Core Genome Multi-locus Sequence Typing Analyses of Leptospira spp. Using the Bacterial Isolate Genome Sequence Database

Linda Grillová, Mathieu Picardeau

Core Genome Multi-locus Sequence Typing Analyses of *Leptospira* spp. Using the Bacterial Isolate Genome Sequence Database

Linda Grillová and Mathieu Picardeau*

Biology of Spirochetes Unit, Institute Pasteur, Paris, France

* Correspondence to be sent to mathieu.picardeau@pasteur.fr

Running Head: cgMLST scheme for *Leptospira*

Abstract

With the advent of whole-genome sequencing (WGS), comparative analysis has led to the use of core genome MLST (cgMLST) schemes for the high-resolution reproducible typing of bacterial isolates. In cgMLST, hundreds of loci are used for gene-by-gene comparisons of assembled genomes for studying the genetic diversity of clinically important pathogens. Combination of the cgMLST data and metadata of the isolates is useful for epidemiological investigations.

Here we present a cgMLST scheme for the high-resolution typing of isolates from the whole *Leptospira* genus, enabling identification at the level of species, clades, clonal groups and sequence types. We show several examples how the cgMLST *Leptospira* database, which is a publicly available web-based database, can be used for the analyses of WGS data of *Leptospira* isolates. This effort was undertaken in order to facilitate international collaborations and support the global surveillance of leptospirosis.

Key Words: Genome, Core Genome Multi-locus Sequence Typing, Bacterial Isolate Genome Sequence Database (BIGSdb)

1. Introduction

Molecular typing of bacterial isolates is a powerful tool for surveillance and epidemiology of diseases. Discrimination of genetic variants and characterization of the predominant *Leptospira* strains in the environment, patients, or animal populations is essential for identifying sources of infection and developing evidence-based infection control and prevention strategies. Different molecular typing schemes are currently available for pathogenic *Leptospira* [1-4] but a harmonized typing tool needs to be established not only for pathogenic species, but for the whole genus. A core genome multi-locus sequence typing (cgMLST; based on 545 core genes) was recently designed based on high-quality genome sequences representing all known *Leptospira* species [5, 6]. This scheme can significantly increase our understanding of the epidemiology and general biology of *Leptospira* spp. First, cgMLST can be applied to pathogenic (sub-clade P1), intermediate (sub-clade P2), and saprophytic isolates (sub-clades S1 and S2) and, thus, has a potential to elucidate the role of intermediates, which is a group of strains phylogenetically related to pathogens of unclear pathogenicity from the clinical perspective. cgMLST has a high discrimination power resulting in the identification of species, clades, clonal groups (CGs), sequencing types (STs) and most probable serogroups. This would allow tracking of *Leptospira* strains and could help, for example, the detection of new genotypes and length of time for which a given genotype persists. The widespread use of this cgMLST scheme should enable the identification of such relationships at the global level and over time. The *Leptospira* cgMLST database (https://bigsdb.pasteur.fr/leptospira/) is a publicly available web-based database hosted at the Institute Pasteur. At present, the database contains data from 1007 *Leptospira* strains (08/28/2019).

The Bacterial Isolate Genome Sequence Database platform (BIGSdb) was initially developed by Jolley and Maiden in 2010 [7] for automatic ST and CG assignments; for determination of new alleles; for storage of sample metadata; for identification of new associations between genotypes and metadata using various tools; and for user-friendly visualization of molecular typing data using breakdown options and external plugins such as interactive tree of life (iTOL) [8] and GrapeTree [9]. This chapter presents some examples of how the WGS data can be used in BIGSdb for user-friendly visualization of phylogenic relationships among STs of *Leptospira*.

2. Materials

2.1. Genome requirements

For submission, the treated WGS data are needed (i.e., low Phred score base, trimming, exogenous oligonucleotide clipping, sequencing error correction and read coverage homogenization) (see Note 1) . Draft genomes with 50× minimum coverage and a minimum N50 of 10,000 nt are required (see Notes 2-4).

2.2. Information on isolates

Relevant information of isolates such as isolate identification name/number; country of origin; biological source of sample; year of isolation; serogroup and serovar; etc. are required.

2.3. Hardware/software requirements

The BIGSdb is an online database. As such, users do not need to install a particular software.

3. Methods

3.1 Data submission

New users can contact the curators by e-mail (leptospiraMLST@pasteur.fr). Subsequently, the curators will create an account and provide the login details to new users, who are then able to submit their data (see Note 5). Whole genome sequence data in FASTA format can be submitted to BIGSdb as i.) a single contig of a closed chromosome; ii.) a multi-fasta file with closed chromosome and plasmids from the same isolate; iii.) multi-contig files (whole genome shotgun) or iv.) scaffold files. There are number of fields that must be filled in so the curators know how the data were obtained; e.g., the sequencing platform used, read length, coverage, and assembly *(de novo* or mapped). Make sure the 'e-mail submission updates' box is checked if you wish to receive e-mail notification of the result of your submission. Subsequently, curators will check the quality of the data and the sequences will be automatically scanned for cgSTs and cgCGs assignments. cgSTs represent profiles that differ by no allele other than for missing data. cgSTs that share all but one or few alleles are considered to be strongly related even if the differing alleles contain multiple single nucleotide variants (SNVs) due to recombination. cgCGs are defined by a single-linkage clustering threshold of 40 allelic mismatches; i.e., CG is defined as a group of cgMLST allelic profiles differing by no more than 40 allelic mismatches, out of 545 gene loci, from at least one other member of the group.

Every sequence entry should be accompanied by metadata of the sample. The researchers are encouraged to upload as much information about patients and isolates as available. The template for *Leptospira* isolate metadata can be downloaded at the Institute Pasteur MLST webpage (https://bigsdb.pasteur.fr/leptospira/). Some fields are mandatory and cannot be left blank. Check

the “Description of database fields” link on the database contents page to see a description of the fields and allowed values where these have been defined.

3.2. Data export

Different data can be exported from BIGSdb. You can export the isolate recordsets by clicking the “Export dataset” link in the Export section of the main contents page or you can export recordsets of isolates returned from a database query by clicking the “Dataset” button in the Export list at the bottom of the results table. You can then download the data in tab-delimited text or Excel formats. In the advanced options, chose the cgMLST scheme in order to export cgSTs and “Test-40” under the “Classification scheme” to obtain the cgCGs.

Similarly, the original submitted data as well as the sequences of core genes extracted from the original data can be exported. By default, the data will be extracted unaligned, but you can also chose to align the sequences by checking the “Align sequences” checkbox (use the MAFFT as the aligner). The export of aligned data is more time consuming than the export of unaligned data, and is restricted for exporting 200 isolates only.

3.3. Data analyses

3.3.1 General overview

An easy way to get a general overview of all or selected data present in BIGSdb, is to use the breakdown option plugins. For example, when you want to know the prevalence of different *Leptospira* species in different geographical areas, you can click the “Two field breakdown” link on the main contents page. This plugin exports a table breaking down one field against another (the breakdown of “species” by “country”) (Fig. 1) (see Note 6).

3.3.2. Interactive Tree of Life (iTOL)

The iTOL [8] plugin incorporated to BIGSdb enables generation and visualization of phylogenetic trees calculated from concatenated sequence alignments of core genes (n=545) using the Neighbour-joining clustering method. It can be assessed from the contents page or following the query by clicking the "iTOL" link. Since this analysis requires the pre-assembly of the data, it is possible to export only 200 isolates or less. The simple neighbor-joining method produces unrooted trees, but it does not assume a constant rate of evolution across lineages [10]. In contrast, the Maximum-likelihood method uses a more complex evolution model, and is known to be stronger than the neighbor-joining method for reconstructing sequence histories [11]. In general, iTOL plugins which generate the phylogenetic tree based on neighbor-joining method are good enough for an initial overview of the phylogeny, however, for publication purposes, we recommend employing the more precise and time-consuming Maximum-likelihood method, which is not a part of the BIGSdb function. Additional fields can be selected to be included as metadata for use in coloring nodes - select any fields you wish to include in the "iTOL datasets" list. For detailed explanation of the iTOL function, see the following link: https://itol.embl.de/help.cgi.

3.3.3 Species identification

If you are not sure which species of *Leptospira* you are working with, you can check using iTOL plugin by generating the phylogenetic tree based on the concatenated core gene sequences of your unknown sample(s) together with the reference sequences of *Leptospira* species (n=64) [6] which are present in BIGSdb (Table 1 and Fig. 2). Selecting the BIGSdb IDs of reference strains and your unknown sample(s) in the iTOL plugin will generate a phylogenetic tree which will cluster your isolates together with one of the reference strains. If the clustering is

not clear, the average nucleotide identity (ANI) of draft genomes should be performed, for example, using the ANI calculator [14] at the following link: https://www.ezbiocloud.net/tools/ani.

3.3.4 Prediction of possible serogroup

Strains belonging to the same serogroups are usually sub-divided into several cgCGs, however, when strains are part of the same clonal group, they should belong to the same serogoup (based on the all available isolates at the time of writing, n=1007); i.e., the branching based on the concatenated cgMLST sequences could be useful for determination of the potential serogroup.

3.3.5 GrapeTree

GrapeTree allows for exploration of the fine-grained population structure and phenotypic properties of large number of genomes (more than 200) in a web browser. It generates and displays the minimum spanning tree (MSTree) based on cgSTs [9]. The GrapeTree algorithm is able to export large datasets and is compatible to handle larger amount of missing sequences thus is perfect for handling cgMLST data. The datasets can include metadata, which allows nodes in the result tree to be colored interactively. It can be accessed from the contents page or following the query by clicking the “GrapeTree” link. In the *Leptospira* setting, it could be very useful in an easy identification of the potential source of infection by determination of cgCGs which are shared among human and animal isolates (Fig. 3). In Fig. 3, it is evident that several cgCGs are unique to particular hosts (e.g. cgCG176 was found only in dogs and cgCG81 was found only in patients), while other cgCGs were shared among multiple hosts (e.g. cgCG6 was shared among humans, dogs, and rats and cgCG5 was found in humans, cows, hedgehogs, dogs, and other

mammals). Another example of how GrapeTree can be used in *Leptospira* molecular epidemiology is the tracking of the different distribution of CGs over a specific time period,.

4. Notes

1. Platforms usually provide some quality control measures and follow protocols for filtering sequencing artifacts.
2. *Leptospira* strains have genomes that are 3.8–4.6 Mb in size with 35–45% GC content.
3. It is important to perform whole-genome sequencing from a clonal culture to avoid comparing mixed genomes. Isolation from an individual colony on an agar plate is, therefore, recommended to recover clonal *Leptospira* cultures [15].
4. A simulation of the effect of missing data (uncalled cgMLST alleles) on the clustering results showed that cluster assignment is robust even with high amounts of missing data (affecting up to 400 loci out of 545) [5]; indicating that even incomplete genomes should be typeable by cgMLST.
5. One of the features of BIGSdb is that the stored data sets have been manually curated to provide researchers with more accurate results.
6. A BIGSdb manual for a detailed description of BIGSdb function is available (https://bigsdb.readthedocs.io/en/latest/).

5. Acknowledgements

This work was supported by a PTR grant (PTR30-17) from the Institut Pasteur. We would like to thank to Julien Guglielmini (Bioinformatics and Biostatistics Hub, Institut Pasteur, Paris, France) for incorporation of the BIGSdb plugins.

6. References

1. Herrmann JL, Bellenger E, Perolat P, Baranton G, Saint Girons I (1992) Pulsed-field gel electrophoresis of NotI digests of leptospiral DNA: a new rapid method of serovar identification. J Clin Microbiol 30(7):1696–1702
2. Majed Z, Bellenger E, Postic D, Pourcel C, Baranton G, Picardeau M (2005) Identification of variable-number tandem-repeat loci in *Leptospira interrogans* sensu stricto. J Clin Microbiol 43(2):539–545
3. Slack AT, Dohnt MF, Symonds ML, Smythe LD (2005) Development of a Multiple-Locus Variable Number of Tandem Repeat Analysis (MLVA) for *Leptospira interrogans* and its application to Leptospira interrogans serovar Australis isolates from Far North Queensland, Australia. Ann Clin Microbiol Antimicrob 4:10
4. Ahmed N, Devi SM, Valverde M de los A, Vijayachari P, Machang'u RS, Ellis WA et al (2006) Multilocus sequence typing method for identification and genotypic classification of pathogenic *Leptospira* species. Ann Clin Microbiol Antimicrob 5:28
5. Guglielmini J, Bourhy P, Schiettekatte O, Zinini F, Brisse S, Picardeau M (2019) Genus-wide *Leptospira* core genome multilocus sequence typing for strain taxonomy and global surveillance. PLoS Negl Trop Dis 13(4):e0007374
6. Vincent AT, Schiettekatte O, Goarant C, Neela VK, Bernet E, Thibeaux R et al. (2019) Revisiting the taxonomy and evolution of pathogenicity of the genus *Leptospira* through the prism of genomics. PLoS Negl Trop Dis 13(5):e0007270
7. Jolley KA, Maiden MC (2010) BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 11(1):595
8. Letunic I, Bork P (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 44(W1):W242-245
9. Zhou Z, Alikhan NF, Sergeant MJ, Luhmann N, Vaz C, Francisco AP et al (2018) GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. Genome Res 28(9):1395–1404
10. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4(4):406–25
11. Fukami K, Tateno Y (1989) On the maximum likelihood method for estimating molecular trees: Uniqueness of the likelihood point. J Mol Evol 28(5):460–4
12. Nally JE, Bayles DO, Hurley D, Fanning S, McMahon BJ, Arent Z (2016) Complete Genome Sequence of *Leptospira alstonii* Serovar Room22 Strain GWTS #1. Genome Announc 4(6)
13. Casanovas-Massana A, Hamond C, Santos LA, de Oliveira D, Hacker KP, Balassiano I et al (2019) *Leptospira yasudae* sp. nov. and *Leptospira stimsonii* sp. nov., two new species of the pathogenic group isolated from environmental sources. Int J Syst Evol Microbiol doi: 10.1099/ijsem.0.003480
14. Yoon SH, Ha SM, Lim J, Kwon S, Chun J (2017) A large-scale evaluation of algorithms to calculate average nucleotide identity. Antonie Van Leeuwenhoek 110(10):1281–1286
15. Thibeaux R, Girault D, Bierque E, Soupé-Gilbert M-E, Rettinger A, Douyère A et al (2018) Biodiversity of Environmental *Leptospira*: Improving Identification and Revisiting the Diagnosis. Front Microbiol 9:816.

Figure Captions

Fig. 1 The prevalence of *Leptospira* species around the globe. The data were generated using the "Two field breakdown" option in the BIGSdb contents page (1st of July, 2019).

Fig. 2 Phylogeny of reference strains representing all known species (n=64) based on Neighbour-joining clustering method extracted from BIGSdb iTOL plugin.

Fig. 3 Identification of potential infection sources. A minimum spanning tree was created using GrapeTree for visualization of core genomic relationships of all available *L. interrogans* strains (n=299) isolated from different hosts around the globe. The numbers inside the tree nodes indicate the cgCGs and colors signify the origins of the samples.

Table Captions

Table 1 Reference strain of *Leptospira* species.

Fig.1

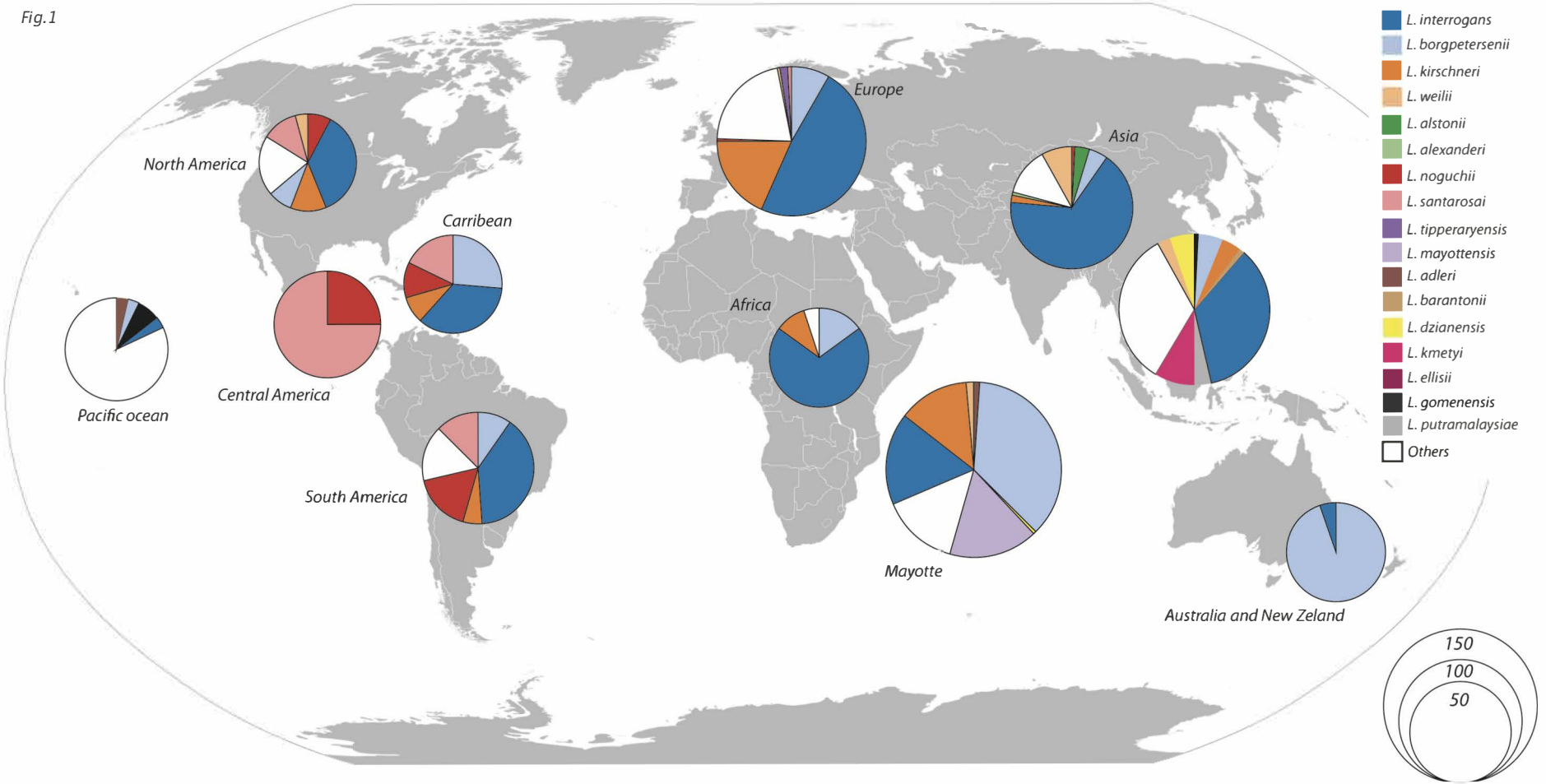
North America
Carribean
Central America
Pacific ocean
South America
Europe
Asia
Africa
Mayotte
Australia and New Zeland
L. interrogans
L. borgpetersenii
L. kirschneri
L. weilii
L. alstonii
L. alexanderi
L. noguchii
L. santarosai
L. tipperaryensis
L. mayottensis
L. adleri
L. barantonii
L. dzianensis
L. kmetyi
L. ellisii
L. gomenensis
L. putramalaysiae
Others
150
100
50

Fig.2

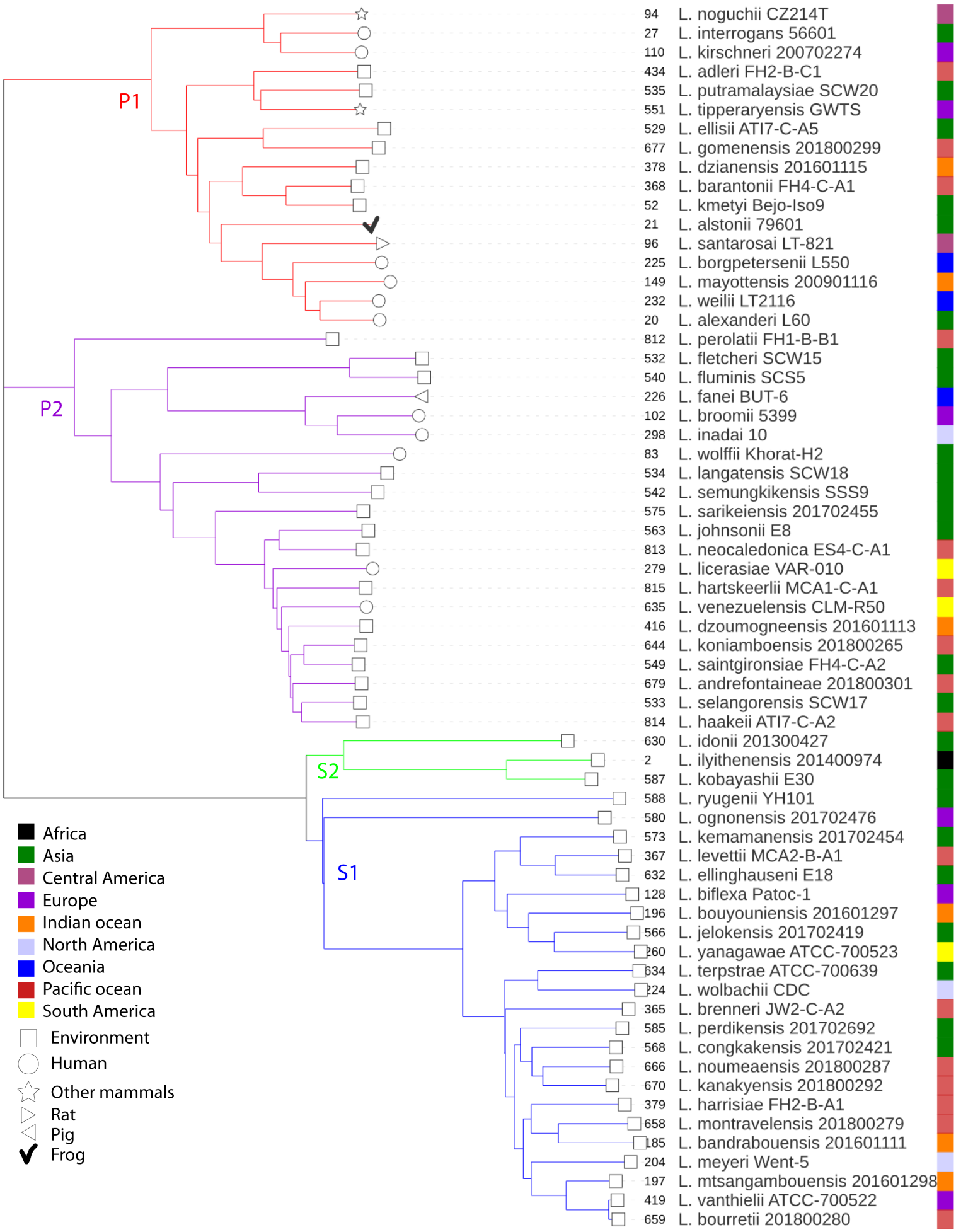
P1
P2
S1
S2
94 L. noguchii CZ214T
27 L. interrogans 56601
110 L. kirschneri 200702274
434 L. adleri FH2-B-C1
535 L. putramalaysiae SCW20
551 L. tipperaryensis GWTS
529 L. ellisii ATI7-C-A5
677 L. gomenensis 201800299
378 L. dzianensis 201601115
368 L. barantonii FH4-C-A1
52 L. kmetyi Bejo-Iso9
21 L. alstonii 79601
96 L. santarosai LT-821
225 L. borgpetersenii L550
149 L. mayottensis 200901116
232 L. weilii LT2116
20 L. alexanderi L60
812 L. perolatii FH1-B-B1
532 L. fletcheri SCW15
540 L. fluminis SCS5
226 L. fanei BUT-6
102 L. broomii 5399
298 L. inadai 10
83 L. wolffii Khorat-H2
534 L. langatensis SCW18
542 L. semungkikensis SSS9
575 L. sarikeiensis 201702455
563 L. johnsonii E8
813 L. neocaledonica ES4-C-A1
279 L. licerasiae VAR-010
815 L. hartskeerlii MCA1-C-A1
635 L. venezuelensis CLM-R50
416 L. dzoumogneensis 201601113
644 L. koniamboensis 201800265
549 L. saintgironsiae FH4-C-A2
679 L. andrefontaineae 201800301
533 L. selangorensis SCW17
814 L. haakeii ATI7-C-A2
630 L. idonii 201300427
2 L. ilyithenensis 201400974
587 L. kobayashii E30
588 L. ryugenii YH101
580 L. ognonensis 201702476
573 L. kemamanensis 201702454
367 L. levettii MCA2-B-A1
632 L. ellinghauseni E18
128 L. biflexa Patoc-1
196 L. bouyouniensis 201601297
566 L. jelokensis 201702419
260 L. yanagawae ATCC-700523
534 L. terpstrae ATCC-700639
224 L. wolbachii CDC
365 L. brenneri JW2-C-A2
585 L. perdikensis 201702692
568 L. congkakensis 201702421
666 L. noumeaensis 201800287
670 L. kanakyensis 201800292
379 L. harrisiae FH2-B-A1
658 L. montravelensis 201800279
185 L. bandrabouensis 201601111
204 L. meyeri Went-5
197 L. mtsangambouensis 201601298
419 L. vanthielii ATCC-700522
659 L. bourretii 201800280
Africa
Asia
Central America
Europe
Indian ocean
North America
Oceania
Pacific ocean
South America
Environment
Human
Other mammals
Rat
Pig
Frog

Fig.3

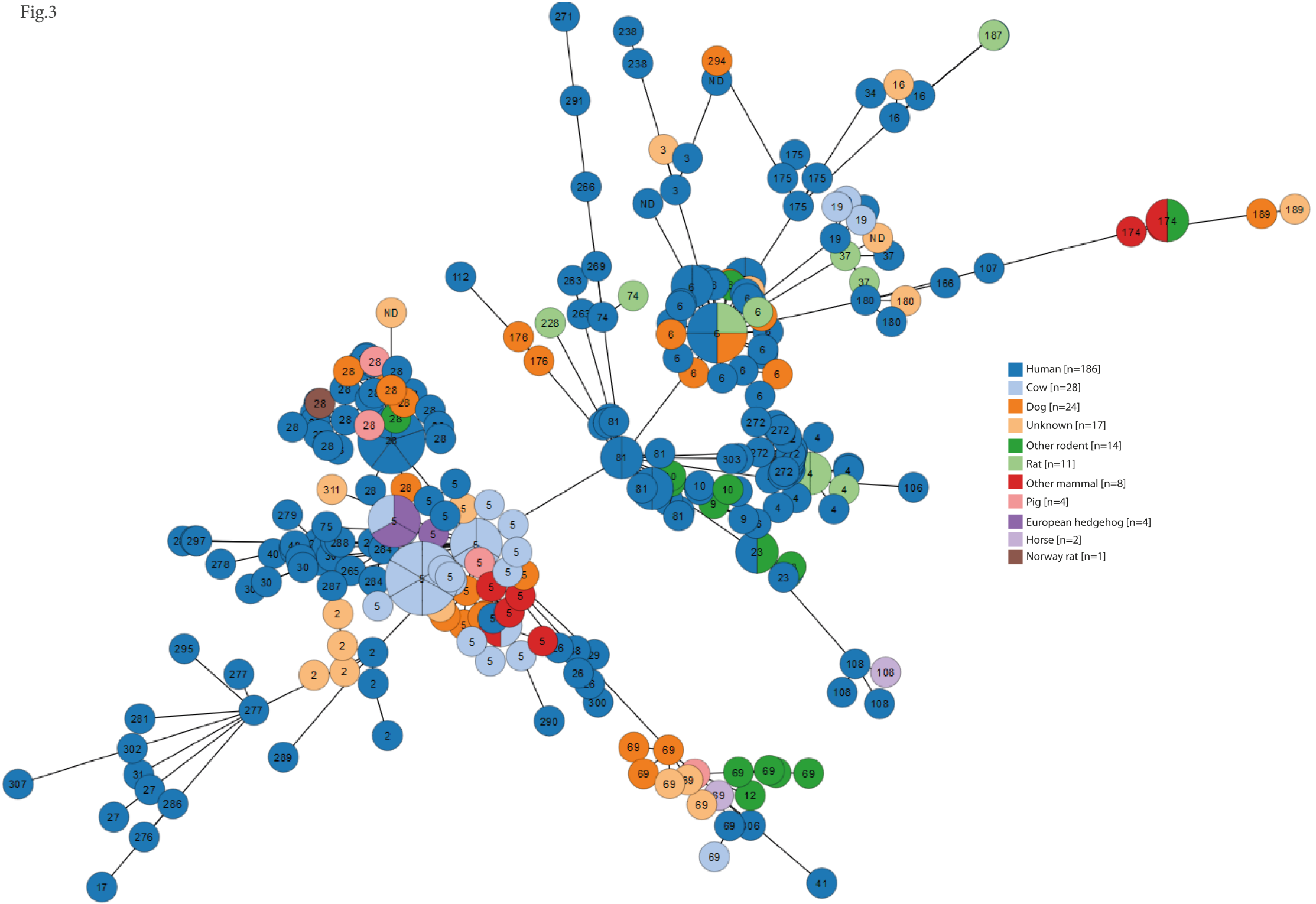
Human [n=186]
Cow [n=28]
Dog [n=24]
Unknown [n=17]
Other rodent [n=14]
Rat [n=11]
Other mammal [n=8]
Pig [n=4]
European hedgehog [n=4]
Horse [n=2]
Norway rat [n=1]

Table 1 Reference strain of *Leptospira* species.

Species	Isolate	BIGSdb id	Phylogenetic group	Country	Source
alexanderi	L 60	20	P1	China	Human
alstonii	79601	21	P1	China	Amphibian
interrogans	56601	27	P1	China	Human
kmetyi	Bejo-Iso9	52	P1	Malaysia	Environment
noguchii	CZ 214T	94	P1	Panama	Other mammal
santarosai	LT821	96	P1	Panama	Rat
kirschneri	200702274	110	P1	France	Human
mayottensis	200901116	149	P1	Mayotte	Human
borgpetersenii	L550	225	P1	Australia	Human
weilii	LT2116	232	P1	Australia	Human
barantonii	201602184	368	P1	New Caledonia	Environment
*dzianensis***	201601115	378	P1	Mayotte	Environment
adleri	201602187	434	P1	New Caledonia	Environment
ellisii	SSW8	529	P1	Malaysia	Environment
*putramalaysiae***	SCW20	535	P1	Malaysia	Environment
*tipperaryensis**	GWTS	551	P1	Ireland	Other mammal
gomenensis	201800299	677	P1	New Caledonia	Environment
wolffii	Khorat-H2	83	P2	Thailand	Human
broomii	5399	102	P2	Denmark	Human
fainei	BUT 6	226	P2	Australia	Pig
licerasiae	VAR010	279	P2	Peru	Human
inadai	10	298	P2	USA	Human
dzoumogneensis	201601113	416	P2	Mayotte	Environment
fletcheri	SCW15	532	P2	Malaysia	Environment
selangorensis	SCW17	533	P2	Malaysia	Environment
langatensis	SCW18	534	P2	Malaysia	Environment
fluminis	SCS5	540	P2	Malaysia	Environment
semungkisensis	SSS9	542	P2	Malaysia	Environment
saintgironsiae	SCS5	549	P2	Malaysia	Environment
johnsonii	E8	563	P2	Japan	Environment
sarikeiensis	201702455	575	P2	Malaysia	Environment
venezuelensis	CLM-U50	635	P2	Venezuela	Human
koniamboensis	201800265	644	P2	New Caledonia	Environment
andrefontaineae	201800301	679	P2	New Caledonia	Environment
perolatii	FH1-B-B1	812	P2	New Caledonia	soil
neocaledonica	ES4-C-A1	813	P2	New Caledonia	soil
haakeii	ATI7-C-A2	814	P2	New Caledonia	soil
hartskeerlii	MCA1-C-A1	815	P2	New Caledonia	soil
biflexa	Patoc 1 (Paris)	128	S1	Italy	Environment

bandrabouensis	201601111	185	S1	Mayotte	Environment
bouyouniensis	201601297	196	S1	Mayotte	Environment
mtsangambouensis	201601298	197	S1	Mayotte	Environment
meyeri	Went 5	204	S1	Canada	Unknown
wolbachii	CDC	224	S1	USA	Environment
yanagawae	Sao Paulo	260	S1	Brazil	Environment
brenneri	201602177	365	S1	New Caledonia	Environment
levettii	201602181	367	S1	New Caledonia	Environment
harrisiae	201602189	379	S1	New Caledonia	Environment
vanthielii	Waz Holland	419	S1	Netherlands	Environment
jelokensis	201702419	566	S1	Malaysia	Environment
congkakensis	201702421	568	S1	Malaysia	Environment
kemamanensis	201702454	573	S1	Malaysia	Environment
perdikensis	201702692	585	S1	Malaysia	Environment
ellinghausenii	201800220	632	S1	Japan	Environment
terpstrae	ATCC 700639	634	S1	China	Environment
montravelensis	201800279	658	S1	New Caledonia	Environment
bourretii	201800280	659	S1	New Caledonia	Environment
noumeaensis	201800287	666	S1	New Caledonia	Environment
kanakyensis	201800292	670	S1	New Caledonia	Environment
ilyithenensis	201400974	2	S2	Algeria	Environment
ognonensis	201702476	580	S2	France	Environment
kobayashii	E30	587	S2	Japan	Environment
ryugenii	YH101	588	S2	Japan	Environment
idonii	201300427	630	S2	Japan	Environment

* *L. tipperaryensis* was originatelly described as *L. alstonii* [12] and later on re-classified as a new species and named *L. tipperaryensis* [6].

**L. yasudae* and *L. stimsonii* [13] are presented here as *L. dzianensis* and *L. putramalaysiae* [6].