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

Roman Thibeaux, Gregorio Iraola, Ignacio Ferrés, Emilie Bierque, Dominique Girault, Marie-Estelle Soupé-Gilbert, Mathieu Picardeau, Cyrille Goarant

▶ **To cite this version:**

Roman Thibeaux, Gregorio Iraola, Ignacio Ferrés, Emilie Bierque, Dominique Girault, et al.. Deciphering the unexplored Leptospira diversity from soils uncovers genomic evolution to virulence. Microbial Genomics, 2018, 4 (1), 10.1099/mgen.0.000144 . hal-01710458

HAL Id: hal-01710458
https://hal.science/hal-01710458

Submitted on 16 Feb 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

MICROBIAL GENOMICS

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$], Cyrille Goarant[1$]

[1] Institut Pasteur in New Caledonia, Institut Pasteur International Network, Leptospirosis Research and Expertise Unit, 11 avenue Paul Doumer, 98800 Noumea, New Caledonia

[2] Institut Pasteur Montevideo, Institut Pasteur International Network, Bioinformatics Unit, Montevideo, Uruguay

[3] Institut Pasteur, Biology of Spirochetes Unit, National Reference Centre and WHO Collaborating Center for Leptospirosis, 28 Rue du Docteur Roux, 75724 Paris Cedex 15, France

[#] Contributed equally to this work

[$] Contributed equally to this work.

ABSTRACT

Despite recent advances in our understanding of *Leptospira* genomics, little is known on how virulence has emerged in this heterogeneous bacterial genus as well as on the lifestyle of pathogenic *Leptospira* outside animal hosts. Here, we isolated 12 novel *Leptospira* species from tropical soils, significantly increasing the number of known species to 35 and evidencing a highly unexplored biodiversity in the genus. Extended comparative phylogenomics and pan-genome analyses at the genus level by incorporating 26 new 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 genus *Leptospira*, but also lay the foundations for refining the classification of the pathogenic species.

DATA SUMMARY

1. *Leptospira* isolates are described in Supplementary Table 1 and locations of isolation is shown on Supplementary Figure 1.
2. Genomes have been deposited in GenBank; accession numbers are given in Supplementary Table 2.
3. Overall Genomic relatedness indices are presented in Supplementary Tables 3 (ANI) and 4 (AAI).
4. Protein domain abundance matrix is listed in Supplementary Table 5 (xls file).

IMPACT STATEMENT

Water-associated exposures are the main risk factors of leptospirosis, a complex disease with a multitude of infecting serovars, broad reservoir host range, nonspecific clinical manifestations and difficult diagnosis. To assess the diversity of environmental *Leptospira*, we isolated and sequenced *Leptospira* from hot spots of leptospirosis. General analysis of these genomes provided unprecedented insight into the diversity of *Leptospira*. We described a total of 12 new 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 evidenced clear differences within this pathogen cluster between virulent species and others failing to cause infection. This sheds new light into evolution and acquisition of virulence in this highly heterogeneous genus.

INTRODUCTION

Pathogenic *Leptospira* spp. 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 economic cost associated with leptospirosis is significant and comparable to that of other important neglected tropical diseases, including schistosomiasis, leishmaniasis or 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 as 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 of *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 16S rRNA gene but later on whole-genome sequences, showed that the genus is separated in 3 monophyletic clusters named “saprophytes”, “intermediates” and“pathogens”. The “saprophytes” are environmental species which are rapidly cleared 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], questioning the ecological coherence of the current classification.

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-virulent strains with remarkable genomic distinctions. Our results provide new insight on the virulence evolution in the genus *Leptospira* and suggest 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 qPCR and notified to the New Caledonian Health Authority, which also investigates cases through interviews. Oral consent was asked by the Health Authority to meet with the patient, visit and collect environmental samples in the suspected infection sites. The detailed procedure was described earlier [8].

Study Sites

Six sites were chosen based on the good acceptance of the project by the patients and custom chiefdom (Koné, Touho (2 sites), Ponerihouen (2 sites) and Yaté). All sites were within Melanesian tribes and 3 (Koné and Touho) were also included in a former study[8]. These sites are indicated on

Supplementary Figure 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).

Collection and processing of environmental samples on site

Environmental investigations were started a few weeks after the human presumed infection dates and after recovery of the patients, between March and June 2016. The soils selected to attempt *Leptospira* culture and isolation were chosen following discussions on site with the patients, based on environmental exposure of the patient the day of probable contamination. The samples mostly included river soils, but also moist soils at distance of any waterway if suggested by patient interviews (muddy walking tracks, agricultural soils). Most samples in the study sites were collected less than 20 meters one from another; 27 soil samples were used to isolate *Leptospira*. Samples were collected and processed on site as follow: approximately 5 g topsoil was collected from riverbanks (from 10 cm below to 1 meter 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 hours of collection and vigorously shaken with 5-10 mL sterile water. The soil particles were allowed to settle for 5-15 minutes and 2 mL of supernatant were filtered through a sterile 0.45 µm filter into a tube filled with 2.5 mL of EMJH 2x. 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 10x concentrated STAFF, a combination of selective agents for isolation of *Leptospira* made of sulfamethoxazole, trimethoprim, amphotericin B, fosfomycin, and 5-fluorouracil [9]. Culture tubes from the field were transported within 12 hours 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 *Leptospira* colonies. One to five characteristic subsurface individual colonies were collected from each plate for confirmation of a *Leptospira* morphology by dark field microscopy before clonal subculture in liquid EMJH (Supplementary Table S1).

Whole-genome sequencing

Genomic DNA was prepared by collection of cells by centrifugation from an exponential-phase culture and extraction with 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 systems (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 Supplementary 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 by 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 membership 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 (*Leptonema illini* DSM 21528 was included as 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 genome as 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 Spirochaetes phylum (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 "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 is scanned against the Pfam database [22] using HMMER3 v3.1b2 hmmsearch [20] and its domain architecture is recorded (presence and order). A primary set of orthologous clusters is generated by grouping genes sharing exactly the same domain architecture. Then, remaining genes without hits against the Pfam database are compared to each other at protein level using

HMMER3 v3.1b2 phmmer and clustered using the MCL algorithm [23]. These coarse clusters are then splitted using a tree-prunning algorithm which allows to discriminate between orthologous and paralogues genes. Functional category assignments to each orthologous cluster was performed with BLAST+ blastp against the Clusters of Ortholog Groups (COGs) database [24]. The Jaccard distance over accessory gene patterns was calculated with 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 package Vegan [27]. Protein domains were extracted by comparing each genome annotations 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 highly contributing genes to the observed clustering we used the PCA loadings and adjusted them to a normal distribution. Then those genes with loadings departing more than 2 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 $2x10^8$ leptospires in pure culture. Similar infections were performed with hamsters and mice, which were similarly, infected with *L. interrogans* serovar Manilae strain L495 ($2x10^6$ / animal) and *L. borgpetersenii* serogroup Ballum strain B3-13S ($2x10^8$ / 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 *Leptospira* strains from tropical soils from six sites in New Caledonia, where the disease is endemic (Supplementary Fig. S1 and Supplementary Table S1).

Whole genome sequences of all 26 isolates were determined (genome statistics are presented in Supplementary Table S2). To primary assign the novel species into the traditional leptospiral phylogenetic clusters we built a full-length 16S rRNA gene phylogeny (Supplementary 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% of 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 *Leptospira* species (see Methods section and Supplementary Tables S3 and S4). Figure 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 *Leptospira* species, thus extending by 55% the number of species within this genus.

As the 16S rRNA gene sequence conservation prevents to precisely separate some well-defined *Leptospira* species [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 section). 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 out of them belonged to 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 (Supplementary Fig. S3).

Accessory gene patterns recapitulate virulence potential

The description of novel species belonging 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]. Figure 2a shows the infection profiles of one representative isolate per novel species belonging to pathogens and intermediates, in comparison to the virulent references *L. 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 suggest the inability of these novel species to establish acute infection nor renal colonization in these animal models. This is in marked contrast to the behavior of virulent pathogens like *L. interrogans* and *L. borgpetersenii*, suggesting the hereinafter denomination of these novel species as "low-virulent pathogens".

These results led us to conduct a detailed comparative analysis of the accessory genomes of virulent pathogens, low-virulent pathogens and intermediates. We first noted that after adding the genomes from novel species belonging to these groups the pan-genome remained open (Supplementary Fig. S4), remarking the divergent and highly diverse attribute 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-virulent pathogens (Fig. 2b). More interestingly, the accessory gene patterns were informative enough to discriminate two clusters that correlate with the subdivision of pathogens in virulent pathogens and low-virulent pathogens, in agreement with the virulence experiments. It is worth mentioning that *L. kmetyi* belongs to the accessory genome cluster containing the low-virulent 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-virulent pathogens that are not evident from the core genome phylogeny, which shows that low-virulent 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-virulent 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-virulent pathogens and intermediates). Figure 3a shows the number of discriminatory genes for each group (n=409),

representing ~1.5% of the accessory genes occurring in intermediates, low-virulent and virulent pathogens. Among these, 18 genes were found to distinguish virulent pathogens from both low-virulent pathogens and from intermediates (Supplementary Table S6). Figure 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. Figure 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 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 of virulent pathogens in comparison to both low-virulent pathogens and intermediates.

To have a more complete description of group-specific molecular functions associated to 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]. Figure 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 to generate the observed clustering (see Methods section), 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 6 domains whose abundance was high in virulent pathogens while almost null in low-virulent 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 are distinguished by an increased equipment of repeat sequence elements in comparison with the low-virulent pathogens and intermediates, suggesting a functional link to virulence. Other Pfam domains allowing this discrimination are presented in Supplementary Table S5.

Given the importance of repeat sequence elements in ecological adaptation of organisms by shaping their genomes [36], a focus was made on the abundance of paralogous genes in the accessory genomes. 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 higher incidence of paralogy in virulent pathogens in comparison with low-virulent 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 (Supplementary Fig. S5). In summary, these results indicate that paralogy has played an important role in the emergence of virulent pathogens.

DISCUSSION

Pathogenic *Leptospira* conform 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 in 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 new uncultured *Leptospira* species has been detected from unrelated sources such as bats [37-41] and the Amazonian soils [42]. The impressive

diversity found in our study suggests 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 evidenced in mouse urine or kidney, suggesting their inability either to infect mice or to colonize kidney tubules, also questioning a need of a mammal reservoir in their biology.. Moreover, only *L. interrogans* and *L. borgpetersenii* have been detected in clinical cases in New Caledonia through an active surveillance system [43, 44]. Together, this suggests that the novel species have no or very limited virulence potential to mammals.

From an evolutionary perspective, genomes of these low-virulent 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 living in the soils. This also suggests 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 exhort 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 some former studies have proposed that virulence may be variable among different species belonging to pathogens [5, 6, 45], our more comprehensive taxonomic coverage combined with infection experiments and accessory genome analyses demonstrated the presence of two groups of *Leptospira* species within the pathogens, correlated with clearly distinctive virulence potentials.

Taken together, our results suggest 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 works in the dissection of the molecular mechanisms and biological role of these genes.

AUTHOR STATEMENTS

Funding information:

The current researcher position of RT is supported by an AXA Research Funds grant "AXA Postdoctoral Fellowship", 15-AXA-PDOC-037. IF is supported by the Agencia Nacional de Investigación e Innovación posgraduation bursaries program from Uruguay grant POS_NAC_2016_1_131079, and GI by the Fondo de Convergencia Estructural del Mercosur (FOCEM) grant COF 04/11. EB 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.

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 asked 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.

Conflicts of interest

The authors declare no conflict of interest.

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 neglected tropical diseases* 2015;9(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 neglected tropical diseases* 2015;9(10):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. *International journal of systematic and evolutionary microbiology* 2009;59(Pt 4):705-708.
4. Murray GL, Morel V, Cerqueira GM, Croda J, Srikram A et al. Genome-wide transposon mutagenesis in pathogenic Leptospira spp. *Infect Immun* 2009;77(2):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 neglected tropical diseases* 2016;10(2):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. *Scientific reports* 2016;6:20020.
7. Picardeau M. Virulence of the zoonotic agent of leptospirosis: still terra incognita? *Nature reviews* 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 neglected tropical diseases* 2017;11(2):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. *Microbiology and immunology* 2011;55(7):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(5):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. *Microbial genomics* 2016;2(8):e000083.
12. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics (Oxford, England)* 2014;30(14):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 research* 2017;45(D1):D535-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. *International journal of systematic and evolutionary microbiology* in press.
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, Perez R, Naya H et al. Campylobacter geochelonis sp. nov. isolated from the western Hermann's tortoise (Testudo hermanni hermanni). *International journal of systematic and evolutionary microbiology* 2016;66(9):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 research* 2002;30(14):3059-3066.
18. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS ONE* 2010;5(3):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 research* 2012;40(Database issue):D284-289.
20. Eddy SR. Accelerated Profile HMM Searches. *PLoS computational biology* 2011;7(10):e1002195.
21. Popescu AA, Huber KT, Paradis E. ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. *Bioinformatics (Oxford, England)* 2012;28(11):1536-1537.
22. Bateman A, Coin L, Durbin R, Finn RD, Hollich V et al. The Pfam protein families database. *Nucleic acids research* 2004;32(Database issue):D138-141.
23. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research* 2002;30(7):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 research* 2015;43(Database issue):D261-269.
25. Dray S, Dufour A-B. The ade4 package: implementing the duality diagram for ecologists. *Journal of statistical software* 2007;22(4):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. *Microbial genomics* 2015;1(1):e000025.
27. Oksanen J, Blanchet F, Kindt R, Legendre P, Minchin P et al. 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 (Oxford, England)* 2011;27(21):3070-3071.
29. R Core Team. R: A language and environment for statistical computing. *R foundation for Statistical Computing* 2014.

30. Merien 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(9):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 colonise the carrier host or cause disease in the acute model of infection. *J Med Microbiol* 2013;62(Pt 10):1601-1608.
33. Eshghi A, Becam J, Lambert A, Sismeiro O, Dillies M-A et al. A putative regulatory genetic locus modulates virulence in the pathogen Leptospira interrogans. *Infect Immun* 2014;82(6):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 not required for the formation of flagella sheath. *Infect Immun* 2012;80(6):2019-2025.
35. Ricaldi JN, Matthias MA, Vinetz JM, Lewis AL. Expression of sialic acids and other nonulosonic acids in Leptospira. *BMC microbiology* 2012;12(1):161.
36. Eme L, Doolittle WF. Microbial Evolution: Xenology (Apparently) Trumps Paralogy. *Curr Biol* 2016;26(22):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. *Environmental microbiology* 2015;17(11):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. *Molecular ecology* 2014;23(11):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 microbiology ecology* 2016;92(4):fiw037.
40. Matthias MA, Diaz 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. *The American journal of tropical medicine and hygiene* 2005;73(5):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(8):e308.
43. Salaün L, Merien 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(11):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(8):926-929.
45. Lehmann JS, Matthias MA, Vinetz JM, Fouts DE. Leptospiral Pathogenomics. *Pathogens*, Review 2014;3(2):280-308.

FIGURES

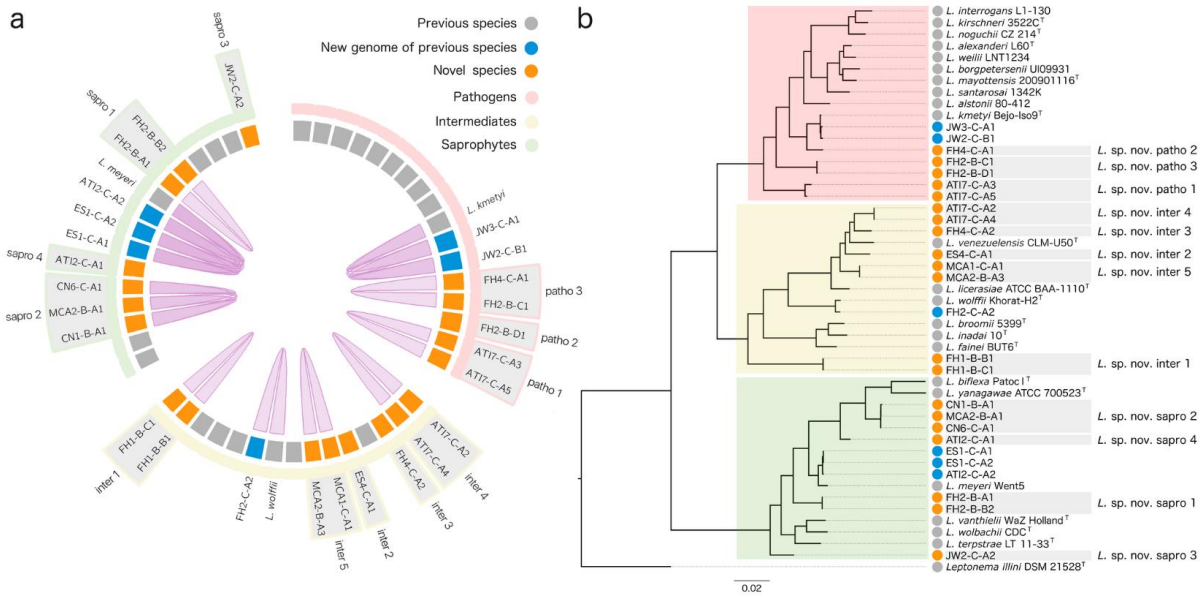

Figure 1. Phylogenetic position of the novel species. a) Circos diagram showing the relationships between leptospiral genomes based on Overall Genomic Relatedness Indices (OGRIs). The inner violet ribbons connect pairs of genomes if they share >95% of Average Nucleotide Identity (ANI) and Average Amino acid Identity (AAI). Blocks represent each genome coloured as explained above. 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. 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 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).

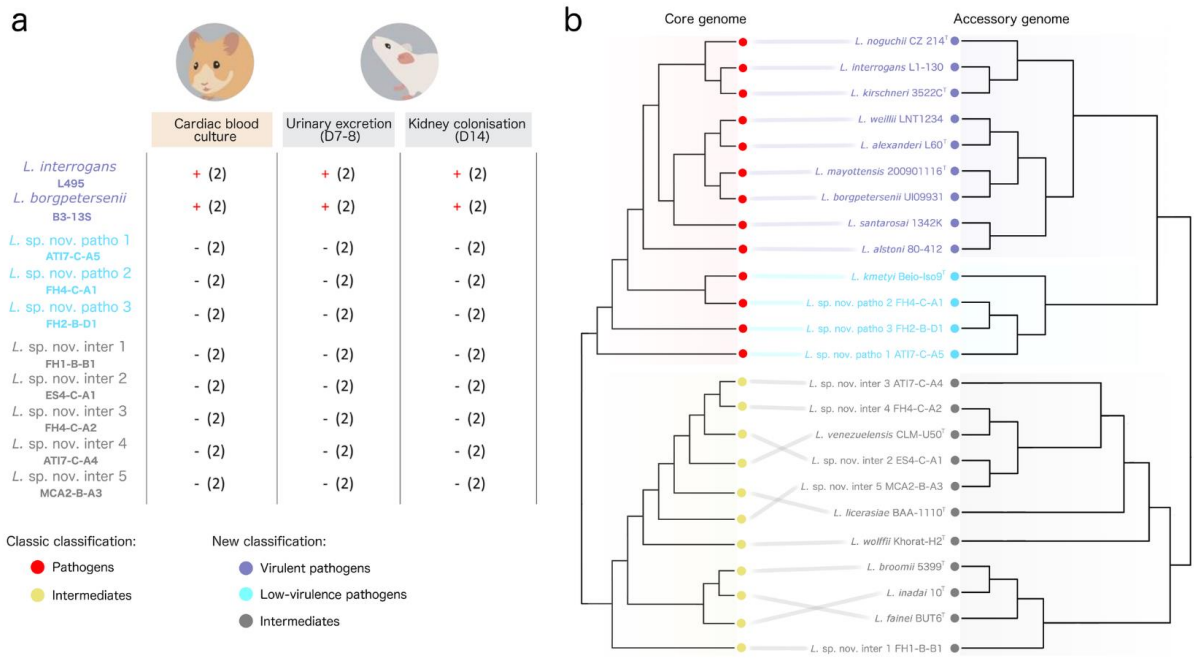

Figure 2:.Virulence in animal models and accessory genome topology. (A) Virulence of novel species in experimental challenge infections. In hamster model, only pathogenic strains *L. interrogans* L495 and *L. borgpeterseni* B3-13S were recover from hamster cardiac blood. (B) Tanglegram comparing the 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). In the left, genomes are coloured according to the classic phylogenetic classification (only pathogens and intermediates are shown here). In the right, genomes are coloured according to the new classification based on accessory gene patterns.

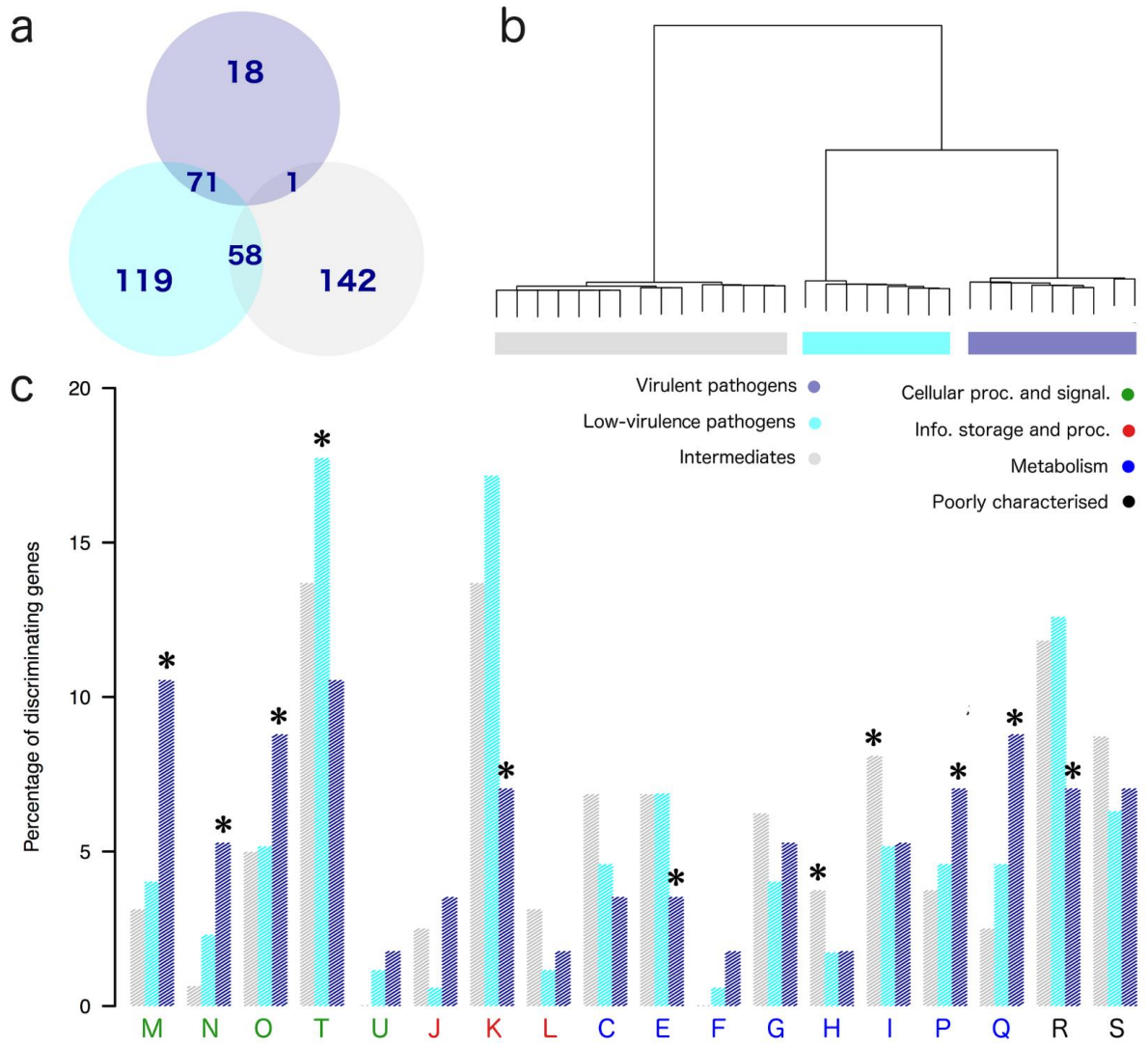

Figure 3. Functional analysis of discriminating accessory genes. a) Venn diagram showing the Bayesian identification of cluster-defining accessory genes from the pan-genome. b) Clustering analysis based on Jaccard distances calculated from the presence/absence vectors of cluster-defining genes. c) Barplots 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.

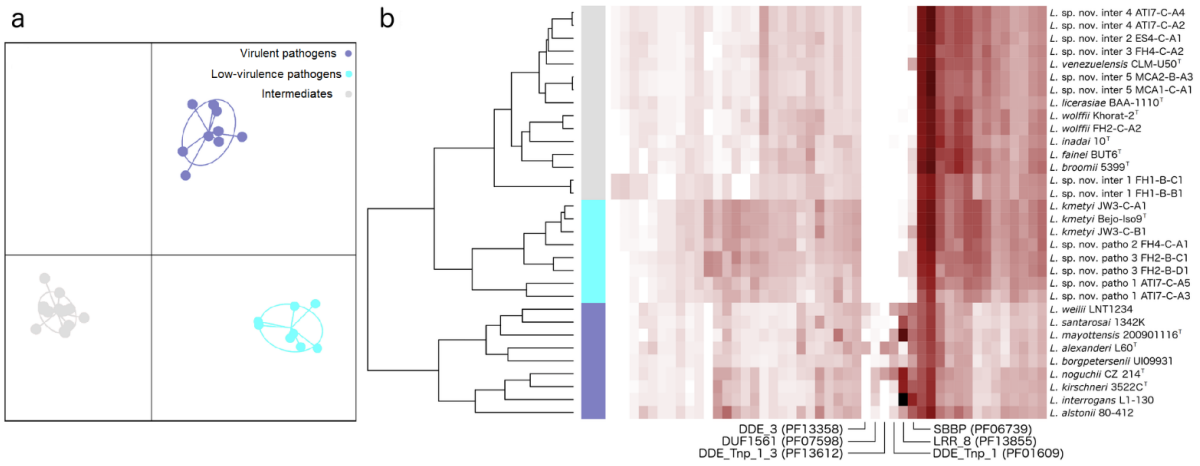

Figure 4. Protein domains analysis. a) Scatterplot showing the first and second discriminant functions obtained from the Discriminant Analysis of Principal Components (DAPC), performed with protein domain abundances extracted from the coding sequences of each genome. Groups are coloured according to the new classification: intermediates (grey), low-virulent pathogens (cyan) and virulent pathogens (purple). b) Heatmap showing the relationships between genomes 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.

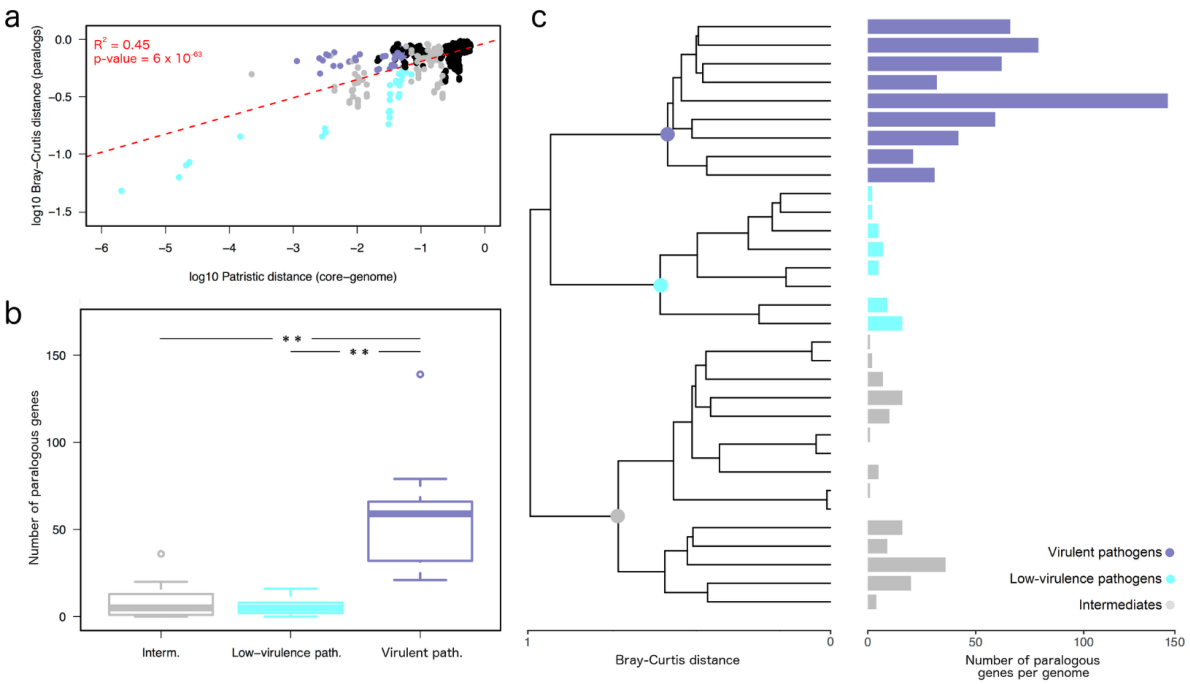

Figure 5. Analysis of paralogous genes. a) Linear regression showing the correlation between patristic distances calculated from the core genome phylogeny and Bray-Curtis distances calculated from the abundance patterns of paralogous genes. Dots are coloured according to virulence clusters when both 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. Asterisk indicates $p < 0.001$ (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.

SUPPLEMENTARY MATERIAL

available online at

http://mgen.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000144#tab5

Supplementary Figure S1: Location of the soil collection areas in New Caledonia (circles) and meteorological data from closest station (red diamonds). The shoreline is black, blue lines are coral reefs. Temperature (mini / maxi) and rainfall data are the normal data calculated as the 30-year averaged data retrieved from the Meteo France public website. Black circles correspond to hotspots of human leptospirosis. The grey circle points to an area of lower incidence.

Supplementary Figure S2: Phylogenetic analysis using the 16S rRNA gene. Phylogenetic tree built with 16S reference sequences from described leptospiral species (black labels) and from the strains described in this work (blue labels). Color shades highlight the three *Leptospira* clades: pathogens (red), intermediates (yellow) and saprophytes (green). Bootstrap values are indicated for relevant nodes. For reference sequences, accession numbers are within brackets. The tree was rooted with *Leptonema ilinii*.

Supplementary Figure S3: (A) A radial view of the core genome topology showing novel species in all three *Leptospira* clusters. (B) Statistical evidence of the basal position of low-virulent species within the cluster “pathogens”.

Supplementary Figure S4: Pan-genome size estimation. Gene abundance cumulative curves showing the estimated pan-genome size for a set of 240 public leptospiral genomes (blue) and for these genomes plus those sequenced in the present study (red). For the public dataset we estimated the pan-genome in 85,645 orthologous groups while when adding the genomes from novel species it increased to 89,868 orthologous groups. Dots show estimations for 10 random samples and solid lines represent the mean.

Supplementary Figure S5: Comparative abundance of paralogous clusters and transposase domains in the genomes of *Leptospira* from the intermediate cluster and two sub-clusters of pathogens.

Supplementary Table S1: Summary of information on samples, isolation process and identification of isolates.

Supplementary Table S2: Genome summary statistics for the novel strains sequenced in the present study.

Supplementary Table S3: Average Nucleotide Identity (ANI) values for the novel genomes against each other and previously identified species.

Supplementary Table S4: Average Amino acid Identity (AAI) values for the novel genomes against each other and previously identified species.

Supplementary Table S5: Protein domain abundance patterns for each genome belonging to intermediates and pathogens groups.

Supplementary Table S6: List of 18 genes found to discriminate virulent pathogens from both low-virulent pathogens and intermediates.