MICROBIAL GENOMICS

RESEARCH ARTICLE
Thibeaux *et al.*, *Microbial Genomics* 2018;4
DOI 10.1099/mgen.0.000144

Deciphering the unexplored *Leptospira* diversity from soils uncovers genomic evolution to virulence

Roman Thibeaux,[1]† Gregorio Iraola,[2]† Ignacio Ferrés,[2] Emilie Bierque,[1] Dominique Girault,[1] Marie-Estelle Soupé-Gilbert,[1] Mathieu Picardeau[3,*]† and Cyrille Goarant[1,*]†

Abstract

Despite recent advances in our understanding of the genomics of members of the genus *Leptospira*, little is known on how virulence has emerged in this heterogeneous bacterial genus as well as on the lifestyle of pathogenic members of the genus *Leptospira* outside animal hosts. Here, we isolated 12 novel species of the genus *Leptospira* from tropical soils, significantly increasing the number of known species to 35 and finding evidence of highly unexplored biodiversity in the genus. Extended comparative phylogenomics and pan-genome analyses at the genus level by incorporating 26 novel genomes, revealed that, the traditional leptospiral 'pathogens' cluster, as defined by their phylogenetic position, can be split in two groups with distinct virulence potential and accessory gene patterns. These genomic distinctions are strongly linked to the ability to cause or not severe infections in animal models and humans. Our results not only provide new insights into virulence evolution in the members of the genus *Leptospira*, but also lay the foundations for refining the classification of the pathogenic species.

DATA SUMMARY

1. *Leptospira* isolates are described in Table S1 (available in the online version of this article) and locations of isolation are shown on Fig. S1.

2. Genomes have been deposited in GenBank; accession numbers are given in Table S2.

3. Overall genomic relatedness indices are presented in Tables S3 (ANI) and S4 (AAI).

4. The protein domain abundance matrix is shown in Table S6 (xls file).

INTRODUCTION

Pathogenic species of the genus *Leptospira* cause leptospirosis, an emerging zoonosis worldwide with high prevalence in tropical low-income countries. Leptospirosis affects 1 million and kills 60 000 people annually, but remains poorly documented and often underestimated [1]. The burden of leptospirosis and its economic cost are significant and similar to that of other important neglected tropical diseases, including schistosomiasis, leishmaniasis and lymphatic filariasis [2]. Pathogenic leptospires are maintained in the renal tubules of asymptomatic reservoir animals, frequently rodents, and are excreted through the urine, contaminating the environment, where they can survive for months. Environment-mediated contamination is considered to be the major source of transmission to humans. Other animals, including livestock and companion animals, can also get infected and develop leptospirosis.

The genus *Leptospira* (phylum *Spirochetes*) is highly heterogeneous and genetically distinct from other bacteria, being currently divided into 22 species and more than 300 serovars. Phylogenetic analysis, initially based on the 16S rRNA gene but later on whole-genome sequences, showed that the genus is separated into three monophyletic clusters named 'saprophytes', 'intermediates' and 'pathogens'. The 'saprophytes' are environmental species which are rapidly cleared

Received 18 September 2017; Accepted 27 November 2017
Author affiliations: [1]Institut Pasteur in New Caledonia, Institut Pasteur International Network, Leptospirosis Research and Expertise Unit, Noumea, New Caledonia; [2]Institut Pasteur Montevideo, Bioinformatics Unit, Uruguay; [3]Institut Pasteur, Unité de Biologie des Spirochètes, 28 rue du docteur Roux, 75724 Paris Cedex 15, France.
***Correspondence:** Mathieu Picardeau, mathieu.picardeau@pasteur.fr; Cyrille Goarant, cgoarant@pasteur.nc
Keywords: *Leptospira*; genomics; evolution; virulence; ecology.
Abbreviations: AAI, average amino acid identity; ANI, average nucletide identity; COG, clusters of ortholog groups; DAPC, discriminant analysis of principal components; EMJH, Ellinghausen McCullough Johnson Harris culture medium; OGRI, overall genomic relatedness indices; SD, standard deviation.
†These authors contributed equally to this work
Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Five supplementary figures and six supplementary tables are available with the online version of this article.

in animal models, and are non-pathogenic to humans and other animals. The 'intermediates' have been recently described in both humans and animals, but infection of the classical animal models for acute leptospirosis with these species cannot reproduce the disease. Life-threatening species like *Leptospira interrogans*, which is the dominant pathogenic species worldwide, are classified within the 'pathogens' cluster and can infect every mammal. Conversely, *Leptospira kmetyi* belongs to the 'pathogens' cluster but has been isolated only from soil and never recovered from animals [3], calling the ecological coherence of the current classification into question.

The molecular bases of leptospiral pathogenicity, virulence and persistence remain at the onset of understanding, mainly because pathogenic species are fastidious and not prone to genetic manipulations, hampering the experimental discovery and validation of virulence determinants [4]. Alternatively, comparative genomics has uncovered key aspects of genomic adaptations to virulence [5, 6] but relevant questions still remain to be answered, fundamentally about the mechanisms that led the leptospiral ancestor to evolve from a saprophytic lifestyle into mammal-adapted pathogens. Consequently, a systematic evaluation of the relationship between genomic traits' evolution and virulence potential requires to be established [7].

In this work we reveal a significant amount of unexplored taxonomic diversity within the genus *Leptospira* by isolating 12 novel species from soils in areas of endemic leptospirosis. Using comparative phylogenetics, pan-genome analyses and *in vivo* models of infection we demonstrate that the 'pathogens' cluster is heterogeneous, being composed of both virulent and low-virulence strains with remarkable genomic distinctions. Our results provide new insights into virulence evolution in the genus *Leptospira* and indicate that the current classification of leptospiral species should be revised.

METHODS

Ethics statement, patient contact and authorization for interview

Institut Pasteur in New Caledonia has been the country reference and only laboratory for the biological diagnosis of human leptospirosis from 1980 to 2016. The patients were identified by a positive diagnostic quantitative PCR and notified to the New Caledonian Health Authority, which also investigates cases through interviews. Oral consent was requested by the Health Authority to meet with the patient, visit and collect environmental samples in the suspected infection sites. The detailed procedure has been described previously [8].

Study sites

Six sites were chosen based on the good acceptance of the project by the patients and custom chiefdom [Koné, Touho (two sites), Ponerihouen (two sites) and Yaté]. All sites were within Melanesian tribal areas and three (Koné and two sites in Touho) were also included in a previous study [8]. These sites are indicated on Fig. S1 together with the 30 year average temperatures (minima and maxima) and rainfall of the closest meteorological stations (retrieved from the Météo France free online public database).

IMPACT STATEMENT

Water-associated exposures are the main risk factors for leptospirosis, a complex disease with a multitude of infecting serovars, a broad reservoir host range, non-specific clinical manifestations and difficult diagnosis. To assess the diversity of environmental members of the genus *Leptospira*, we isolated and sequenced members of the genus *Leptospira* from hot spots of leptospirosis. General analysis of these genomes provided unprecedented insight into the diversity of the genus *Leptospira*. We described a total of 12 novel species, including species belonging to the cluster of potentially infectious leptospires. Surprisingly, novel species from the pathogenic cluster failed to produce an infection in animal models. A detailed analysis of accessory genomes revealed clear differences within this pathogen cluster between virulent species and others failing to cause infection. This sheds new light into the evolution and acquisition of virulence in this highly heterogeneous genus.

Collection and processing of environmental samples on site

Leptospira collection permits were obtained from the North (# 60912-2002-2017/JJC) and South (Arrêté 1689-2017/ARR/DENV) Provinces of New Caledonia. Environmental investigations were started a few weeks after the presumed human infection dates and after recovery of the patients, between March and June 2016. The soils selected to attempt culture of members of the genus *Leptospira* and isolation were chosen following discussions on site with the patients, based on environmental exposure of the patient on the day of probable contamination. The samples mostly included river soils, but also moist soils at a distance from any waterway if suggested by patient interviews (muddy walking tracks, agricultural soils). Most samples from the study sites were collected less than 20 meters one from one another; 27 soil samples were used to isolate members of the genus *Leptospira*. Samples were collected and processed on site as follows: approximately 5 g topsoil was collected from riverbanks (from 10 cm below to 1 m above water level), walking track or culture fields from a core sample (3 cm large by 5–7 cm height). Each soil sample was placed into a 15 ml sterile Falcon tube within 2 h of collection and vigorously shaken with 5–10 ml sterile water. The soil particles were allowed to settle for 5–15 min and 2 ml of supernatant were filtered through a sterile 0.45 µm filter into a tube filled with 2.5 ml of 2× EMJH medium. Alternatively, the process was repeated the next day at the laboratory, leaving more time for particles to settle, then culturing without filtration. Finally, we added 500 µl of 10× concentrated STAFF, a combination of selective agents for isolation of members of the genus *Leptospira* made of sulfamethoxazole, trimethoprim, amphotericin B, fosfomycin, and 5-fluorouracil [9]. Culture tubes from the field were transported within 12 h at ambient temperature to the laboratory, where they were put in an incubator at 30 °C. Alternatively, they were directly placed in the incubators when prepared in the laboratory.

Leptospira isolation

Cultures were checked daily by dark-field microscopy for the growth of spirochetes. When contaminants were observed, the cultures tubes were subcultured with STAFF after filtration through a 0.45 µm membrane filter. When spirochetes were observed, a 50 µl and a 200 µl volume of the culture at various dilutions was plated onto EMJH agar and incubated at 30 °C until individual subsurface colonies were visible. Most of individual colonies started to appear after 3 days of incubation, and at day 10 all plates were positive with 10 to 100 colonies of members of the genus *Leptospira*. One to five characteristic subsurface individual colonies were collected from each plate for confirmation of a morphology typical of members of the genus *Leptospira* by dark field microscopy before clonal subculture in liquid EMJH (Table S1).

Whole-genome sequencing

Genomic DNA was prepared by collection of cells by centrifugation from an exponential-phase culture and extraction with a MagNA Pure 96 Instrument (Roche). Next-generation sequencing was performed by the Mutualized Platform for Microbiology (P2M) at Institut Pasteur, using the Nextera XT DNA Library Preparation kit (Illumina), the NextSeq 500 sequencing system (Illumina) and the CLC Genomics Workbench 9 software (Qiagen) for analysis. The quality of the initial assemblies was improved with SPAdes [10] and a post-assembly improvement pipeline [11], the resulting draft genomes were automatically annotated with Prokka [12]. Draft genomes were submitted to Genbank, accession numbers are available in Table S2.

Taxonogenomics, pan-genome and phylogenetic analyses

A comprehensive set of draft and closed genomes that represents the currently described leptospiral species was retrieved from the PATRIC database [13]. Most of these genomes have been previously used to study the genomic evolution of the genus *Leptospira* [5]. The final dataset was composed of available genomes (*n*=22, because whole genome sequences for *Leptospira idonii* were not publicly available, but including the reference genome of *Leptospira venezuelensis* sp. nov. currently under description by members of our group [14]) and those sequenced in this study (*n*=26).

To determine the relationship of each sequenced genome to previously described or novel leptospiral species, we calculated two Overall Genetic Relatedness Indices (OGRIs): the Average Nucleotide Identity (ANI) and the Average Amino acid Identity (AAI). Both indices were automatically calculated using two-way BLAST + blastn and blastp [15] comparisons as previously implemented [16], using the Taxxo R package (https://github.com/giraola/taxxo).

To build a standard phylogeny the 16S rRNA gene sequences were extracted from whole genomes (the genome of *Leptonema illini* DSM 21528^T was included as an outgroup) using BLAST + blastn against the 16S ribosomal RNA sequence database at the NCBI. Sequences were aligned with MAFFT [17] and phylogenetic reconstruction was performed with FastTree v.2.1 [18] using the GTR substitution model and 1000 replicates to calculate bootstrap values.

To build a genome-wide high-resolution phylogeny of the whole genus *Leptospira* and using the *Leptonema illini* DSM 21528^T genome as the outgroup (*n*=49), a set of highly conserved core genes (present in at least 95 % of the genomes) was identified by comparing each genome against the eggNOG v3.0 database [19] specifically customized for the phylum *Spirochaetes* (spiNOG) using HMMER v3.1b2 [20]. A set of 671 genes were identified, concatenated and aligned with MAFFT [17] (total alignment length was 778 190 bp). Phylogenetic reconstruction was performed as described above. Pairwise patristic distances were calculated from the resulting tree using the APE package [21].

Comparative pan-genome analyses were performed over the set of genomes belonging to the 'intermediates' (*n*=15) and 'pathogens' (*n*=17) clusters. The pan-genome was reconstructed using an in-house pipeline (available at https://github.com/iferres/pewit). Briefly, for every genome, each annotated gene was scanned against the Pfam database [22] using HMMER3 v3.1b2 hmmsearch [20] and its domain architecture was recorded (presence and order). A primary set of orthologous clusters was generated by grouping genes sharing exactly the same domain architecture. Then, remaining genes without hits against the Pfam database were compared with each other at protein level using HMMER3 v3.1b2 phmmer and clustered using the MCL algorithm [23]. These coarse clusters were then split using a tree-prunning algorithm which allows discrimination between orthologous and paralogous genes. Functional category assignments to each orthologous cluster were performed with BLAST + blastp against the Clusters of Ortholog Groups (COGs) database [24]. The Jaccard distance over accessory gene patterns was calculated with the package ade4 [25]. Cluster-defining accessory genes were identified with K-pax2 [26] by running its Bayesian clustering method over the pan-genome matrix with default parameters. Paralogous genes were defined as those orthologous clusters with more than one gene copy in at least one genome and the Bray–Curtis distance was calculated with the package Vegan [27]. Protein domains were extracted by comparing each genome annotation against the Pfam database [23] using HMMER3 v3.1b2 hmmsearch [20]. A domain abundance matrix was created by recording the number of occurrences of each domain in each genome and this was used to perform a Discriminant Analysis of Principal Components (DAPC) as implemented in package adegenet [28]. To identify genes contributing highly to the observed clustering we used the PCA loadings and adjusted them to a normal distribution. Then those genes with loadings departing more than two standard deviations (SD) from the mean were selected. Bray–Curtis distances from the abundance patterns of selected genes were calculated as explained above. Tests of proportions and Mann–Witney U were calculated in R [29].

Virulence of novel species

To evaluate if isolates of the novel pathogenic and intermediate species were virulent, 7–8-week-old golden Syrian hamsters (males and females) and 8-week-old Oncins-France 1 (OF-1, outbred) mice (males and females) whose progenitors originate from Charles River Laboratories, were infected by intraperitoneal injection of 2×10^8 leptospires in pure culture. Similar infections were performed with hamsters and mice, which were similarly infected with *Leptospira interrogans* serovar Manilae strain L495 (2×10^6 per animal) and *Leptospira borgpetersenii* serogroup Ballum strain B3-13S (2×10^8 per animal). Hamsters were euthanized by carbon dioxide inhalation 4.5 days after infection and the blood collected from heart puncture was cultured in EMJH. The urine of mice was collected 7–8 days after infection, DNA was extracted and analyzed by a real-time PCR targeting the conserved regions of the 16S rRNA gene *rrs* [30]. After two weeks, mice were euthanized by carbon dioxide inhalation. One kidney was collected and its DNA extracted and analyzed using the same PCR. All experiments were replicated twice on different days and using independent bacterial cultures. Animal experiments were conducted according to the guidelines of the Animal Care and Use Committees of the Institut Pasteur of Paris and of New Caledonia, and followed European Recommendation 2007/526/EC. Protocols and experiments were approved by the Animal Care and Use Committees of the Institut Pasteur in New Caledonia.

RESULTS

Culture isolation, identification and phylogenetic position of novel species

Using a previously described [9] combination of selective agents that facilitates the isolation of leptospires from complex environmental samples, we isolated 26 strains of members of the genus *Leptospira* from tropical soils from six sites in New Caledonia, where the disease is endemic (Fig. S1 and Table S1).

Whole-genome sequences of all 26 isolates were determined (genome statistics are presented in Table S2). To assign the novel species to the traditional leptospiral phylogenetic clusters we built a full-length 16S rRNA gene phylogeny (Fig. S2). By comparing the identity of full-length 16S rRNA gene sequences, we could assign a few isolates to previously described leptospiral species (100 % nucleotide identity). However, the remaining isolates had unique sequences, suggesting the presence of unknown species. We calculated both the average nucleotide identity (ANI) and the average amino acid identity (AAI) of the 26 isolates against each other and to the 22 previously described species of the genus *Leptospira* (see Methods and Tables S3 and S4). Fig. 1a shows the relationship between genomes according to the standard ANI and/or AAI threshold >95±1 %. This analysis confirmed the presence of 12 novel species of the genus *Leptospira*, thus extending by 55 % the number of species within this genus.

As 16S rRNA gene sequence conservation prevents precise separation of some well-defined species of the genus *Leptospira* [5], we then built a high-resolution phylogenetic tree based on the concatenated coding sequences of 671 leptospiral core genes (also occurring in *Leptonema illini*) (see Methods). This phylogeny not only reproduced the typical topology with the three main clusters designated as pathogens, intermediates and saprophytes but also confirmed the position of the 12 novel species as separate branches (Fig. 1b). Three of them were classified with the pathogens (later designated sp. nov. patho 1–3), five novel species were identified as intermediates (later designated sp. nov. inter 1–5) and four novel species were assigned to the saprophytes (later designated sp. nov. sapro 1–4). Interestingly, the three novel species assigned to the pathogens presented a basal position with respect to the previously identified species within this group and were closer to the tree root (Fig. S3).

Accessory gene patterns recapitulate virulence potential

The description of novel species assigned to both pathogens and intermediates prompted us to evaluate their relative virulence using animal models. Infection with virulent strains is associated with systemic infection with bacteremia and usually with severe acute disease in susceptible animals such as hamsters [31], and with an asymptomatic infection leading to renal colonization and urinary shedding in mice and rats [32]. Fig. 2a shows the infection profiles of one representative isolate per novel species identified as a member of the pathogens and intermediates, in comparison with the virulent references *Leptospira interrogans* strain L495 and *Leptospira borgpetersenii* strain B3-13S. The hamsters infected with these virulent strains showed signs of acute infection 3–4 days after infection (decreased activity, anorexia, ruffled fur and jaundice visible at the oral mucosa and skin levels) and renal colonization was evidenced in mice one and two weeks after infection. In contrast, hamsters infected with the novel species displayed no alteration in behavior, aspect or appetite and no culture was obtained from their blood and no leptospiral DNA was detected in the urine or the kidney of mice infected with the same strains. These results indicate the inability of these novel species to establish acute infection or renal colonization in these animal models. This is in marked contrast to the behavior of virulent pathogens like *Leptospira interrogans* and *Leptospira borgpetersenii*, suggesting the hereinafter

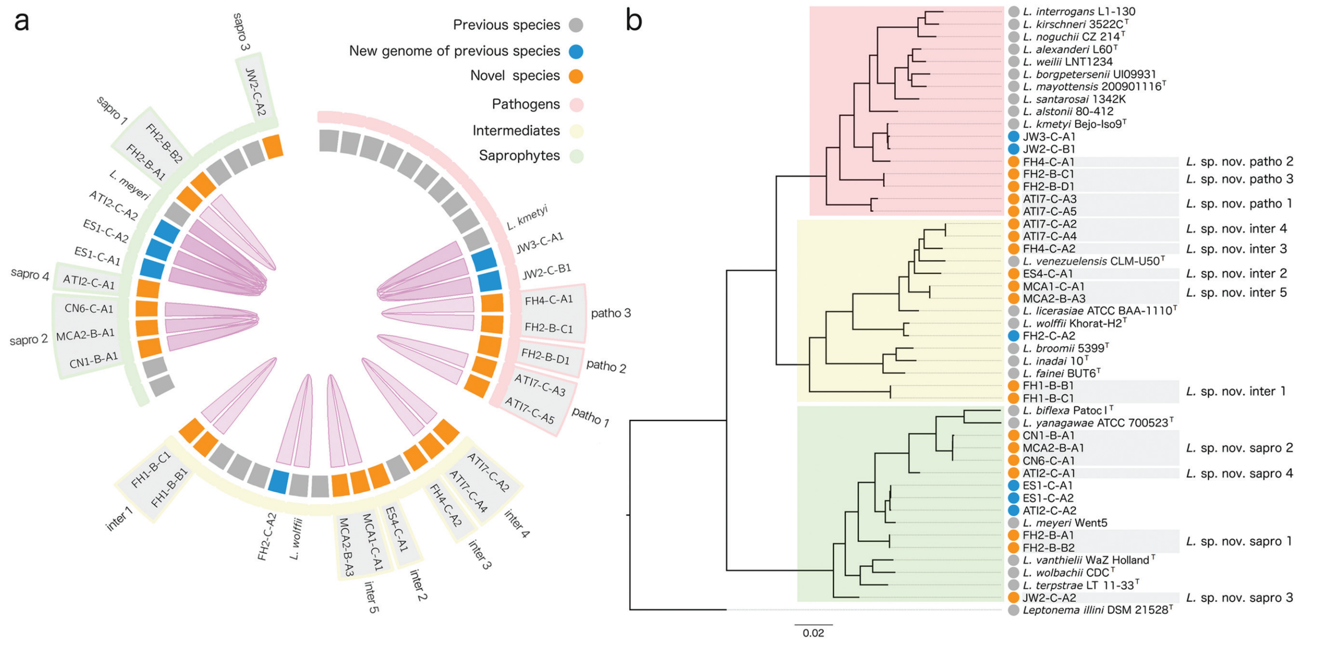

Fig. 1. Phylogenetic position of the novel species. (a) Circos diagram showing the relationships between leptospiral genomes based on overall genomic relatedness indices (OGRIs). The inner violet ribbons connect pairs of genomes if they share >95 % average nucleotide identity (ANI) and average amino acid identity (AAI). Blocks represent each genome coloured as explained below. The outer highlights show the leptospiral clades. (b) Maximum-likelihood phylogeny for the genus *Leptospira* based on the core genome alignment. The tree is rooted with *Leptonema illini* DSM 21528^T. The three classic leptospiral clades historically associated with differential pathogenicity are highlighted in red ('pathogens'), yellow ('intermediates') and green ('saprophytes'). Coloured circles at species labels indicate a public genome from a previously described species (grey), a genome sequenced in this study assigned to a previously described species (blue) or a genome sequenced in this study from a novel species (orange).

denomination of these novel species as 'low-virulence pathogens'.

These results led us to conduct a detailed comparative analysis of the accessory genomes of virulent pathogens, low-virulence pathogens and intermediates. We first noted that after adding the genomes from novel species belonging to these groups the pan-genome remained open (Fig. S4), revealing the divergent and highly diverse attributes of the members of the genus *Leptospira*. Then, a comparison of accessory gene patterns using the Jaccard distance showed a clear separation of intermediates from virulent and low-virulence pathogens (Fig. 2b). More interestingly, the accessory gene patterns were informative enough to discriminate two clusters that correlate with the subdivision of pathogens into virulent pathogens and low-virulence pathogens, in agreement with the virulence experiments. It is worth mentioning that *Leptospira kmetyi* belongs to the accessory genome cluster containing the low-virulence pathogens, which is also coherent with the unknown virulence potential of this species whose isolation has been only reported from soils. Hence, this analysis revealed clear genomic distinctions in the accessory genome of virulent and low-virulence pathogens that are not evident from the core genome phylogeny, which shows that low-virulence pathogens are a paraphyletic group (Fig. 2b).

Genomic features associated with leptospiral virulence

To provide a functional overview of the evolutionary adaptations associated with leptospiral virulence, we identified the genes that are associated with virulent or low-virulence pathogens. First, we used a Bayesian probabilistic framework [26] to detect those accessory genes with high discriminatory power for the three virulence groups (virulent pathogens, low-virulence pathogens and intermediates). Fig. 3a shows the number of discriminatory genes for each group (*n*=409), representing approximately 1.5 % of the accessory genes occurring in intermediates, low-virulence and virulent pathogens. Among these, 18 genes were found to distinguish virulent pathogens from both low-virulence pathogens and from intermediates (Table S5). Fig. 3b shows that using the presence/absence patterns of this small subset of genes completely recapitulates the three virulence groups, showing that very specific accessory genes can explain the evolution of virulence in the genus *Leptospira*. To gain insight into the biological functions related to this discrimination, we assigned COG [24] annotations to detect any functional enrichment. Fig. 3c shows that many functional categories are differentially represented in the accessory gene subsets defining each virulence group. Virulent pathogens are mainly distinguished from others by a significantly

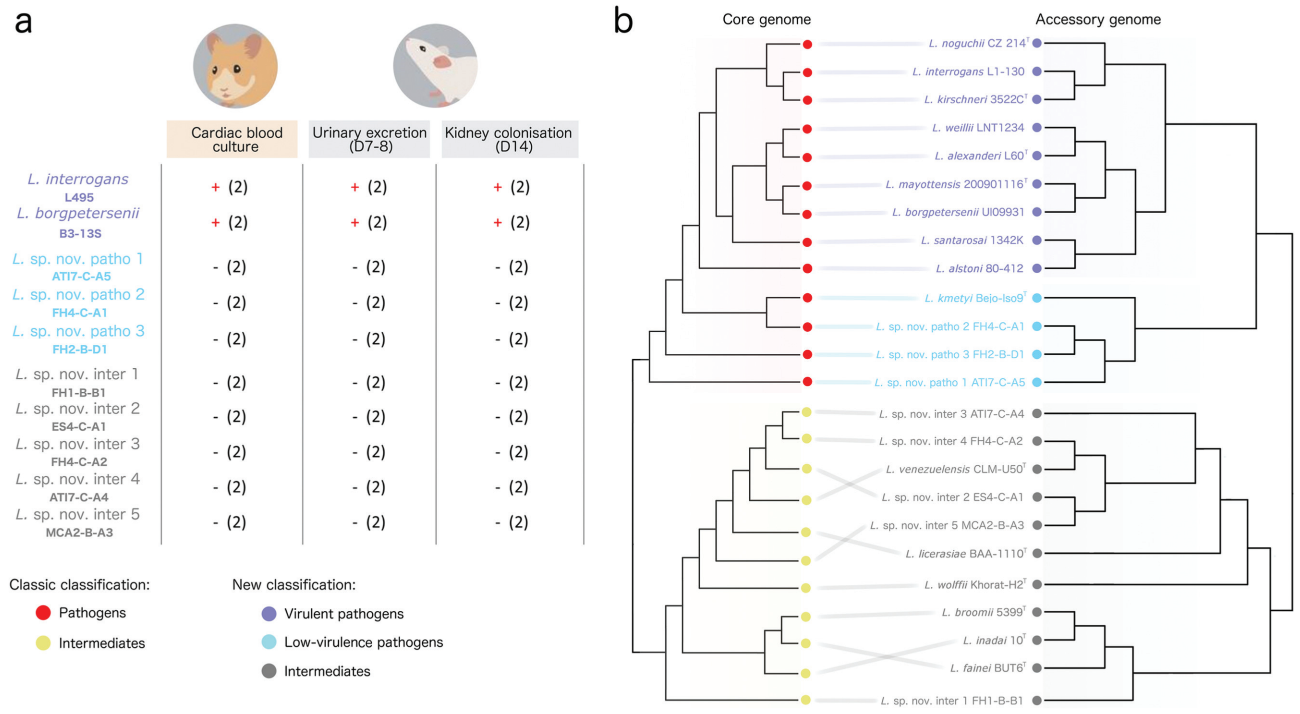
a
Cardiac blood culture
Urinary excretion (D7-8)
Kidney colonisation (D14)
L. interrogans L495 + (2) + (2) + (2)
L. borgpetersenii B3-13S + (2) + (2) + (2)
L. sp. nov. patho 1 ATI7-C-A5 - (2) - (2) - (2)
L. sp. nov. patho 2 FH4-C-A1 - (2) - (2) - (2)
L. sp. nov. patho 3 FH2-B-D1 - (2) - (2) - (2)
L. sp. nov. inter 1 FH1-B-B1 - (2) - (2) - (2)
L. sp. nov. inter 2 ES4-C-A1 - (2) - (2) - (2)
L. sp. nov. inter 3 FH4-C-A2 - (2) - (2) - (2)
L. sp. nov. inter 4 ATI7-C-A4 - (2) - (2) - (2)
L. sp. nov. inter 5 MCA2-B-A3 - (2) - (2) - (2)
Classic classification:
● Pathogens
● Intermediates
New classification:
● Virulent pathogens
● Low-virulence pathogens
● Intermediates
b
Core genome
Accessory genome
L. noguchii CZ 214T
L. interrogans L1-130
L. kirschneri 3522CT
L. weillii LNT1234
L. alexanderi L60T
L. mayottensis 200901116T
L. borgpetersenii UI09931
L. santarosai 1342K
L. alstoni 80-412
L. kmetyi Bejo-Iso9T
L. sp. nov. patho 2 FH4-C-A1
L. sp. nov. patho 3 FH2-B-D1
L. sp. nov. patho 1 ATI7-C-A5
L. sp. nov. inter 3 ATI7-C-A4
L. sp. nov. inter 4 FH4-C-A2
L. venezuelensis CLM-U50T
L. sp. nov. inter 2 ES4-C-A1
L. sp. nov. inter 5 MCA2-B-A3
L. licerasiae BAA-1110T
L. wolffii Khorat-H2T
L. broomii 5399T
L. inadai 10T
L. fainei BUT6T
L. sp. nov. inter 1 FH1-B-B1

Fig. 2. Virulence in animal models and accessory genome topology. (a) Virulence of novel species in experimental challenge infections (*n*=2 for each strain and animal model). Only the pathogenic strains *Leptospira interrogans* L495 and *Leptospira borgpetersenii* B3-13S were recovered from hamster cardiac blood or evidenced from mouse urine and kidney. (b) Tanglegram comparing the topology of the core genome phylogeny (left) and the topology obtained by clustering the genomes using Jaccard distance calculated over the accessory gene patterns (right). On the left, genomes are coloured according to the classic phylogenetic classification (only pathogens and intermediates are shown here). On the right, genomes are coloured according to the new classification based on accessory gene patterns.

higher number of genes related to cell wall/membrane biogenesis (M), cell motility and chemotaxis (N), both known or suspected to be involved in virulence [33, 34], post-translational modification (O), also suspected to be involved in virulence [5, 35], as well as a lower number of genes related to amino acid metabolism and transport (E) and transcription (K). These differences reflect functional distinctions that are specific to virulent pathogens in comparison with both low-virulence pathogens and intermediates.

To have a more complete description of group-specific molecular functions associated with virulence, and considering the observed bias in COG annotations where a substantial proportion (44 %) of genes is not assigned to any known function, we analyzed the abundance patterns of protein domains by comparing each genome against the Pfam database [22]. Fig. 4a shows a Discriminant Analysis of Principal Components (DAPC) [28] that completely discriminates the three virulence groups using protein domain patterns. Furthermore, when considering only those domains that are highly informative for generating the observed clustering (see Methods), we were able to reproduce the three virulence groups using a different clustering analysis based on the Bray–Curtis distance (Fig. 4b). Interestingly, we noticed a group of six domains whose abundance was high in virulent pathogens while almost null in low-virulence pathogens and intermediates. Most of these domains belong to repeated elements such as mobile elements (DDE endonuclease superfamily) and proteins of paralogous families (Beta-propeller repeat- and leucine-rich repeat-containing proteins). This indicates that virulent pathogens can be distinguished by increased repeat sequence elements in comparison with the low-virulence pathogens and intermediates, suggesting a functional link to virulence. Other Pfam domains allowing this discrimination are presented in Table S6.

Given the importance of repeat sequence elements in ecological adaptation of organisms by shaping their genomes [36], an analysis focused on the abundance of paralogous genes in the accessory genomes was performed. First, we evidenced that patristic distances obtained from the core genome phylogeny were highly correlated with Bray–Curtis distances calculated from the abundance of paralogous genes (Fig. 5a), indicating that phylogenetically closer species share more similar paralogy patterns. Also, when observing just the abundance distributions of paralogous genes in each virulence group we detected a significantly

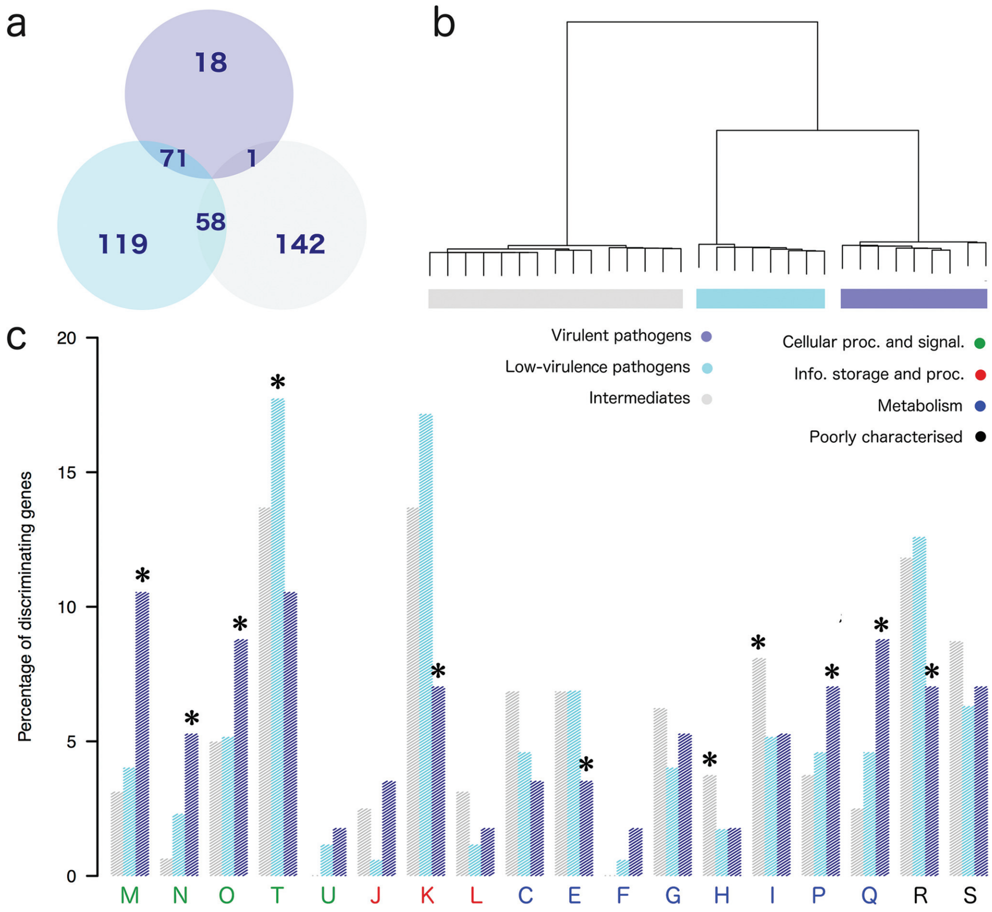

Fig. 3. Functional analysis of discriminating accessory genes. (a) Venn diagram showing the Bayesian identification of cluster-defining accessory genes from the pan-genome. (b) Clustering analysis based on Jaccard distances calculated from the presence/absence vectors of cluster-defining genes. (c) Barplots showing the percentage of cluster-defining genes assigned to each COG functional category in each cluster. Statistical significance (*P*<0.001, test of proportions) is indicated with asterisks.

higher incidence of paralogy in virulent pathogens in comparison with low-virulence pathogens and intermediates (*P*<0.001, Mann–Witney U test) (Fig. 5b). The same trend was observed in Fig. 5c, where Bray–Curtis distances were used to perform a cluster analysis that reconstructed the three virulence groups. Additionally, a significant and positive correlation was found between the number of transposase domains and the number of paralogous genes encoded in each genome (Fig. S5). In summary, these results indicate that paralogy has played an important role in the emergence of virulent pathogens.

DISCUSSION

Pathogenic species of the genus *Leptospira* are a unique group of highly fastidious bacteria, difficult to isolate in pure cultures. In this study, 12 novel species were successfully isolated from a relatively small number of soil samples from New Caledonia, highlighting a greatly unexplored biodiversity in the genus that is probably only the tip of the iceberg, and the number of recognized species may explode in a near future. Indeed, the presence of putatively novel uncultured species of the genus *Leptospira* has been detected from unrelated sources such as bats [37–41] and Amazonian soils [42]. The impressive diversity found in our study indicates that soils may not only be considered as a secondary passive reservoir of leptospirosis, but also the birthplace of the genus *Leptospira* as previously suggested [6].

From a medical point of view, the description of novel species of intermediates and pathogens may have implications for public health. However, in our infection experiments and despite high infectious doses, none of these novel species could induce signs or symptoms of infection in the hamster model, and isolates could not be recovered from hamster blood. Similarly, these isolates could not be detected in mouse urine or kidney, suggesting their inability either to infect mice or to colonize kidney tubules, also calling into question the need for a mammal reservoir in their biology. Moreover, only *Leptospira interrogans* and *Leptospira borgpetersenii* have been detected in clinical cases in

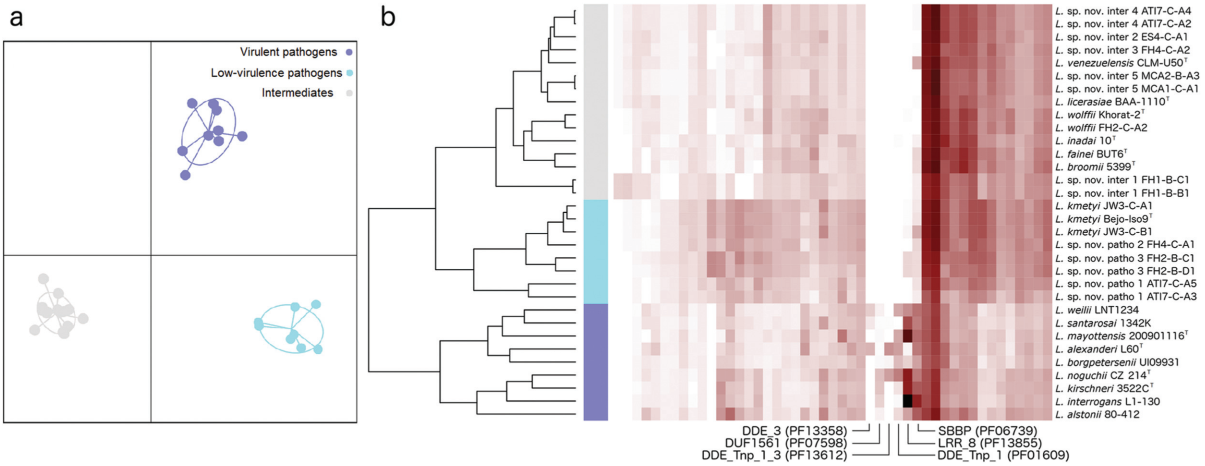

Fig. 4. Protein domains analysis. (a) Scatterplot showing the first and second discriminant functions obtained from the Discriminant Analysis of Principal Components (DAPC), performed with protein domain abundances extracted from the coding sequences of each genome. Groups are coloured according to the new classification: intermediates (grey), low-virulence pathogens (cyan) and virulent pathogens (purple). (b) Heatmap showing the relationships between genomes obtained by calculating the Bray–Curtis distances from abundance patterns of a subset of highly discriminating domains obtained from the DAPC analysis. Redness indicates increasing domain copy number.

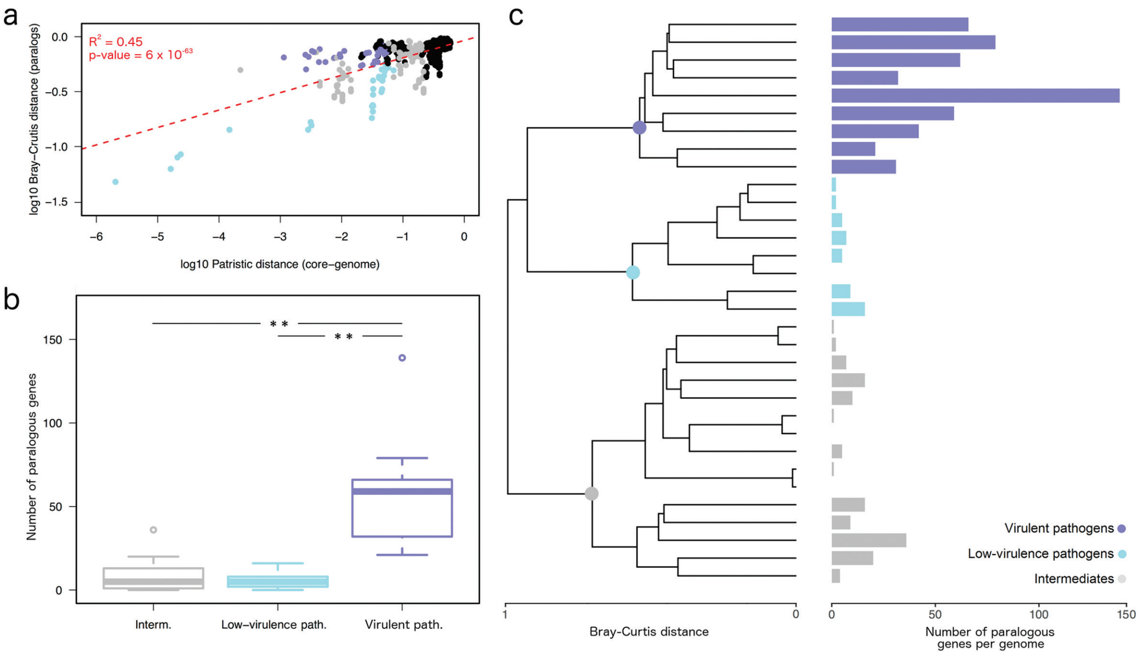

Fig. 5. Analysis of paralogous genes. (a) Linear regression showing the correlation between patristic distances calculated from the core genome phylogeny and Bray–Curtis distances calculated from the abundance patterns of paralogous genes. Dots are coloured according to virulence clusters when both genomes in the pair belong to the same cluster, black dots represent pairs of genomes belonging to different groups. (b) Boxplots showing the distribution of paralogous genes in the three virulence clusters. Asterisks indicate *P*<0.001 (Mann–Witney U test). (c) Clustering analysis using the Bray–Curtis distances calculated from the abundance patterns of paralogous genes. Horizontal bars indicate the number of paralogous genes per genome and are coloured according to virulence clusters.

New Caledonia through an active surveillance system [43, 44]. Together, these results indicate that the novel species have no or very limited virulence potential to mammals.

From an evolutionary perspective, genomes of these low-virulence species present an ancestral phylogenetic position with respect to the virulent pathogens, supporting the current hypothesis for explaining the emergence of leptospiral pathogens from free-living ancestral species inhabiting soils. This also indicates that virulence has evolved independently in pathogens and intermediates, as evidenced by different accessory gene and domain patterns in virulent pathogens and intermediates. More importantly, our results support the need to refine the classification of pathogens, which today are assumed to be an ecologically coherent group by sharing a higher virulence potential in comparison with intermediates. Despite the authors of some previous studies having proposed that virulence may be variable among different species classed as pathogens [5, 6, 45], our more comprehensive taxonomic coverage combined with infection experiments and accessory genome analyses demonstrated the presence of two groups of species of the genus *Leptospira* within the pathogens, correlated with clearly distinctive virulence potentials.

Taken together, our results indicate that virulent pathogens have adapted their genomes from a soil free-living to a mammal-associated virulent lifestyle mainly by expanding particular groups of protein families through gene duplication. These genomic distinctions should be used to establish more adequate criteria for the classification of pathogenic leptospires and to focus future work on the dissection of the molecular mechanisms and biological role of these genes.

Funding information

The current researcher position of R. T. is supported by an AXA Research Funds grant 'AXA Postdoctoral Fellowship', 15-AXA-PDOC-037. I. F. is supported by the Agencia Nacional de Investigación e Innovación posgraduation bursaries program from Uruguay grant POS_NAC_2016_1_131079 and G. I. by the Fondo de Convergencia Estructural del Mercosur (FOCEM) grant COF 04/11. E. B. is supported by a doctoral grant of the Institut Pasteur International Network.

Acknowledgements

Thanks are due to the patients and their relatives, who kindly visited the suspected contamination sites with us for the study and to the New Caledonian Health Authority who gained the consent for site visits. We thank Vincent Enouf and his platform team [Institut Pasteur, Pasteur International Bioresources network (PIBnet), Mutualized Platform for Microbiology (P2M)] for the next-generation sequencing analysis.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Ethical statement

The Institut Pasteur in New Caledonia has been the country reference and only laboratory for the biological diagnosis of human leptospirosis from 1980 to 2016. The patients were identified by a positive diagnostic qPCR and notified to the New Caledonian Health Authority, which also investigates cases through interviews. Oral consent was requested by the Health Authority to meet with the patient, visit and collect environmental samples in the suspected infection sites. *Leptospira* collection permits were obtained from the North (# 60912-2002-2017/JJC) and South (Arrêté 1689-2017/ARR/DENV) Provinces of New Caledonia. Animal experiments were conducted according to the guidelines of the Animal Care and Use Committees of the Institut Pasteur of Paris and of New Caledonia, and followed European Recommendation 2007/526/EC. Protocols and experiments were approved by the Animal Care and Use Committees of the Institut Pasteur in New Caledonia.

Data bibliography

1. Bateman A, Coin L, Durbin R, Finn RD, Hollich V *et al*. The Pfamprotein families database. *Nucleic Acids Res* 2004;32:138D–141.

2. Fouts DE, Matthias MA, Adhikarla H, Adler B, Amorim-Santos L *et al*. What makes a bacterial species pathogenic?: comparative genomic analysis of the genus *Leptospira*. *PLoS Negl Trop Dis* 2016;10: e0004403.

3. Galperin MY, Makarova KS, Wolf YI, Koonin EV. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res* 2015;43:D261–D269.

4. Puche R, Ferres I, Caraballo L, Rangel Y, Picardeau M *et al*. *Leptospira venezuelensis* sp. nov., a new member of the intermediates group isolated from rodents, cattle and humans. *IJSEM*. doi:10.1099/ijsem.0.002528 [Epub ahead of print].

References

1. Costa F, Hagan JE, Calcagno J, Kane M, Torgerson P *et al.* Global morbidity and mortality of leptospirosis: a systematic review. *PLoS Negl Trop Dis* 2015;9:e0003898.
2. Torgerson PR, Hagan JE, Costa F, Calcagno J, Kane M *et al.* Global burden of leptospirosis: estimated in terms of disability adjusted life years. *PLoS Negl Trop Dis* 2015;9:e0004122.
3. Slack AT, Khairani-Bejo S, Symonds ML, Dohnt MF, Galloway RL *et al. Leptospira kmetyi* sp. nov., isolated from an environmental source in Malaysia. *Int J Syst Evol Microbiol* 2009;59:705–708.
4. Murray GL, Morel V, Cerqueira GM, Croda J, Srikram A *et al.* Genome-wide transposon mutagenesis in pathogenic *Leptospira* species. *Infect Immun* 2009;77:810–816.
5. Fouts DE, Matthias MA, Adhikarla H, Adler B, Amorim-Santos L *et al.* What makes a bacterial species pathogenic?:comparative genomic analysis of the genus *Leptospira*. *PLoS Negl Trop Dis* 2016;10:e0004403.
6. Xu Y, Zhu Y, Wang Y, Chang YF, Zhang Y *et al.* Whole genome sequencing revealed host adaptation-focused genomic plasticity of pathogenic *Leptospira*. *Sci Rep* 2016;6:20020.
7. Picardeau M. Virulence of the zoonotic agent of leptospirosis: still terra incognita? *Nat Rev Microbiol* 2017;15:297–307.
8. Thibeaux R, Geroult S, Benezech C, Chabaud S, Soupé-Gilbert ME *et al.* Seeking the environmental source of leptospirosis reveals durable bacterial viability in river soils. *PLoS Negl Trop Dis* 2017; 11:e0005414.
9. Chakraborty A, Miyahara S, Villanueva SY, Saito M, Gloriani NG *et al.* A novel combination of selective agents for isolation of *Leptospira* species. *Microbiol Immunol* 2011;55:494–501.
10. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.
11. Page AJ, de Silva N, Hunt M, Quail MA, Parkhill J *et al.* Robust high-throughput prokaryote *de novo* assembly and improvement pipeline for Illumina data. *Microb Genom* 2016;2:e000083.
12. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.
13. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T *et al.* Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Res* 2017; 45D535–D542.
14. Puche R, Ferrès I, Caraballo L, Rangel Y, Picardeau M *et al. Leptospira venezuelensis* sp. nov., a new member of the intermediates group isolated from rodents, cattle and humans. *IJSEM*. doi: 10.1099/ijsem.0.002528 [Epub ahead of print].

15. **Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J *et al.*** BLAST+: architecture and applications. *BMC Bioinformatics* 2009; 10:421.

16. **Piccirillo A, Niero G, Calleros L, Pérez R, Naya H *et al.*** *Campylobacter geochelonis* sp. nov. isolated from the western Hermann's tortoise (*Testudo hermanni hermanni*). *Int J Syst Evol Microbiol* 2016;66:3468–3476.

17. **Katoh K, Misawa K, Kuma K, Miyata T.** MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res* 2002;30:3059–3066.

18. **Price MN, Dehal PS, Arkin AP.** FastTree 2-approximately maximum-likelihood trees for large alignments. *PLoS One* 2010;5: e9490.

19. **Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M *et al.*** eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res* 2012;40:D284–D289.

20. **Eddy SR.** Accelerated profile HMM searches. *PLoS Comput Biol* 2011;7:e1002195.

21. **Popescu AA, Huber KT, Paradis E.** ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. *Bioinformatics* 2012;28:1536–1537.

22. **Bateman A, Coin L, Durbin R, Finn RD, Hollich V *et al.*** The Pfam protein families database. *Nucleic Acids Res* 2004;32:138D–141.

23. **Enright AJ, van Dongen S, Ouzounis CA.** An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 2002; 30:1575–1584.

24. **Galperin MY, Makarova KS, Wolf YI, Koonin EV.** Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res* 2015;43:D261–D269.

25. **Dray S, Dufour A-B.** The **ade4** package: implementing the duality diagram for ecologists. *J Stat Softw* 2007;22:1–20.

26. **Pessia A, Grad Y, Cobey S, Puranen JS, Corander J.** K-Pax2: Bayesian identification of cluster-defining amino acid positions in large sequence datasets. *Microb Genom* 2015;1:e000025.

27. **Oksanen J, Blanchet F, Kindt R, Legendre P, Minchin P.** *Vegan: community Ecology Package*. R Package 2.0. 3 CRAN R-project org/package= vegan. 2012

28. **Jombart T, Ahmed I.** adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* 2011;27:3070–3071.

29. **R Core Team.** *R: A language and environment for statistical computing*. R foundation for Statistical Computing. 2014

30. **Mérien F, Amouriaux P, Perolat P, Baranton G, Saint Girons I.** Polymerase chain reaction for detection of *Leptospira* spp. in clinical samples. *J Clin Microbiol* 1992;30:2219–2224.

31. **Haake DA.** Hamster model of leptospirosis. *Current Protocols in Microbiology*; 2006. Chapter :12:Unit 12E 12.

32. **Marcsisin RA, Bartpho T, Bulach DM, Srikram A, Sermswan RW *et al.*** Use of a high-throughput screen to identify *Leptospira* mutants unable to colonize the carrier host or cause disease in the acute model of infection. *J Med Microbiol* 2013;62:1601–1608.

33. **Eshghi A, Becam J, Lambert A, Sismeiro O, Dillies MA *et al.*** A putative regulatory genetic locus modulates virulence in the pathogen *Leptospira interrogans*. *Infect Immun* 2014;82:2542–2552.

34. **Lambert A, Picardeau M, Haake DA, Sermswan RW, Srikram A *et al.*** FlaA proteins in *Leptospira interrogans* are essential for motility and virulence but are not required for formation of the flagellum sheath. *Infect Immun* 2012;80:2019–2025.

35. **Ricaldi JN, Matthias MA, Vinetz JM, Lewis AL.** Expression of sialic acids and other nonulosonic acids in *Leptospira*. *BMC Microbiol* 2012;12:161.

36. **Eme L, Doolittle WF.** Microbial evolution: Xenology (apparently) trumps paralogy. *Curr Biol* 2016;26:R1181–R1183.

37. **Dietrich M, Wilkinson DA, Benlali A, Lagadec E, Ramasindrazana B *et al.*** *Leptospira* and paramyxovirus infection dynamics in a bat maternity enlightens pathogen maintenance in wildlife. *Environ Microbiol* 2015;17:4280–4289.

38. **Dietrich M, Wilkinson DA, Soarimalala V, Goodman SM, Dellagi K *et al.*** Diversification of an emerging pathogen in a biodiversity hotspot: *Leptospira* in endemic small mammals of Madagascar. *Mol Ecol* 2014;23:2783–2796.

39. **Gomard Y, Dietrich M, Wieseke N, Ramasindrazana B, Lagadec E *et al.*** Malagasy bats shelter a considerable genetic diversity of pathogenic *Leptospira* suggesting notable host-specificity patterns. *FEMS Microbiol Ecol* 2016;92:fiw037.

40. **Matthias MA, Díaz MM, Campos KJ, Calderon M, Willig MR *et al.*** Diversity of bat-associated *Leptospira* in the Peruvian Amazon inferred by bayesian phylogenetic analysis of 16S ribosomal DNA sequences. *Am J Trop Med Hyg* 2005;73:964–974.

41. **Ogawa H, Koizumi N, Ohnuma A, Mutemwa A, Hang'ombe BM *et al.*** Molecular epidemiology of pathogenic *Leptospira* spp. in the straw-colored fruit bat (*Eidolon helvum*) migrating to Zambia from the Democratic Republic of Congo. *Infect Genet Evol* 2015;32:143–147.

42. **Ganoza CA, Matthias MA, Collins-Richards D, Brouwer KC, Cunningham CB *et al.*** Determining risk for severe leptospirosis by molecular analysis of environmental surface waters for pathogenic *Leptospira*. *PLoS Med* 2006;3:e308.

43. **Salaün L, Mérien F, Gurianova S, Baranton G, Picardeau M.** Application of multilocus variable-number tandem-repeat analysis for molecular typing of the agent of leptospirosis. *J Clin Microbiol* 2006;44:3954–3962.

44. **Goarant C, Laumond-Barny S, Perez J, Vernel-Pauillac F, Chanteau S *et al.*** Outbreak of leptospirosis in New Caledonia: diagnosis issues and burden of disease. *Trop Med Int Health* 2009; 14:926–929.

45. **Lehmann JS, Matthias MA, Vinetz JM, Fouts DE.** Leptospiral pathogenomics. *Pathogens* 2014;3:280–308.