Universidad de la República
PEDECIBA - Bioinformática

Tesis de Maestría

Desarrollo, implementación y optimización de herramientas de genómica comparativa para el género *Leptospira*

Ignacio Ferrés

Tutor
Dr. Gregorio Iraola

Co-tutor
Dr. Hugo Naya

18 de marzo de 2019

(...) welcome to the machine.
Where have you been?
It's alright we know where you've been.
You've been in the pipeline, filling in time.

— ***Roger Waters***

Agradecimientos

A la ANII, por apoyar mi investigación.

Al tribunal, José, Laura, y Alejandro, por sus valiosos aportes.

A Hugo por permitirme realizar mis estudios de posgrado en la Unidad de Bioinformática.

A los compañeros del Laboratorio de Genómica Microbiana, especialmente a Gregorio por guiarme en todo este proceso, valorar en todo momento mi esfuerzo, y permitirme ser parte de su equipo.

A toda la Unidad de Bioinformática, en general, por el apoyo recibido siempre, el inmejorable y siempre cálido ambiente laboral, y las instancias de formación recibidas.

A Cecilia, Leticia y Alejandro, por permitirme investigar con datos que iba generando su laboratorio.

A los amigos, de la facultad y del liceo, que siempre estuvieron.

A familia, siempre atrás.

Gracias.

Resumen

La leptospirosis es una enfermedad zoonónica con alta prevalencia en países tropicales de bajos ingresos provocada por bacterias del género *Leptospira*. Gracias a los avances en secuenciación, en lo últimos años las bases de datos genómicos han crecido exponencialmente, y con ellas el número de genomas secuenciados de cepas del género, lo cual ha permitido un entendimiento más profundo de este grupo de bacterias. La generación de nuevas herramientas que permitan analizar la diversidad existente resulta fundamental en la era genómica, en especial si se quiere entender la dinámica poblacional, la evolución, y los factores genéticos que determinan la patogenicidad de organismos de interés sanitario como *Leptospira*. En este contexto, el objetivo de este trabajo fue desarrollar herramientas que faciliten los estudios en genómica comparativa de *Leptospira* y otras bacterias de relevancia. En esta tesis se describen tres paquetes bioinformáticos desarrollados en lenguaje R que, utilizando el caso de *Leptospira* como hilo conductor, pretenden ser un aporte al conjunto de herramientas que utiliza la comunidad científica para estudiar la genómica comparativa de bacterias relevantes para la salud humana y animal. En primer lugar se presenta `MLSTar`, un paquete que permite tipificar genomas microbianos utilizando la base de datos PubMLST, y que fue utilizado para describir la diversidad de aislamientos de *Leptospira* en Uruguay. En segundo lugar se presenta `Phylen`, el cual utiliza la base de datos EggNOG para identificar grupos de ortólogos del genoma núcleo y con ellos realizar una reconstrucción filogenética, y se incluyen también casos de estudio donde `Phylen` fue aplicado exitosamente en la descripción de especies nuevas del género *Leptospira*. Por último se introduce `Pewit`, una metodología novedosa que permite reconstruir pangenomas bacterianos de linajes divergentes, y se muestran ejemplos de aplicación en *Leptospira*. Los dos primeros paquetes fueron publicados, y el manuscrito del tercero se encuentra en fase de escritura. En resumen, las tres herramientas facilitan el análisis integrativo de genomas bacterianos y aportan nuevas estrategias para el estudio de la epidemiología, la filogenia y la evolución de grupos microbianos, todo implementado para el entorno R.

Abstract

Leptospirosis is a zoonotic disease with high prevalence in tropical low-income countries, caused by bacteria that belong to the *Leptospira* genus. Advances in Next Generation Sequencing promoted the growth of genomic databases, and with them the number of sequenced genomes of the genus, which has prompted major advances in understanding this bacterial clade. The development of new tools which allow to analyze the existing diversity have become key in the genomic era, particularly if population dynamics, evolution, and genetic determinants of virulence want to be understood in the case organisms of interest for public and animal health. The main goal of this thesis was to develop tools that facilitate comparative genomics studies in *Leptospira* and other relevant bacterial species. More specifically, this dissertation presents three bioinformatics packages written in R which contribute to performing comparative genomics of pathogenic bacteria. `MLSTar`, a software that enables microbial genome typing using the PubMLST database, and that was used to describe *Leptospira* diversity in Uruguay. `Phylen` is described, a package that uses the EggNOG database to identify core genes in a given dataset of genomes, and with those core genes infer a phylogeny; cases where phylen was successfully applied in describing new *Leptospira* species are also included. `Pewit` is introduced, a novel approach to reconstruct highly divergent bacterial pangenomes; applications of this pipeline using *Leptospira* datasets are also shown. The former two packages were published in peer-reviewed journals, and in the case of the latter a manuscript is being written. In short, the three software packages allow integrative analysis of bacterial genomes, and contribute with novel strategies for studying epidemiology, phylogenetics, and evolution of microbial groups, all three implemented for the R environment.

Índice general

Índice de figuras

Índice de cuadros

Lista de acrónimos

EMJH Ellinghausen-McCullough-Johnson-Harris. 12

LPS lipopolisacárido. 12, 14

CAAT ensayo de adsorción de agutinina cruzada. 12

ADN Ácido desoxirribonucleico. 13, 21

NGS tecnologías de secuenciación de nueva generación. 13, 22

ARNr ARN ribosomal. 14, 22, 41

BSR *Blast Score Ratio*. 14

MLST genotipado de secuencias multilocus. 14, 19, 20, 28, 40

TCS sistemas de dos componentes. 17, 18

HGT transferencia horizontal de genes. 17

PF familia proteica. 17

MRCA ancestro común más reciente. 17, 18

LRR_8 repetido rico en leucina 8. 18

ST *sequence-type*. 19

MLEE genotipado multilocus enzimático por electroforesis. 19

UPGMA *Unweighted Pair Group Method using Arithmetic Averages*. 20, 21

NJ *Neighbor-Joining*. 20, 21

ML *Maximum Likelihood*. 20, 21

HMM Modelos Ocultos de Markov. 25, 26, 41

INIA Instituto Nacional de Investigación Agropecuaria. 28

UMPI Unidad Mixta INIA-Pasteur. 28

cgMLST genotipado de secuencias multilocus del genoma núcleo. 40, 46

Glosario

pangenoma Del inglés: *pangenome* ('pan' - $\pi\alpha\nu$: 'todo' en griego). Repertorio global de genes de una especie bacteriana. 17, 22–26, 47–49

pangenoma abierto La secuenciación de nuevos genomas aporta nuevos genes a una tasa promedialmente constante. 17, 22–24

genoma núcleo Del inglés: *coregenome* (*core*: núcleo). Regiones genómicas compartidas por todas las cepas de determinada especie. 17, 22–24, 41

homología Existencia de un ancestro común en relación a dos o más secuencias.. 17, 25, 26, 47

genoma accesorio Regiones genómicas presentes en algunas, pero no en todas, las cepas de determinada especie. 22, 47

pangenoma cerrado El número de nuevos genes con cada nueva secuenciación tiende a cero. 22–24

fluidez Del inglés: *fluidity.* Promedio del ratio de familias génicas únicas entre la suma de familias génicas en pares de genomas seleccionados al azar de los N genomas que conforman el pangenoma. 24

ortólogos Dos secuencias son ortólogas so derivan de un solo gen presente en el último ancestro común de ambos organismos. La historia evolutiva de dos secuencias ortólogas reflejan la historia evolutiva de los organismos que la contienen. 24, 25, 41, 47, 48

parálogos Dos secuencias son parálogas si se crearon de un evento de duplicación, evolucionando en paralelo. 24, 25, 46

parálogos internos Del inglés: *in-paralogs.* Dos secuencias son parálogas internas si la duplicación se produjo después de la especiación. 24

parálogos externos Del inglés: *out-paralogs.* Dos secuencias son parálogas externas si la duplicación se produjo previo a la especiación. 24

Capítulo 1

Introducción

1.1. Leptospirosis y *Leptospira*

Esta sección pretende realizar una concisa reseña sobre la leptospirosis, dar una descripción muy general del organismo, y presentar el estado del arte en cuanto a la genómica comparada y taxonomía del género al momento que comencé a realizar el posgrado en la Unidad de Bioinformática del Institut Pasteur de Montevideo, en 2015.

1.1.1. Leptospirosis, enfermedad emergente

La leptospirosis es una enfermedad zoonótica emergente con alta prevalencia en regiones tropicales de bajos ingresos que se estima que afecta a 1 millón de personas y su tasa de mortandad podría alcanzar el 6 % anualmente [1]. Se presenta como una enfermedad sistémica que afecta a animales domésticos como perros, ganado y suinos, además del humano[2].

Las infecciones, tanto en humanos como en otros animales, ocurren del contracto directo con orina o indirectamente con agua contaminada. Las bacterias entran al hospedero a través de heridas en la piel o a través de las mucosas, y se establecen en el hígado y riñones [3, 4]. Algunas especies de animales actúan como portadoras asintomáticas albergando leptospiras virulentas en los túbulos renales por períodos de tiempo largos, esparciendo bacterias infecciosas al ambiente [5]. Otras especies de *Leptospira* son incluso oportunistas de vida libre, pudiendo transmitirse a través del contacto con agua o suelo contaminado.

Los signos clínicos de la enfermedad son muy variables, dependiendo tanto del serovar de *Leptospira* así como de la especie infectada. En perros la clínica es típicamente aguda, pudiéndose presentar con fiebre, ictericia, vómitos, diarrea, coagulación intravascular diseminada, falla renal, hemorragias y muerte [6]. En ganado y cerdos los signos incluyen falla reproductiva, abortos, momificación fetal, debilidad en las crías, y agalactia. Infecciones crónicas en equinos se manifiestan como uveítis recurrentes [7]. Varios roedores como la rata y el ratón usualmente no exhiben la enfermedad pero muestran colonización renal crónica y excretan la bacteria en la orina, sirviendo como reservorio y parte del ciclo vital del huésped.

Las vacunas disponibles se basan en células enteras inactivadas o preparaciones de leptospiras patogénicas, las cuales confieren respuestas protectivas a través de la inducción de anticuerpos neutralizantes, mayoritariamente contra el lipopolisacárido de membrana. Estas formulaciones, sin embargo, no generan protección de largo plazo (linfocitos B de memoria) ni proveen protección cruzada contra leptospiras de otros serovares que los que estén presentes en la preparación.

1.1.2. *Leptospira*, breve descripción

Las leptospiras son bacterias aerobias o microaerofílicas obligadas, pertenecientes al phylum *Spirochaetes*, con una temperatura de crecimiento óptima de 28-30°C. Entre la variedad de medios desarrollados para cultivar la bacteria el más comúnmente utilizado es el medio Ellinghausen-McCullough-Johnson-Harris (EMJH)[8, 9]. En general el crecimiento es lento pudiendo tardar hasta 13 semanas si se trata de un aislamiento primario, aunque una vez obtenidos los sub cultivos puros en medio líquido pueden crecer en sólo 10-14 días [5].

Miden 0,1 μm de diámetro y entre 6-20 μm de largo y poseen una forma distintiva alargada en espiral con extremos en forma de gancho [10-12]. Dos flagelos periplásmicos con inserciones en los polos son los responsables de la motilidad. Tienen una doble membrana típica en la cual la membrana citoplasmática y la pared de peptidoglicano están estrechamente asociadas, algo único en las espiroquetas, y envueltas por una membrana externa [13]. Poseen proteínas de adhesión a componentes de la matriz extracelular y del plasma, lo cual es fundamental para la colonización e infección, algunas de las cuales podrían contribuir a la evasión de la respuesta inmune.

Una caraterística que distingue a las leptospiras de otros géneros del phylum *Spirochaetes* es la expresión de lipopolisacárido (LPS) en su superficie, el cual juega un rol clave en la virulencia [14-16]. A su vez, el LPS se expresa en un mosaico de antígenos que permiten la clasificación serológica en 29 serogrupos y en más de 300 serovares [5, 17] basada en ensayo de adsorción de agutinina cruzada (CAAT). Esta variabilidad fenotípica también se ve expresada en la capacidad de causar enfermedad en las distintas especies del género, ya que existen leptospiras saprofíticas y patogénicas, siendo el último grupo capaz de infectar un amplio rango de huéspedes y sobrevivir tanto en ambientes marinos como en ambientes terrestres, aunque en algunas especies la capacidad de vivir en el ambiente se perdió por completo siendo patógenas obligadas.

1.1.3. Taxonomía y genómica comparativa del género *Leptospira*

Si bien los reportes de enfermedades cuyos síntomas sugieren que se trataba de leptospirosis datan de varios siglos atrás, tanto en China como en Japón y Europa, la historia moderna de la enferemedad comienza con Adolph Weil en 1886 que describe un tipo de ictericia acompañada de esplenomegaglia, falla renal, conjuntivitis, y erupciones en la piel, que posteriormente se la conocería como enfermedad de Weil [18]. La primera identificación de la espiroqueta la da Stimson en 1907, quien observa la bacteria en tejido renal de un paciente al que se le había diagnosticado fiebre amarilla, y nombra este organismo como *Spirocheta interrogans* dada la similitud que encuentra entre la forma de la bacteria y el signo de interrogación [19]. Años más tarde, en Japón, Inada y colaboradores (1916) logran aislar por primera vez la bacteria y son capaces de inducir la enfermedad en cobayos inyectándoles sangre de pacientes enfermos de la enfermedad de Weil, generando un modelo de estudio de la misma. El mismo grupo también demuestra que lisados de leptospira proveen protección inmunológica contra infecciones, proponiendo la primer vacuna contra la enfermedad [20]. Un año más tarde, Ido y colaboradores (1917) identifican a ratas salvajes como portadoras asintomáticas de estas bacterias, y las denominan *Spirocheta icterohaemorrhagiae* basados en la enfermedad que provocaba en trabajadores de minas de carbón [21]. Al siguiente año, Noguchi (1918) propone el nombre *Leptospira* para distinguir al género de espiroquetas que provocaba la enfermedad de Weil de otras conocidas hasta el momento [22]. En la medida que nuevos serovares eran aislados, se los clasificaba como nuevas especies (como por ejemplo *Leptospira pomona*, *Leptospira hardjo*, o *Leptospira copenhageni*). En 1982 el Subcomité sobre la Taxonomía de *Leptospira* del Comité Internacional de Bacteriología Sistemática corrige la taxonomía del género determinando la existencia de 2 especies dentro del mismo: *L. interrogans* y *L. biflexa*, para designar a los organismos con capacidad patogénica y a aquellos saprófitos, respectivamente, aunque se reconocen 3 subgrupos dentro de cada especie basados en ensayos de hibridación ADN-ADN [23]. Más adelante, en 1987, se subdivide *L. interrogans* en 5 especies y *L. biflexa* en 2 especies, utilizando nuevamente técnicas de hibridación del ADN [24]. Posteriormente, gracias a nuevos aislamientos y mejoras en las técnicas de discriminación utilizando datos de tecnologías de secuenciación de nueva generación (NGS), el Subcomité sobre la Taxonomía de *Leptospiraceae* en 2007 reconoce la existencia de 13 especies patogénicas y 6 especies saprófitas, contabilizando un total de 17 genomoespecies [2]. Al momento de iniciar esta tesis se reportaban un total de 22 especies y una estructura filogenética que comprende 3 subclados principales, como se observa en la figura 1.1, tomada de Lehmann y colaboradores (2014) [25]. Dicha estructura sugiere una evolución desde bacterias saprófitas basales (Grupo III) hacia la patogeneicidad (Grupo I), pasando por un clado intermedio en el que se encuentran bacterias saprófitas de vida libre pero capaces de provocar leptospirosis de forma oportunista (Grupo II), y es apoyada por análisis filogenéticos utilizando datos de 39 proteínas universales concatenadas, genes *core* identificados por *Blast Score Ratio* (BSR), genotipado de secuencias multilocus (MLST), el gen de la subunidad 16S del ARNr, y el gen *secY* [26].

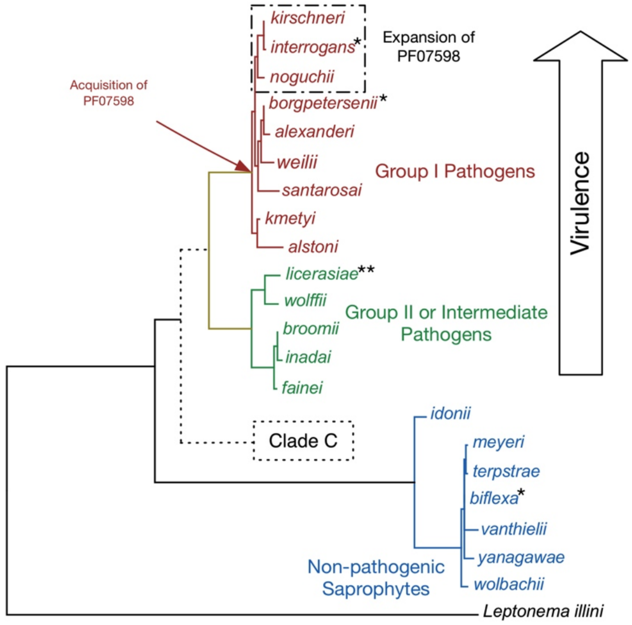


Figura 1.1: Árbol filogenético de *Leptospira* inferido a partir del gen de la subunidad 16S del ARNr. Tomada de Lehmann y colaboradores, 2014. Se muestran los tres clados en colores distintos: clado patogénico (rojo), el clado intermedio (verde) y el clado saprófito (azul). Como grupo externo se muestra *Leptonema illini*.

Como se mencionó anteriormente, los aislamientos del género se clasifican en más de 300 serovares que interesantemente se distribuyen entre las distintas especies en muchos casos de forma no vertical. Dado que el serovar se ha asociado al locus de biosíntesis del LPS (locus *rfb*), la presencia de un mismo serovar en distintas especies sugiere transferencia horizontal de genes, la cual se ha reportado tanto para los genes del locus *rfb* [27] como para los genes codificantes de proteínas de membrana [28].

El tamaño de los genomas varía entre 3,85 Mb y 4,9 Mb, y un contenido GC de entre 35 % y 42 % (Cuadro 1.1). En general presentan al menos 2 replicones [29], de los cuales el mayor de ellos, cI (>3,6Mb), es el cromosoma bacteriano y contiene secuencias que codifican genes *housekeeping*. Un segundo replicón más pequeño, cII (de entre 0.28 Mb y 0.35 Mb), contiene genes esenciales y se encuentra presente en casi todas las cepas. En *L. biflexa* serovar Patoc (serogrupo Semaranga, cepa Ames) se ha reportado la existencia de un tercer replicón (p74, 74Kb) que también contiene genes *core* que se encuentran localizados en el cromosoma cI en otras especies de *Leptospira* [30]. Debido a que los dos últimos replicones, además de contener genes esenciales, tienen un contenido GC % similar al del cromosoma, se los considera "crómidos" (del inglés, *chromids*) [30, 31]. (Nótese que la designación de "crómidos" a estos 2 replicones es tomada de la bibliografía, y puede no coincidir con la información que aparece en el Cuadro 1.1, el cual fue tomado directamente de la base de datos PATRIC[32]. En este cuadro el replicón cII aparece contabilizado como un cromosoma mientras que p74 aparece como plásmido). Además del cromosoma y los crómidos, se han encontrado hasta 3 plásmidos adicionales en una misma cepa (Cuadro 1.1). En cuanto a número de secuencias codificantes de proteínas (CDS), el género fluctúa entre 3.877 y 6.309 (Cuadro 1.1), mostrando una tendencia a la reducción genómica aquellas cepas que presentan una restricción de nicho por ser patógenas obligadas [33].

Cuadro 1.1: Genomas completos en la base de datos PATRICbrc al 14 de Junio de 2018

Genome ID	Genome Name	Chromosomes	Plasmids	Contigs	Length (pb)	GC %	CDS
1141102.3	*L. borgpetersenii* str. 4E	2	0	2	3916189	40.19	4723
1279460.3	*L. interrogans* serovar Hardjo str. Norma	2	0	2	4762150	35.02	4914
1395589.3	*L. interrogans* serovar Linhai str. 56609	2	3	5	4915652	35.05	4886
189518.3	*L. interrogans* serovar Lai str. 56601	2	0	2	4691184	35	4665
211880.11	*L. interrogans* serovar Canicola	2	0	2	4570695	34.88	5109
211880.12	*L. interrogans* serovar Canicola	2	0	2	4560160	34.77	6309
214675.23	*L. interrogans* serovar Manilae strain L495	1	0	1	4614703	35	4700
214675.5	*L. interrogans* serovar Manilae strain UP-MMC-NIID LP	2	1	3	4667405	34.99	4558
214675.6	*L. interrogans* serovar Manilae strain UP-MMC-NIID HP	2	1	3	4667354	34.99	4571
267671.5	*L. interrogans* serovar Copenhageni str. Fiocruz L1-130	2	0	2	4627366	35	4550
280505.15	*L. borgpetersenii* serovar Ballum strain 56604	2	2	4	4037579	40.19	4443
28183.17	*L. santarosai* strain DU92	2	0	2	3858945	41.85	4516
28452.8	*L. alstonii* strain GWTS #1	2	0	2	4591888	42.42	4779
328971.1	*L. borgpetersenii* serovar Hardjo strain BK-9	2	0	2	3949086	39.01	4558
328971.11	*L. borgpetersenii* serovar Hardjo strain NVSL S 1343	2	0	2	3932487	39	4554
328971.7	*L. borgpetersenii* serovar Hardjo strain BK-30	2	0	2	3947069	39.05	4563
328971.8	*L. borgpetersenii* serovar Hardjo strain NVSL S 818	2	0	2	3884697	39.34	4550
328971.9	*L. borgpetersenii* serovar Hardjo strain BK-6	2	0	2	3967801	38.77	4556
338215.3	*L. interrogans* serovar Bratislava strain PigK151	2	0	2	4721584	35.03	4580
355276.3	*L. borgpetersenii* serovar Hardjo-bovis str. L550	2	0	2	3931782	40.2	4262
355277.4	*L. borgpetersenii* serovar Hardjo-bovis str. JB197	2	0	2	3876235	40.2	4197
355278.4	*L. biflexa* serovar Patoc strain 'Patoc 1 (Ames)'	2	1	3	3956089	38.89	3881
38347.3	*L. interrogans* serovar Hardjo-prajitno strain Hardjoprajitno	2	0	2	4692322	34.78	4663
44275.5	*L. interrogans* serovar Copenhageni strain FDAARGOS_203	1	1	2	4630574	35.05	4780
456481.4	*L. biflexa* serovar Patoc strain 'Patoc 1 (Paris)'	2	1	3	3951448	38.9	3877
573825.3	*L. interrogans* serovar Lai str. IPAV	2	0	2	4708530	35.03	4825

Recientemente, dos trabajos aportaron datos sobre la variabilidad del pangenoma del género *Leptospira* [26, 34] (Por definición de “pangenoma” ver glosario, página 10. Más detalles en sección 1.2.3, página 22). Si bien presentaron diferencias metodológicas, las conclusiones son concordantes y complementarias. En ambos casos se determinó que el pangenoma se encuentra abierto; esto significa que no es posible saturar el número observado de genes mediante el muestreo de nuevos genomas (Curva de rarefacción, ver figura 1.2). Esta propiedad es típica de especies que se encuentran colonizando nuevos ambientes [35]. A nivel intraespecífico también se observó un pangenoma abierto para las especies predominantes (*L. interrogans* y *L. borgpetersenii*) [34].

Ambos trabajos difirieron significativamente en los tamaños del genoma núcleo y del pangenoma obtenidos. En el trabajo de Xu y colaboradores (2016) hallaron un pangenoma de 57.765 genes y un genoma núcleo de 1.023 genes; mientras que Fouts y colaboradores (2016) reportaron un pangenoma de 17.477 genes (13.822 genes con parálogos colapsados) y un genoma núcleo de 1.764 genes (1.592 con parálogos colapsados) [26, 34]. Estas diferencias pueden deberse a la selección de cepas sobre las que realizaron el análisis, a la calidad de las anotaciones, aunque el factor más relevante probablemente se deba a diferencias metodológicas utilizadas para determinar grupos de ortólogos ya que, como se verá más adelante, los algoritmos y parámetros utilizados pueden tener un efecto significativo sobre el resultado.

Las especies patogénicas presentaron más genes especie-específicos que las intermedias o saprófitas, lo cual sugiere una adaptación específica a los hospederos. Contienen genes de síntesis de vitamina B12, que permiten sobrevivir en nichos con limitaciones nutritivas como en los mamíferos hospederos, así como también metaloproteasas y genes asociados a la virulencia como la lipoproteína LigB [26].

En el pasaje de saprófitas a patógenas hubo una pérdida neta de genes, entre ellos genes del metabolismo de carbohidratos y energía, y genes codificantes de sistemas de dos componentes (TCS), aunque se ganaron otros genes de este último sistema que presentan homología a *Legionella pneumoniae* y *Yersinia pestis*, dos patógenos humanos, en suma resultando en una pérdida de diversidad global de TCS [34]. Esto puede explicarse como una pérdida de presión selectiva a mantener sistemas de censado de entornos muy variables, y a eventos de transferencia horizontal de genes (HGT).

Los análisis de familias proteicas (PF) revelaron la adquisición de motivos y dominios proteicos en patogénicas, con 204 PF exclusivas de este grupo, incluidas 15 PF clasificadas como peptidasas algunas de las cuales se cree que podrían ser capaces de clivar proteínas del complemento [26, 34].

A su vez se observó la expansión masiva de 139 PFs heredadas del ancestro

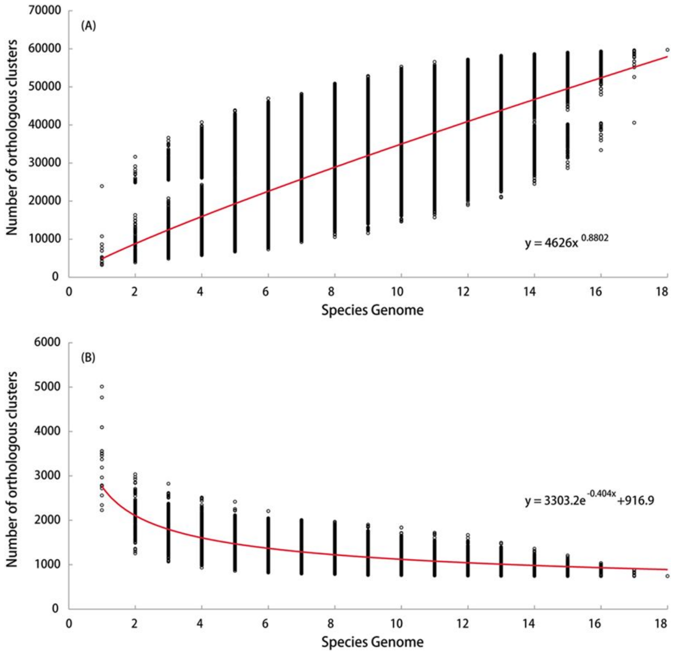


Figura 1.2: Curvas de rarefacción del género *Leptospira*, tomada de Xu y colaboradores (2016). La curva de arriba muestra el número de genes totales del pangenoma, mientras que la de abajo muestra el tamaño del genoma núcleo. Ambas curvas se construyen muestreando un número creciente de genomas al azar y contando el número de genes pertenecientes a cada categoría (pangenoma o genoma núcleo) para cada caso. En el caso de la curva del pangenoma (arriba), se observa que no parece tender hacia una estabilización en la medida que se agregan más genomas, evidenciando un pangenoma abierto. En el caso del genoma núcleo se observa una estabilización cercana a los 1000 genes que corresponderían a aquellos que se encuentran presentes en todos las especies del género.

común más reciente (MRCA) incluido el motivo estructural LRR_8 (repetido rico en leucina 8, PF13855) de 1 en especies saprófitas, a 8 en *L. borgpetersenii* y 21 en *L. interrogans*, resultado de la duplicación génica [34]. Estas expansiones de dominios asociadas a duplicaciones génicas también se detectaron para otros dominios asociados a virulencia, como exo_endo_phos, la familia endonucleasa/exonucleasa/fosfatasa (PF03372), HTH_19, el dominio hélice-giro-hélice (PF12844), y el dominio bacteriano tipo Ig (PF13205) [34].

Las especies intermedias son similares a las patogénicas pero tienen un set diverso de TCS, como en especies saprófitas, un locus *rfb* pequeño, y menos genes asociados a virulencia. En contraste, presentan profagos de la misma forma que las especies patógenas, casi ausentes en saprófitas [26].

1.2. Herramientas bioinformáticas para el estudio de la genómica comparada

A continuación se describen brevemente algunas metodologías que se utilizan para analizar la genómica comparada de bacterias. En particular me centro en aquellos abordajes en los que decidimos investigar e implementar nuevas aproximaciones durante la maestría.

La justificación sobre las herramientas que desarrollamos y su relación con *Leptospira* se encuentran en los capítulos 2 (página 27), 3 (página 40), y 4 (página 46).

1.2.1. Genotipado de secuencias multilocus (*MLST*)

El genotipado de secuencias multilocus o MLST, como se lo conoce por sus siglas en inglés, fue propuesto en 1998 como una herramienta para caracterizar las relaciones intraespecíficas de distintas cepas [36]. Este abordaje utiliza la información combinada de aproximadamente 500 pb de secuencias de ADN de un conjunto de típicamente 7 genes *housekeeping* para resumir la variabilidad génica de todo el genoma [36]. Consiste en indexar cada alelo existente de cada uno de los loci seleccionados asignándole un número arbitrario a cada variante, y de la misma forma asignar un número distinto, el *sequence-type* (ST), a cada posible combinación de alelos observada (perfiles). El genotipado se realiza comparando la combinación de alelos observada en un aislamiento dado con las combinaciones del índice, y asignando correspondientemente el ST del perfil con la misma combinación al aislamiento.

Desde su introducción el MLST ha ofrecido mayor resolución y reproducibilidad al estudio de la dinámica y estructura poblacional de patógenos microbianos respecto a técnicas precedentes como el genotipado multilocus enzimático por electroforesis (MLEE) [37, 38]. Este tipo de análisis, sin embargo, puede presentar la desventaja de que provee una resolución limitada para ciertos patógenos genómicamente muy diversos, y puede ser riesgoso hacer inferencias evolutivas y filogenéticas basadas en un número limitado de secuencias génicas. Estas limitaciones han sido abordadas mediante la mejora de la metodología original, ya sea generando nuevos esquemas con nuevos conjuntos de genes que recuperen mejor la diversidad, aumentando el largo de los fragmentos, o aumentando el número de loci. Este último abordaje en particular ha ganado terreno gracias al avance de las tecnologías de secuenciación permitiendo crear esquemas que tienen en cuenta toda la variablidad genética de aquellos genes compartidos por un conjunto de genomas [39].

La disponibilidad de miles de genomas, con decenas nuevos cada día, ha generado incentivos a la comunidad científica para centralizar y organizar toda esta información de modo de hacerla lo más útil y accesible posible. La base de datos PubMLST ha logrado efectivamente organizar, estandarizar y hacer disponible los esquemas de MLST y sus definiciones alélicas.

1.2.2. Análisis filogenéticos

El análisis filogenético es la forma más común de estudiar las estructuras poblacionales. Las relaciones evolutivas entre los miembros se vuelven aparentes si se infiere correctamente la filogenia [40]. Todos los métodos modernos de análisis filogenéticos se basan en la comparación de secuencias de ADN y, en general, parten de alineamientos génicos.

De más rudimentario a más complejo, los métodos más populares de análisis basados en secuencias son: *Unweighted Pair Group Method using Arithmetic Averages* (UPGMA), *Neighbor-Joining* (NJ), parsimonia, *Maximum Likelihood* (ML), e inferencia bayesiana.

El método UPGMA es el más simple y fue propuesto en 1958 por Sokal y Michener [41]. Es un método basado en distancias, por lo que requiere partir de una matriz de distancias entre cada par de secuencias alineadas. El algoritmo considera a las unidades taxonómicas como *clusters*, luego agrupa a los *clusters* más cercanos basados en la matriz de distancias y actualiza esta matriz estableciendo como distancia el promedio de distancias entre los *clusters* que conforman. Se repite el proceso de modo que en cada paso la matriz va disminuyendo sus dimensiones de a 1 fila y 1 columna, hasta que todos los miembros se clusterizan juntos en un mismo grupo. Cada vez que dos *clusters* son agrupados, su distancia define el tiempo desde su ancestro común [40].

Un algoritmo similar es el NJ, propuesto en 1987 [42]. Utiliza una matriz de distancias entre las secuencias, y en cada paso los nodos más cercanos son definidos como vecinos. Esto es repetido de forma recursiva partiendo desde un árbol en forma de estrella hasta que todos los nodos son agrupados de a pares. Detalles en cuanto a la determinación de distancias entre nodos por el algoritmo pueden ser encontrados en la bibliografía. Este método funciona bien cuando se puede asumir el modelo de sitios infinitos de Kimura (1969), esto es, que no existe saturación de sustituciones. Esto se da cuando existe una gran cantidad de sitios en relación con la tasa de mutación y los tiempos de divergencia entre cepas.

Tanto UPGMA como NJ están limitados por el hecho de usar una matriz de distancias para resumir la información contenida en las secuencias de ADN y no tener en cuenta un modelo evolutivo, pudiendo potencialmente estar descartando información valiosa sobre la historia evolutiva de las muestras [40].

La idea general de los métodos por parsimonia fue mencionada por primera vez por Edwards y Cavalli-Sforza en 1963 [43], aunque el primer desarrollo teórico formal fue propuesto por Camin y Sokal 2 años después [44]. Se basa en buscar aquella/s filogenia/s en la cual/es, cuando se reconstruyen los eventos evolutivos, existe el menor número de eventos posibles. Esto conlleva 2 implicancias: en primer lugar ha de ser posible realizar una reconstrucción de los eventos invocando el menor número de eventos posibles para cualquier filogenia propuesta; en segundo lugar es necesario poder buscar entre todas las filogenias posibles a aquellas que minimicen el número de eventos [45]. A diferencia de UPGMA y NJ, los métodos por parsimonia no están basados en distancias pareadas sino que utilizan la información contenida en las secuencias directamente, aunque no utiliza un modelo explícito de evolución genética como los métodos que se mencionarán a continuación [40].

El método por ML para la reconstrucción de filogenias fue introducido por Edwards y Cavalli-Sforza en 1964 pero sólo fue aplicable en la práctica luego de 1981 gracias al trabajo de Felsenstein [46, 47]. La idea es encontrar el mejor árbol que maximice la probabilidad de observar los datos, en este caso un alineamiento, bajo cierto modelo evolutivo. El cálculo de la probabilidad de observar cierto alineamiento dada una filogenia es computacionalmente sencillo, el problema surge de tener que realizar el cálculo para un espacio muestral de árboles exponencialmente creciente con el número de secuencias del alineamiento. Este problema es similar al que presentan los métodos de parsimonia, y diversas heurísticas se han desarrollado para realizar una búsqueda más eficiente [47]. La ventaja que presenta frente a los métodos de parsimonia es que permite realizar inferencias teniendo en cuenta modelos de sustitución más complejos, permitiendo, por ejemplo, asumir distintas tasas de evolución a lo largo de las ramas del árbol [40].

Los metodos bayesianos para la inferencia filogenética aparecen de forma madura a mediados de la década de los 90 [48, 49]. Aplicada a los análisis filogenéticos, la topología del árbol y el modelo de sustitución especifican el model estadístico para los datos (los cuales suelen ser alineamientos). Distintas topologías corresponden a distintos modelos mientras que los largos de rama y los parámetris de sustitución son parámetros en el modelo [50]. Una propiedad atractiva de la inferencia Bayesiana es que permite realizar afirmaciones probabilistas directas acerca del modelo o el parámetro desconocido. Presenta 2 diferencias fundamentales respecto a la inferencia por ML: 1) La conveniencia de un árbol no es evaluada en base a su verosimilitud sino en base a su probabilidad posterior, la cual involucra la verosimilitud pero también la probabilidad anterior (*prior*) del árbol. 2) Devuelve una lista de árboles, lo cual es útil para evaluar la probabilidad de determinada topología [40]. La desventaja de este tipo de aproximaciones es que son computacionalmente demandantes y teóricamente complejos, lo cual puede llevar a un mal uso si no se tiene un conocimiento profundo en el tema.

Como ya se mencionó, los datos de entrada para utilizar estas metodologías suele ser un alineamiento nucleotídico o proteico.

Previo al desarrollo de las tecnologías de secuenciación de nueva generación lo habitual era partir del alineamiento de un gen presente en todas las muestras a analizar. En microbiología, una elección muy popular ha sido el gen que codifica para la subunidad 16S del ARNr dada su universalidad, bajo número de copias (1 sola copia en la gran mayoría de los casos), baja tasa de transferencia horizontal entre especies, presenta regiones de alta y baja variabilidad lo que permite estudiar las filogenias a baja y alta profundidad taxonómica, respectivamente, y que su evolución es mayormente neutral lo cual permite asumir la hipótesis de reloj molecular.

El advenimiento de las tecnologías de secuenciación de nueva generación ha permitido ampliar el catálogo de genes utilizados en un mismo análisis, permitiendo inferir filogenias más precisas. El uso de genes *housekeeping* [51, 52], los genes que conforman el genoma núcleo, o los bloques genómicos localmente colineares [53], son hoy en día la materia prima para los análisis filogenéticos en los estudios de genómica comparada en bacterias.

1.2.3. Pangenómica

El término “pangenoma” fue acuñado por Tettelin y colaboradores en 2005, en pleno auge de la secuenciación masiva y cuando las bases de datos de genomas microbianos comenzaban a crecer de forma exponencial [54]. Se definió como el repertorio global de genes de una especie bacteriana, esto es, la suma de aquellos genes compartidos por todas las cepas de dicha especie (genoma núcleo) y de aquellos genes presentes en algunas, pero no en todas, las cepas (genoma dispensable o genoma accesorio) [54].

Una propiedad intrínseca que describe a los pangenomas es su estado de cerrado o abierto. El primer caso se da cuando al secuenciar nuevos genomas el número de nuevos genes encontrados tiende a cero; o lo que es lo mismo, que es posible caracterizar completamente el repertorio de genes de la especie, asumiendo que no existe un sesgo en el muestreo, si se secuencian un número relativamente bajo de genomas. En la publicación original, Tettelin y colaboradores ponen el ejemplo del pangenoma de *Bacillus anthracis* en el cual el número de genes específicos agregados al mismo converge rápidamente a cero luego de agregar tan sólo cuatro genomas [54]. En el caso de un pangenoma abierto, nuevos genomas aportan nuevos genes, de modo que sin importar el número de nuevos genomas muestreados, alcanzado cierto punto el número de nuevos genes tiende a crecer en promedio a una tasa constante. Esto último genera la pregunta de si el ambiente contiene suficientes genes (virtualmente infinitos) para ajustarse a la predicción, lo cual ya ha sido abordado por estudios posteriores y se comentará más adelante [55, 56]. Un ejemplo de pangenoma abierto es el de *Streptococcus agalactiae* en el cual se estimó que un promedio de 33 genes específicos se agregan con cada genoma secuenciado [54].

En un trabajo posterior, Tettelin y colaboradores (2008) desarrollan el marco teórico para la determinación del estado de los pangenoma [57]. En el mismo proponen la utilización de la Ley de Heaps, una ley potencial originalmente formulada para la disciplina de búsqueda y recuperación de información [58]. La misma determina que:

$$n = \kappa N^{-\alpha}$$

donde n es el número de atributos distintos, en este caso genes, N es el número de entidades, o sea genomas, y κ y α ($\alpha = 1 - \gamma$) son parámetros libres determinados empíricamente.

Si α>1, el pangenoma se considera cerrado, y la adición de nuevos genomas no aumentará el número de genes de forma significativa. Por el contrario, si α<1, el pangenoma se considera abierto y por cada nuevo genoma el número de genes crecerá de forma significativa.

La estimación del tamaño del pangenoma y del genoma núcleo también son parámetros de interés, relacionados a lo anterior. Como ya se vio, los pangenomas pueden estar abiertos y teóricamente ser capaces de incorporar un número infinito de nuevos genes [54]. Esta afirmación se ha visto cuestionada más recientemente con el desarrollo teórico de métodos para estimar dichos tamaños [55, 56], aunque sin dejar de ser válida ya que el estado de un pangenoma ofrece información valiosa sobre la flexibilidad genómica del clado estudiado y permite obtener indicios sobre si se trata de un organismo que está ocupando nuevos nichos y emergiendo en nuevos ambientes [35].

El método más sencillo utilizado para estimar el tamaño del pangenoma es el propuesto por Chao en 1987 para estimar tamaños poblacionales basado en el número de capturas y re capturas de individuos [59]. El mismo fue adaptado e implementado por Snipen and Liland en el paquete de R `micropan` para estimar el tamaño de pangenoma [60].

Un método más complejo basado en modelos binomiales mixtos fue desarrollado por Hogg y colaboradores [55] y posteriormente refinado por Snipen y colaboradores [56]. Este permite no solo estimar el tamaño del pangenoma sino también el del genoma núcleo [56].

Por último, otro estadístico muy utilizado para describir un pangenoma es el de fluidez (del inglés *fluidity*, ϕ), introducida por Kislyuk y colaboradores [61] para describir la disimilitud entre los genomas a nivel génico. Se define como el promedio de la razón entre familias génicas únicas con respecto a la suma de familias génicas en pares de genomas seleccionados al azar de los N genomas que conforman el pangenoma:

$$\phi = \frac{2}{N(N-1)} \sum_{\substack{k,l=1\dots N \\ k<l}} \frac{U_k + U_l}{M_k + M_l}$$

donde U_k y U_l son el número de familias génicas encontradas sólo en los genomas k y l, respectivamente, y M_k y M_l son el número total de familias génicas encontradas en k y l respectivamente.

Genomas más variables tienden a tener mayor fluidez, observándose esa misma propiedad en pangenomas considerados abiertos. Lo contrario ocurre en pangenomas, cerrados. Recientemente se ha observado también que la fluidez está muy relacionada al tamaño poblacional efectivo [62].

El primer paso para estudiar pangenomas radica en identificar los grupos de genes ortólogos. Dos secuencias son ortólogas si derivan de un solo gen presente en el último ancestro común de ambos organismos, vía eventos de especiación. Dos secuencias son parálogas si se crearon de un evento de duplicación. De otro modo, la historia evolutiva de dos secuencias ortólogas reflejan la historia evolutiva de los organismos que las contienen, mientras que en el caso de dos secuencias parálogas ambas evolucionaron en paralelo [63]. Dentro de las relaciones de paralogía se pueden distinguir a su vez dos tipos: parálogos internos y parálogos externos (del inglés, *in-paralogs* y *out-paralogs*, respectivamente). Dos secuencias son parálogas internas si la duplicación se produjo antes de la especiación, y son parálogas externas si se produjo posteriormente a la especiación [64].

Una aproximación clásica para detectar ortólogos es utilizar `BLASTp` [65] para comparar todas las secuencias proteicas contra todas y posteriormente agrupar las mismas en familias utilizando algún criterio de *clustering* en base al *bit score*, o en base al criterio de mejor *hit* recíproco [66] aunque se le han encontrado críticas al uso de esta última metodología como “patrón de oro” en lo que respecta a la búsqueda de ortólogos [67].

`OrthoMCL` [68] fue uno de los primeros programas de análisis de familias génicas y utilizaba el algoritmo `MCL` [69] para agrupar las secuencias en base a los *scores* del resultado de la búsqueda de todas contra todas mediante `BLASTp`.

El problema de este tipo de abordajes es que su complejidad aumenta de forma cuadrática con el número de secuencias a comparar, por lo que puede volverse inviable con apenas unas pocas decenas de genomas.

Por lo anterior se han desarrollado heurísticas que permiten mejorar la eficiencia de este paso evitando utilizar `BLASTp` para realizar todas las comparaciones posibles. Un ejemplo notable en este sentido es `roary` [70] el cual utiliza el algoritmo `CD-Hit` [71] para pre agrupar muy rápidamente las secuencias, luego compara una sola secuencia representativa de cada pre grupo con las demás mediante `BLASTp`, y los re agrupa utilizado `MCL`, logrando de este modo reducir la complejidad a una forma casi lineal perdiendo muy poca sensibilidad. `roary` utiliza también información del contexto genético para distinguir entre parálogos y verdaderos ortólogos.

Si bien estas estrategias han sido muy exitosas por su eficiencia y son altamente fiables cuando se trabaja con cepas de una misma especie, presentan la desventaja de ser poco sensibles cuando se trabaja con cepas muy divergentes, como por ejemplo especies distintas dentro de un mismo género. En esos casos se dificulta la detección de relaciones de homología ya sea porque las secuencias son muy distintas para los límites de detección de los algoritmos de búsqueda basados en similaridad de secuencia, o porque el contexto génico perdió toda relación debido a rearreglos genómicos impidiendo así la distinción de ortólogos y parálogos. Esto es un problema si se busca estudiar el pangenoma de clados a un nivel taxonómico superior al de especie, que aunque si bien la definición de pangenoma fue originalmente pensada para referirse a este nivel, la extensión del concepto a otros órdenes taxonómicos resulta casi inevitable.

Por lo anterior se ha sugerido utilizar métodos que permitan detectar relaciones de homología remota basada en perfiles proteicos, como es el caso de los Modelos Ocultos de Markov (HMM) [72]. Se ha propuesto que la utilización de dominios proteicos como parámetro para detectar relaciones de homología remotas entre genes es particularmente robusta, ya que por un lado la variación en las zonas conservadas están más asociadas a cambios funcionales que variación en zonas menos conservadas de las proteínas y se ha visto que los ortólogos tienden a ser funcionalmente más similares entre sí respecto a los parálogos, y por otro que proteínas ortólogas retienen niveles mayores de similitud a nivel estructural respecto a proteínas parálogas, conservando la arquitectura de dominios entendida como la secuencia lineal de dominios no solapantes que aparecen en la proteína [73-75]. La base de datos Pfam [76] aparece como una excelente alternativa ya que provee perfiles HMM de dominios proteicos altamente curados. La estrategia de utilizar la información de dominios proteicos para generar las familias génicas y con ellas el pangenoma de una especie ya ha sido propuesta y llevada a cabo [60, 77, 78]. Sin embargo no han considerado casos en los que la arquitectura proteica no es capaz de distinguir parálogos de ortólogos por sí solas, o casos en que no se detectan dominios proteicos en absoluto.

A su vez, dado que el contexto genético se pierde entre cepas que divergieron de forma temprana, no es aconsejable utilizar esa información para distinguir las relaciones de homología. Es necesario apelar a las definiciones dadas para ortólogos y parálogos, y aplicar algoritmos filogenéticos para resolver este problema [79-81].

1.3. R como entorno de análisis bioinformáticos

La popularidad del lenguaje y entorno R [82] ha explotado en los últimos años entre la comunidad dedicada al análisis de datos, especialmente en la academia [83].

En bioinformática y biología computacional, dicha popularidad puede evidenciarse con los miles de paquetes dedicados a las ciencias de la vida que se encuentran depositados en CRAN (*The Comprehensive R Archive Network*, `https://cran.r-project.org/`), sumados a los más de 1.500 en el repositorio Bioconductor (`http://bioconductor.org/`) [84, 85], y probablemente otros cientos exclusivamente en repositorios como GitHub [86].

Entre las ventajas que se pueden mencionar sobre el uso de este lenguaje se encuentran la facilidad de uso, la calidad de *software* libre, la robustez y riqueza de análisis estadísticos disponibles, la capacidad de generar gráficas y visualizaciones, y la extensa comunidad académica que se encuentra utilizándolo donde es posible encontrar soporte y múltiples recursos didácticos.

1.4. Objetivo general

Como objetivo general nos planteamos el desarrollo de herramientas bioinformáticas en R que permitan el estudio de la genómica comparada del género *Leptospira.* Los objetivos específicos y la justificación de por qué decidimos implementar cada herramienta se encuentran contextualizados en cada capítulo dedicado a cada una de ellas.

Capítulo 2

Genotipado bacteriano automático en R

2.1. Justificación y objetivo específico

El Laboratorio de Microbiología Molecular y Estructural del Institut Pasteur de Montevideo comienza a desarrollar una línea de investigación en leptospirosis en 2013, vinculándose con el Ministerio de Ganadería, Agricultura y Pesca (MGAP).

Posteriormente, en 2015, nace la Unidad Mixta INIA-Pasteur (UMPI), con el objetivo de “(...) potenciar y aportar valor agregado al conocimiento, conjugando las áreas de investigación de ambas instituciones” [87]. La leptospirosis animal es un área fundamental de investigación en dicho laboratorio, siendo uno de sus objetivos crear una colección de aislamientos de *Leptospira spp.* y caracterizarlos. Cumpliendo con la propuesta, en la UMPI se han venido aislando, analizando bioquímicamente y secuenciando múltiples cepas del género reportadas localmente [88].

Como colaboradores, se nos planteó desde la UMPI la necesidad de contar con una herramienta que permita tipificar los aislamientos a partir de las secuencias de forma fácil y eficiente, por lo que como objetivo nos propusimos desarrollar un *software* relativamente amigable al usuario que simplifique dicha tarea.

2.2. Métodos e implementación

Se desarrolló `MLSTar`[1], un paquete de R que permite consultar la base de datos PubMLST, descargar perfiles y secuencias de los distintos esquemas MLST disponibles, y evaluar genomas de forma local para determinar los *sequence type* de los mismos. Detalles de la implementación se encuentran especificados en la publicación

[1] Juego de palabras utilizando las siglas “MLST” propias de la técnica, “ar” por la pronunciación en inglés del nombre del lenguaje R. A su vez la unión de “ST” con “ar” forman la palabra “STar” (“estrella” en inglés). Comúnmente pronunciado “em-el-star”, aunque sin cánones al respecto.

adjunta (sección 2.3.1, página 28).

2.3. Resultados

El paquete se encuentra depositado en la dirección `https://github.com/iferres/MLSTar`, distribuido bajo una licencia *open source* MIT©. Fue publicado en la revista *PeerJ* (doi: 10.7717/peerj.5098). En el artículo se detalla la implementación, y se evalúa el desempeño frente a otros *software* similares y el grado de precisión de las asignaciones comparándolo contra una base de datos de referencia. Una aplicación del *software* se encuentra en la tesis de maestría "Estudios genómicos y moleculares de bacterias del género *Leptospira*: análisis de la variabilidad genética y contribución en diagnóstico y tipificación", de la Mag. Cecilia Nieves, investigadora del Laboratorio de Microbiología Molecular y Estructural (IPMont) cuyo trabajo se centró en la caracterización de aislamientos locales del género *Leptospira*.

2.3.1. Publicación

MLSTar: automatic multilocus sequence typing of bacterial genomes in R

Ignacio Ferrés[1] and Gregorio Iraola[1,2]

[1] Bioinformatics Unit, Institut Pasteur de Montevideo, Montevideo, Uruguay
[2] Center for Integrative Biology, Universidad Mayor, Santiago de Chile, Chile

ABSTRACT

Multilocus sequence typing (MLST) is a standard tool in population genetics and bacterial epidemiology that assesses the genetic variation present in a reduced number of housekeeping genes (typically seven) along the genome. This methodology assigns arbitrary integer identifiers to genetic variations at these loci which allows us to efficiently compare bacterial isolates using allele-based methods. Now, the increasing availability of whole-genome sequences for hundreds to thousands of strains from the same bacterial species has allowed us to apply and extend MLST schemes by automatic extraction of allele information from the genomes. The PubMLST database is the most comprehensive resource of described schemes available for a wide variety of species. Here we present MLSTar as the first R package that allows us to (i) connect with the PubMLST database to select a target scheme, (ii) screen a desired set of genomes to assign alleles and sequence types, and (iii) interact with other widely used R packages to analyze and produce graphical representations of the data. We applied MLSTar to analyze more than 2,500 bacterial genomes from different species, showing great accuracy, and comparable performance with previously published command-line tools. MLSTar can be freely downloaded from http://github.com/iferres/MLSTar.

Subjects Bioinformatics, Ecology, Genomics, Microbiology, Population Biology
Keywords MLST, PubMLST, R package, Microbial genomics, Multilocus genotyping, Bacterial genomes

Submitted 6 March 2018
Accepted 5 June 2018
Published 15 June 2018

Corresponding author
Gregorio Iraola,
giraola@pasteur.edu.uy

Academic editor
Keith Crandall

Additional Information and Declarations can be found on page 8

DOI 10.7717/peerj.5098

INTRODUCTION

Multilocus sequence typing (MLST) was introduced in 1998 as a portable tool for studying epidemiological dynamics and population structure of bacterial pathogens based on PCR amplification and capillary sequencing of housekeeping gene fragments (*Maiden et al., 1998*). In most MLST schemes, seven loci are indexed with arbitrary and unique allele numbers that are combined into an allelic profile or sequence type (ST) to efficiently summarize genetic variability along the genome. Rapidly, MLST demonstrated enhanced reproducibility and convenience in comparison with previous methods such as multilocus enzyme electrophoresis or pulsed-field gel electrophoresis, allowing us to perform global epidemiology and surveillance studies (*Urwin & Maiden, 2003*). For example, MLST has been applied to elucidate the global epidemiology of *Burkholderia multivorans* in cystic fibrosis patients (*Baldwin et al., 2008*) or to understand the dissemination of antibiotic-resistant enterobacteria (*Castanheira et al., 2011*).

How to cite this article Ferrés and Iraola (2018), MLSTar: automatic multilocus sequence typing of bacterial genomes in R. **PeerJ 6:e5098; DOI 10.7717/peerj.5098**

However, as MLST started to be massively applied two main drawbacks were uncovered: (i) the impossibility of establishing a single universal MLST scheme applicable to all bacteria; and (ii) the lack of high resolution of seven-locus MLST schemes required for some purposes.

These problems pushed the development of improved alternatives to the original methodology. The extended MLST approach which is based on the analysis of longer gene fragments (*Chen et al., 2011*) or increased number of loci (*Dingle et al., 2008*; *Crisafulli et al., 2013*) proved to improve resolution, and the scheme based on 53 ribosomal protein genes (rMLST) was proposed as an universal approach since these loci are conserved in all bacteria (*Jolley et al., 2012*). Beyond these improvements, the advent of high-throughput sequencing and the increasing availability of hundreds to thousands whole-genome sequences (WGS) for many bacterial pathogens caused a paradigmatic change in clinical microbiology, making it possible to use nearly complete genomic sequences to enhance typing resolution. This revolution allowed the transition from standard MLST schemes testing a handful of genes to core genome approaches that scaled to hundreds of loci common to a set of bacterial genomes (*Maiden et al., 2013*).

The generation of this massive amount of genetic information required the accompanying development of database resources to effectively organize and store typing schemes and allele definitions. Rapidly, the PubMLST database (http://pubmlst.org) turned into the most comprehensive and standard resource storing today schemes and allelic definitions for more than 100 microorganisms. Subsequently, the shift to WGS motivated the development of the Bacterial Isolate Genome Sequence Database (BIGSdb) (*Jolley & Maiden, 2010*), which now encompasses all the software functionalities used for the PubMLST. Also, many tools for automatic MLST analysis from WGS have been developed using web servers like MLST-OGE (*Larsen et al., 2012*) or EnteroBase (http://enterobase.warwick.ac.uk), pay-walled tools like BioNumerics or SeqSphere+, and open source tools like mlst (http://github.com/tseemann/mlst) or MLSTcheck (*Page, Taylor & Keane, 2016*). Here, we present MLSTar as the first tool for automatic MLST of bacterial genomes written in R (*R Development Core Team, 2008*), allowing us to expand the application of MLST tools within this very popular and useful environment for data analysis and visualization.

METHODS

Implementation

MLSTar is written in R and contains all data processing steps and command line parameters to call external dependencies wrapped in the package. MLSTar depends on BLAST+ (*Camacho et al., 2009*) that is used as sequence search engine, and must be installed locally. MLSTar is designed to work on Unix-based operating systems and is distributed as an open source software (MIT license) stored in GitHub (http://github.com/iferres/MLSTar). MLSTar contains four main functions that (i) takes genome assemblies or predicted genes in FASTA format from any number of strains, (ii) performs sequence typing using a previously selected scheme from PubMLST, and (iii) applies

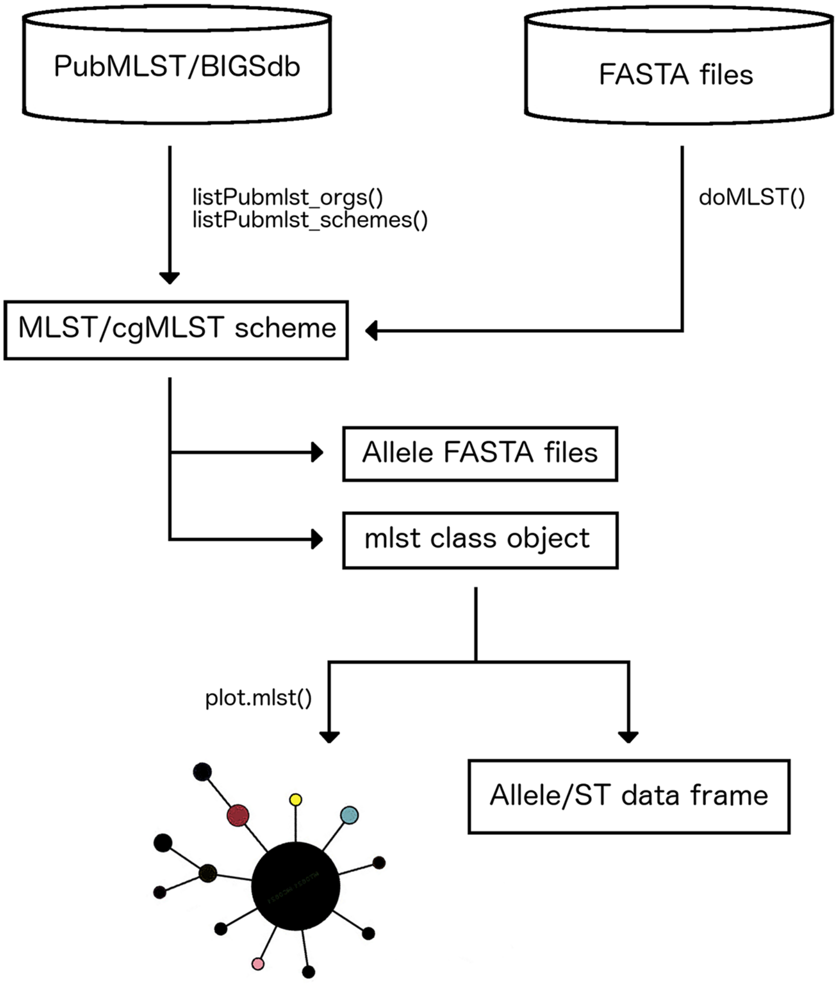


Figure 1 **Main steps in** `MLSTar` **workflow.** Full-size DOI: 10.7717/peerj.5098/fig-1

standard phylogenetic approaches to analyze the data. An overview of the overall workflow has been outlined in Fig. 1.

Interaction with PubMLST

First step in `MLSTar` workflow involves to interact with the PubMLST database to select a target scheme. This interaction requires Internet connection because is performed using the RESTful web application programming interface provided by PubMLST. The `listPubmlst_orgs()` function allows us to list the names of all microorganisms that have any scheme stored in PubMLST. Then, as some microorganisms have more than one scheme (i.e., one classical seven-loci and one core genome scheme), the `listPubmlst_schemes()` function lists the available schemes for any selected species. Additionally, `MLSTar` is not restricted only to the MLST definitions present in PubMLST since schemes stored in other databases can be manually downloaded and analyzed with `MLSTar`.

Calling and storing alleles and sequence types

MLSTar make allele and ST calls from FASTA files containing closed genomes or contigs using BLAST+ blastn comparisons implemented by the doMLST() function. Parallelization is available as internally implemented in R by the parallel package. Also, the doMLST() function can be run at the same time for different schemes using internal R functions like lapply(). Results are stored in a S3 class object named mlst that contains two data.frame objects: one containing allele and ST assignments for the analyzed genomes (unknown alleles or STs are labeled as "u"), and the other storing known allele profiles for the selected scheme. If required, nucleotide sequences for known or novel alleles can be written as multi FASTA files for downstream analyses.

Post analysis

Allele profiles are frequently used to reconstruct phylogenetic relationships among strains. Function plot.mlst() directly takes the mlst class object to compute distances assuming no relationships between allele numbers, so each locus difference is treated equally. Then, identical isolates have a distance of 0; those with no alleles in common have a distance of 1 and, for example, in a seven-loci scheme two strains with five differences would have a distance of 0.71 (5/7). The resulting distance matrix is used to build a minimum spanning tree using igraph (*Csardi & Nepusz, 2006*) that returns an object of class igraph or a neighbor-joining tree as implemented in APE package (*Paradis, Claude & Strimmer, 2004*) that returns an object of class phylo. The package also contains a specific method defined as plot.mlst that recognizes the mlst class object and plots the results using the generic plot() function. Additionally, a better resolution analysis based on the variability of the underlying sequences using more sophisticated Maximum-Likelihood or Bayesian phylogenies, can be achieved externally by aligning the allele sequences that are automatically retrieved by MLSTar.

RESULTS AND DISCUSSION

Comparison with capillary sequencing data

Multilocus sequence typing analysis based on capillary sequencing has been considered as the gold standard. Hence, we used a previously reported dataset (*Page et al., 2017*) consisting in 72 *Salmonella* samples originally tested by capillary sequencing and deposited in the EnteroBase (*Alikhan et al., 2018*), that were posteriorly whole-genome sequenced. This dataset covers a wide host range and isolation dates of *Salmonella* strains comprising 32 different STs (Table S1). In average, MLSTar assignments at ST level matched in 92% of cases when compared with capillary sequencing. Additionally, ST calls for five samples that were distinct between capillary sequencing and genome-derived inferences using several software tools (*Page et al., 2017*), were also discordant in the same way when using MLSTar. This is expected since capillary sequencing is not error free (*Liu et al., 2012*), in spite of being considered as the gold standard. By the contrary, the result for sample 139K matched between capillary sequencing and MLSTar but most other software tools, except stringMLST (*Gupta, Jordan & Rishishwar, 2016*), failed to assign confident STs. MLSTar results on the same dataset but in comparison

Table 1 Accuracy of MLSTar against reference alleles and STs obtained from BIGSdb, measured as the percentage of correct calls in seven-locus MLST schemes from 11 different pathogens comprising a total of 3,021 genomes.

Species	Genomes	Scheme							
Bordetella spp.	66	*adk*	*fumC*	*glyA*	*tyrB*	*icd*	*pepA*	*pgm*	ST
		96.7	96.7	96.7	96.7	96.7	95	96.7	95
Staphylococcus aureus	72	*gdh*	*gyd*	*pstS*	*gki*	*aroE*	*xpt*	*yqiL*	ST
		94.4	94.4	94.5	95.3	94.4	95.2	99.4	93.1
Helicobacter pylori	79	*atpA*	*efp*	*mutY*	*ppa*	*trpC*	*ureI*	*yphC*	ST
		97.5	96.2	98.7	97.5	98.7	97.5	97.5	93.7
Bacillus cereus	115	*glp*	*gmk*	*ilv*	*pta*	*pur*	*pyc*	*tpi*	ST
		98.3	100	100	100	100	96.5	98.2	93.9
Campylobacter jejuni/coli	176	*aspA*	*glnA*	*gltA*	*glyA*	*pgm*	*tkt*	*uncA*	ST
		100	99	100	100	100	100	100	99
Burkholderia pseudomallei	225	*ace*	*gltB*	*gmhD*	*lepA*	*lipA*	*narK*	*ndh*	ST
		98.7	96	93	96	96.9	95.6	96	93
Streptococcus agalactiae	258	*adhP*	*pheS*	*atr*	*glnA*	*sdhA*	*glcK*	*tkt*	ST
		99.2	99.6	99.2	99.2	99.2	99.6	99.6	98.1
Klebsiella pneumoniae	284	*gapA*	*infB*	*mdh*	*pgi*	*phoE*	*rpoB*	*tonB*	ST
		100	100	100	100	100	100	100	100
Pseudomonas aeruginosa	604	*acs*	*aro*	*gua*	*mut*	*nuo*	*pps*	*trp*	ST
		96.4	98.8	98.1	98.3	98.1	98.3	98.8	95.9
Acinetobacter baumannii	847	*cpn60*	*fusA*	*gltA*	*pyrG*	*recA*	*rplB*	*rpoB*	ST
		98.6	97.4	99.3	99.2	97.3	99.1	98.7	94.9

with other softwares designed to screen whole-genome assemblies such as mlst (http://github.com/tseemann/mlst) and MLSTcheck (*Page, Taylor & Keane, 2016*) matched in 89% and 92% of cases, respectively. These results demonstrate that MLSTar and other software have comparable performance when testing against standard MLST results based on capillary sequencing.

Comparison against BIGSdb

We retrieved 2,726 genomes from the BIGSdb belonging to 10 species most of which are very well-known pathogens (Table S2). For these datasets, reference allele, and ST assignments based on the corresponding standard MLST schemes were extracted from the BIGSdb and compared with results obtained running MLSTar. The concordance at allele and ST levels is shown in Table 1, measured as the percentage of identical assignments between BIGSdb and MLSTar. In average, assignments were 97.9% (SD = 1.95) and 95.6% (SD = 2.5) coincident for alleles and STs, respectively. These results evidence a very good performance of MLSTar in comparison with the reference assignments from the BIGSdb. Additionally, we tested MLSTar using the ribosomal MLST scheme (*Jolley et al., 2012*) over the same 354 genomes belonging to *Staphylococcus aureus* and *Streptococcus agalactiae.* This scheme was conceived as an universal approach for discrimination of bacterial species. Accordingly, the automatic phylogenetic analysis

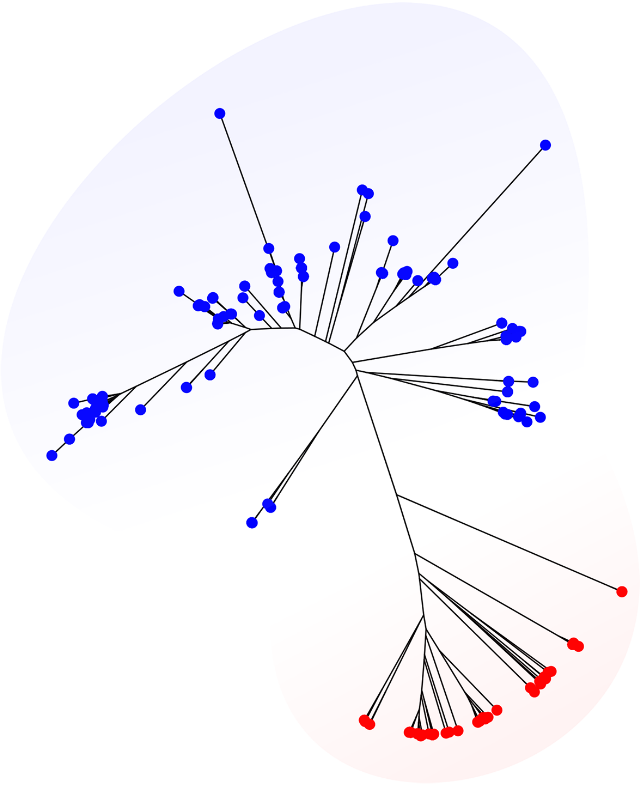


Figure 2 Phylogeny based on ribosomal alleles. *Staphylococcus aureus* (red) and *Streptococcus agalactiae* (blue) genomes from the BIGSdb (*n* = 356) were characterized using the universal rMLST scheme (based on 53 ribosomal genes). The phylogenetic tree was automatically generated with the plot.mlst() function using the Neighbor-Joining algorithm from a distance matrix obtained from allele patterns.

Full-size DOI: 10.7717/peerj.5098/fig-2

implemented in MLSTar was able to discriminate both species using ribosomal alleles (Fig. 2).

Comparison with MLST schemes of close species

The PubMLST database stores schemes for 10 different species within the genus *Campylobacter*, hence we used this case as negative control to test the specificity of MLSTar. We chose the 172-*C. jejuni*/*coli* dataset from BIGSdb and 150 randomly selected *C. fetus* genomes from a previously published study (*Iraola et al., 2017*) to run MLSTar against the schemes defined for the remaining *Campylobacter* species, in order to detect potential false positive calls when analyzing closely related taxa. False positives at both allele and ST levels were not detected neither for *C. jejuni*/*coli* nor for *C. fetus* against the rest (Table S3), indicating that MLSTar is highly specific when working with genetically related bacteria.

Comparison of variable coverage depths and number of genomes

Variable depths of sequencing coverage have been shown to affect the accuracy of different softwares to achieve confident ST calls. In general, most softwares require over than 10× to ensure optimal performance (*Page et al., 2017*). Here, we tested MLSTar by

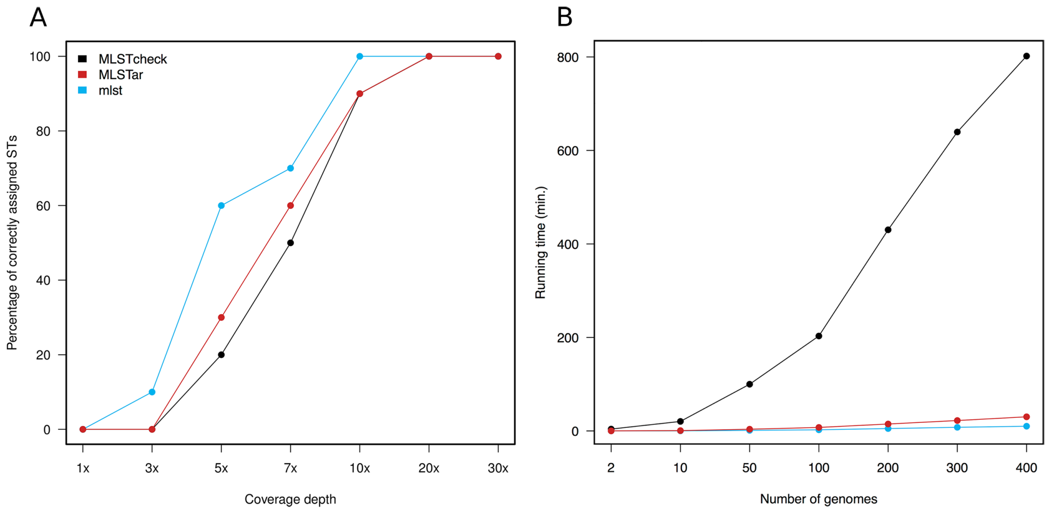


Figure 3 Comparison of MLSTar performance. (A) Comparison of MLSTar, MLSTcheck and mlst softwares using a dataset of 10 *Salmonella* genomes de novo assembled at variable coverage depths. (B) Comparison of MLSTar, MLSTcheck, and mlst running times on a single CPU using increasing number of genomes. Full-size DOI: 10.7717/peerj.5098/fig-3

sampling reads at gradual depths from 10 genomes (representing different STs) from the *Salmonella* dataset and measured the percentage of correctly assigned STs. Figure 3A shows that MLSTar produce good-enough results when sequencing depth is greater than 10×, and its performance is comparable to similar tools such as MLSTcheck and mlst. Considering that nowadays bacterial genome sequencing experiments typically ensure at least 30× of coverage depth, our results evidence that MLSTar is appropriate for analyzing WGS with average or even slightly lower coverage depths. Additionally, we used a random set of genomes ($n = 400$) from the BIGSdb dataset to compare the running time between MLSTar, MLSTcheck, and mlst softwares in a single AMD Opteron 2.1 GHz processor, by gradually increasing the number of analyzed genomes from 2 to 400 (Fig. 3B). These results showed that MLSTar is 26-fold faster than MLSTcheck but is threefold slower than mlst (Table S4).

CONCLUSION

The advent of WGS has now allowed to type bacterial strains directly from their whole genomes avoiding to repeat tedious PCR amplifications and fragment capillary sequencing for multiple loci. Today MLST is a valid tool which is frequently used as a first-glimpse approach to explore genetic diversity and structure within huge bacterial population sequencing projects. This incessant availability of genomic information has motivated a constant effort to develop efficient analytical tools from multilocus typing data (*Page et al., 2017*). Here, we developed a new software package called MLSTar that expands the possibilities of performing allele-based genetic characterization within the R environment. We demonstrate that MLSTar has comparable performance with previously validated software tools and can be applied to analyze hundreds of genomes in a reasonable time.

ACKNOWLEDGEMENTS

We thank Daniela Costa and Cecilia Nieves for testing `MLSTar`.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the Agencia Nacional de Investigación e Innovación of Uruguay post-graduation program (No. POS_NAC_1_2016_131079) granted to Ignacio Ferrés. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Agencia Nacional de Investigación e Innovación of Uruguay post-graduation program: No. POS_NAC_1_2016_131079.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Ignacio Ferrés conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.
- Gregorio Iraola conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.

Data Availability

The following information was supplied regarding data availability:

Github: https://github.com/iferres/MLSTar.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.5098#supplemental-information.

REFERENCES

Alikhan N, Zhou Z, Sergeant M, Achtman M. 2018. A genomic overview of the population structure of salmonella. *PLOS Genetics* **14(4)**:e1007261 DOI 10.1371/journal.pgen.1007261.

Baldwin A, Mahenthiralingam E, Drevinek P, Pope C, Waine DJ, Henry DA, Speert DP, Carter P, Vandamme P, LiPuma JJ, Dowson CG. 2008. Elucidating global epidemiology of *Burkholderia multivorans* in cases of cystic fibrosis by multilocus sequence typing. *Journal of Clinical Microbiology* **46(1)**:290–295 DOI 10.1128/jcm.01818-07.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. Blast+: architecture and applications. *BMC Bioinformatics* **10(1)**:421 DOI 10.1186/1471-2105-10-421.

Castanheira M, Deshpande LM, Mathai D, Bell JM, Jones RN, Mendes RE. 2011. Early dissemination of ndm-1-and oxa-181-producing enterobacteriaceae in indian hospitals: report from the sentry antimicrobial surveillance program, 2006–2007. *Antimicrobial Agents and Chemotherapy* **55**(**3**):1274–1278 DOI 10.1128/aac.01497-10.

Chen Y, Zhen Q, Wang Y, Xu J, Sun Y, Li T, Gao L, Guo F, Wang D, Yuan X, Yuan J, Huang L, Chen Z, Yu Y. 2011. Development of an extended multilocus sequence typing for genotyping of brucella isolates. *Journal of Microbiological Methods* **86**(**2**):252–254 DOI 10.1016/j.mimet.2011.05.013.

Crisafulli G, Guidotti S, Muzzi A, Torricelli G, Moschioni M, Masignani V, Censini S, Donati C. 2013. An extended multi-locus molecular typing schema for streptococcus pneumoniae demonstrates that a limited number of capsular switch events is responsible for serotype heterogeneity of closely related strains from different countries. *Infection, Genetics and Evolution* **13**:151–161 DOI 10.1016/j.meegid.2012.09.008.

Csardi G, Nepusz T. 2006. The igraph software package for complex network research. *InterJournal, Complex Systems* **1695**(**5**):1–9.

Dingle KE, McCarthy ND, Cody AJ, Peto TE, Maiden MC. 2008. Extended sequence typing of campylobacter spp., United Kingdom. *Emerging Infectious Diseases* **14**(**10**):1620–1622 DOI 10.3201/eid1410.071109.

Gupta A, Jordan IK, Rishishwar L. 2016. stringMLST: a fast k-mer based tool for multilocus sequence typing. *Bioinformatics* **33**(**1**):119–121 DOI 10.1093/bioinformatics/btw586.

Iraola G, Forster SC, Kumar N, Lehours P, Bekal S, Garca-Peña FJ, Paolicchi F, Morsella C, Hotzel H, Hsueh P-R, Vidal A, Lévesque S, Yamazaki W, Balzan C, Vargas A, Piccirillo A, Chaban B, Hill JE, Betancor L, Collado L, Truyers I, Midwinter AC, Dagi HT, Mégraud F, Calleros L, Pérez R, Naya H, Lawley TD. 2017. Distinct campylobacter fetus lineages adapted as livestock pathogens and human pathobionts in the intestinal microbiota. *Nature Communications* **8**(**1**):1367 DOI 10.1038/s41467-017-01449-9.

Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, Wimalarathna H, Harrison OB, Sheppard SK, Cody AJ, Maiden MCJ. 2012. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology* **158**(**4**): 1005–1015 DOI 10.1099/mic.0.055459-0.

Jolley KA, Maiden MC. 2010. BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* **11**(**1**):595 DOI 10.1186/1471-2105-11-595.

Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-Pontén T, Ussery DW, Aarestrup FM, Lund O. 2012. Multilocus sequence typing of total-genome-sequenced bacteria. *Journal of Clinical Microbiology* **50**(**4**):1355–1361 DOI 10.1128/jcm.06094-11.

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. 2012. Comparison of next-generation sequencing systems. *BioMed Research International.* **2012:**251364 DOI 10.1155/2012/251364.

Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America* **95**(**6**):3140–3145 DOI 10.1073/pnas.95.6.3140.

Maiden MC, Van Rensburg MJJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND. 2013. Mlst revisited: the gene-by-gene approach to bacterial genomics. *Nature Reviews Microbiology* **11**(**10**):728–736 DOI 10.1038/nrmicro3093.

Page AJ, Alikhan N-F, Carleton HA, Seemann T, Keane JA, Katz LS. 2017. Comparison of classical multi-locus sequence typing software for next-generation sequencing data. *Microbial Genomics* **3**(**8**):e000124 DOI 10.1099/mgen.0.000124.

Page AJ, Taylor B, Keane JA. 2016. Multilocus sequence typing by blast from de novo assemblies against pubmlst. *Journal of Open Source Software* **1**(**8**):118 DOI 10.21105/joss.00118.

Paradis E, Claude J, Strimmer K. 2004. Ape: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**(**2**):289–290 DOI 10.1093/bioinformatics/btg412.

R Development Core Team. 2008. *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing. *Available at https://www.r-project.org/*.

Urwin R, Maiden MC. 2003. Multi-locus sequence typing: a tool for global epidemiology. *Trends in Microbiology* **11**(**10**):479–487 DOI 10.1016/j.tim.2003.08.006.

2.4. Perspectivas

El explosivo crecimiento de las bases de datos de genomas completos ha permitido el desarrollo de esquemas de MLST extendidos que utilizan genes presentes en todos los aislamientos de un clado dado recuperando así la diversidad presente de forma más detallada. A estos esquemas se los ha designado como cgMLST, por sus siglas en inglés "*core genome MLST*" [39]. Se planea optimizar los algoritmos de búsqueda para que el paquete sea capaz de tipificar decenas de genomas utilizando esquemas cgMLST con decenas o cientos de *loci* distintos con varios alelos posibles cada uno, en tiempos razonables.

Recientemente se ha propuesto e implementado la tipificación de comunidades en metagenomas [89], lo cual se considerará incorporar en futuras versiones de `MLSTar`, junto con nuevos métodos y funciones que permitan analizar y visualizar los resultados.

Capítulo 3

Reconstrucción filogenética utilizando la base de datos EggNOG

3.1. Justificación y objetivo específico

El estudio de cualquier género microbiano requiere como análisis fundamental contar con inferencias filogenéticas robustas. Históricamente, como ya se mencionó en la introducción (sección 1.2.2), se utilizó el gen de la subunidad 16S del ARNr como punto de partida para realizar dichan inferencias. Los métodos de secuenciación masiva y la mejora del poder computacional, sin embargo, han posibilitado el uso de alineamientos de genes del genoma núcleo para mejorar esas inferencias.

Dos colaboraciones internacionales en las que se pretendía describir nuevas especies del género *Leptospira* plantearon la necesidad de desarrollar una metodología capaz de identificar genes núcleo en organismos divergentes, alinearlos e inferir una filogenia a partir del alineamiento concatenado de dichos genes.

3.2. Métodos e implementación

Se implementó `Phylen`[2], un paquete en R que permite consultar la base de datos de ortólogos EggNOG [90] y descargar modelos HMM específicos del clado de interés. Cada modelo HMM disponible en EggNOG es construido a partir del alineamiento curado de secuencias proteicas ortólogas de un clado determinado. Esos modelos son usados por `Phylen` para identificar genes presentes en todos los genomas de forma eficiente, extraerlos, alinearlos, e inferir una filogenia a partir del concatenado de los alineamientos obtenidos. Detalles de la imlpementación se encuentran detallados en

[2] Acrónimo de “PHYlogeny” y “EggNog”, haciendo referencia a la funcionalidad y a que utiliza la base de datos EggNOG.

la publicación adjunta (sección 3.3.1, página 41)

3.3. Resultados

El paquete se encuentra disponible en la dirección `https://github.com/iferres/phylen`, distribuido bajo una licencia *open source* MIT©, y fue publicado en la revista *Journal of Open Source Software* (doi: 10.21105/joss.00593). Como se mencionó, `phylen` fue aplicado en dos publicaciones donde se describieron especies nuevas del género *Leptospira* [91, 92] (Ver apéndices A y B).

3.3.1. Publicación

The Journal of Open Source Software

Phylen: automatic phylogenetic reconstruction using the EggNOG database

Ignacio Ferrés[1] **and Gregorio Iraola**[1]

1 Unidad de Bioinformática, Institut Pasteur de Montevideo, Uruguay

DOI: 10.21105/joss.00593

Software

- Review
- Repository
- Archive

Submitted: 19 February 2018
Published: 04 May 2018

Summary

High-throughput sequencing is dramatically increasing the amount of genetic data available from all domains of life, but particularly from bacteria. The smaller size of bacterial genomes allows to sequence large collections of strains, mainly from species that deserve interest for their importance as human or farm animal pathogens. Phylogenetic analysis has become a standard tool to understand the evolutionary history, epidemiology and virulence of these bacteria, and the availability of genomic information has allowed to move from single-gene (e.g. using the 16S rRNA gene) to multilocus or core genome trees that bring us closer to a more reliable reconstruction of phylogenetic structure.

The EggNOG database (Powell et al. 2014) is an excellent resource providing orthologous groups shared at different taxonomic ranks including several prokaryotes. Here we present Phylen, a simple and automated software package written in R that reconstructs phylogenies by interacting with the EggNOG database. First, a set of orthologous groups available at the EggNOG database is selected and automatically downloaded or, alternatively, an external set of orthologous groups can be provided formatted as a Hidden Markov Model (HMM) file. Second, genome annotations in GFF3 format (such as those from Prokka annotation software (Seemann 2014)) are parsed to extract translated coding sequences. Third, genomes are screened against these orthologous groups using HMMER3 (Eddy 2011). Forth, "core" coding sequences are extracted and multiple sequence alignment is performed over each recovered gene set using MAFFT (Katoh and Standley 2013). Fifth, alignments are concatenated into a single supergene and phylogenetic reconstruction is performed using Maximum-Likelihood or distance methods (Fig. 1A). Phylen outputs one multi-fasta alignment per gene, one supergene multi-fasta alignment file, one tree file in Newick format and an object of class "phylo" which can be further analysed using the R packeges ape (Paradis, Claude, and Strimmer 2004) and phangorn (Schliep 2011).

Phylen has been already used by our group for building the *Helicobacter* genus phylogeny (Fresia et al. 2017) from a set of 40 universal marker genes (Mende et al. 2013), and to reconstruct core genome phylogenies of *Leptospira* genus (Puche et al. 2017; Thibeaux et al. 2018) from orthologous groups defined in the EggNOG database (spiNOG) (Powell et al. 2014). Additionaly, here we screened 93 *Epsilonproteobacteria* genomes against 4513 orthologous groups from the EggNOG database (eproNOG) to obtain the phylogenetic tree shown in Fig. 1B. In the near future we plan to add more functionalities such as different multiple sequence alignment algorithms and tools for alignment quality check and trimming.

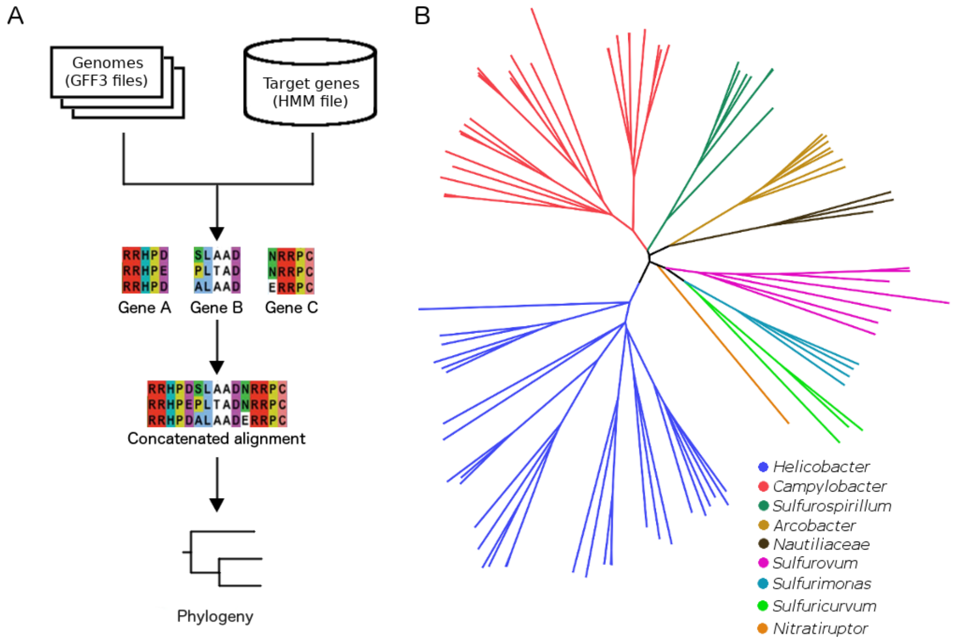


Figure 1: A) Schematic workflow of Phylen. B) Phylogeny of *Epsilonproteobacteria* obtained with the eggNOG database (eproNOG orthologs).

Phylen has been designed to facilitate the reconstruction of high-resolution phylogenies at any desired taxonomic rank, and from any set of genes like taxon-specific markers or the whole core genome. Phylogenetic reconstruction is a standard kick-off analysis in almost every comparative genomics project and despite many methods have been developed, Phylen is unique as it integrates the highly accessed EggNOG database (for phylogenetic marker genes) with the R environment as a widely used programming interface for microbial genomics and data analysis. Phylen depends on the R package phangorn (Schliep 2011) for phylogenetic reconstruction and external tools including HMMER3 (Eddy 2011) as gene search engine and MAFFT (Katoh and Standley 2013) for multiple sequence alignment.

Acknowledgements

I.F. was supported by ANII (Uruguay) postgraduation grant POS_NAC_2016_1_131079. We thank Pablo Fresia and Daniela Costa for testing Phylen.

References

Eddy, Sean R. 2011. “Accelerated Profile Hmm Searches.” *PLoS Computational Biology* 7 (10). Public Library of Science:e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

Fresia, Pablo, Ronald Jara, Rafael Sierra, Ignacio Ferrés, Gonzalo Greif, Gregorio Iraola, and Luis Collado. 2017. “Genomic and Clinical Evidence Uncovers the Enterohepatic Species Helicobacter Valdiviensis as a Potential Human Intestinal Pathogen.” *Helicobacter* 22 (5). Wiley Online Library. https://doi.org/10.1111/hel.12425.

Katoh, Kazutaka, and Daron M Standley. 2013. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” *Molecular Biology and Evolution* 30 (4). Society for Molecular Biology; Evolution:772–80. https://doi.org/10.1093/molbev/mst010.

Mende, Daniel R, Shinichi Sunagawa, Georg Zeller, and Peer Bork. 2013. “Accurate and Universal Delineation of Prokaryotic Species.” *Nature Methods* 10 (9). Nature Publishing Group:881. https://doi.org/10.1038/nmeth.2575.

Paradis, Emmanuel, Julien Claude, and Korbinian Strimmer. 2004. “APE: Analyses of Phylogenetics and Evolution in R Language.” *Bioinformatics* 20 (2). Oxford University Press:289–90. https://doi.org/10.1093/bioinformatics/btg412.

Powell, Sean, Kristoffer Forslund, Damian Szklarczyk, Kalliopi Trachana, Alexander Roth, Jaime Huerta-Cepas, Toni Gabaldon, et al. 2014. “EggNOG V4. 0: Nested Orthology Inference Across 3686 Organisms.” *Nucleic Acids Research* 42 (D1). Oxford University Press:D231–D239. https://doi.org/10.1093/nar/gkt1253.

Puche, Rafael, Ignacio Ferrés, Lizeth Caraballo, Yaritza Rangel, Mathieu Picardeau, Howard Takiff, and Gregorio Iraola. 2017. “Leptospira Venezuelensis Sp. Nov., a New Member of the Intermediates Group Isolated from Rodents, Cattle and Humans.” *International Journal of Systematic and Evolutionary Microbiology*. https://doi.org/10.1099/ijsem.0.002528.

Schliep, Klaus Peter. 2011. “Phangorn: Phylogenetic Analysis in R.” *Bioinformatics* 27 (4). Oxford University Press:592. https://doi.org/10.1093/bioinformatics/btq706.

Seemann, Torsten. 2014. “Prokka: Rapid Prokaryotic Genome Annotation.” *Bioinformatics* 30 (14). Oxford University Press:2068–9. https://doi.org/10.1093/bioinformatics/btu153.

Thibeaux, Roman, Gregorio Iraola, Ignacio Ferrés, Emilie Bierque, Dominique Girault, Marie-Estelle Soupé-Gilbert, Mathieu Picardeau, and Cyrille Goarant. 2018. “Deciphering the Unexplored Leptospira Diversity from Soils Uncovers Genomic Evolution to Virulence.” *Microbial Genomics* 4 (1). Microbiology Society. https://doi.org/10.1099/mgen.0.000144.

3.4. Perspectivas

Se planea mejorar la implementación permitiendo un control más exhaustivo de los parámetros utilizados para la inferencia filogenética. A su vez se propone que es posible hacer una selección más inteligente de los genes que integran el súper alineamiento, teniendo en cuenta, por ejemplo, el largo o la presencia de parálogos. La implementación actual fue pensada en aras de la eficiencia y usabilidad, por lo que muchas decisiones son tomadas por el *software* utilizando criterios simples y aproximados. Mejorar esta característica permitiría, por ejemplo, seleccionar genes que puedan ser aptos para, eventualmente, proponer esquemas de cgMLST.

Capítulo 4

Pangenómica extendida a taxones superiores

4.1. Justificación y objetivo específico

Un análisis del genoma accesorio es requisito para estudiar la genómica comparada de un clado. Es en este sentido que se han desarrollado diversos métodos para la reconstrucción de pangenomas, aunque en buena medida se han centrado en optimizar los procesos para analizar clados pertenecientes al taxón especie, dejando de lado que los mismos conceptos que describen a los pangenomas de especies pueden aplicarse a niveles taxonómicos superiores. Un problema que surge de lo anterior es que los métodos clásicos para determinar si dos o más secuencias (nucleotídicas o aminoacídicas) son ortólogas se basan en similaridad, lo cual dificulta la detección de homología remota en casos donde dos clados presenten mucha divergencia, como por ejemplo especies distintas de un mismo género.

Nos propusimos desarrollar un *software* capaz de reconstruir pangenomas bacterianos utilizando métodos que permitan la detección de homología remota y que, de esa forma, habilite el estudio clados de niveles taxonómicos superiores al de especie.

4.2. Métodos e implementación

Se implementó `Pewit`[3], un paquete de R que utiliza información de dominios proteicos y un algoritmo de "poda" (*prunning*) de árboles filogenéticos, para detectar

[3] En inglés se denomina *pewit* a una especie de pájaro oriundo de la zona templada de eurasia, *Vallenus vallenus*, pariente cercano al tero regional, *Vallenus chilensis*. El nombre del software es un acrónimo de "Pangenome Estimation Walks Inside Taxonomy" que intenta, de forma rebuscada, dar cuenta de su propósito y a su vez homenajear a la versión criolla del animal.

ortólogos remotos y así reconstruir el pangenoma de clados taxonómicamente divergentes. Detalles de la imlpementación se encuentran detallados en el manuscrito (sección 4.3.1, página 47)

4.3. Resultados

El paquete se encuentra disponible en un estado *beta* en la dirección `https://github.com/iferres/pewit`, (Licencia GNU v3.0©, *open source*). Un manuscrito presentando la metodología y comparando su desempeño con otros *software* similares se encuentra en preparación, y se adjunta en la siguiente sección un bosquejo del trabajo. Se presenta, a su vez, como ejemplo la aplicación de `Pewit` sobre un sub clado de especies del género *Leptospira*, mostrando que logra resolver su pangenoma de forma biológicamente congruente, en contraste con otro *software* muy utilizado por la comunidad científica dedicada al estudio de la evolución microbiana. Cabe mencionar que el conjunto de datos de *Leptospira* utilizado pertenece a una publicación en la cual `Pewit` ya había sido aplicado exitosamente (ver apéndice B).

4.3.1. Manuscrito

Pewit: pangenome estimation walks inside taxonomy

Ignacio Ferrés[1], Pablo Fresia[1], Daniela Costa[1], Hugo Naya[1,2] & Gregorio Iraola[1,3,*]

[1] *Unidad de Bioinformática, Institut Pasteur de Montevideo, Montevideo, Uruguay.*

[2] *Departamento de Producción Animal y Pasturas, Facultad de Agronomía, Universidad de la República, Montevideo, Uruguay.*

[3] *Center for Integrative Biology, Universidad Mayor, Santiago de Chile, Chile.*

Inferring functional patterns from bacterial genomes is essential for taxonomy, pathogen evolution and epidemiology. Pangenome reconstruction is today the gold-standard for comparartive analysis of multiple bacterial genomes. However, most pangenome reconstruction strategies have been designed to deal with phylogenetically close genomes, such as those belonging to the same bacterial species, hampering the straightforward identification of common and distinctive gene repertories at higher taxonomic ranks. To fill this gap, here we present Pewit (https://github.com/iferres/pewit), a tool that comprehensively expands the pangenome concept towards the inside of bacterial taxonomy. Our method can automatically and efficiently reconstruct pangenomes of phylogenetically close and increasingly distant bacteria, maximising the recovery of core genes and minimising the spurious clustering at higher ranks. As a case study, we identify virulence determinants distinguishing environmental from pathogenic species belonging to the divergent genus *Leptospira*. Pewit provides a novel approach that is appropriate for characterizing gene repertories in bacterial populations across taxonomy.

The bacterial pangenome concept was coined by Tettelin et al. in 2005 [1] to describe the union of genes present in a set of genomes representing the same species [2]. This notion was then formalised in the distributed genome hypothesis, which states that the fundamental action of gene gain/loss, mainly directed by horizontal gene transfer mechanisms, determines that the gene pool of a bacterial taxon is more complex than that found in the genome of any individual strain [3]. The consolidation of the pangenome concept was fed by the increasing availability of whole-genome sequences, that have allowed to perform extensive comparative pangenome analyses to gain insight on the evolutionary forces shaping bacterial phenotypes like niche-adaptation, transmission or virulence [4–6].

In fact, this has had a tremendous impact on our understanding of evolutionary and ecological processes occurring in bacterial populations and communities. In particular, horizontal gene transfer has been described not only between components of the same species but also between higher taxonomic ranks [7], such as in the phylum *Proteobacteria*, whose members have pervasively exchanged genes at different taxonomic depths [8]. Accordingly, the definition of bacterial pangenomes could be expanded to consider genome plasticity occurring beyond the species level. Indeed, several recent works have focused on studying the pangenome variability at taxonomic depths different from species, such as the genus *Clostridium* [9], the family *Bifidobacteriaceae* [10] or even at higher ranks like the phylum *Latescibacteria* [11]. However, most available tools for automatic pangenome reconstruction use heuristics that efficiently compare phylogenetically close genomes, since they have been designed to work at the species level, but are not optimal to deal with unrelated genomes as evolutionary distance erodes sequence signals typically used to infer homology.

It is well known that the use of profile-to-sequence comparison methods performs better than classic sequence-to-sequence similarity methods in detecting highly divergent homologous relationships, particularly Hidden Markov Models (HMMs) [12]. It has also been suggested that protein domains retain key information about the function of a protein, and that protein functionality is more conserved among true orthologues rather than between paralogues sequences [13]. Domain architecture, understood as the linear sequence of non-overlapping domains of a protein, appear to be key in detecting functional and then orthology relationships between diverging sequences [14–16]. Accordingly to the above, some strategies have been proposed in order to use domain architecture information to detect gene families and build pangenomes [17–19].

On the other hand, homology relationships between genes which still keep the same domain architecture even if they are paralogues and not true orthologous, or proteins which do not have known domains at all, can not use the above criteria and other strategies have to be applied. Best reciprocal blast hit has been shown to not be an optimal methodology to detect true orthologues [20], and gene context cannot be used neither to distinguish between them as proposed by some of the cited software [21,22] because genomic rearrangements are usual and the context get lost easily between distantly related strains. As consequence, it became necessary to apply orthology and paralogy definitions [23,24] in the design of a method to split paralogues from orthologues, as has been already suggested [13,25,26].

We thus present Pewit (Pangenome Estimation Walks Inside Taxonomy), a tool specifically designed to accurately reconstruct bacterial pangenomes at higher taxonomic ranks. Pewit builds the pangenome from a set of bacterial assemblies representing any taxon from phylum to species by implementing the following main steps: i) all genes are extracted from GFF3 files, ii) genes sharing the same protein domain architectures are considered as functionally equivalent, hence they are clustered together as homologs, iii) protein-to-protein comparisons are performed to cluster remaining genes lacking protein domains, iv) resulting cluster sets are then splited applying a tree-pruning algorithm that is used to split paralogs by identifying minimum subtrees, and v) singletons (genes occurring in a single genome) are refined by comparing against previously built clusters. Pewit outputs a panmatrix that describes gene cluster counts in each genome and optionally writes gene cluster sequences and a core genome alignment. Also, companion methods are provided to automatically explore pangenome summary statistics and standard data visualizations. Pewit is an open-source package written in R available at `http://github.com/iferres/pewit`.

We first evaluated Pewit using simulated datasets. To address this, we ran Pewit and other softwares (Roary[22], Micropan [19] and FindMyFriends[27]) on pangenome datasets generated under both the infinitely many genes model and the neutral model. This allowed us to simulate taxonomic depth by increasing the number of generations, mutation rate, and gene gain/loss rates. Figure **??**A shows that Pewit...

We then used Pewit to reconstruct the pangenomes of real life datasets comprising gram-negative and gram-positive bacteria from phylum to species. Figure 1 shows that all softwares produce comparable results in the recovery of core genes at the species level, but Pewit systematically outperforms the rest when increasing taxonomic depth from genus to phylum. Pewit also produces moderate pangenome sizes in comparison with softwares designed to work with phylogenetically close genomes, like Roary. This is expected since spurious cluster splitting due to sequence divergence at higher taxonomic ranks can cause an increase in the total number of gene clusters, an issue that Pewit resolves using domain architectures instead of sequence identity. The same tendency is observed for the number of singletons, where Roary and FindMyFriends return more singletons at higher ranks due to the impossibility of grouping them based on sequence identity. Pewit running time scales linear according to the number of analyzed genomes (Fig. 2a) and genome size (Fig. 2b). Despite Pewit is slower than Roary and FindMyFriends, which implement fast heuristics to reduce the number of sequence comparisons, it performs better than approaches implemented in Micropan.

As a working example of how Pewit improves the discovery of functional patterns in distant pangenomes, we analyzed the pathogenic clade of the genus *Leptospira*. In a previous work, we reported 12 novel *Leptospira* species [28] some of which presented a low-virulence phenotype using *in vivo* models of infection, but were phylogenetically placed within the pathogenic clade which encloses most virulent species reported so far [5]. In that work, a pre-release version of Pewit was applied to uncover substantial differences in accessory gene repertories of low-virulent species in comparison with virulent species. Here, we re-analyzed this dataset with Pewit and Roary to show how our approach optimizes the discovery of accessory genes associated to the observed low-virulent or virulent phenotypes. Figure 3a shows a discriminant analysis of principal components (DAPC) evidencing that accessory gene patterns obtained with Pewit can significantly discriminate low-virulent from virulent species (p = 6.4e-4, Wilcoxon), while Roary fails to recover these patterns (p = 0.41, Wilcoxon). Additionally, Pewit is unique in its capacity to find individual genes that completely discriminate between both groups of species (Fig. 3b, 3c, ,3d).

By optimizing pangenome reconstruction at taxonomic levels different from species with Pewit, we enable the automatic comparison of distant bacterial genomes with unprecedented comprehensiveness and accuracy. The increasing availability of whole-genome sequences and the recent advances in understanding of functional interactions between members of complex microbial communities, require the development and refinement of tools for the identification of genomic features associated to bacterial taxa. Accordingly, Pewit can characterize shared gene sets as well as taxon-specific gene content being resilient to taxonomic depth, which allows better understanding of genomic evolution and niche adaptation.

Methods

General workflow. Pewit is a tool for reconstruction of bacterial pangenomes optimized to characterize shared and taxon-specific gene contents at higher taxonomic ranks. As input, Pewit requires GFF3 files for each genome of interest, as typically returned by Prokka annotation software [29]. Pewit is written in R [30] but also requires HMMER [31], MAFFT [32] and MCL [33] as external software dependencies. The pipeline also depends on a collection of Hidden Markov Models (HMMs) of protein domains obtained from Pfam [34]. Pewit consists in four main steps: (i) clustering of genes based on shared domain architectures, (ii) clustering of genes lacking protein domains based on sequence similarity, (iii) identification of paralogous genes, and (vi) refinement of singletons against previously built gene clusters. These steps generate abundance profiles and extract gene sequences of each cluster. The result is stored in a pangenome object that can be post-processed with many provided functions to perform statistical and comparative analyses and produce standard data visualization.

Pangenome generation. First step involves the screening of every gene (translated into amino acid sequences) from each genome against protein domain Hidden Markov Models (HMMs) from Pfam [34] using HMMER v3.1b2 [31]. For each gene, its domain architecture is recorded (presence and linear order of protein domains). If overlapping protein domains are reported, both are kept if they belong to different functional clans, alternatively, for overlapping domains that belong to the same functional clan only the best match is kept. Then, coarse clusters are generated by grouping genes with identical architectures. Second, Pewit attempts to group genes lacking conserved protein domains using an heuristic based on portein-to-protein comparisons with phmmer (HMMER 3.1b2) and MCL. After these two steps, a tree-pruning algorithm is implemented using ape [35] and phangorn [36] to identify and split paralogous genes. Briefly, each gene cluster is automatically aligned and a mid-point-rooted Neighbor-Joining tree is generated. Then, paralogues are identified in each gene-tree by i) detecting nodes whose descendants all belong to the same genome and ii) splitting the tree in many subtrees as necessary to achieve the minimum set of subtrees with just one tip per genome. Finally, singletons generated in the previous steps are refined by trying to reallocate them to previously generated clusters by comparing each singleton against cluster-specific HMMs using hmmsearch (HMMER 3.1b2). Supplementary figure 1a shows a schematic representation of main steps in Pewit pipeline and suppl. figure 1b describes main steps of the tree-pruning algorithm.

Pangenome simulation. To test and compre the performance of Pewit we simulated different pangenomes using pansimulatoR (`http://github.com/iferres/pansimulatoR`), an in-house R package that generates sequences following the infinitely many genes model (IMG) described in Bumdicker et al. (2012)[3], and the neutral model described by Kimura (1983)[37]. Briefly, a random coalescent tree is generated and the expected per branch gene gain/loss events are simulated according to an effective population size (interpreted as the number of generations), branch length, gene gain rate per generation and gene loss rate per generation. Since the IMG model is used, genes are only transmitted vertically from generation to generation. Point mutations are simulated following a similar process, according to the effective population size, branch length, and mutation rate. Mutations are then distributed along sites with uniform probability. Mutations producing stop codons are avoided. Finally, sequences are sampled from a reference set of genes as the starting point (MRCA), and a pangenome is generated following the simulated evolutionary history over those sampled sequences. Pangenomes were simulated sampling 10 individuals and varying the number of generations from 25 million to 250 million, at a constant mutation rate of $5x10^{-10}$ mutations per site per generation, fixing the gene gain rate per generation at $2x10^{-6}$ and varying the gene loss rate per generation from $1x10^{-6}$ to $2x10^{-6}$.

Pangenome reconstruction using real datasets. To further test and compare Pewit we built real genomic datasets considering both gram-positive and gram-negative bacteria (**Supplementary Table ??**). For gram-positives, we randomly chose 50 public genomes at each taxonomic rank belonging to phylum *Actinobacteria*, order *Actinomycetales*, family *Corynebacteraceae*, genus *Corynebacterium* and species *C. diphteriae*. For gram-negatives, we followed the same criteria but using genomes of phylum *Proteobacteria*, order *Campylobacteriales*, family *Campylobacteraceae*, genus *Campylobacter* and species *C. fetus*. At each taxonomic rank, pangenomes were reconstructed from five independent samples of ten genomes. Results obtained using Pewit were compared to FindMyFriends [27] with default parameters, Micropan [19] based on BLAST+ [38] or HMMER [31], and Roary [22] using the sequence identity parameter set to 95% (default) or 70% to allow clustering of more divergent sequences. Additional comparisons were performed to evaluate running time using 10 to 150 previously reported *C. fetus* genomes [6], increasing sample size by 10 genomes, and sampling 10 times on each step. The BLAST+ all versus all approach used by Micropan was computationally intractable with increasing number of genomes, hence it is not presented. Running times were also tested against genome size. Each software was ran over datasets consisting of 10 randomly chosen genomes from different species with distinct average genome size (**Supplementary Table ??**). All evaluations were performed on a single processor on an Intel Xeon CPU E7-8837 (2.67 GHz), 250 Gb of RAM.

Identification of virulence-associated genes in *Leptospira*. To show the capabilities of Pewit, we apply the methodology on a *Leptospira spp.* sub dataset used in Thibeaux et al. (2018)[5]. This subset consists in the *Leptospira* pathogenic clade, which includes two sub-clades that are associated with differential virulence potential. Eight species conform the so-called “low-virulence” sub-clade, and nine the “virulent” one. Both Pewit and Roary (`-i 70`) were run against this dataset. Since Roary is expected to retrieve less divergent clusters of orthologous, also less discriminative genes between the two sub-clades are expected to be found. To check this hypothesis a discriminant analysis of principal components (DAPC) was performed using adegenet R package [39,40]. Genes contribution to the principal component were analyzed, and clusters identified by Pewit which original variables contribute 0.0005 or more to the principal component were arbitrarily selected and plotted as a heatmap asociated to a clustering dendrogram with the presence or absence of each selected gene on each genome. A complete list of identified discriminative genes can be found in **Supplementary Table ??**. Pangenome rarefaction curves (Supplementary figure 2a) and gene content distribution barplots (Supplementary figure 2b) for this case were also generated, as well as a pangenome openess analysis [41,42].

References

1. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: implications for the microbial pan-genome. *Proceedings of the National Academy of Sciences* **102**, 13950–13955 (2005).

2. Vernikos, G., Medini, D., Riley, D. R. & Tettelin, H. Ten years of pan-genome analyses. *Current opinion in microbiology* **23**, 148–154 (2015).

3. Baumdicker, F., Hess, W. R. & Pfaffelhuber, P. The infinitely many genes model for the distributed genome of bacteria. *Genome biology and evolution* **4**, 443–456 (2012).

4. McNally, A. *et al.* Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. *PLoS*

genetics **12**, e1006280 (2016).

5. Thibeaux, R. *et al.* Deciphering the unexplored leptospira diversity from soils uncovers genomic evolution to virulence. *Microbial genomics* **4** (2018).

6. Iraola, G. *et al.* Distinct campylobacter fetus lineages adapted as livestock pathogens and human pathobionts in the intestinal microbiota. *Nature Communications* **8**, 1367 (2017).

7. Smillie, C. S. *et al.* Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**, 241 (2011).

8. Kloesges, T., Popa, O., Martin, W. & Dagan, T. Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. *Molecular biology and evolution* **28**, 1057–1074 (2010).

9. Udaondo, Z., Duque, E. & Ramos, J.-L. The pangenome of the genus clostridium. *Environmental microbiology* **19**, 2588–2603 (2017).

10. Lugli, G. A. *et al.* Comparative genomic and phylogenomic analyses of the bifidobacteriaceae family. *BMC genomics* **18**, 568 (2017).

11. Farag, I. F., Youssef, N. H. & Elshahed, M. S. Global distribution patterns and pangenomic diversity of the candidate phylum latescibacteria(ws3). *Applied and environmental microbiology* AEM–00521 (2017).

12. Park, J. *et al.* Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *Journal of Molecular Biology* **284**, 1201–1210 (1998).

13. Gabaldn, T. & Koonin, E. V. Functional and evolutionary implications of gene orthology. *Nature Reviews. Genetics* **14**, 360–366 (2013).

14. Fong, J. H., Geer, L. Y., Panchenko, A. R. & Bryant, S. H. Modeling the Evolution of Protein Domain Architectures Using Maximum Parsimony. *Journal of Molecular Biology* **366**, 307–315 (2007).

15. Lee, B. & Lee, D. Protein comparison at the domain architecture level. *BMC bioinformatics* **10 Suppl 15**, S5 (2009).

16. Forslund, K., Pekkari, I. & Sonnhammer, E. L. Domain architecture conservation in orthologs. *BMC Bioinformatics* **12**, 326 (2011).

17. Snipen, L.-G. & Ussery, D. W. A domain sequence approach to pangenomics: applications to Escherichia coli. *F1000Research* **1** (2013).

18. Lukjancenko, O., Thomsen, M. C., Voldby Larsen, M. & Ussery, D. W. PanFunPro: PAN-genome analysis based on FUNctional PROfiles. *F1000Research* (2013).

19. Snipen, L. & Liland, K. H. Micropan: An r-package for microbial pan-genomics. *BMC bioinformatics* **16**, 79 (2015).

20. Koski, L. B. & Golding, G. B. The closest BLAST hit is often not the nearest neighbor. *Journal of Molecular Evolution* **52**, 540–542 (2001).

21. Fouts, D. E., Brinkac, L., Beck, E., Inman, J. & Sutton, G. PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic Acids Research* **40**, e172 (2012).

22. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).

23. Fitch, W. M. Distinguishing Homologous from Analogous Proteins. *Systematic Biology* **19**, 99–113 (1970).

24. Sonnhammer, E. L. L. & Koonin, E. V. Orthology, paralogy and proposed classification for paralog subtypes. *Trends in Genetics* **18**, 619–620 (2002).

25. Gabaldn, T. Large-scale assignment of orthology: back to phylogenetics? *Genome Biology* **9**, 235 (2008).

26. Tekaia, F. Inferring Orthologs: Open Questions and Perspectives. *Genomics Insights* **9**, 17–28 (2016).

27. Pedersen, T. L. Findmyfriends: microbial comparative genomics in r. *R package version 1.0.2* (2015). URL `http://bioconductor.org/packages/FindMyFriends/`.

28. Thibeaux, R. *et al.* Biodiversity of environmental leptospira: improving identification and revisiting the diagnosis. *Frontiers in microbiology* **9** (2018).

29. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

30. Team, R. C. *et al.* R: A language and environment for statistical computing (2014).

31. Eddy, S. R. Accelerated profile hmm searches. *PLoS computational biology* **7**, e1002195 (2011).

32. Katoh, K. & Standley, D. M. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* **30**, 772–780 (2013).

33. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research* **30**, 1575–1584 (2002).

34. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic acids research* **42**, D222–D230 (2013).

35. Paradis, E., Claude, J. & Strimmer, K. Ape: analyses of phylogenetics and evolution in r language. *Bioinformatics* **20**, 289–290 (2004).

36. Schliep, K. P. phangorn: phylogenetic analysis in r. *Bioinformatics* **27**, 592–593 (2010).

37. Kimura, M. *The neutral theory of molecular evolution* (Cambridge University Press, Cambridge [Cambridgeshire] ; New York, 1983).

38. Camacho, C. *et al.* Blast+: architecture and applications. *BMC bioinformatics* **10**, 421 (2009).

39. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC genetics* **11**, 94 (2010).

40. Jombart, T. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics (Oxford, England)* **24**, 1403–1405 (2008).

41. Heaps, H. S. *Information retrieval, computational and theoretical aspects* (Academic Press, 1978).

42. Tettelin, H., Riley, D., Cattuto, C. & Medini, D. Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology* **11**, 472–477 (2008).

Author contributions

G.I. supervised the work and originally conceived the idea. I.F., P.F., D.C. and H.N. contributed to the design of the work. I.F. implemented, validated and documented the software. P.F., D.C. and H.N. tested the software. G.I. and I.F. wrote the manuscript. All authors provided feedback, edited and approved the manuscript.

Acknowledgements Put acknowledgements here.

Competing Interests The authors declare that they have no competing financial interests.

Supplementary Information All supplementary tables and figures can be found at ....

Correspondence Correspondence and requests for materials should be addressed to G.I. (email: giraola@pasteur.edu.uy).

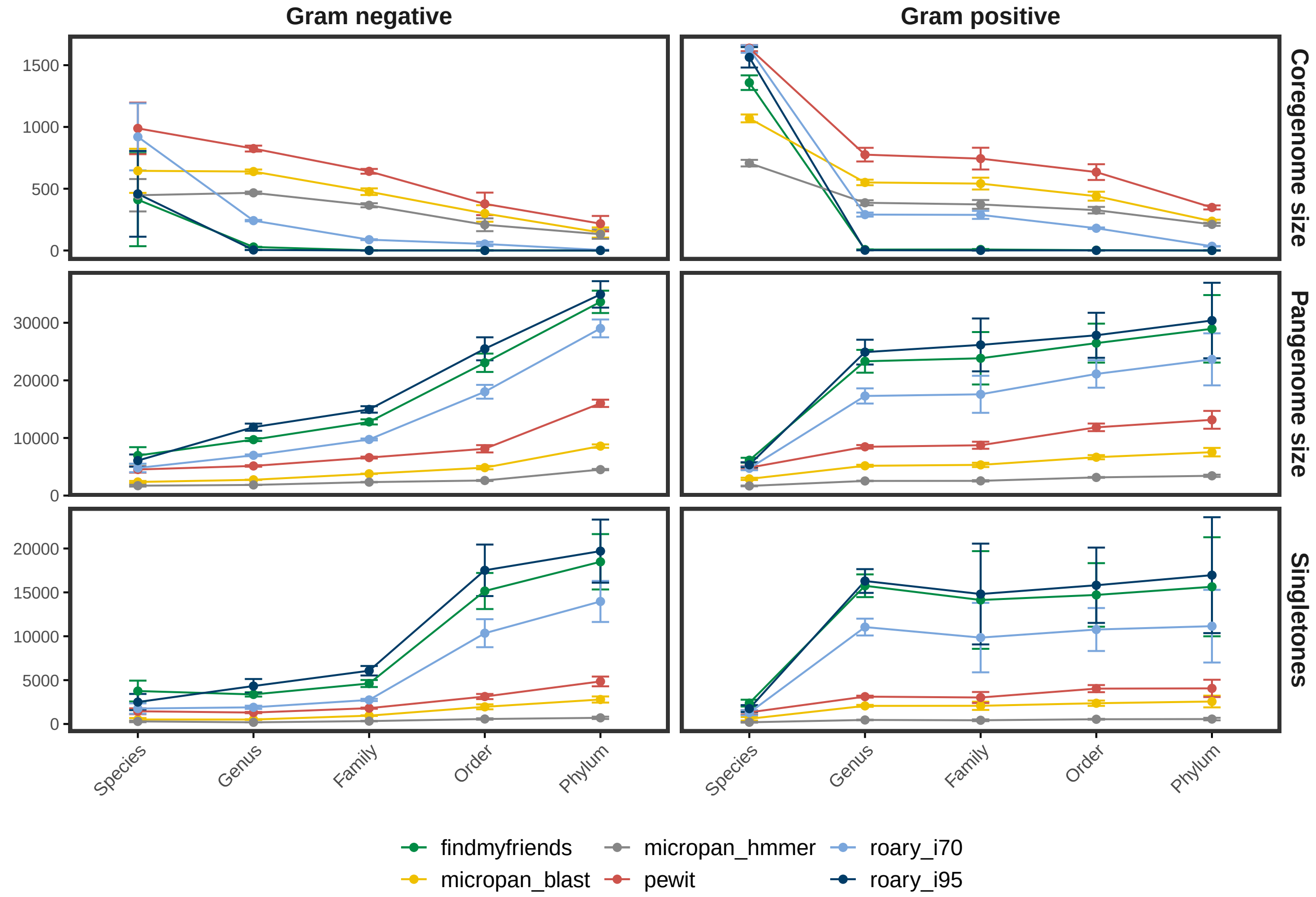
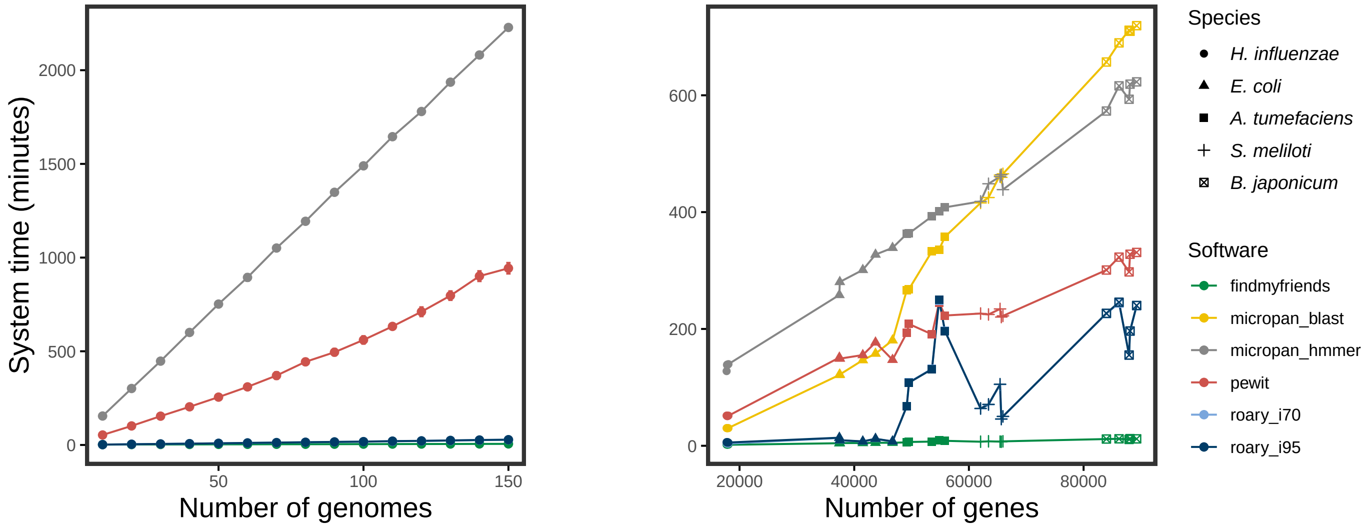
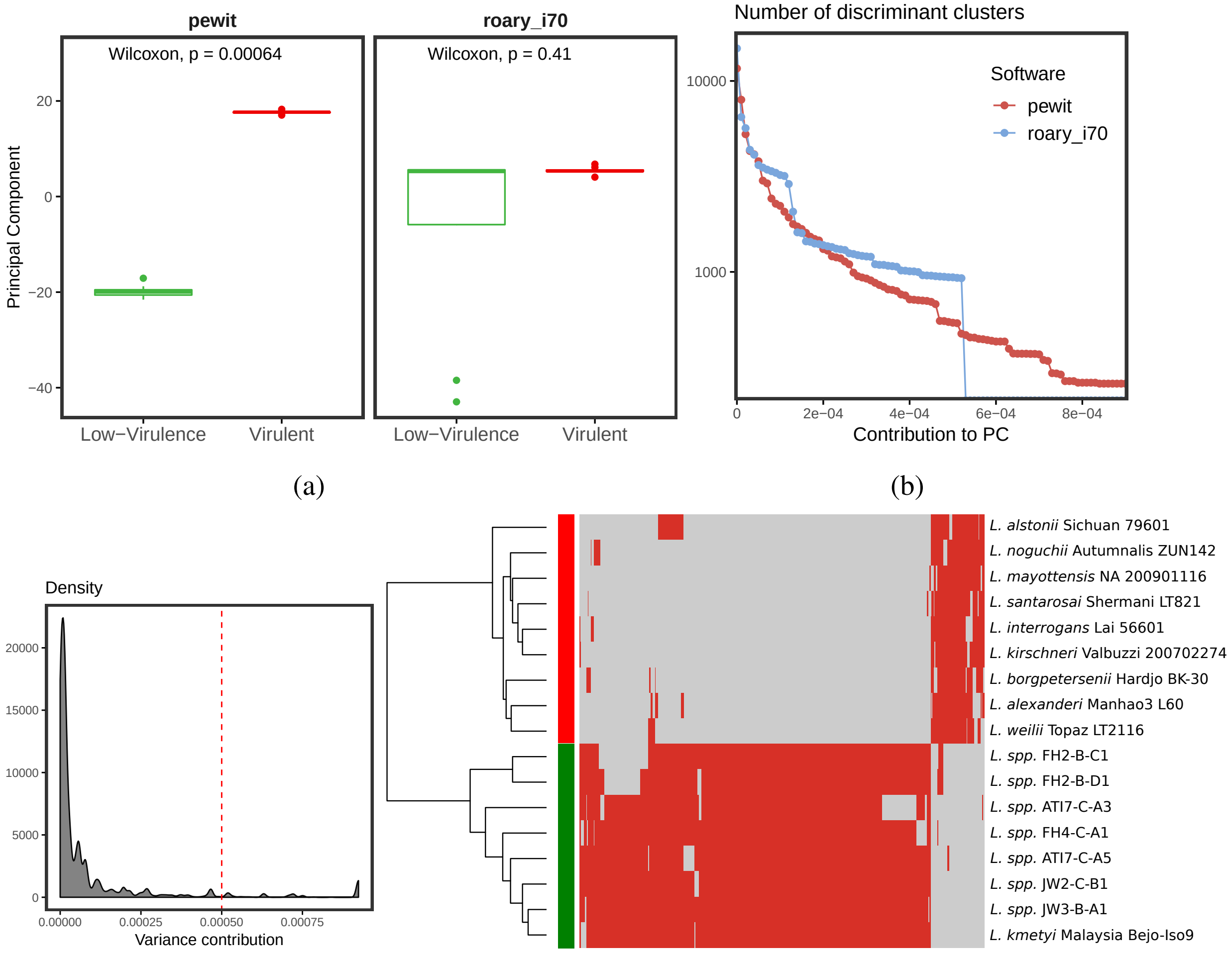


Figure 1: Pangenome reconstruction software comparison. Evaluation of coregenome size, pangenome size, and number of singletons, as a function of taxonomic rank. The analysis was done for two datasets, one corresponding to a gram negative clade (phylum *Proteobacteria*, order *Campylobacteriales*, family *Campylobacteraceae*, genus *Campylobacter* and species *C. fetus*), and the other for a gram positive dataset (phylum *Actinobacteria*, order *Actinomycetales*, family *Corynebacteraceae*, genus *Corynebacterium* and species *C. diphteriae*). The mean (points) asociated with a standard deviation (bars) are shown.



(a) (b)

Figure 2: Running time evaluation. (a) Running time (minutes) of different algorithms (except micropan_blast) over a growing number of *C. fetus* genomes. At each step, 10 samples of the corresponding number of genomes were taken. Mean (points) and standard deviation (bars) are shown. (b) Running time (minutes) as a function of pangenome size (number of genes). The pangenome of five bacterial species with different mean genome size were evaluated with each software, sampling ten genomes of the same species five times. Different point shapes correspond to different species.



(a)

(b)

(c)

(d)

Figure 3: *Leptospira* case study. (a) Distribution of genomes in the first principal component of DAPC, comparing Pewit and Roary (-i 70) results. (b) Scatter-plot of the number of discriminant clusters as a function of its variance contribution to the first principal component, comparing Pewit and Roary (-i 70) results. (c) Density-plot of the variance contribution to the first principal component of using the panmatrix generated by Pewit. (d) Heatmap of a subset of the panmatrix generated by Pewit showing only those genes (columns) which variance contributed more than $5x10^{-4}$ (dashed red line in 3c) to the first principal component.

4.4. Perspectivas

El paquete incluirá funciones para el análisis y visualización de pangenoma, además del método descrito, las cuales están en desarrollo activo al momento de escribir esta tesis.

Por otro lado, se encuentra en desarrollo un *software* (`https://github.com/iferres/pansimulatoR`) que simula pangenomas permitiendo obtener conjuntos de datos de acuerdo a una tasa de sustitución, tasas de ganancia y pérdida de genes, y un número de generaciones, a partir de un árbol coalescente simulado. Se planea poner a prueba los *software* utilizados ensayándolos contra conjuntos de datos simulados a distintos tiempos de coalescencia (número de generaciones) para contrastar el grado de sensibilidad y especificidad de los mismos a la hora de agrupar genes ortólogos a tiempos de coalescencia cada vez más largos. El manuscrito de `Pewit` incorporará estos resultados previo a someterse a revisión por pares.

En futuras versiones proyectamos mejorar la eficiencia en tiempo incorporando métodos de *preclustering*, así como desarrollar y proponer nuevos métodos estadísticos para describir pangenomas bacterianos.

Capítulo 5

Conclusión y perspectivas generales

Se presentaron tres herramientas desarrolladas en el lenguaje R, dos de las cuales se encuentran publicadas en revistas internacionales arbitradas [93, 94], y una tercera cuyo manuscrito se encuentra en estado avanzado previo a someterse a revisión.

Las tres herramientas han sido utilizadas con éxito para aportar al conocimiento del género *Leptospira*: `MLSTar` fue aplicado en un trabajo de tesis de maestría sobre la variabilidad genética de aislamientos bovinos de cepas del género en Uruguay, y que posteriormente se espera que madure en publicaciones arbitradas; también viene siendo utilizado por investigadores del grupo para la descripción genómica y poblacional de la especie *Campylobacter hyointestinalis*. Por otro lado, `Phylen` y `Pewit` fueron utilizados en trabajos donde nuevas especies del género *Leptospira* fueron descritas por primera vez [91, 92], aunque también vienen siendo utilizados internamente por el grupo en investigación sobre grupos bacterianos diversos, como *Sutterella*, *Mycobacterium* y *Helicobacter*. Cabe destacar, como puede entenderse de lo anterior, que las herramientas implementadas son genéricas y pueden aplicarse a casi cualquier clado bacteriano. En los tres casos los *software* son de libre uso y su código se encuentra público en el repositorio GitHub.

Para el caso de `Pewit`, como se mencionó en la sección 4.4, resta evaluar el desempeño frente a otros *software* utilizando datos simulados. Para dicho fin, se encuentra en etapas de desarrollo un paquete que permite simular pangenomas a partir de un árbol coalescente, una tasa de mutación, tasas de ganancia y pérdida de genes, y un número poblacional efectivo. Sin entrar en mayores detalles, el paquete se encuentra en fase de prueba en el sitio `https://github.com/iferres/pansimulatoR`.

Como perspectiva a mediano plazo se plantea el desarrollo de nuevas herramientas y la paulatina integración de las mismas en una *suite* que facilite los estudios de genómica comparativa microbiana a la comunidad científica. Pasos en esa dirección ya han sido dados con el desarrollo de "`mirecipe`: Microbial Genomics Pipeline" (url: `https://github.com/iferres/mirecipe`), un paquete aún en etapas muy tempranas de implementación, cuya filosofía se basa en el encapsulado de los archivos de salida de *software* populares de genómica microbiana en objetos estandarizados que sean fácilmente consultables y utilizables como entrada por sucesivos programas, todo usando el entorno de R. Dicho de otro modo, la idea es facilitar los análisis de genómica microbiana y generar un marco estandarizado que permita a otros bioinformáticos agregar "pasos" a la cadena de procesos que tomen los objetos de salida de pasos anteriores como entrada para generar nuevas salidas.

Por otro lado, una comunicación personal de un colaborador del Institut Pasteur de París da cuenta de una investigación en la cual se describen aislamientos que supondrían alrededor de 30 especies nuevas de *Leptospira*, casi duplicando la cantidad conocida hasta el momento. Esto, sumado a las dos publicaciones que se mencionan en esta tesis (apéndices A y B), sugiere que se trata de un género muy diverso y que podría estar aún en buena medida inexplorado. Una estrategia que nos planteamos implementar para explorar esta diversidad es recuperar genomas de *Leptospira* a partir de metagenomas de distintos ambientes que se encuentren disponibles en las bases de datos públicas. Una primera aproximación sería identificar los *reads* de los metagenomas que mappeen contra los genomas disponibles de *Leptospira spp*, y luego ensamblarlos *de novo*. Sobre esa metodología base pueden desarrollarse métodos más sofisticados utilizando las estrategias de *binning* aplicadas en metagenómica basadas generalmente en frecuencias de palabras ("*k-mers*"), de modo de no limitar la búsqueda a la similaridad de secuencia entre los *reads* y los genomas, y permitir la reconstrucción de genomas de especies no descritas. Así buscamos analizar la diversidad de *Leptospira spp.* de una forma menos dependiente de cultivo ya que son organismos en general difíciles de aislar y manipular si se quieren obtener sus secuencias genómicas. Esto nos permitiría también analizar su distribución, sus nichos, y las adaptaciones a éstos por parte de las distintas especies, de una forma mucho más extensiva.

Bibliografía

[1] Costa, F. y col. (2015). "Global Morbidity and Mortality of Leptospirosis: A Systematic Review". En: *PLOS Neglected Tropical Diseases* 9.9, e0003898.

[2] Adler, B. y Peña Moctezuma, A. de la (2010). "Leptospira and leptospirosis". En: *Veterinary Microbiology* 140.3, págs. 287-296.

[3] Ellis, W. A. (2015). "Animal leptospirosis". En: *Current Topics in Microbiology and Immunology* 387, págs. 99-137.

[4] Haake, D. A. y Levett, P. N. (2015). "Leptospirosis in humans". En: *Current Topics in Microbiology and Immunology* 387, págs. 65-97.

[5] Levett, P. N. (2001). "Leptospirosis". En: *Clinical Microbiology Reviews* 14.2, págs. 296-326.

[6] Bolin, C. A. (1996). "Diagnosis of leptospirosis: a reemerging disease of companion animals." En: *Seminars in veterinary medicine and surgery (small animal)* 11.3, págs. 166-171.

[7] Rohrbach, B. W. y col. (2018). "Effect of vaccination against leptospirosis on the frequency, days to recurrence and progression of disease in horses with equine recurrent uveitis". En: *Veterinary Ophthalmology* 8.3, págs. 171-179.

[8] Ellinghausen, H. C. y Mccullough, W. G. (1965). "Nutrition of Leptospira pomona and growth of 13 other serotypes: Fractionation of oleic albumin complex and a medium of bovine albumin and polysorbate 80". En: *American Journal of Veterinary Research* 26, págs. 45-51.

[9] Johnson, R. C. y Harris, V. G. (1967). "Differentiation of pathogenic and saprophytic letospires. I. Growth at low temperatures". En: *Journal of Bacteriology* 94.1, págs. 27-31.

[10] Carleton, O. y col. (1979). "Helix handedness of Leptospira interrogans as determined by scanning electron microscopy". En: *Journal of Bacteriology* 137.3, págs. 1413-1416.

[11] Faine (1999). *Leptospira and Leptospirosis.* Google-Books-ID: MU1WAAAAYAAJ. MediSci. 308 págs.

[12] Goldstein, S. F. y Charon, N. W. (1988). "Motility of the spirochete Leptospira". En: *Cell Motility and the Cytoskeleton* 9.2, págs. 101-110.

[13] Cullen, P. A., Haake, D. A. y Adler, B. (2004). "Outer membrane proteins of pathogenic spirochetes". En: *FEMS Microbiology Reviews* 28.3, págs. 291-318.

[14] Murray, G. L. y col. (2010). “Mutations affecting Leptospira interrogans lipopolysaccharide attenuate virulence”. En: *Molecular Microbiology* 78.3, págs. 701-709.

[15] Nahori, M.-A. y col. (2005). “Differential TLR recognition of leptospiral lipid A and lipopolysaccharide in murine and human cells”. En: *Journal of Immunology (Baltimore, Md.: 1950)* 175.9, págs. 6022-6031.

[16] Werts, C. y col. (2001). “Leptospiral lipopolysaccharide activates cells through a TLR2-dependent mechanism”. En: *Nature Immunology* 2.4, págs. 346-352.

[17] Picardeau, M. (2017). “Virulence of the zoonotic agent of leptospirosis: still terra incognita?” En: *Nature Reviews. Microbiology* 15.5, págs. 297-307.

[18] Weil, A. (1886). “Ueber einer eigenhuemliche, mit Milztumor, Icterus un Nephritis einhergehende, acute Infektionskrankheit”. En: *Deutsch Arch Klin Med* 39.209.

[19] Stimson, A. (1907). “Note on an organism found in yellow-fever tissue.” En: *Pub Health Rep (Washington)* 22.541.

[20] Inada, R. y col. (1916). “The etiology, mode of infection, and specific therapy of Weil’s disease (Spirochaetosis Icterohaemorrhagica).” En: *J Exp Med* 23, págs. 377-402.

[21] Ido, Y. y col. (1917). “The rat as a carrier of Spirochaeta icterohaemorrhagiae, the causative agent of Spirochaetosis icterohaemorrhagica.” En: *J Exp Med* 26, págs. 341-353.

[22] Noguchi, H. (1918). “Morphological characteristics and nomenclature of Leptospira (Spirochaeta) icterohaemorrhagiae (Inada and Ido).” En: *J Exp Med* 27, págs. 575-592.

[23] Faine, S. y Stallman, N. D. (1982). “Amended Descriptions of the Genus Leptospira Noguchi 1917 and the Species L. interrogans (Stimson 1907) Wenyon 1926 and L. biflexa (Wolbach and Binger 1914) Noguchi 1918”. En: *International Journal of Systematic and Evolutionary Microbiology* 32.4, págs. 461-463.

[24] Yasuda, P. H. y col. (1987). “Deoxyribonucleic Acid Relatedness between Serogroups and Serovars in the Family Leptospiraceae with Proposals for Seven New Leptospira Species”. En: *International Journal of Systematic and Evolutionary Microbiology* 37.4, págs. 407-415.

[25] Lehmann, J. S. y col. (2014). “Leptospiral Pathogenomics”. En: *Pathogens* 3.2, págs. 280-308.

[26] Fouts, D. E. y col. (2016). “What Makes a Bacterial Species Pathogenic?:Comparative Genomic Analysis of the Genus Leptospira”. En: *PLoS neglected tropical diseases* 10.2, e0004403.

[27] Zuerner, R. y col. (2000). “Technological advances in the molecular biology of Leptospira”. En: *Journal of Molecular Microbiology and Biotechnology* 2.4, págs. 455-462.

[28] Haake, D. A. y col. (2004). “Molecular evolution and mosaicism of leptospiral outer membrane proteins involves horizontal DNA transfer”. En: *Journal of Bacteriology* 186.9, págs. 2818-2828.

[29] Zuerner, R. L. (1991). “Physical map of chromosomal and plasmid DNA comprising the genome of Leptospira interrogans”. En: *Nucleic Acids Research* 19.18, págs. 4857-4860.

[30] Picardeau, M. y col. (2008). “Genome sequence of the saprophyte Leptospira biflexa provides insights into the evolution of Leptospira and the pathogenesis of leptospirosis”. En: *PloS One* 3.2, e1607.

[31] Harrison, P. W. y col. (2010). “Introducing the bacterial ’chromid’: not a chromosome, not a plasmid”. En: *Trends in Microbiology* 18.4, págs. 141-148.

[32] Snyder, E. E. y col. (2007). “PATRIC: The VBI PathoSystems Resource Integration Center”. En: *Nucleic Acids Research* 35 (Database issue), págs. D401-D406.

[33] Bulach, D. M. y col. (2006). “Genome reduction in Leptospira borgpetersenii reflects limited transmission potential”. En: *Proceedings of the National Academy of Sciences of the United States of America* 103.39, págs. 14560-14565.

[34] Xu, Y. y col. (2016). “Whole genome sequencing revealed host adaptation-focused genomic plasticity of pathogenic Leptospira”. En: *Scientific Reports* 6, pág. 20020.

[35] Medini, D. y col. (2005). “The microbial pan-genome”. En: *Current Opinion in Genetics & Development* 15.6, págs. 589-594.

[36] Maiden, M. C. y col. (1998). “Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms”. En: *Proceedings of the National Academy of Sciences of the United States of America* 95.6, págs. 3140-3145.

[37] Selander, R. K. y Levin, B. R. (1980). “Genetic diversity and structure in Escherichia coli populations”. En: *Science (New York, N.Y.)* 210.4469, págs. 545-547.

[38] Achtman, M. (2004). “Population structure of pathogenic bacteria revisited”. En: *International Journal of Medical Microbiology* 294.2, págs. 67-73.

[39] Maiden, M. C. J. y col. (2013). “MLST revisited: the gene-by-gene approach to bacterial genomics”. En: *Nature Reviews. Microbiology* 11.10, págs. 728-736.

[40] Didelot, X. (2010). “Sequence-Based Analysis of Bacterial Population Structures”. En: *Bacterial Population Genetics in Infectious Disease*. Wiley-Blackwell, págs. 37-60.

[41] Sokal, R. R. y Michener, C. D. (1958). *A Statistical Method for Evaluating Systematic Relationships*. Google-Books-ID: o1BlHAAACAAJ. University of Kansas. 30 págs.

[42] Saitou, N. y Nei, M. (1987). “The neighbor-joining method: a new method for reconstructing phylogenetic trees.” En: *Molecular Biology and Evolution* 4.4, págs. 406-425.

[43] Edwards y Cavalli-Sforza (1963). *The reconstruction of evolution.*

[44] Camin, J. H. y Sokal, R. R. (1965). “A Method for Deducing Branching Sequences in Phylogeny”. En: *Evolution* 19.3, págs. 311-326.

[45] Felsenstein, J. (2004). *Inferring phylogenies.* Sunderland, Mass: Sinauer Associates. 664 págs.

[46] Edwards, A. W. F. y Cavalli-Sforza, L. L. (1964). “Reconstruction of Evolutionary Trees”. En: *Phenetic and Phylogenetic Classification* 6, págs. 67-76.

[47] Felsenstein, J. (1981). “Evolutionary trees from DNA sequences: A maximum likelihood approach”. En: *Journal of Molecular Evolution* 17.6, págs. 368-376.

[48] Rannala, B. y Yang, Z. (1996). “Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference”. En: *Journal of Molecular Evolution* 43.3, págs. 304-311.

[49] Mau, B. y Newton, M. A. (1997). “Phylogenetic Inference for Binary Data on Dendograms Using Markov Chain Monte Carlo”. En: *Journal of Computational and Graphical Statistics* 6.1, págs. 122-131.

[50] Nascimento, F. F., Reis, M. d. y Yang, Z. (2017). “A biologist’s guide to Bayesian phylogenetic analysis”. En: *Nature Ecology & Evolution* 1.10, págs. 1446-1454.

[51] Ciccarelli, F. D. y col. (2006). “Toward automatic reconstruction of a highly resolved tree of life”. En: *Science (New York, N.Y.)* 311.5765, págs. 1283-1287.

[52] Sorek, R. y col. (2007). “Genome-wide experimental determination of barriers to horizontal gene transfer”. En: *Science (New York, N.Y.)* 318.5855, págs. 1449-1452.

[53] Darling, A. E., Mau, B. y Perna, N. T. (2010). “progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement”. En: *PloS One* 5.6, e11147.

[54] Tettelin, H. y col. (2005). “Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial ”pan-genome””. En: *Proceedings of the National Academy of Sciences of the United States of America* 102.39, págs. 13950-13955.

[55] Hogg, J. S. y col. (2007). “Characterization and modeling of the Haemophilus influenzae core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains”. En: *Genome Biology* 8.6, R103.

[56] Snipen, L., Almøy, T. y Ussery, D. W. (2009). “Microbial comparative pangenomics using binomial mixture models”. En: *BMC genomics* 10, pág. 385.

[57] Tettelin, H. y col. (2008). “Comparative genomics: the bacterial pan-genome”. En: *Current Opinion in Microbiology* 11.5, págs. 472-477.

[58] Heaps, H. S. (1978). *Information retrieval, computational and theoretical aspects.* Academic Press.

[59] Chao, A. (1987). "Estimating the Population Size for Capture-Recapture Data with Unequal Catchability". En: *Biometrics* 43.4, págs. 783-791.

[60] Snipen, L. y Liland, K. H. (2015). "micropan: an R-package for microbial pangenomics". En: *BMC bioinformatics* 16, pág. 79.

[61] Kislyuk, A. O. y col. (2011). "Genomic fluidity: an integrative view of gene diversity within microbial populations". En: *BMC Genomics* 12, pág. 32.

[62] Andreani, N. A., Hesse, E. y Vos, M. (2017). "Prokaryote genome fluidity is dependent on effective population size". En: *The ISME Journal* 11.7, págs. 1719-1721.

[63] Fitch, W. M. (1970). "Distinguishing Homologous from Analogous Proteins". En: *Systematic Biology* 19.2, págs. 99-113.

[64] Sonnhammer, E. L. L. y Koonin, E. V. (2002). "Orthology, paralogy and proposed classification for paralog subtypes". En: *Trends in Genetics* 18.12, págs. 619-620.

[65] Altschul, S. F. y col. (1990). "Basic local alignment search tool". En: *Journal of Molecular Biology* 215.3, págs. 403-410.

[66] Rivera, M. C. y col. (1998). "Genomic evidence for two functionally distinct gene classes". En: *Proceedings of the National Academy of Sciences of the United States of America* 95.11, págs. 6239-6244.

[67] Koski, L. B. y Golding, G. B. (2001). "The closest BLAST hit is often not the nearest neighbor". En: *Journal of Molecular Evolution* 52.6, págs. 540-542.

[68] Li, L., Stoeckert, C. J. y Roos, D. S. (2003). "OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes". En: *Genome Research* 13.9, págs. 2178-2189.

[69] Enright, A. J., Van Dongen, S. y Ouzounis, C. A. (2002). "An efficient algorithm for large-scale detection of protein families". En: *Nucleic Acids Research* 30.7, págs. 1575-1584.

[70] Page, A. J. y col. (2015). "Roary: rapid large-scale prokaryote pan genome analysis". En: *Bioinformatics* 31.22, págs. 3691-3693.

[71] Li, W. y Godzik, A. (2006). "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences". En: *Bioinformatics (Oxford, England)* 22.13, págs. 1658-1659.

[72] Park, J. y col. (1998). "Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods". En: *Journal of Molecular Biology* 284.4, págs. 1201-1210.

[73] Fong, J. H. y col. (2007). "Modeling the Evolution of Protein Domain Architectures Using Maximum Parsimony". En: *Journal of Molecular Biology* 366.1, págs. 307-315.

[74] Lee, B. y Lee, D. (2009). "Protein comparison at the domain architecture level". En: *BMC bioinformatics* 10 Suppl 15, S5.

[75] Forslund, K., Pekkari, I. y Sonnhammer, E. L. (2011). "Domain architecture conservation in orthologs". En: *BMC Bioinformatics* 12, pág. 326.

[76] Sonnhammer, E. L., Eddy, S. R. y Durbin, R. (1997). "Pfam: a comprehensive database of protein domain families based on seed alignments". En: *Proteins* 28.3, págs. 405-420.

[77] Snipen, L.-G. y Ussery, D. W. (2013). "A domain sequence approach to pangenomics: applications to Escherichia coli". En: *F1000Research* 1.

[78] Lukjancenko, O. y col. (2013). "PanFunPro: PAN-genome analysis based on FUNctional PROfiles". En: *F1000Research*.

[79] Gabaldón, T. (2008). "Large-scale assignment of orthology: back to phylogenetics?" En: *Genome Biology* 9.10, pág. 235.

[80] Gabaldón, T. y Koonin, E. V. (2013). "Functional and evolutionary implications of gene orthology". En: *Nature Reviews. Genetics* 14.5, págs. 360-366.

[81] Tekaia, F. (2016). "Inferring Orthologs: Open Questions and Perspectives". En: *Genomics Insights* 9, págs. 17-28.

[82] R Core Team (2016). *R: A Language and Environment for Statistical Computing.* bibtex[organization=R Foundation for Statistical Computing]. Vienna, Austria.

[83] Robinson, D. (2017). *The Impressive Growth of R.* Stack Overflow Blog. URL: `https://stackoverflow.blog/2017/10/10/impressive-growth-r/` (visitado 03-08-2018).

[84] Gentleman, R. C. y col. (2004). "Bioconductor: open software development for computational biology and bioinformatics". En: *Genome Biology* 5.10, R80.

[85] Huber, W. y col. (2015). "Orchestrating high-throughput genomic analysis with Bioconductor". En: *Nature Methods* 12.2, págs. 115-121.

[86] Russell, P. y col. (2018). "A large-scale analysis of bioinformatics code on GitHub". En: *bioRxiv*, pág. 321919.

[87] Portal INIA (2015). *Inauguración del Laboratorio de la Unidad Mixta del Institut Pasteur e INIA.* URL: `http://www.inia.uy:80/estaciones-experimentales/direcciones-regionales/inia-direcci%C3%B3n-nacional/inauguraci%C3%B3n-del-laboratorio-de-la-unidad-mixta-del-institut-pasteur-e-inia` (visitado 31-07-2018).

[88] Zarantonelli, L. y col. (2018). "Isolation of pathogenic Leptospira strains from naturally infected cattle in Uruguay reveals high serovar diversity, and uncovers a relevant risk for human leptospirosis". En: *PLOS Neglected Tropical Diseases* 12.9, e0006694.

[89] Zolfo, M. y col. (2017). “MetaMLST: multi-locus strain-level bacterial typing from metagenomic samples”. En: *Nucleic Acids Research* 45.2, e7.

[90] Powell, S. y col. (2014). “eggNOG v4.0: nested orthology inference across 3686 organisms”. En: *Nucleic Acids Research* 42 (Database issue), págs. D231-239.

[91] Puche, R. y col. (2018). “Leptospira venezuelensis sp. nov., a new member of the intermediate group isolated from rodents, cattle and humans”. En: *International Journal of Systematic and Evolutionary Microbiology* 68.2, págs. 513-517.

[92] Thibeaux, R. y col. (2018). “Deciphering the unexplored Leptospira diversity from soils uncovers genomic evolution to virulence”. En: *Microbial Genomics* 4.1.

[93] Ferrés, I. e Iraola, G. (2018a). “MLSTar: automatic multilocus sequence typing of bacterial genomes in R”. En: *PeerJ* 6, e5098.

[94] Ferrés, I. e Iraola, G. (2018b). “Phylen: automatic phylogenetic reconstruction using the EggNOG database”. En: *Journal of Open Source Software* 3.25, pág. 593.

Apéndice A

Leptospira venezuelensis sp. nov., a new member of the intermediate group isolated from rodents, cattle and humans.

INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY

TAXONOMIC DESCRIPTION
Puche *et al*., *Int J Syst Evol Microbiol*
DOI 10.1099/ijsem.0.002528

MICROBIOLOGY SOCIETY

Leptospira venezuelensis sp. nov., a new member of the intermediate group isolated from rodents, cattle and humans

Rafael Puche,[1]† Ignacio Ferrés,[2]† Lizeth Caraballo,[1] Yaritza Rangel,[1] Mathieu Picardeau,[3] Howard Takiff[1,4,]* and Gregorio Iraola[2,]*

Abstract

Three strains, CLM-U50^T, CLM-R50 and IVIC-Bov1, belonging to the genus *Leptospira*, were isolated in Venezuela from a patient with leptospirosis, a domestic rat (*Rattus norvegicus*) and a cow (*Bos taurus*), respectively. The initial characterisation of these strains based on the *rrs* gene (16S rRNA) suggested their designation as a novel species within the 'intermediates' group of the genus *Leptospira*. Further phylogenomic characterisation based on single copy core genes was consistent with their separation into a novel species. The average nucleotide identity between these three strains was >99 %, but below 89 % with respect to any previously described leptospiral species, also supporting their designation as a novel species. Given this evidence, these three isolates were considered to represent a novel species, for which the name *Leptospira venezuelensis* sp. nov. is proposed, with CLM-U50^T (=CIP 111407^T=DSM 105752^T) as the type strain.

Leptospirosis is an emerging and re-emerging worldwide distributed disease caused by spirochetes of the genus *Leptospira* [1]. Based on the clinical presentation of leptospirosis, the species within the genus have been historically divided in three groups: 'pathogens' that cause the most severe cases; 'intermediates' that cause a milder disease; and non-pathogenic 'saprophytes' that do not cause disease in humans or animals [2]. The application of 16S rRNA sequencing, multilocus sequence typing and comparative genomics has revealed a strong correlation between the three groups defined by disease severity and the phylogenetic position of the species within the genus. Accordingly, 'pathogens' were renamed as group I, 'intermediates' as group II and 'saprophytes' as group III [3]. At the time of writing this article, the genus *Leptospira* comprised 22 different species; 10 belonging to group I (*Leptospira interrogans*, *Leptospira kirschneri*, *Leptospira noguchii*, *Leptospira borgpetersenii*, *Leptospira alexanderi*, *Leptospira weilii*, *Leptospira santarosai*, *Leptospira kmetyi*, *Leptospira alstoni* and *Leptospira mayottensis*), five to group II (*Leptospira licerasiae*, *Leptospira wolffii*, *Leptospira fainei*, *Leptospira broomii* and *Leptospira inadai*) and seven to group III (*Leptospira idonii*, *Leptospira meyeri*, *Leptospira terpstrae*, *Leptospira biflexa*, *Leptospira vanthielii*, *Leptospira yanagawae* and *Leptospira wolbachii*).

Venezuela is located in the tropical region of South America, a suitable environment for the development of leptospirosis. Leptospires can survive longer in warm and humid conditions [4], making the disease particularly prevalent in wet tropical and subtropical regions [5]. We isolated three strains: (i) CLM-U50^T from the urine of a patient suffering a moderately severe leptospirosis characterised by fever, icteric and elevated liver enzymes, with vomiting, myalgia and arthralgia but no evidence of renal or pulmonary involvement. The patient was treated with antibiotics and recovered. (ii) CLM-R50 from the kidney of a rat (*Rattus norvegicus*) captured in the same region where the patient resided and, (iii) IVIC-Bov1 from the urine of a cow (*Bos taurus*) at a geographically close (within 40 km) farm from where the patient resided. These strains were grown in Ellinghausen–McCullough–Johnson–Harris medium (EMJH), and when examined under dark-field microscopy, showed the motility and helix shape characteristic of the members of the genus *Leptospira*.

The whole-genome sequences of CLM-U50^T, CLM-R50 and IVIC-Bov1 were obtained with an Illumina MiSeq platform. Libraries were reconstructed with the Nextera XT DNA Library Preparation Kit and sequenced with the MiSeq Reagent Kit v3 (pair-end reads of 150 bp). Reads were *de novo* assembled with SPAdes version 3.10.1 [6] and

Author affiliations: [1]Unidad de Estudios Genéticos y Forenses, Instituto Venezolano de Investigación Científica, Venezuela; [2]Unidad de Bioinformática, Institut Pasteur Montevideo, Uruguay; [3]Biology of Spirochetes Unit, Institut Pasteur Paris, France; [4]Unité de Genetique Mycobacterienne, Institut Pasteur, Paris, France.
***Correspondence:** Howard Takiff, htakiff@gmail.com; Gregorio Iraola, giraola@pasteur.edu.uy

Keywords: *Leptospirosis*; intermediates; Venezuela; whole-genome sequencing.

†These authors contributed equally to this work.

NCBI accession numbers for sequenced genomes: CLM-U50 T : NETS00000000, Ivic-Bov1: NFUQ00000000, CLM-R50: NFUP00000000.

scaffolded with a post-assembly improvement pipeline [7]. The quality of the final assemblies was analysed with QUAST version 4.0 [8] and then annotated with Prokka v1.12 [9]. The three genome assemblies were deposited in the GenBank under the accession numbers NETS00000000, NFUQ00000000 and NFUP00000000 for strains CLM-U50^T, IVIC-Bov1 and CLM-R50, respectively.

An initial phylogenetic characterisation of full-length 16S rRNA sequences was performed to determine the phylogenetic position of the three strains. The 16S rRNA gene sequences of CLM-U50^T, CLM-R50 and IVIC-Bov1 were extracted from their genomes and aligned with those of other leptospiral species. A neighbour-joining tree was built with MEGA6 [10] using 1000 repetitions to determine bootstrap values. The 16S rRNA gene sequence similarity between the three strains was 100 %, while the similarity with respect to the closest species, *L. licerasiae*, was 93 %. The three strains formed a distinct monophyletic clade within the 'intermediates' (group II) with *L. licerasiae* as a sister species (Fig. 1). To further confirm the phylogenetic position of the three strains within the genus *Leptospira*, we screened the genomes of CLM-U50^T, CLM-R50 and IVIC-Bov1, and available genomes of other leptospiral species

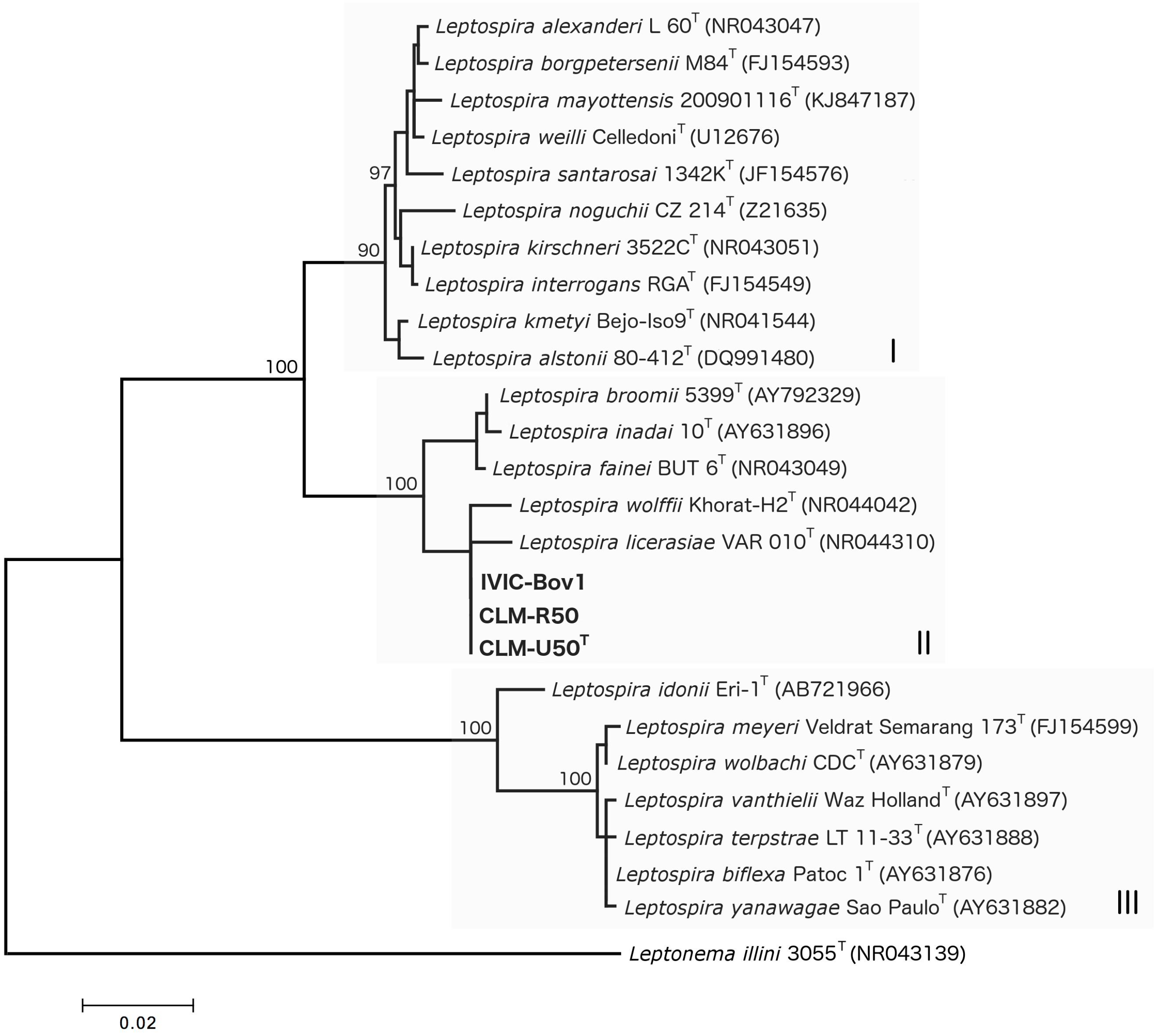

Leptospira alexanderi L 60T (NR043047)
Leptospira borgpetersenii M84T (FJ154593)
Leptospira mayottensis 200901116T (KJ847187)
Leptospira weilli CelledoniT (U12676)
97
Leptospira santarosai 1342KT (JF154576)
Leptospira noguchii CZ 214T (Z21635)
90
Leptospira kirschneri 3522CT (NR043051)
Leptospira interrogans RGAT (FJ154549)
Leptospira kmetyi Bejo-Iso9T (NR041544)
Leptospira alstonii 80-412T (DQ991480)
I
100
Leptospira broomii 5399T (AY792329)
Leptospira inadai 10T (AY631896)
Leptospira fainei BUT 6T (NR043049)
100
Leptospira wolffii Khorat-H2T (NR044042)
Leptospira licerasiae VAR 010T (NR044310)
IVIC-Bov1
CLM-R50
CLM-U50T
II
Leptospira idonii Eri-1T (AB721966)
100
Leptospira meyeri Veldrat Semarang 173T (FJ154599)
Leptospira wolbachi CDCT (AY631879)
100
Leptospira vanthielii Waz HollandT (AY631897)
Leptospira terpstrae LT 11-33T (AY631888)
Leptospira biflexa Patoc 1T (AY631876)
Leptospira yanawagae Sao PauloT (AY631882)
III
Leptonema illini 3055T (NR043139)
0.02

Fig. 1. 16S rRNA phylogeny. Phylogenetic tree built with 16S rRNA gene full-length sequences using the neighbour-joining method. Bootstrap values (1000 replicates) are displayed for most important internal nodes. The tree was rooted with *Leptonema illini* 3055^T. Phylogenetic groups I ('pathogens'), II ('intermediates') and III ('saprophytes') are shaded in light grey.

against the eggNOG version 3.0 database [11] specifically customised for spirochetes (spiNOG), using HMMER version 3.1b2 [12]. A total of 530 single copy genes were recovered as the core genome, a concatenated alignment was generated with MUSCLE [13], and phylogenetic reconstruction was performed with phangorn [14]. Fig. 2 shows the resulting phylogeny evidencing the membership of the three strains in the 'intermediates' group, and demonstrating their phylogenetic position with *L. licerasiae* as a sister species.

The average genomic relatedness between CLM-U50^T, CLM-R50 and IVIC-Bov1, and with respect to the other available sequenced *Leptospira* genomes was calculated using the average nucleotide identity (ANI) value [15] as previously implemented [16]. ANI can be used to replace DNA–DNA hybridisation (DDH), whereby the DDH species threshold of 70 % corresponds to an ANI of 95 % [17]. Strains CLM-U50^T, CLM-R50 and IVIC-Bov1 were extremely similar, showing an ANI >99 % to each other, confirming that they belong to the same bacterial species and could represent a single clone circulating in Venezuela. The ANI value was <89 % with respect to the genome of any other leptospiral species, thereby justifying their designation as representatives of a distinct new species (Table 1). The mean G+C content of the genomic DNA was 39.5 %, which falls within the range reported for the members of the genus *Leptospira* [18].

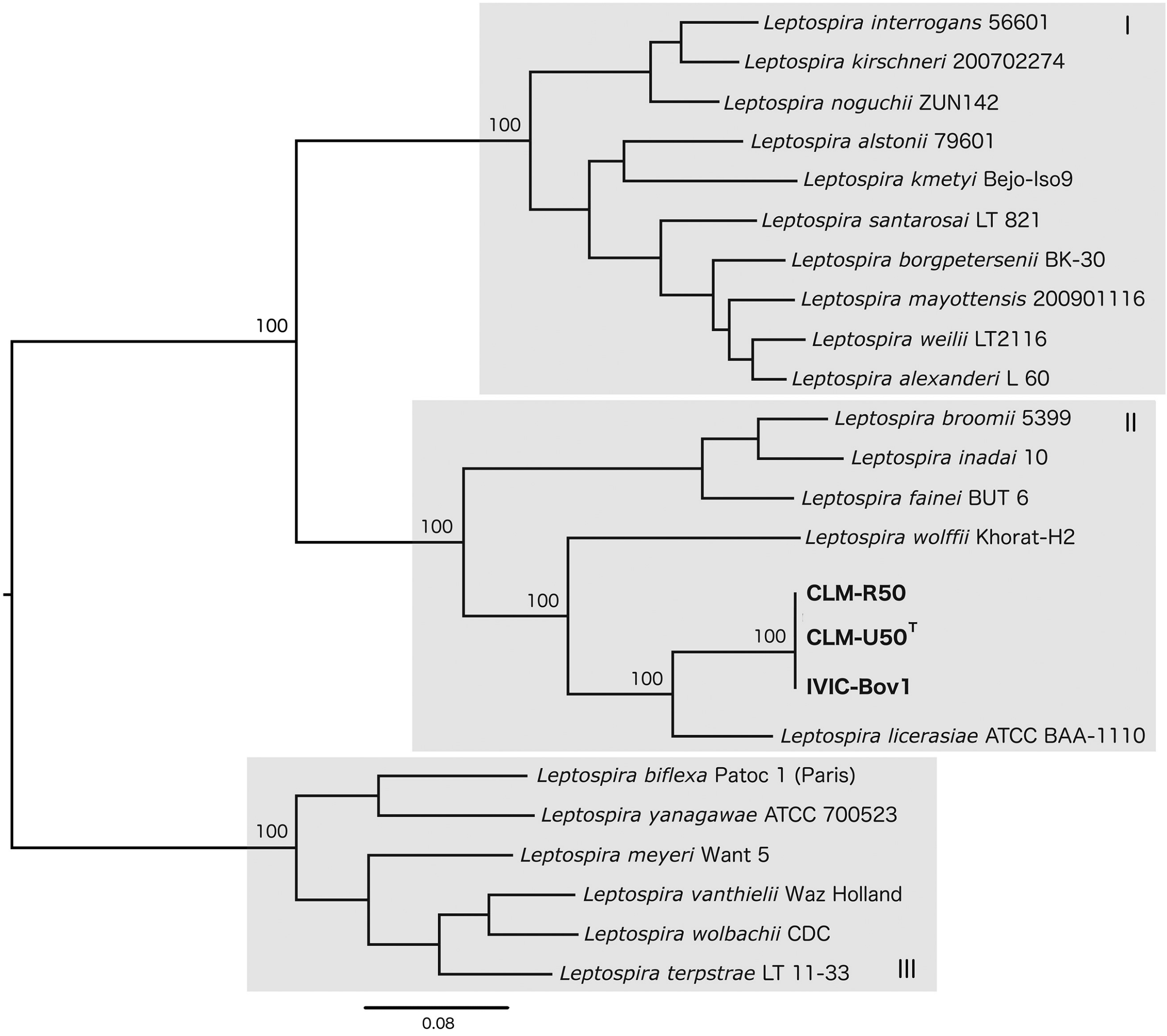

Leptospira interrogans 56601
I
Leptospira kirschneri 200702274
Leptospira noguchii ZUN142
100
Leptospira alstonii 79601
Leptospira kmetyi Bejo-Iso9
Leptospira santarosai LT 821
Leptospira borgpetersenii BK-30
Leptospira mayottensis 200901116
100
Leptospira weilii LT2116
Leptospira alexanderi L 60
Leptospira broomii 5399
II
Leptospira inadai 10
Leptospira fainei BUT 6
100
Leptospira wolffii Khorat-H2
CLM-R50
100
100
CLM-U50T
IVIC-Bov1
100
Leptospira licerasiae ATCC BAA-1110
Leptospira biflexa Patoc 1 (Paris)
Leptospira yanagawae ATCC 700523
100
Leptospira meyeri Want 5
Leptospira vanthielii Waz Holland
Leptospira wolbachii CDC
Leptospira terpstrae LT 11-33
III
0.08

Fig. 2. Single copy core genes phylogeny. Phylogenetic tree built with a concatenated alignment of 530 single copy genes present in all the analysed genomes, using the maximum-likelihood method. Bootstrap values (100 replicates) are displayed for most important internal nodes. Phylogenetic groups I ('pathogens'), II ('intermediates') and III ('saprophytes') are shaded in light grey.

Table 1. ANI analysis

ANI values based on reciprocal BLAST+ blastn comparisons of CLM-U50^{T}, CLM-R50 and IVIC-Bov1 genomes and the remaining sequenced 'intermediates' (group II) species.

		1	2	3	4	5	6	7	8
1	**CLM-U50^{T}**	100							
2	**CLM-R50**	**99.7**	100						
3	**IVIC-Bov1**	**99.7**	**99.7**	100					
4	*L. broomii* 5399^{T}	78.4	78.4	78.4	100				
5	*L. inadai* 10^{T}	78.5	78.5	78.5	90.3	100			
6	*L. fainei* BUT 6^{T}	78.1	78.1	78.1	87.6	86.9	100		
7	*L. wolffii* Khorat-H2^{T}	79.6	79.6	79.6	78.4	78.8	78.1	100	
8	*L. licerasiae* ATCC BAA-1110	87.5	87.5	87.5	78.7	78.6	77.9	79.8	100

Because CLM-U50^{T} was isolated from a human patient with leptospirosis, which is infrequent for an 'intermediate' species, we screened the genomes for previously characterised leptospiral virulence factors [18] using BLAST+ blastp [19] and Pfam [20]. We found putative distant homologs of *L. interrogans* proteins LipL32/LIC11352 (amino acid identity=67 %) and the collagenase ColA/LIC12760 (amino acid identity=75 %). Also, *L. venezuelensis* has a gene encoding a leptospiral immunoglobulin-like (Lig) protein, which contains bacterial immunoglobulin-like (Big) domains. This family of surface-exposed lipoproteins, including LigA, LigB and LigC, is present in pathogenic species, but not in saprophytes.

In conclusion, the genotypic and genomic analyses strongly support the designation of the three *Leptospira* strains isolated in Venezuela as representatives of a novel species within this genus. The phylogenetic position and the genetic relatedness measured by 16S, single copy core genes and ANI values are sufficient to define strains CLM-U50^{T}, CLM-R50 and IVIC-Bov1 as members of a new species distinct from all others currently described for the genus *Leptospira*. The finding that a novel species belonging to the 'intermediates' group can cause leptospirosis in humans, and the identification of key leptospiral virulence genes in its genome, highlight the need to further study this group of leptospires whose actual incidence, epidemiology and pathogenic potential in humans and animals remain largely unknown. Furthermore, the wide host range of this novel species provides a suitable scenario for its dissemination, requiring a deeper analysis of its epidemiological characteristics in Venezuela and a broader sampling effort to determine its presence and incidence in the rest of the world. The proposed name for this species is *Leptospira venezuelensis* sp. nov., with CLM-U50^{T} (=CIP 111407^{T}=DSM 105752^{T}) as the type strain.

DESCRIPTION OF *LEPTOSPIRA VENEZUELENSIS* SP. NOV.

Leptospira venezuelensis (ve.ne.zu.e.len′sis. N.L. fem. adj. *venezuelensis*, belonging to the country of Venezuela in South America).

Cells have helical morphology, finely thin and usually curved at each end, forming a semicircular hook. Cells are 6–20 µm long and 0.1 µm in diameter. Motile and strict anaerobic. Optimal growth temperature is 30 °C, although growth is also observed at 13 and 37 °C. Cells grow well in liquid and semi-solid medium (0.1–0.3 % agar) used for growth of *Leptospira*, such as the EMJH medium, Stuart supplemented with rabbit serum and Fletcher at pH from 7.2 to 7.4. Growth in the presence of antibiotics, such us 5-fluorouracil (50 $\mu g\,ml^{-1}$) and amphotericin B (20 $\mu g\,ml^{-1}$) was also observed. The genomic G+C content of the type strain is 39.5 %. The type strain, CLM-U50^{T} (=CIP 111407^{T} =DSM 105752^{T}), was isolated from the urine of a patient with moderately severe leptospirosis in Venezuela in 2010. Strains CLM-R50 and IVIC-Bov1 belong to the same species and were isolated from a rat (*Rattus norvegicus*) and a cow (*Bos taurus*), respectively. The ability to colonize kidneys in the Syrian golden hamster model of infection is observed.

Funding information

I.F. is supported by the Agencia Nacional de Investigación e Innovación (ANII, Uruguay) grant FSSA_X_2014_1_105252. This work was partially funded by the Fondo de Convergencia Estructural del Mercosur (FOCEM) grant COF 04/11.

Acknowledgements

The authors thank Pablo Fresia for providing insightful comments on the manuscript. The authors also thank to Armando Reyna for collection of bovine samples.

Conflicts of interest

The authors declare that there are no conflicts of interest.

References

1. **Levett PN.** Leptospirosis. *Clin Microbiol Rev* 2001;14:296–326.
2. **Adler B.** History of leptospirosis and *Leptospira*. *Curr Top Microbiol Immunol* 2015;387:1–9.
3. **Lehmann JS, Matthias MA, Vinetz JM, Fouts DE.** Leptospiral pathogenomics. *Pathogens* 2014;3:280–308.
4. **Hartskeerl RA, Collares-Pereira M, Ellis WA.** Emergence, control and re-emerging leptospirosis: dynamics of infection in the changing world. *Clin Microbiol Infect* 2011;17:494–501.

5. **Bharti AR, Nally JE, Ricaldi JN, Matthias MA, Diaz MM *et al.*** Leptospirosis: a zoonotic disease of global importance. *Lancet Infect Dis* 2003;3:757–771.
6. **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.
7. **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.
8. **Gurevich A, Saveliev V, Vyahhi N, Tesler G.** QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013;29:1072–1075.
9. **Seemann T.** Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.
10. **Tamura K, Stecher G, Peterson D, Filipski A, Kumar S.** MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* 2013;30:2725–2729.
11. **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.
12. **Eddy SR.** Accelerated profile HMM searches. *PLoS Comput Biol* 2011;7:e1002195.
13. **Edgar RC.** MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004;32:1792–1797.
14. **Schliep KP.** phangorn: phylogenetic analysis in R. *Bioinformatics* 2011;27:592–593.
15. **Konstantinidis KT, Tiedje JM.** Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci USA* 2005; 102:2567–2572.
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 Sys Evol Microbiol* 2016;66:3468–3476.
17. **Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P *et al.*** DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* 2007;57:81–91.
18. **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.
19. **Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J *et al.*** BLAST+: architecture and applications. *BMC Bioinformatics* 2009; 10:421.
20. **Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J *et al.*** The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* 2016;44:D279–D285.

Apéndice B

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

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

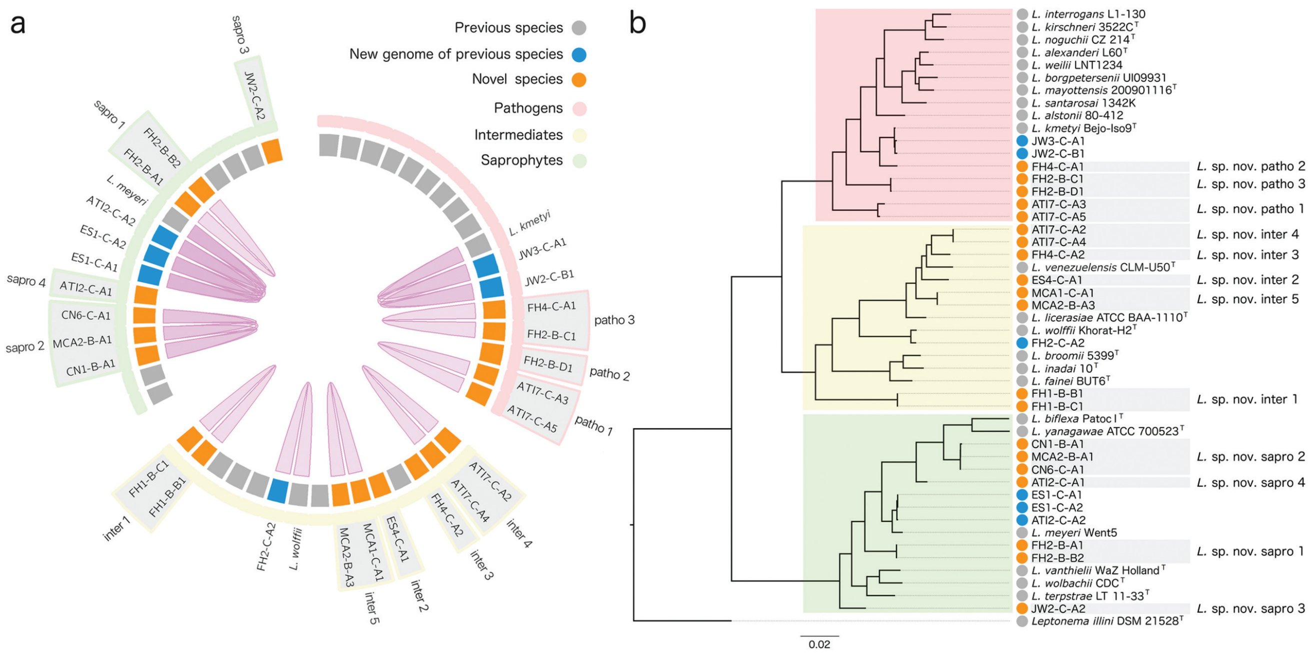


Fig. 1. Phylogenetic position of the novel species. (a) Circos diagram showing the relationships between leptospiral genomes based on overall genomic relatedness indices (OGRIs). The inner violet ribbons connect pairs of genomes if they share >95 % average nucleotide identity (ANI) and average amino 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 only been 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

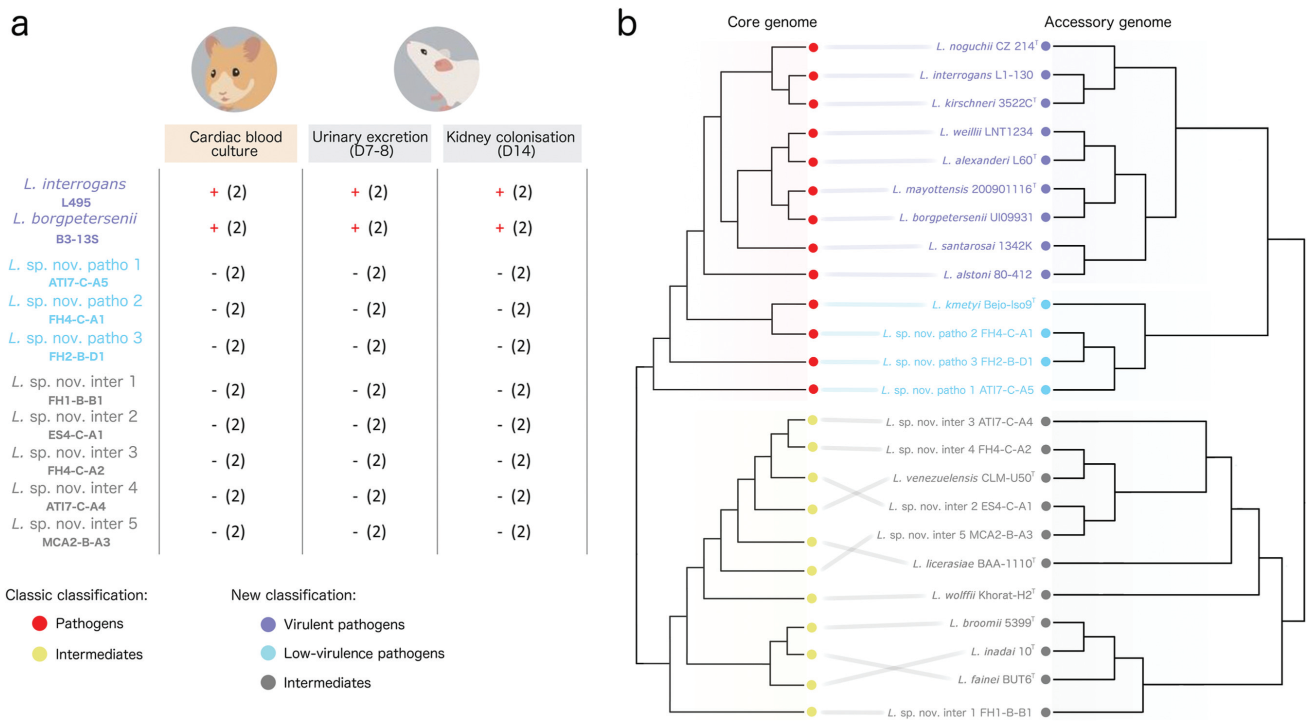

a
Cardiac blood culture
Urinary excretion (D7-8)
Kidney colonisation (D14)
L. interrogans L495 + (2) + (2) + (2)
L. borgpetersenii B3-13S + (2) + (2) + (2)
L. sp. nov. patho 1 ATI7-C-A5 - (2) - (2) - (2)
L. sp. nov. patho 2 FH4-C-A1 - (2) - (2) - (2)
L. sp. nov. patho 3 FH2-B-D1 - (2) - (2) - (2)
L. 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 borgpetersenI* 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

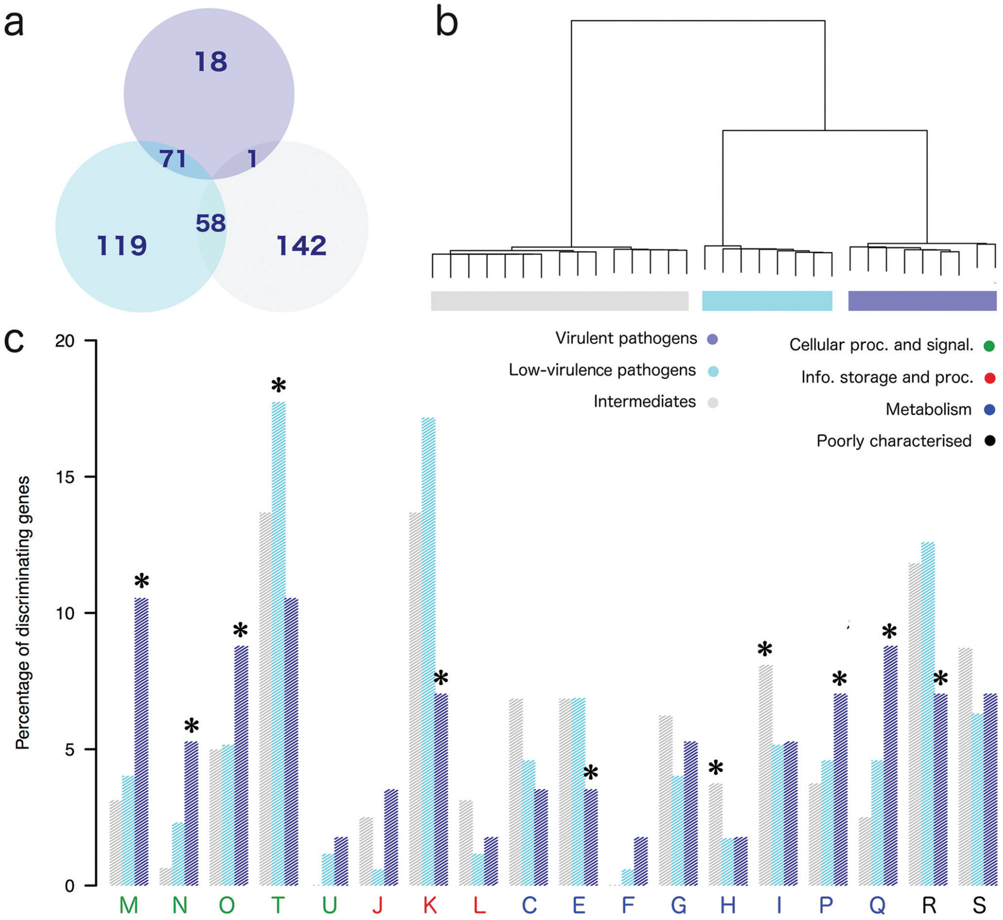


Fig. 3. Functional analysis of discriminating accessory genes. (a) Venn diagram showing the Bayesian identification of cluster-defining accessory genes from the pan-genome. (b) Clustering analysis based on Jaccard distances calculated from the presence/absence vectors of cluster-defining genes. (c) Barplots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

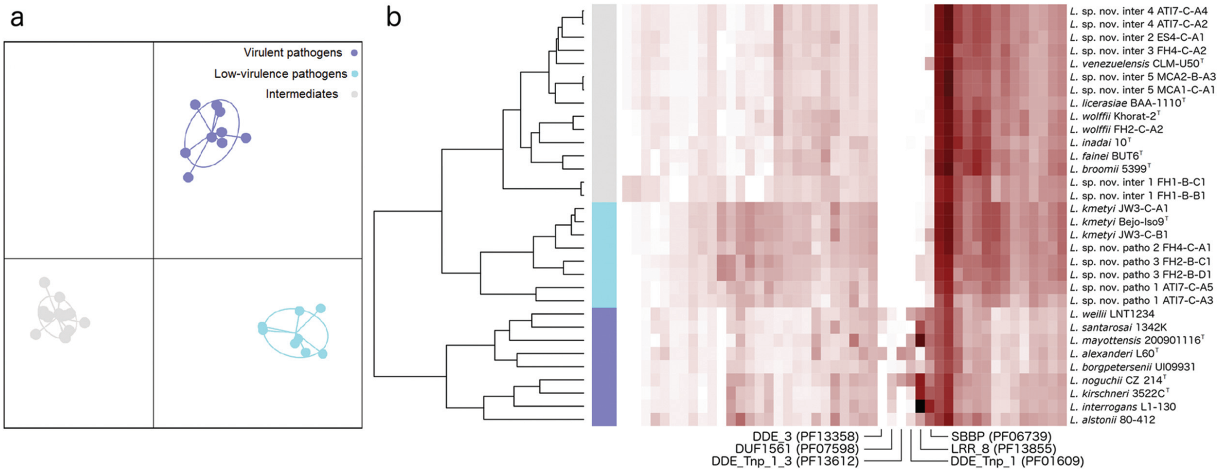


Fig. 4. Protein domains analysis. (a) Scatterplot showing the first and second discriminant functions obtained from the Discriminant Analysis of Principal Components (DAPC), performed with protein domain abundances extracted from the coding sequences of each genome. Groups are coloured according to the 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.

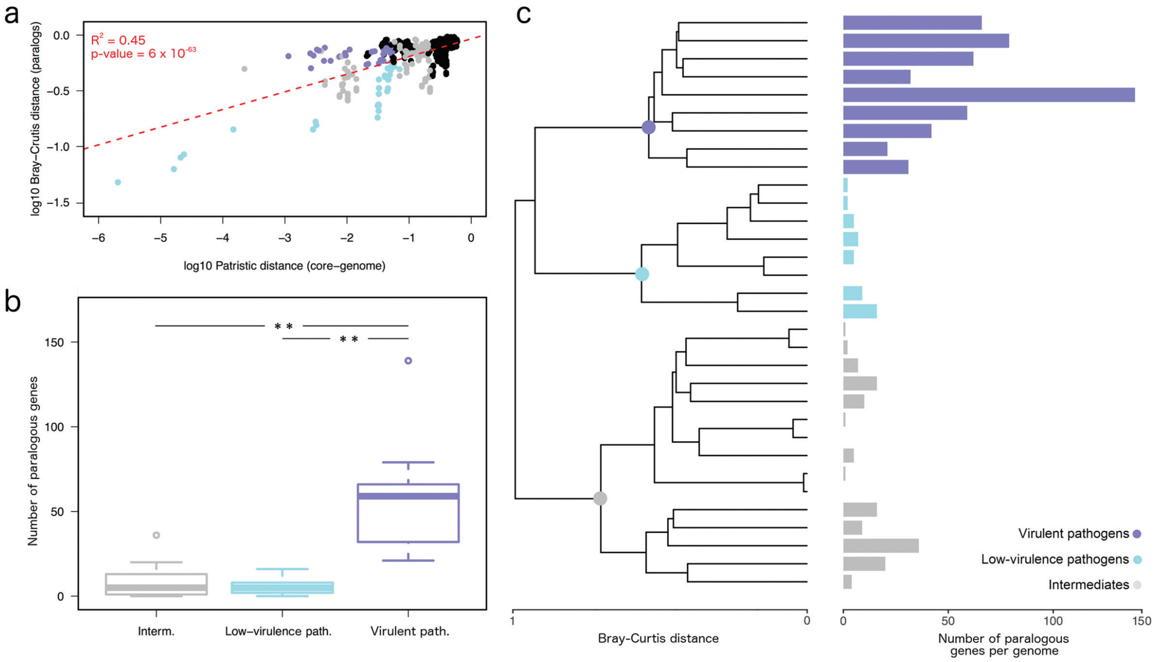


Fig. 5. Analysis of paralogous genes. (a) Linear regression showing the correlation between patristic distances calculated from the core genome phylogeny and Bray–Curtis distances calculated from the abundance patterns of paralogous genes. Dots are coloured according to virulence clusters when both genomes 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.