayed in the Plugins tab of the Plugins and Scripts dialog box at the bottom of the list and is preceded by a database icon \equiv (see Figure 13).

Plugins Scripts Resistance detection Resistance detection RiboPrinter Surues functional genotyping Salmonella functional genotyping SARS-CoV-2 SNP calling Spa typing User management tools wgMLST admin wgMLST admin wgMLST validation WGS tools WGS tools WGS tools Show only activated Show valways active' plugins Manage database plugins	Plugins & Scripts			×
Resistance detection Version 1.00 RiboPrinter Saureus functional genotyping Salmonella functional genotyping This plugin provides additional functionality to make the WGS tools fully local. SARS-CoV-2 SNP calling Spa typing User management tools wgMLST admin wgMLST biological Validation wgMLST curator wgMLST validation ✓ WGS tools local ✓ Show only activated Show valways active' plugins Install Show Manual	Plugins Scripts			
	Resistance detection RiboPrinter S. aureus functional genotyping Salmonella functional genotyping SARS-CoV-2 SNP calling Spa typing User management tools wgMLST admin wgMLST biological Validation wgMLST curator wgMLST validation ✓ WGS tools WGS tools local Show only activated Show valways active' plugins Install Show Manual Manage database plugins	*	Version 1.00 Database installation This plugin provides additional functionality to make the WQS tools fully local.	

Figure 13: The *Plugins and Scripts* dialog box with the *WGS tools local plugin* high-lighted.

8. Click on the *WGS tools local plugin* to highlight it, press <*Install*> and confirm the plugin installation.

The plugin automatically creates and initiates an experiment type wgMLST_Local by performing

a synchronization with the Calculation Engine.



The WGS tools local plugin will use the experiment type that is specified to contain the wgMLST allele calls in the *Experiment types panel* of the *Calculation engine settings* dialog box and will add the **_Local** suffix to the experiment name. Even though the actual name might be different, the experiment type will be referred to as **wgMLST_Local** for reasons of conciseness.

9. When the synchronization is complete, press < OK > to close the message box.

A message box pops up with the question "Do you want to initialize the local allele nomenclature with the accepted alleles?".

Answering "No" will create a local wgMLST nomenclature from scratch: allele IDs are assigned starting from 1 for each locus.

Answering "Yes" will start the local nomenclature from the current accepted alleles in the central nomenclature and new alleles will be added by increasing integer identifiers.



When the assembly-based accepted alleles are next updated, there might be newly accepted alleles with IDs that have meanwhile already been assigned to (other) local alleles. Inevitably, from the point of initialization on, the allele IDs will start to diverge!

If the assembly-based wgMLST search data are not up-to-date, an update of the search data will be performed first. This may take several minutes.

10. Select < *Yes*>.

This action will create a file wgMLST_Local_Nomenclature.txt in the source files directory with the local nomenclature.

A notification appears when the process is completed, indicating that the plugin is installed and prompting to restart the database.

11. Press < *Close*> in the *Plugins and Scripts* dialog box and close and restart the BIONUMERICS database.

5 (Re-)running assembly-based wgMLST calls

From the *Experiment presence* panel in the *Main* window, it can be seen that there are currently no experiments present for **wgMLST_Local**. To populate **wgMLST_Local** experiments for all entries in the database, it is required to re-run all assembly-based wgMLST calls. On the local calculation engine, this is a relatively quick process that does not require CE credits. For extremely large databases, we recommend submitting the jobs in batches of maximum 1000 entries.

- 1. Select all 51 entries in the *Listeria monocytogenes* demonstration database using *Edit* > *Select all* (Ctrl+A).
- 2. Select *WGS tools* > *Submit jobs...* (\triangleright).
- 3. In the *Submit jobs* dialog box (see Figure 14), make sure *Own computer* is checked to run the jobs on the local calculation engine, which does not require CE credits.
- 4. Check *wgMLST assembly-based calls* under *Algorithms* and *Re-submit already processed data* to make sure that the jobs are re-run.
- 5. Press the < OK > button to submit the jobs.

ubmit jobs	? ×
Submit jobs to:	
Algorithms De novo assembly wgMLST assembly-based calls	wgMLST assembly-based calls Performs a BLAST-based wgMLST allele calling on the assembled genome sequence Settings
Jobs Submitting 51 jobs for 51 entries. Re-submit already processed data Open jobs overview window Credits available for job submission:	0
Credits needed to submit selected jobs: Buy extra credits	0
	OK Cancel

Figure 14: The *Submit jobs* dialog box, ready to re-run all wgMLST assembly-based allele calling jobs on the local calculation engine.

If the local assembly-based wgMLST search data are not up-to-date, they will be updated first. This might take several minutes. Next, the jobs are submitted to the local calculation engine.

6. If not already open, select *WGS tools* > *Jobs overview...* (*) to open the *Job overview* window.

The assembly-based allele calling jobs complete relatively fast. In our test setup, the 51 jobs finished in about 25 minutes and it took another 5 minutes to retrieve the job results and store them in the BIONUMERICS database.

- 7. Select View > Refresh (v, F5) to check the execution progress.
- 8. When the jobs are completed, highlight them in the *Job overview* window and select *Jobs* > *Get results* () to retrieve the results.

The *Experiment presence* panel now shows a green dot for the **wgMLST_Local** experiment type with each entry in the database.

Click on a green dot for the wgMLST_Local experiment of a random entry and on one for the wgMLST experiment of the same entry to display their *Experiment card* window (see Figure 15).

Since we started the local nomenclature from the accepted alleles of the central nomenclature (see 4), both wgMLST profiles look almost identical. Small differences can occur (as illustrated in Figure 15 for locus LMO_49) because assembly-free allele calls are taken into account for **wgMLST**, while **wgMLST_Local** uses exclusively assembly-based allele calls. As time progresses, both nomenclatures will start to diverge and more differences will be observed between both experiments because of new alleles that are assigned different IDs. When the Calculation Engine goes offline, new alleles will not be called for **wgMLST_Local** will remain complete because the hash-based allele calling will continue to call new alleles.

B WGS_demo_database_for_Listeria_monocytogenes	- BioNumerics			- 🗆 X
File Edit Database Analysis Scripts WGS tools	Window Help			
	°a 8.			
	• @W *V			
Experiment types	Database entries			Comparisons
	PL P O O I O		1.85	
			10	
# Name Type 🔻	Key Collected	dBy CollectionDate	GeoLocName 🔻 1 2 3 4 5 6 7 8	Name Mod 🔻
⊇ ≩ 1 wgs Sequence rea ^	SRR1112181 CDC	missing	USA ••••••	^
2 wgMLST Character type	SRR1378346 missing	2014-01-21	Indonesia • • • • • •	
3 denovo Sequence typ	SRR1535729 FDA	2014-06-04	Mexico	
4 quality Character type	SRR1562163 CDC	2014-08	USA ••••••	
	SRR1566202 FDA	1993-06	italy	
→ work_ong Services to	SRR1506203 SRR1657003		SRR1657003	
8 woll ST Local Character tro	Spp1623013 Character	Value Mapping	Character Value Mapping	-
o wgmest_cocar character typ v	SRR1623016 LMO 1	1 (1)		
< >	SBB1623030 LMO_13	2 (+>	LMO_13 2 <+>	··· < >
Entry fields Database design	SRR1654984 I MO 19	2 <+>	LMO 19 2 <+>	Identification projects Decis, NW.
	SRR1656973 LMO 21	1 <+>	LMO 21 1 <+>	
	SRR1657003 LMO 32	4 <+>	LMO 32 4 <+>	23 + □ ∞ € Ε
Name Field type 🤝	SRR1695809 LMO_46	2 <+>	LMO_46 2 <+>	Name Moc 🤝
All CollectedBy Fixed	SRR1695811 LMO_48	2 <+>	LMO_48 2 <+>	
All CollectionDate Fixed	SRR1695816 LMO_49	131 <+>	LMO_104 2 <+>	
GeoLocName Fixed	SRR1695823 LMO_104	2 <+>	LMO_106 1 <+>	
All IsolationSource Fixed	SRR1695834 LMO_106	1 <+>	LMO_107 1 <+>	
All Sampleid Fixed	SRR1695836 LMO_107	1 <+>	LMO_125 1 <+>	
All serovar Fixed	SRR1709560 LMO_125	1 <+>	LMO_130 2 <+>	
ANC Source Fixed	SRR1709629 LMO_130	2 <+>	LMO_131 3 <+>	
All: MLST PubMLST ST Flexible	SRR1709681 LMO_131	3 <+>	LMO_132 1 <+>	
MLST PubMLST ST Flexible	SRR1745438 LMO_132	1 <+>	LMO_134 1 <+>	< >
< >	SRR1745445 LMO_134	1 <+>	LMO_135 2 <+>	Alexandra RLACT Character Come
	SRR1745458 LMO_135	2 <+>	LMO_143 2 <+>	Augminents BLAST Chrom. Comp.
Fingerprint files Power assemblies Annotations	SRR1745464 LMO_143	2 <+>	LMO_147 2 <+>	+ [ੈ ⊗ 🔩 🛍
月 + 7→ ⊗ 昆 品 ∞.	SRR1745479 LMO_147	2 <+>	LMO_154 3 <+>	Name Mod 🔻
File name Experiment type	SRR1/45466 LMO_154	3 <+>	LMO_16/ 2 <+>	×
	Press Insert to add cha	iracter	Press Insert to add character	
	SRR1763858 FDA	2015-01-06		
	SRR1767762 FDA	2015-01-00	USA:CA	
	SRR1767763 FDA	2015-01-01	USA:CA	
	SRR1767799 FDA	2015-01-01	USA:CA	
	SRR1767823 FDA	2015-01-01	USA:CA • • • • • •	
	SRR1770492 FDA	2015-01-12	USA:NE • • • • • •	
< >	< <		, · · · · · ·	v < >
		-		
Database: WGS_demo_database_for_Listeria_monocyto			ts C:\Users\Public\Documents\homedir8\WGS_demo_da	

Figure 15: The *Main* window with an *Experiment card* window for **wgMLST** (left) and **wgMLST_Local** (right) for entry SRR1657003. Note that locus LMO_49 is called differently in both profiles.

The **wgMLST_Local** experiments can be analyzed with the same methods and the results interpreted in the same way as **wgMLST** experiments.

6 Import and export of hash-based calls

Because a *local* nomenclature is used, **wgMLST_Local** allelic profiles from different BIONUMER-ICS databases cannot be directly compared with each other. In order to exchange and compare data with other databases or labs, the allele IDs need to be converted into something that can be uniquely referenced, i.e. the hash values obtained from the allele sequence.

Proceed as follows to export allele hashes for a set of entries:

- 1. Select the entries that you want to export and use *File* > *Export...*.
- 2. In the *Export* dialog box, under the topic *Character type data*, highlight *Export fields and hashed wgMLST calls* and press <*Next*>.
- 3. In the *Export* dialog box that appears, select the *Fields* and the subset of the **wgMLST_Local** experiment (under *Characters*) that you wish to include in the export. Press <*Next*>.
- 4. In the *Settings* dialog box, specify how absent values are denoted and whether or not the export should be limited to the active characters only. Press <*Finish*>.

This action will create the export.csv in the database directory and will open the file in your computer's default editor for *.csv files.



MS Excel, which is the default CSV editor on most PCs, displays the allele hashes in scientific notation (e.g. 5,83E+16). For categorical data, this does in fact not make sense.



By specifying "No" for the preference *Export table files in CSV format*, allele hashes files will be exported in tab-delimited text format (see the Reference manual, Chapter The BIONUMERICS user interface).

Import should be done in another BIONUMERICS database in which the *WGS tools plugin* and the *WGS tools local plugin* are both installed. The installation procedure for the *WGS tools plugin* is covered in the corresponding plugin manual and the *WGS tools local plugin* is described in 4, so the procedures will not be repeated here.

Proceed as follows to import entries with their allele hashes from an exported text file:

- 5. Select *File* > *Import...* (\Box , Ctrl+I).
- Browse for the text file (*.csv or *.txt) exported earlier, highlight the option *Import fields and hashed wgMLST calls (text file)* and press < *Finish*>.
- 7. Press <*Next*>.

Since no import template is available yet, we will need to create one first. We assume in this tutorial that the other database also uses SRA accessions as database key:

Double-click on 'Key' in the *Import rules* dialog box, select *Key* as destination in the *Edit data destination* dialog box (see Figure 16) and press < OK >.

Edit data destination	?	×
 <none></none> Entry info field Entry attachment Hashed wgMLST call Character set info field 		
ок	Car	icel

Figure 16: The Edit data destination dialog box for the 'Key' field.

- 9. Highlight all wgMLST loci in the *Import rules* dialog box and press the <*Edit destination...*> button.
- In the *Edit data destination* dialog box, select *wgMLST_Local* under *Hashed wgMLST call* (see Figure 17).
- 11. Press < *Next* > in the *Import rules* dialog box.
- 12. Leave *Key* checked and press < *Finish*> in the *Import links* dialog box.
- 13. Enter e.g. "Hashes" as *Name* for the import template and press < *OK* >.

Now that an import template is created, we can use it in this and future imports of hashed wgMLST allele calls.



Figure 17: The *Edit data destination* dialog box for all wgMLST loci in the export file.

- 14. Press < *Next*> in the *Import template* dialog box to apply the **Hashes** import template.
- 15. Press < *Finish*> to import the hash-based allele calls.