

BIONUMERICS Tutorial: Phylogenomics with RAxML and FastTree

1 Introduction

RAxML [2] and FastTree [1] are two open source maximum likelihood clustering methods which can be launched on aligned sequence data in BIONUMERICS. Aligned sequence data includes:

- Nucleic acid or amino acid sequences in the *Comparison* window on which a multiple alignment was calculated (see the "Cluster analysis of sequences" tutorial).
- A SNP matrix generated by a wgSNP analysis (see the "wgSNP with mapping performed locally on your own computer" and "wgSNP with mapping performed on the cloud calculation engine" tutorials), stored as an aspect of a reference mapped sequence type.
- A SNP matrix generated by the CFSAN SNP pipeline (see the "CFSAN SNP pipeline" tutorial), stored as an aspect of a sequence read set.

Both algorithms are available through the *WGS tools plugin* and can be launched on your own computer as well as on the Cloud Calculation Engine.

This tutorial will demonstrate a RAxML and FastTree analysis on a multiple alignment of concatenated MLST loci.

2 Preparing the database

The RAxML and FastTree analyses can only be performed in BIONUMERICS after installation of the *WGS tools plugin* in the BIONUMERICS database (*File* > *Install* / *remove plugins...* (\square)).

In this tutorial the **WGS demo database for the** *Burkholderia cepacia* complex will be used in which the *WGS tools plugin* is already installed. The Calculation engine option requires credits for running jobs on the Applied Maths cloud calculation engine. Credits are linked to credentials that you need to enter when installing the *WGS tools plugin*. No credits are assigned to the demo project so no RAxML or FastTree jobs can be launched on the external calculation engine. Please contact Applied Maths to obtain more information.

The demo database can be downloaded directly from the *BIONUMERICS Startup* window (see 2.1), or restored from the back-up file available on our website (see 2.2). The database contains the NCBI reference genome sequences for the *Burkholderia cepacia* complex and the genome sequences of the *Burkholderia cepacia* complex type strains if available at the time of database creation.

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).



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

- 2. Select WGS_demo_database_for_Bcc from the list and select Database > Download (2).
- 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_Bcc appears in the BIONUMERICS Startup window.

5. Double-click the **WGS_demo_database_for_Bcc** 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 demo database for the *Burkholderia cepacia* complex is also available on our website. This backup can be restored to a functional database in BIONUMERICS.

6. Download the file WGS_Bcc.bnbk file from https://www.applied-maths.com/download/ sample-data, under 'WGS_demo_database_for_Bcc'.



In contrast to other browsers, some versions of Internet Explorer rename the WGS_Bcc.bnbk database backup file into WGS_Bcc.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_Bcc_demobase".
- 10. Click < OK > to start restoring the database from the backup file.
- 11. Once the process is complete, click < Yes> to open the database.

The *Main* window is displayed (see Figure 2).

WGS_demo_database_for_Bcc - BioNumerics											-	
File Edit Database Analysis Scripts WGS tool	ls Window He	elp										
☞ 글 ▦ ☞ ᢀ ▙, ▷ 《	s, 8⊗ ∞,											
Experiment types		Datal	base entries								Comparisons	
g + P ⊗ € €	⊴ ↑	ţ.	+ 🖻 ⊗ 🛢	L 6	<all entries=""></all>	ڻ					+ 🖒 ⊗ 🗞 €	<u> </u>
# Name Type			Key		Assembly_accession	Species	Strain 🗨	1 2 3	4 5 6	5 7	Name M	odified da 🔷
1 denovo Sequence to			WGS DEMO DATABASE	FOR BCC0000001	GCE 000012945 1	Burkholderia lata	383			• ^		_
⊇ ≩ 2 was Sequence re	ead set types		WGS DEMO DATABASE	FOR BCC0000002	GCF 000203915.1	Burkholderia ambifaria	AMMD		• •			
□ ≩ 3 wgsLong Sequence re	ead set types		WGS DEMO DATABASE	FOR BCC0000003	GCF_000959505.1	Burkholderia dolosa	AU0158					
4 wgMLST Character ty	rpes		WGS_DEMO_DATABASE_	FOR_BCC0000004	GCF_000959525.1	Burkholderia multivorans	ATCC BAA-247					
5 quality Character ty	rpes		WGS_DEMO_DATABASE	FOR_BCC0000005	GCF_001029145.1	Burkholderia contaminans	MS14		• •			
□ ≩ 6 wgs_TrimmedStats Sequence re	ead set types		WGS_DEMO_DATABASE_	FOR_BCC0000006	GCF_001532085.1	Burkholderia pseudomultivorans	MSMB2199WGS					
7 wgMLST_CalTypes Character ty	pes		WGS_DEMO_DATABASE_	FOR_BCC0000007	GCF_001718315.1	Burkholderia diffusa	RF2-non-BP9		• •			
	~		WGS_DEMO_DATABASE_	FOR_BCC000008	GCF_001718335.1	Burkholderia territorii	RF8-non-BP5					
<	,		WGS_DEMO_DATABASE_	FOR_BCC0000009	GCF_001718535.1	Burkholderia seminalis	FL-5-4-10-S1-D7				< c	,
Entry fields Database design			WGS_DEMO_DATABASE_	FOR_BCC0000010	GCF_001718555.1	Burkholderia metallica	FL-6-5-30-S1-D7				Identification projects Decision pr	tworks
			WGS_DEMO_DATABASE_	FOR_BCC0000011	GCF_001718775.1	Burkholderia vietnamiensis	AU1233					
+ / ⊗ €, El ∝ 1	t ↓ <ai< td=""><td></td><td>WGS_DEMO_DATABASE_</td><td>FOR_BCC0000012</td><td>GCF_001718795.1</td><td>Burkholderia latens</td><td>AU17928</td><td></td><td></td><td></td><td> 2 + 🗗 ⊗ 🔩</td><td></td></ai<>		WGS_DEMO_DATABASE_	FOR_BCC0000012	GCF_001718795.1	Burkholderia latens	AU17928				2 + 🗗 ⊗ 🔩	
Name Field type	-		WGS_DEMO_DATABASE_	FOR_BCC0000013	GCF_001718955.1	Burkholderia stagnalis	MSMB735WGS		• •	•	Name M	odified da 🔫
All Assembly accession Fixed	0		WGS_DEMO_DATABASE_	FOR_BCC0000014	GCF_001718975.1	Burkholderia ubonensis	MSMB1471WGS					_
Att Species Fixed			WGS_DEMO_DATABASE_	FOR_BCC0000015	GCF_001883705.1	Burkholderia catarinensis	89					
All Strain Fixed			WGS_DEMO_DATABASE_	FOR_BCC0000016	GCF_001992495.1	Burkholderia cenocepacia	VC13600		• •			
All Type strain			WGS_DEMO_DATABASE_	FOR_BCC0000017	GCF_002097715.1	[Pseudomonas] mesoacidophila	ATCC 31433					
Als Comment Fixed			WGS_DEMO_DATABASE_	FOR_BCC0000018	GCF_003330765.1	Burkholderia pyrrocinia	mHSR5	•	• •			
AK MI ST DubMI ST ST Flexible			WGS_DEMO_DATABASE_	FOR_BCC0000019	GCF_009586235.1	Burkholderia cepacia	BC16		• •			
			WGS_DEMO_DATABASE_	FOR_BCC0000020	GCF_902498995.1	Burkholderia anthina	LMG 20980					
			WGS_DEMO_DATABASE_	FOR_BCC0000021	GCF_902499105.1	Burkholderia paludis	LMG 30113	•	• •	•		
	Ŷ		WGS_DEMO_DATABASE_	FOR_BCC0000022	GCF_902499125.1	Burkholderia arboris	LMG 24066		• •		č	,
Eingerprint files Power assemblies Appotations			WGS_DEMO_DATABASE_	FOR_BCC0000023	GCF_902833185.1	Burkholderia stabilis	BCC1308	•	• •		Alignments BLAST projects 1 Ch	rom Comp
			WGS_DEMO_DATABASE_	FOR_BCC0000024	GCF_902859845.1	Burkholderia puraquae	LMG 29660	•	• •			
	<all finge<="" td=""><td></td><td>WGS_DEMO_DATABASE</td><td>FOR_BCC0000025</td><td>GCF_000732615.1</td><td>Burkholderia paludis</td><td>MSh1</td><td>•</td><td>• •</td><td></td><td>+ 🗁 🛛 🗟 🖣</td><td></td></all>		WGS_DEMO_DATABASE	FOR_BCC0000025	GCF_000732615.1	Burkholderia paludis	MSh1	•	• •		+ 🗁 🛛 🗟 🖣	
File name Experiment type	Link 🔫		WGS_DEMO_DATABASE_	FOR_BCC0000026	GCF_000959445.1	Burkholderia vietnamiensis	LMG 10929		• •	•	Name M	odified da
			WGS_DEMO_DATABASE_	FOR_BCC0000027	GCF_001028665.1	Burkholderia pyrrocinia	DSM 10685	•	• •	•		
			WGS_DEMO_DATABASE	FOR_BCC0000028	GCF_001411495.1	Burkholderia cepacia	UCB 717		• •			
			WGS_DEMO_DATABASE_	FOR_BCC0000029	GCF_001742165.1	Burkholderia stabilis	ATCC BAA-67	•	• •			
			WGS_DEMO_DATABASE_	FOR_BCC0000030	GCF_001758385.2	Burkholderia contaminans	LMG 23361	•	• •			
			WGS_DEMO_DATABASE	FOR_BCC0000031	GCF_008802145.1	Burkholderia diffusa	CCUG 54558		• •			
			WGS_DEMO_DATABASE_	FOR_BCC0000032	GCF_902499045.1	Burkholderia latens	LMG 24064	•	• •			
			WGS_DEMO_DATABASE_	FOR_BCC0000033	GCF_902499075.1	Burkholderia pseudomultivorans	LMG 26883	•	• •			
			WGS_DEMO_DATABASE_	FOR_BCC0000034	GCF_902499135.1	Burkholderia dolosa	LMG 18943			• •		
<	>		<				>	<		>	<	>
Database: WGS_demo_database_for_Bcc (_DefaultUs	ser_) Entries: L	oaded=3	38, View=38, Selected=0	7 experiments	C:\Users\10029961\Desk	top\Demos\Databases\WGS_demo_	database_for_Bcc					

Figure 2: The *Burkholderia cepacia* complex demonstration database: the *Main* window.

3 Generate aligned sequence data

3.1 Sequence extraction

An assembly-based wgMLST analysis has already been performed on the 38 genome sequences in the database (for more information on wgMLST analysis on imported genomes in BIONUMER-ICS see the "wgMLST typing: routine workflow starting from imported genomes" tutorial). To

demonstrate the RAxML and FastTree maximum likelihood analyses we will extract the allele sequences of the 7 MLST loci from the 38 genome sequences and concatenate the obtained sequences.

- 1. Select the *Database entries* panel to make it the active panel and select all entries in the database with *Edit* > *Select all* (Ctrl+A).
- 2. Extract the allele sequences of the 7 MLST loci from the 38 genome sequences by selecting *WGS tools* > *Store wgMLST locus sequences...*.

The Store sequences dialog box will pop up.

3. Select the **MLST PubMLST (7 loci)** subscheme and select **LocusTag** as sequence experiment type identifier (see Figure 3).

Store sequenc	es	?	×						
Subschores									
Subschema									
For each locus in the subschema, a sequence experiment type will be created:									
wgMLST (4	5465 loci)								
MLST PubML	ST (7 loci)								
Sequence ex	periment type								
Select the ide	entifier vou want tr	use for th	e						
sequence ex	periment type nam	les:							
LocusTag			\sim						
LocusTag			\sim						

Figure 3: The Store sequences dialog box.

 Press < OK> and < Yes> to confirm the creation of the 7 new sequence experiment types (see Figure 4).



Figure 4: Confirmation dialog box.

Seven new sequence experiment types will be created and listed in the *Experiment types* panel. The 7 MLST loci sequences will be extracted from the 38 genome sequences and stored in the respective sequence experiment types. Each of these sequences can be viewed in the *Sequence editor* window by clicking on the respective green dot in the *Experiment presence* panel.

3.2 Concatenate sequences

As we want to include the sequence information from the seven loci in our phylogenetic analyses, we will first concatenate the 7 MLST loci sequences and store the concatenated sequences in a new sequence experiment type.

5. To create a new sequence experiment type, click on the *Experiment types* panel to activate it and select *Edit* > *Create new object...* (+).

- 6. From the *Create a new experiment type* dialog box that pops up, select **Sequence type** to start the *New sequence type* wizard.
- 7. The first page prompts you to enter a name for the new sequence experiment type. Enter a name (for example "Concatenated loci" and press <*Next*>.
- In the second page of the New sequence type wizard leave the check box for Nucleic acid sequences selected and press the < Finish > button to complete the setup of the new sequence experiment type.

The sequence experiment type will be listed in the *Experiment types* panel of the *Main* window (see Figure 5).

Experim	ient ty	/pes					
କ୍ଷ		+ 🗗 🛛 🗟	Ên ⊽… ↑↓				
	#	Name	Туре 🔻				
🗆 🛟	1	denovo	Sequence types				
≣	2	2 wgs	Sequence read set types				
≣	3	8 wgsLong	Sequence read set types				
	4	wgMLST	Character types				
	5	quality	Character types				
⊒ ≵	e	8 wgs_TrimmedStats	Sequence read set types				
	7	wgMLST_CallTypes	Character types				
🗆 🛟	8	3 atpD	Sequence types				
	9	9 gltB	Sequence types				
	10) gyrB	Sequence types				
	11	recA	Sequence types				
	12	2 lepA	Sequence types				
	13	3 phaC	Sequence types				
□ 🛱	14	trpB	Sequence types				
C AC	15	5 Concatenated loci	Sequence types				
			¥				
	<		>				

Figure 5: The *Experiment types* panel.

To concatenate the MLST loci sequences and store the concatenated sequences in the newly created sequence experiment type an online script can be used.

- 9. Select *Scripts* > *Browse internet...* to open the *Browse for scripts* window.
- 10. Select Concatenate sequences under Sequence related tools.
- 11. In the dialog box that pops up, select the 7 sequence experiment types containing the 7 MLST loci sequences as sequence experiment types which should be merged and select the newly created sequence experiment type as the merged sequence type (see Figure 6). Press < OK >.

The sequences of the 7 loci will be concatenated and stored in the merged sequence experiment type for the 38 selected entries. The *Main* window will now look like Figure 7.

3.3 Calculate a multiple alignment

A multiple sequence alignment of the concatenated sequences can be calculated in the *Comparison* window.

- 12. In the Database entries panel of the Main window make sure all entries are selected.
- 13. Click on *Edit* > *Create new object...* (+) in the *Comparisons* panel to open the *Comparison* window for the 38 selected entries.
- 14. In the *Experiments* panel highlight the merged sequence experiment which contains the concatenated sequences of the entries in the comparison (i.e. "Concatenated loci") and select *Layout* > *Show image* (◎) to visualise the concatenated sequences in the *Experiment data* panel.

Merge sequences				?	×
Concatenated loci denovo	> < >> < < < V	atpD gltB gyrB lepA phaC recA trpB			
Select merged sequence type: Concatenated loci	~				
		C	ж	Can	cel

Figure 6: The Merge sequences dialog box.

🛃 wo	S_demo_database_for_Bcc - I	BioNumerics								- 0	ı ×
File Fr	lit Database Analysis Se	crinte WGS toole Winde	w Hel								
		t	1.0	٢							
₩ 7 :	≠ ⊞ /₽ ©	Nº D @ 90	₩⁄								
Export	rimont tunce		Dat	abase estrice						Compadaana	
Expor	ninencippes	0	Uat							Companisons	
6	∄ + 12 ⊗ €		1] + 岱 ⊗ ಟ, 印	All Entries>	10				+ /) ⊗ ಟ,∣ ಟ	∽
	# Name	Туре		Key	Assembly_accession			Туре 👻	8 9 10 11 12 13 14 <mark>15</mark>	Name Modified date	te 💌
	1 denovo	Sequence types	. 🗹	WGS_DEMO_DATABASE_FOR_BCC0000001	GCF_000012945.1	Burkholderia lata	383	Yes			
⊒≩	2 wgs	Sequence read set types	Z	WGS_DEMO_DATABASE_FOR_BCC0000002	GCF_000203915.1	Burkholderia ambifaria	AMMD	Yes			
⊒≩	3 wgsLong	Sequence read set types		WGS_DEMO_DATABASE_FOR_BCC0000003	GCF_000959505.1	Burkholderia dolosa	AU0158	No			
	4 wgMLST	Character types	~	WGS_DEMO_DATABASE_FOR_BCC0000004	GCF_000959525.1	Burkholderia multivorans	ATCC BAA-247	Yes			
	5 quality	Character types	_	WGS_DEMO_DATABASE_FOR_BCC0000005	GCF_001029145.1	Burkholderia contaminans	MS14	No			
⊒ ≩	6 wgs_TrimmedStats	Sequence read set types	Z	WGS_DEMO_DATABASE_FOR_BCC0000006	GCF_001532085.1	Burkholderia pseudomultivorans	MSMB2199WGS	No	• • • • • • • •		
	7 wgMLST_CalTypes	Character types	~	WGS_DEMO_DATABASE_FOR_BCC0000007	GCF_001718315.1	Burkholderia diffusa	RF2-non-BP9	No			
🗌 🗆 👬	8 atpD	Sequence types	~	WGS_DEMO_DATABASE_FOR_BCC0000008	GCF_001718335.1	Burkholderia territorii	RF8-non-BP5	No		1	
	9 gitB	Sequence types	~	WGS_DEMO_DATABASE_FOR_BCC0000009	GCF_001718535.1	Burkholderia seminalis	FL-5-4-10-S1-D7	No			
	10 gyrB	Sequence types	~	WGS_DEMO_DATABASE_FOR_BCC0000010	GCF_001718555.1	Burkholderia metallica	FL-6-5-30-S1-D7	No		Identification projects Decision networ	rks
	11 recA	Sequence types	~	WGS_DEMO_DATABASE_FOR_BCC0000011	GCF_001718775.1	Burkholderia vietnamiensis	AU1233	No			A
	12 lepA	Sequence types	~	WGS_DEMO_DATABASE_FOR_BCC0000012	GCF_001718795.1	Burkholderia latens	AU17928	No		?∃ + [] ⊗ E,	E
	13 phaC	Sequence types	~	WGS_DEMO_DATABASE_FOR_BCC0000013	GCF_001718955.1	Burkholderia stagnalis	MSMB735WGS	No		Name Modified date	• •
	14 trpB	Sequence types	~	WGS_DEMO_DATABASE_FOR_BCC0000014	GCF_001718975.1	Burkholderia ubonensis	MSMB1471WGS	No			
	15 Concatenated loci	Sequence types	~	WGS_DEMO_DATABASE_FOR_BCC0000015	GCF_001883705.1	Burkholderia catarinensis	89	No			
			′ <u> </u>	WGS_DEMO_DATABASE_FOR_BCC0000016	GCF_001992495.1	Burkholderia cenocepacia	VC13600	No			
	<	,	~	WGS_DEMO_DATABASE_FOR_BCC0000017	GCF_002097715.1	[Pseudomonas] mesoacidophila	ATCC 31433	No			
Entry f	Felds Database design			WGS_DEMO_DATABASE_FOR_BCC0000018	GCF_003330765.1	Burkholderia pyrrocinia	mHSR5	No			
			∠	WGS_DEMO_DATABASE_FOR_BCC0000019	GCF_009586235.1	Burkholderia cepacia	BC16	No			
+		∽ ↑ ↓	<u> </u>	WGS_DEMO_DATABASE_FOR_BCC0000020	GCF_902498995.1	Burkholderia anthina	LMG 20980	Yes			
	Name F	ield type	. 👱	WGS_DEMO_DATABASE_FOR_BCC0000021	GCF_902499105.1	Burkholderia paludis	LMG 30113	No		<	>
AK	Assembly accession E	ixed	. 🞽	WGS_DEMO_DATABASE_FOR_BCC0000022	GCF_902499125.1	Burkholderia arboris	LMG 24066	Yes			
ARC	Species F	ixed	Ľ.	WGS_DEMO_DATABASE_FOR_BCC0000023	GCF_902833185.1	Burkholderia stabilis	BCC1308	No		Alignments BLAST projects Chrom.	Comp.
AIC	Strain F	ixed	¥	WGS_DEMO_DATABASE_FOR_BCC0000024	GCF_902859845.1	Burkholderia puraquae	LMG 29660	Yes			~
ARC	Type strain F	ixed		WGS_DEMO_DATABASE_FOR_BCC0000025	GCF_000732615.1	Burkholderia paludis	MSh1	Yes			Na.
AIC	Comment F	ixed .		WGS_DEMO_DATABASE_FOR_BCC0000028	GCF_000959445.1	Burkholderia vietnamiensis	LMG 10929	Yes		Name Modified dat	ite 🔻
AK				WGS_DEMO_DATABASE_FOR_BCC0000027	GCF_001028665.1	Burkholdena pyrrocinia	DSM 10685	Yes			
		,		WGS_DEMO_DATABASE_FOR_BCC0000028	GCF_001411495.1	Burkholderia cepacia	008717	res			
Finger	print files Power assemblies	Annotations		WGS_DEMO_DATABASE_FOR_BCC0000029	GCF_001742165.1	Burkholderia stabilis	ATCC BAA-67	res			
L.E.		A 7		WGS_DENO_DATABASE_FOR_BCC0000030	OCE 009902145.4	Burkholderia diffusa	0010 54559	Vee			
B		E Y. <ai< td=""><td></td><td>WGS_DENO_DATABASE_FOR_BCC0000031</td><td>GCE_002400045.4</td><td>Burkholderia lateos</td><td>LMQ 24064</td><td>Ver</td><td></td><td></td><td></td></ai<>		WGS_DENO_DATABASE_FOR_BCC0000031	GCE_002400045.4	Burkholderia lateos	LMQ 24064	Ver			
	File name Exp	periment type 🛛 🛛 🗨		WGS DEMO DATABASE FOR BCC0000032	CCF 002409075 1	Burkholderis neeudomultivorane	LMG 26883	Vae			
		1		WGS DEMO DATABASE FOR BCC0000034	GCF_002409135.1	Burkholderis dolosa	LMG 18943	Vee			
	<			<	001_002400100.1	barmolacita colosa	200 10040	>	< >	<	>

Figure 7: The Main window.

- 15. Select *Sequence* > *Multiple alignment...* (i) to open the *Multiple alignment* dialog box (see Figure 8).
- 16. Leave the settings as default and press < OK >.

A multiple alignment is calculated. The Comparison window now looks like Figure 9.

D

For more information on multiple sequence alignment in BIONUMERICS see the "Cluster analysis of sequences" tutorial. Note that for MLST loci a good multiple alignment can be obtained on the concatenated sequences. However, when non-MLST loci are considered it is good practice to concatenate the multiple sequence alignments of the loci instead to reduce the amount of misaligned positions.

Multiple alignment	? ×
Alignment settings Open gap penalty: 0 × % Unit gap penalty: 0 × % Match score: 100%	Fast alignment Use fast algorithm Minimum match sequence: 2 Maximum number of gaps: 98
Save to database as new defaults	OK Cancel

Figure 8: The Multiple alignment dialog box.



Figure 9: The Comparison window.

4 Launch a maximum likelihood analysis

A maximum likelihood analysis with RAxML and FastTree can be launched on your own computer or on the calculation engine for the entries in the created comparison. Following the same principles as cluster analyses in comparisons, comparison jobs apply on the whole comparison and the data stored in the *active* experiment and (where applicable) the active aspect in the *Experiments* panel.

1. Save the comparison by selecting *File* > *Save* (ℍ, Ctrl+S). Specify a name for the comparison for example "ML analyses for Bcc strains" and press < *OK* >.

2. To launch a maximum likelihood comparison job, select *File* > *Launch comparison jobs...* (▷) from the menu in the *Comparison* window.

In case the comparison was not saved to the database yet, you will be prompted to save first. Comparisons should be saved before any jobs can be launched.

Subsequently, the Submit comparison jobs dialog box will open (see Figure 10).



If the active experiment does not support any comparison jobs, an error message is shown and the *Submit comparison jobs* dialog box will not appear.

Submit comparison jobs	? ×
Submit jobs to:	
Algorithms	RAxML Performing RAxML pipeline Settings
Jobs	
Submitting 2 jobs	
Open jobs overview window	
Credits available for job submission:	0
Credits needed to submit selected jobs:	0
Buy extra credits	
	OK Cancel

Figure 10: The Submit comparison jobs dialog box.

The RAxML and FastTree algorithms are available on the local calculation engine and on the Cloud Calculation Engine. In this tutorial the jobs will be launched on the local calculation engine (i.e. will be launched on your own computer).

- 3. In the *Submit jobs to* panel select *Own computer* and in the *Algorithms* panel select both the *RAxML* and *FastTree* algorithm (see Figure 10).
- 4. With the *RAxML* algorithm highlighted, press the <*Settings...*> to open the *RAxML pipeline* settings dialog box in which the evolutionary model and other settings for the RAxML job can be defined (see Figure 11).

In the *Evolutionary models* panel an evolutionary model for maximum likelihood analysis can be selected. The available models in the drop-down list next to the *Model* option consist of a combination of a basic evolutionary model (e.g. GTR, DAYHOFF,...) with a model for rate heterogeneity among sites (+GAMMA or +CAT) and an allowance for the presence of invariant sites (+I).

The GTR model is the most common substitution model for nucleotide sequence analysis and the only model available through RAxML. For amino acid sequence analysis twelve general amino acid substitution models can be selected, i.e. BLOSUM62, DAYHOFF, DCMUT, GTR, JTT, JTTDCMUT, LG, LG4M, LG4X, PMB, VT and WAG.

The GAMMA and CAT models account for variable rates of evolution across sites. The GAMMA model assumes a gamma distribution of rates across sites with four discrete rate categories. The CAT model optimizes the individual per-site substitution rates and classifies these individual rates

AxML pi	peline		?	×
Evolution	ary models			
Model:	GTR+GAMMA		\sim	
	Evolutionary rate	categories		
	among sites (CAT	model):		
Options				
Perf	orm bootstrap analy	sis		
🗹 Num	ber of replicates	100		
Boot	stop criteria	autoFC		\sim
Save a	Igorithm settings as	default		
		01	0.00	

Figure 11: The RAxML pipeline settings dialog box.

into the number of rate categories specified by the *Evolutionary rate categories among sites* option. The default number of rate categories is set to 25. Note that this option is specific for the CAT model and is therefore not available when the GAMMA model has been chosen from the drop-down list.

The *Options* panel allows you to set the criteria for bootstrap analysis. When the *Perform boot-strap analysis* option is checked, bootstrap analysis is enabled and two options for bootstrap analysis become available:

- *Number of replicates*: This option allows the user to set the number of bootstrap replicates. Default the number of bootstrap replicates is set to 100.
- **Bootstop criteria**: When this option is checked the optimal number of bootstrap replicates to obtain stable support values will be determined automatically. In the drop-down list a RAxML bootstop criterion (i.e. threshold to decide when enough replicates have been computed) can be selected to determine convergence of the bootstrap estimates.

The bootstrap support values are drawn on the best-scoring maximum likelihood tree.

When altering the RAxML settings, one can save the updated settings as defaults to the database with *Save algorithm settings as default*.

- 5. Leave the settings as default and press < OK >.
- 6. In the Submit comparison jobs dialog box highlight the FastTree algorithm and press the <Settings...> to open the FastTree ML pipeline settings dialog box in which the evolutionary model and other settings for the FastTree job can be defined (see Figure 12).

In the *Evolutionary models* panel an evolutionary model for maximum likelihood analysis can be selected. Two models are available in the drop-down list for nucleotide sequence analysis (i.e. *Generalized time-reversible* and *Jukes-Cantor*), while three models are available for amino acid analysis (i.e. *Jones-Taylor-Thorton*, *Le-Gascuel (2008)* and *Whelan-And-Goldman (2001)*).

In the *Options* panel four additional options for FastTree analysis can be enabled:

• **Speed up**: This option can be checked to speed up the neighbor joining phase and to reduce memory usage.

FastTree M	?	×						
Evolutiona	ary models							
Model:	Generalized time-reversible \lor							
Options								
Speed	l up							
Use p	seudocounts							
Evolut	ionary rate categ	ories among sites	s (CAT mo	del)				
Amou	nt of categories	20						
Resca	le lengths to opti	mize the Gamma2	0 likelihoo	d				
Save al	gorithm settings a	as default						
		ОК	Can	cel				

Figure 12: The FastTree ML pipeline settings dialog box.

- **Use pseudocounts**: This option is recommended if the alignment has sequences with little or no overlap. A pseudocount (weight of 1.0) will be used to estimate the distances between these sequences.
- Evolutionary rate categories among sites (CAT model): If this option is checked the evolutionary model will be run under the CAT model for rate heterogeneity among sites. The CAT model optimizes the individual per-site substitution rates and classifies these individual rates into the number of rate categories specified by **Amount of categories**. The default amount of categories is set to 20.
- *Rescale lengths to optimize the Gamma20 likelihood*: After the final round of optimizing branch lengths with the CAT model, report the likelihood under the discrete gamma model with the same number of categories. FastTree uses the same branch lengths but optimizes the gamma shape parameter and the scale of the lengths. The final tree will have rescaled lengths.

Local support values computed with the Shimodaira-Hasegawa test are drawn on the resulting maximum likelihood tree and provide an estimate of the reliability of each split in the tree. The FastTree support values range from 0 to 1 but are multiplied by 100 in BIONUMERICS.

When altering the FastTree settings, one can save the updated settings as defaults to the database with *Save algorithm settings as default*.

- 7. Leave the settings as default and press < OK >.
- 8. Press < *OK* > in the *Submit comparison jobs* dialog box to launch the RAxML and FastTree jobs on your computer.

The jobs are submitted to your computer and the *Job overview* window opens. In the *Job overview* window, the job type, job name, time of submission, job status, a description of the job, its progress and much more can be monitored.

5 Import and analyse the maximum likelihood job results

Once the jobs have been finished (see Figure 13), the results can be imported in the database by selecting Jobs > Get results () from the *Job overview* window.

1. When the jobs are finished, select the jobs and select Jobs > Get results ($\$) to import the results in the comparison window.

_													
	ć 0	verview of Com	parison: ML analysis for B	cc strains								-	×
F	ile	Jobs View W	indow Help										
	×	- 1 🔍 d	ິ ບ 🗼 Alljo	bs	8								
1	_												_
	Ove	rview of submit	ted jobs								1		
		Туре	Name	Submitted time (UTC)	Status	Message	Progress	Job type	Description	User	Jobid		
	1	Comparison, local	ML analysis for Bcc strai	ns 2020-09-30 10:27:54	Finished	Done	100%	RAxML pipeline	Performing RAxML pipeline	_DefaultUser_	827cc09a-62be-4b69-8b12-8c6		
	2 1	comparison, local	ML analysis for BCC strail	ns 2020-09-30 10:27:54	rinished	Done	100%	rastiree ML pipe	Performing Pasti ree ML pipeline	_DefaultUser_	40013507-0060-4014-8071-609		
A	l jobs	: 2 All queued	jobs: 0 All running jobs:	0 All finished jobs: 2 A	II failed jobs: 0								

Figure 13: The *Job overview* window listing a finished RAxML and FastTree pipeline job.

2. Close the *Job overview* window.

The Comparison window now looks like Figure 14.



Figure 14: The *Comparison* window after import of the RAxML and FastTree pipeline job results.

The *Analyses* panel lists the performed maximum likelihood analyses. The name of the analysis includes the analysis type (RAxML or FastTree), the aspect and the experiment type on which the analysis was performed and the selected maximum likelihood settings. If the dendrogram is not automatically displayed in the *Dendrogram* panel, double-click on the analysis or use *File* > *Analysis components* > *Open* (P) to display.

The toolbar of the *Dendrogram* panel (see Figure 15) provides different options to manipulate the visualised dendrogram (i.e. changing the display settings, rearranging the branches, collapsing/-expanding the branches and rerooting the tree).



Figure 15: The toolbar of the *Dendrogram* panel.

The log file of the maximum likelihood pipeline jobs can be consulted in the Job log panel (see Figure 16).

Groups Job log								
Job description	Job ID	Start	Stop	Status	User	Message	Job type	-
Performing RAxML pipeline	827cc09a-62be-4b69-8b12-8c600ceeff05	2020-09-30 10:	2020-09-30 10:	Finished	_DefaultUser_	Job completed successfully and resul	RAxML	^
Performing FastTree ML pipeline	46c135c7-66eb-4cf4-a071-e69420a08276	2020-09-30 10:	2020-09-30 10:	Finished	_DefaultUser_	Job completed successfully and resul	FastTree	
								~
	<							>

Figure 16: The Job log panel.

Bibliography

- [1] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree 2–approximately maximumlikelihood trees for large alignments. *PloS one*, 5(3), 2010.
- [2] Alexandros Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.