UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Genetic and genomic variability of *Legionella pneumophila*: applications to molecular epidemiology and public health

Leonor Sánchez Busó

PhD thesis

Director
Fernando González Candelas

June 2015

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Genetic and genomic variability of *Legionella pneumophila*: applications to molecular epidemiology and public health

Thesis submitted by

Leonor Sánchez Busó

Valencia, June 2015

DIRECTOR
Prof. Fernando González Candelas
TUTOR
Dr. Javier Forment Millet

This thesis has been developed at the *Fundación para la Investigación Sanitaria y Biomédica de la Comunidad Valenciana* (FISABIO - Salud Pública), *Genomics and Health* (recently *Infection and Health*) joint unit with the *Instituto Cavanilles de Biodiversidad y Biología Evolutiva* (Universidad de Valencia). Funding has been received from the Spanish Ministry of Education, Culture and Sport through the FPU program (Ref. AP2009-5234) and projects BFU2008-03000 and BFU2011-24112 from the Spanish Ministry of Science and Innovation.

Agradecimientos/Acknowledgements

Quisiera agradecer a todos los que han hecho posible que esta tesis doctoral haya ido adelante, a los que han creído en mí y me han apoyado incondicionalmente desde el principio y a los que he conocido en el camino y me han dado fuerza.

En primer lugar, a mis padres, hermanos y abuelos, porque esto no es sólo resultado de cuatro años de investigación, sino de casi 28 dándome todo el cariño, el apoyo y los medios para llegar lejos. Por supuesto a Pedro, porque 11 años no son pocos, siempre ha creído en mí y me ha animado a seguir adelante.

A todos los que dejé atrás en el cole y que también me hicieron crecer. A Sor Milagros, porque sabemos que algo tuvo que ver en todo esto, y a Sor Carmen que siempre se acordaba de mí y creía que yo podría llegar lejos.

A Fernando, porque como director de tesis, sin él esto no habría sido posible. Por acogerme en su grupo, confiar en mí desde el primer momento y enseñarme tanto.

A Javi Forment, por aceptar tutorizarme y ser siempre tan atento con mis progresos.

A Iñaki, porque también ha sido muy importante la confianza que ha tenido en mí desde el principio. Sin él no habría *Nature Genetics* ni viaje a los Alpes Suizos en enero.

A todos mis compañeros del área de Genómica y Salud del CSISP/FISABIO, investigadores principales, postdocs, técnicos, predocs y estudiantes. No me atrevo a nombraros a todos y dejarme algún nombre, pero tampoco puedo evitar hacer mención especial a mis compañeros más

cercanos, fundadores de la sala 1 y que me acogieron desde el principio: Ana Elena, Pedro, Raúl, Ana D, Peris y Bea. Y las que fueron llegando con el tiempo, Ana DJ y Anny. Con vosotros he pasado momentos inolvidables y sé que seremos amigos para toda la vida. Pero no me puedo olvidar de los nuevos relevos de la sala, que me han dado ánimo en esta última etapa y que van a dejar el listón muy alto. Mil gracias a Manoli, mi salvadora.

A Belén Picó, Javier Forment, Ximo Cañizares y Santi Vilanova, por confiar en mí y darme la oportunidad de iniciarme en la docencia.

I would like to thank specially Julian Parkhill, Simon Harris, Sophia David and their group at the Wellcome Trust Sanger Institute, because of a fruitful and unforgettable experience during my visit in 2014. For accepting me in their group and giving me the opportunity of a Postdoctoral position.

A Nacho, por ser un poco nuestro consejero profesional y amigo. Porque de la estancia en Cambridge, que tanto me costó decidirme a pedir, me llevo una experiencia inolvidable, un postdoc en el que tengo muchas expectativas, y también grandes amigos, que ya son una colonia española que en breve sumará dos más.

A Jose y Eva, porque con vosotros he subido hasta lo más alto, me habéis dado fuerza para continuar y nos habéis acogido como uno (dos) más desde el principio. Porque sí, en el gimnasio se pueden hacer amigos. Con vosotros y el resto de compis he podido superarme a mí misma.

A Coby, por la tranquilidad que me da.

Todos vosotros juntos me habéis hecho crecer profesionalmente y también como persona. Han sido unos años inolvidables, la tesis pasará y seguiremos rumbos distintos durante unos años pero sé que vosotros seguiréis ahí. Muchas gracias a todos, de corazón.

Contents

Summary

Legionella pneumophila is a strictly environmental and opportunistic pathogen that can cause severe pneumonia after inhalation of aerosols with enough bacterial load. Outbreaks and sporadic cases are usually localized in temperate environments, and the reservoirs are often water-related sources where biofilms are created. The existence of non-cultivable forms of the bacteria increases the risk for public health, as culture-based methods may miss them, thus complicating the environmental investigations of the sources.

Genetic classification through the Sequence-Based Typing (SBT) technique allowed an increased discrimination among *L. pneumophila* strains compared to previous methods. SBT data can also be used for genetic variability and population structure studies, but a more exhaustive analysis can be performed using high-throughput genome sequencing strategies.

This thesis describes the use of both SBT and genomic sequencing to evaluate and provide solutions to different public health needs in *L. pneumophila* epidemiology. We have focused in the Comunidad Valenciana (CV), the second region in Spain with the highest incidence of Legionellosis, with special interest in the city of Alcoy, where recurrent outbreaks have occurred since 1998.

Firstly, SBT data were used to gain a deeper insight into the genetic variability and distribution of the most abundant Sequence Types (ST) in the CV area. We have shown that the level of variability in this region is comparable to that from other countries, revealing the existence of both locally and broadly extended profiles. Approximately half of the observed genetic diversity was found to result from geographical and temporal structure.

Secondly, *L. pneumophila* detection from environmental sources remains a challenge for public health. A comparison between water and biofilm samples using a sensitive touchdown PCR (TD-PCR) strategy revealed that the use of biofilms increased by ten-fold the detection rate. This method allowed evaluating the hidden uncultivable *L. pneumophila* diversity in the locality of Alcoy and the real-time investigation of a Legionellosis outbreak affecting a hotel in Calpe (Southeast of Spain) in 2012.

Thirdly, genomic sequencing was applied to a set of 69 strains isolated during 13 outbreaks occurred in Alcoy in the period 1999-2010, mainly the recurrent ST578. Higher intra-outbreak variability than expected was observed, pointing to the potential existence of multiple sources in this endemic area or high environmental diversity. Interestingly, above 98% of the genomic variability in this ST was found as being incorporated through recombination processes rather than through point mutations.

Finally, a metagenomic analysis of environmental biofilms from Alcoy revealed a microbial community dominated by *Proteobacteria*, *Cyanobacteria, Actinobacteria* and *Bacteroidetes*. Despite the known endemism of *Legionella* in this area, the genus was only found in a relative abundance ranging 0.01-0.07%, which explains the low recovery from environmental sources.

In summary, the results from this thesis can benefit public health efforts to control this pathogen in the environment, as we provide new insight into its molecular epidemiology, with immediate applications to surveillance and outbreak investigations.

Resumen

Legionella pneumophila es un patógeno oportunista estrictamente ambiental capaz de causar neumonía debido a la inhalación de aerosoles con suficiente carga bacteriana. Los brotes y casos esporádicos suelen producirse en ambientes templados y los reservorios encontrarse en zonas con agua donde pueden crearse biopelículas microbianas. La existencia de formas no cultivables de la bacteria aumenta el riesgo para la salud pública, ya que los métodos estándar basados en cultivo microbiológico no pueden detectarlas, complicando las investigaciones ambientales.

La clasificación genética basada en el método *Sequence-Based Typing* (SBT) permite un mayor poder de discriminación entre cepas de *L. pneumophila* en comparación con métodos previos. Los datos derivados del SBT pueden utilizarse para estudios de variabilidad genética y estructura poblacional. Sin embargo, puede llevarse a cabo un análisis más exhaustivo mediante técnicas de secuenciación genómica de alto rendimiento.

Esta tesis describe la utilización tanto de SBT como de secuenciación genómica para evaluar e incluso proponer soluciones a diferentes necesidades en salud pública relacionadas con la epidemiología de *L. pneumophila*. Nos centramos en la Comunidad Valenciana (CV), la segunda región en España con mayor incidencia de Legionelosis, con especial interés en la localidad de Alcoy, donde ocurren brotes de forma recurrente.

En primer lugar, utilizamos datos derivados de SBT para conocer mejor la variabilidad y la distribución de los perfiles genéticos (*Sequence Types*, ST) en el área de la CV. Mostramos que el nivel de variabilidad en sólo esta región es comparable a la de otros países, con perfiles extendidos local y globalmente. Aproximadamente la mitad de la diversidad genética observada se estima que procede de estructuración geográfica y temporal.

En segundo lugar, la detección de *L. pneumophila* a partir de fuentes ambientales sigue suponiendo un reto para la salud pública. En esta tesis realizamos una comparación entre la detección mediante *touchdown PCR* (TD-PCR) a partir de muestras de agua y biopelículas microbianas y mostramos que estas últimas proporcionan un aumento de 10 veces en la tasa de detección de la bacteria. Este método permitió evaluar la diversidad no cultivable de *L. pneumophila* en la localidad de Alcoy y la investigación a tiempo real de un brote en un hotel en Calpe (Sudeste de España) en 2012.

A continuación, aplicamos la secuenciación genómica a 69 cepas aisladas durante 13 brotes ocurridos en Alcoy en el período 1999-2010, principalmente el recurrente ST578. Se observó mayor variabilidad entre cepas de un mismo brote que la esperada, lo cual apunta a la existencia potencial de múltiples fuentes en este área, o alta diversidad ambiental. Además, se observó que más del 98% de la variabilidad genómica fue introducida por procesos de recombinación y no de mutación puntual.

Finalmente, se realizó un análisis metagenómico de biopelículas ambientales recogidas en Alcoy. Se encontró que la comunidad está dominada por *Proteobacteria, Cyanobacteria, Actinobacteria* y *Bacteroidetes*. A pesar del conocido endemismo de *Legionella* en el área, este género sólo se encontró en una abundancia relativa entre 0.01-0.07%, lo cual explica su baja tasa de recuperación a partir de muestras ambientales.

En resumen, los resultados de esta tesis pueden ser de utilidad para los programas de control de este patógeno llevados a cabo por las autoridades de salud pública, ya que proporcionan una nueva percepción de su epidemiología molecular, con aplicación inmediata a la vigilancia e investigación de brotes.

Resum

Legionella pneumophila és un patogen oportunista estrictament ambiental capaç d'ocasionar pneumònia degut a la inhalació d'aerosols amb la suficient carga bacteriana. Els brots i casos esporàdics solen ocórrer en ambients temperats, i els reservoris solen trobar-se en zones amb aigua on poden crear-se biopel·lícules microbianes. La existència de formes no cultivables del bacteri augmenten el risc per a la salut pública, ja que els mètodes estàndard basats en el cultiu microbiològic no poden detectar-les, complicant les investigacions ambientals.

La classificació genètica basada en el mètode *Sequence-Based Typing* (SBT) permet un major poder de discriminació entre soques de *L. pneumophila* en comparació amb previs mètodes. Les dades derivades del SBT poden utilitzar-se per a estudis de variabilitat genètica i estructura poblacional, però un anàlisis més exhaustiu pot dur-se a terme a través de tècniques de seqüenciació genòmica d'alt rendiment.

Esta tesis descriu la utilització tant del SBT com de la seqüenciació genòmica per a avaluar i proposar solucions a diferents necessitats en salut pública relacionades amb l'epidemiologia de *L. pneumophila*. Ens centrem en la Comunitat Valenciana (CV), la segona regió d'Espanya amb la major incidència de Legionel·losi, amb especial interès en la localitat d'Alcoi, on els brots ocorren de forma recurrent des de 1998.

Primer, hem utilitzat dades derivades del SBT per a conèixer millor la variabilitat i la distribució dels perfils genètics (*Sequence Types*, ST) en l'àrea de la CV. Mostrem que el nivell de variabilitat en només aquesta regió és comparable a la d'altres països, amb perfils estesos tant de forma local com més amplia. Aproximadament la meitat de la diversitat genètica observada s'estima que procedeix d'estructuració geogràfica i temporal.

Segon, la detecció de *L. pneumophila* a partir de fonts ambientals continua suposant un repte per a la salut pública. En aquesta tesis realitzem una comparació entre la detecció mitjançant *touchdown PCR* (TD-PCR) a partir de mostres d'aigua i biopel·lícules microbianes i mostrem que aquestes últimes proporcionen un augment de deu vegades en la tassa de detecció. A més, aquest mètode ens va permetre avaluar la diversitat no cultivable de *L. pneumophila* a la localitat d'Alcoi i la investigació a temps real d'un brot de Legionelosis que va afectar a un hotel en Calp (Sud-est d'Espanya) a l'any 2012.

Tercer, vam aplicar la seqüenciació genòmica a 69 soques aïllades durant 13 brots ocorreguts a Alcoi en el període 1999-2010, principalment el recurrent ST578. Es va observar una major variabilitat entre soques d'un mateix brot de l'esperada, apuntant a l'existència potencial de múltiples fonts en aquesta àrea, considerada endèmica, o alta diversitat ambiental. A més, es va observar que més del 98% de la variabilitat genòmica havia sigut introduïda a partir de processos de recombinació i no de mutació puntual.

Finalment, es va realitzar una anàlisi metagenòmica de biopel·lícules ambientals recollides a Alcoi. Varem trobar que la comunitat està dominada per *Proteobacteria, Cyanobacteria, Actinobacteria* i *Bacteroidetes*. A pesar del conegut endemisme de *Legionella* en l'àrea, aquest gènere només es va trobar en una abundància relativa entre 0.01-0.07%, el qual explica la seua baixa tassa de recuperació a partir de mostres ambientals.

En resum, els resultats d'aquesta tesi poden ser d'utilitat per als programes de control d'aquest patogen duts a terme per les autoritats de salut pública, ja que proporcionen una nova percepció de la seua epidemiologia molecular, amb aplicació immediata a la vigilància i la investigació de brots.

Abbreviations

AMOVA	Analysis of Molecular Variance
AU	Approximately Unbiased
BAS	Bronchoalveolar Aspirate
BCYE	Buffered Charcoal-Yeast Extract
BLAST	Basic Local Alignment Search Tool
Bp	Base pairs
CV	Comunidad Valenciana
DNA	Deoxyribonucleic Acid
ELW	Expected Likelihood Weight
EPS	Extracellular Polymeric Substances
ESGLI	European Study Group for *Legionella* Infections
FDR	False Discovery Rate
GOS	Global Ocean Sampling
GVPC	Glycine-Vancomycin-Polymyxin-Cycloheximide
HGT	Horizontal Gene Transfer
HPD	Highest Posterior Density
Kb	Kilobases
LCA	Last Common Ancestor
LCV	*Legionella*-Containing Vacuoles
LD	Legionnaries’ Disease
LPS	Lipopolysaccharide
mAb	Monoclonal Antibody
Mb	Megabases (millons of bases)
ML	Maximum Likelihood
MLST	Multi-Locus Sequence Typing
MRCA	Most Recent Common Ancestor

OTU	Operational Taxonomic Unit
PCR	Polymerase Chain Reaction
qPCR	Quantitative PCR
RNA	Ribonucleic Acid
rRNA/rDNA	Ribosomal RNA/DNA
SBT	Sequence-Based Typing
SH	Shimodaira-Hasegawa
SLV	Single-Locus Variant
ST	Sequence Type
TD-PCR	Touchdown PCR
TMRCA	Time for the Most Recent Common Ancestor
TV	Taxonomic Vote
VBNC	Viable Non-Cultivable
WGS	Whole Genome Sequencing

Introduction

1. Biology of *L. pneumophila*, an opportunistic pathogen

1.1. Taxonomy and biology

Legionella pneumophila is a Gram-negative *Proteobacteria* from the *Gamma* subgroup, genus *Legionella*[1]*.* Up to now, 59 different species have been described for the genus *Legionella* (www.dsmz.de), and at least three subspecies have been found within *L. pneumophila* (subsp. *pneumophila*, subsp. *fraseri* and subsp. *pascullei*)[2]. Fifteen different serogroups have been described within this *Legionella* species, although serogroup (sg) 1 has been the most frequently found both in clinical and environmental sources[3–6]. Because of *L. pneumophila* ubiquity, a new subgrouping scheme based on an immunofluorescent antibody reaction using monoclonal antibodies was proposed[7,8] for this species. New techniques, using genetic and even genomic-based procedures, reviewed in more detail in Section 3, can be used to obtain an enhanced discrimination between strains at a finer scale.

Legionella species are characterized morphologically for being non-encapsulated bacilli of 0.3 – 0.9 µm wide and 2 – 20 µm long[9]. Unlike other Gram-negatives, their cell walls have a high amount of fatty acids and ubiquinones that make cell staining difficult. *Legionella* species are aerobic, heterotrophic, urease-negative, catalase-positive and can be motile, for which they use one or more flagella (Fig. 1). They use amino acids as sources of energy and carbon, requiring L-cysteine and iron salts to grow. However, they do not oxidize or ferment carbohydrates[10].

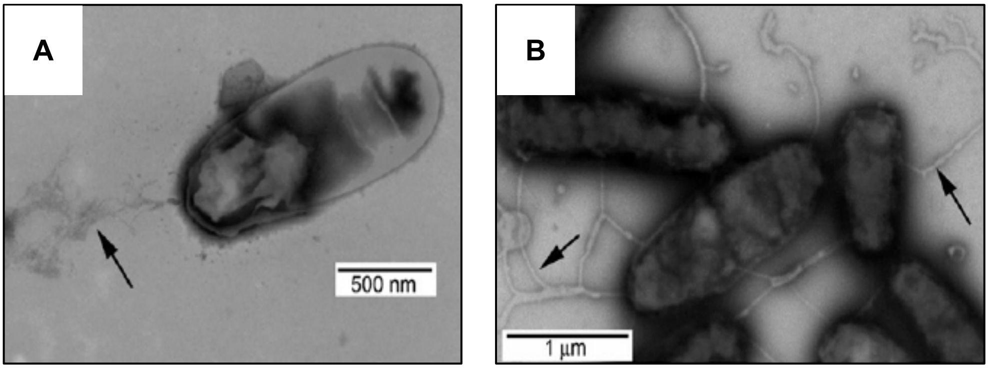


Figure 1 | Electron microscopy images of *L. pneumophila* cells in stationary phase. (A) A single cell with an arrow highlighting a flexible pilus. (B) A group of cells grown on agar; arrows mark strings of excreted material (adapted from Al-Bana *et al*, 2014)[11].

1.2. Microbial ecology

The main reservoir for most *Legionella* species (including *L. pneumophila*) is environmental freshwater, where it has been found distributed worldwide[5,12]. However, they have also been found in soils and composted material[13–17], especially in Australia and New Zealand, where *L. longbeachae* is one of the species with the highest prevalence in this type of environments[13,18,19].

From natural sources, *L. pneumophila* can colonize potable water supply systems and man-made facilities and use them as reservoirs, usually associated to amoeba taking part in biofilm communities[20,21]. It has an optimal growth temperature of 35ºC, which goes up to 37ºC when taking into account the environmental conditions, motility and adherence to host cells[22]. However, it can grow in the range 25-42ºC[23], at which many human-made aquatic environments work. In those facilities in which water temperature is higher than the environmental temperature, a shift in the balance between protozoa and bacteria can result in increased multiplication of *Legionella*[5].

It is currently accepted that *Legionella* species can survive in both aquatic and moist soil environments by parasitizing free-living or biofilm-attached amoeba[24]. The capability of *L. pneumophila* to infect and live inside freshwater and soil amoeba of the genera *Acanthamoeba* and *Naegleria* was already hypothesized in the 80's[25]. During an outbreak investigation[26], ciliates and amoeba were isolated from *L. pneumophila*-containing water taken from a cooling tower system. This revealed the ability of these bacteria to replicate within protozoa under laboratory conditions. These authors already postulated that a very low number of *L. pneumophila* cells (1 - 30 bacterial cells in a suspension of at least 100 cells/mL of the ciliate *Tetrahymena* sp.) was enough to result in intracellular multiplication (Fig. 2).

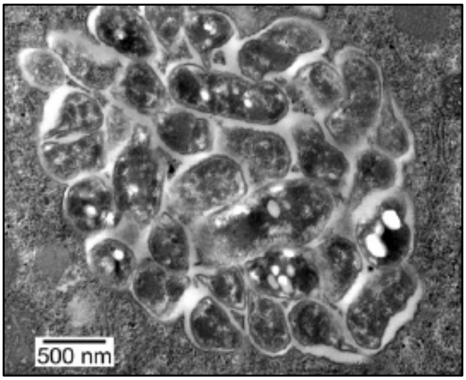


Figure 2 | Electron microscopy image of *L. pneumophila* cells grown on water and ingested by the ciliate *Tetrahymena*, forming a vacuole (adapted from Al-Bana *et al*, 2014)[11].

In response to environmental stresses, such as water treatment using chlorine or a rapid temperature change, amoeba differentiate into cysts, in which intracellular *L. pneumophila* cannot replicate but can survive[27]. This leads to an increased resistance to extracellular stresses and, consequently, persistence in the environment[28,29]. At least fourteen species of amoeba (being *Hartmannellae* and *Acanthamoeba* the most frequent ones), as well as two species of ciliated protozoa and one slime-mold, have been described to support intracellular replication of *L. pneumophila*[5,27].

1.3. Biofilms as global reservoirs

Biofilms are complex microbial communities in which cells are attached to a substratum or phase in contact with each other and to a matrix formed by self-produced extracellular polymeric substances (EPS). The cells involved in biofilms show altered growth rates and gene transcription compared to their planktonic counterparts[30–32]. These communities can change over time and space so that the associated microorganisms can better adapt to changing environments and ensure increased survival and growth. This is why the most frequent microbial lifestyle is dependent on these complex structures and the planktonic phase is primarily seen as a mechanism of transport[33].

Biofilms develop mainly on solid-water interfaces (substrate-associated biofilms), but also at the water-air interface (floating biofilms)[21]. Their growth is favoured by an increased water flow rate[30,34,35] as opposed to the common belief that stagnation is a predisposing factor for colonization[36]. Specifically, it has been shown that, for *Legionella,* turbulent flow results in faster biofilm growth. This is due to a higher overall mass transfer that would lead to a greater particle deposition onto the surface[35]. Besides, higher bacterial recovery was found in turbulent than in laminar flow, being stagnant water where the lowest concentration of *Legionella* was recovered. Apart from natural environments, biofilms can also develop in man-made facilities and devices, where bacteria, dictated by nutrient availability, can adhere to surfaces enriched by organic molecules[37]. Bacteria living in these microbial structures follow a 'biofilm life cycle' that starts with the attachment to a substratum, continues with biofilm maturation and finishes with detachment of the bacteria from the biofilm and dispersal into the environment[21,31,32].

Different pathogens have been found in water-related environments, such as *Pseudomonas aeruginosa*, *Escherichia coli* or different species of *Mycobacterium, Salmonella, Campylobacter, Shigella, Yersinia*, *Vibrio, Listeria,* etc.[38–40], and also in biofilms formed on medical devices, such as *Candida albicans*, *Klebsiella pneumoniae*, *P. aeruginosa* or *Staphylococcus aureus* [31], with the subsequent risk for patients.

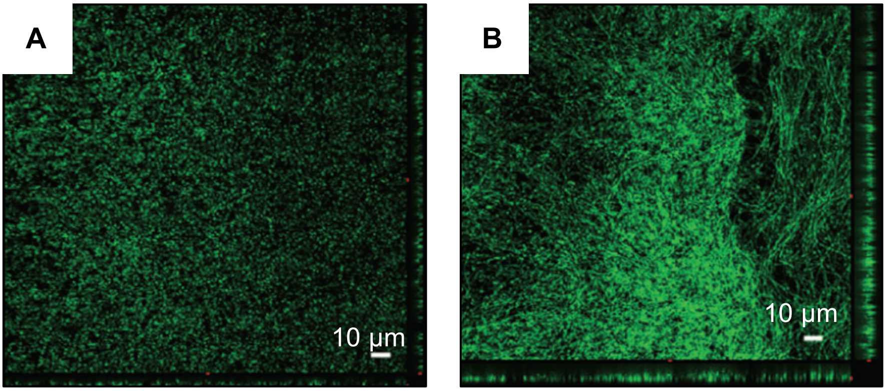


Figure 3 | *L. pneumophila* Philadelphia-1 biofilms expressing green fluorescent protein (GFP) shown under a confocal laser-scanning microscope. (A) shows the rod-shaped cells of a 6-day-old biofilm at 25ºC and (B) shows the filamentous cells within a thicker structure of a 3-day-old biofilm at 37ºC (adapted from Declerk, 2010)[21].

Some of these pathogens are capable of producing monospecies biofilms but, in the case of *L. pneumophila*, this has been reported only under strict nutrient-rich experimental conditions[41,42]. *Legionella* grows preferentially within protozoa, although previous intracellular replication allows its free proliferation within biofilms (Fig. 3)[42].

1.3.1. Environmental spread

Legionella spreads mainly through aerosols[5]. These become loaded with bacteria when the water surface is disturbed or the flow velocity changed, resulting in the detachment of biofilm cells and subsequent dispersal[31,32,43,44]. Storey and colleagues (2004)[45] showed that, in detached biofilms, the number of *Legionella* cells could vary from a few in small clusters to hundreds in larger aggregates. Moreover, it has been shown that infected amoebae release small breathable-size vesicles containing from 20 to approximately 200 *L. pneumophila* cells (Fig. 4)[46].

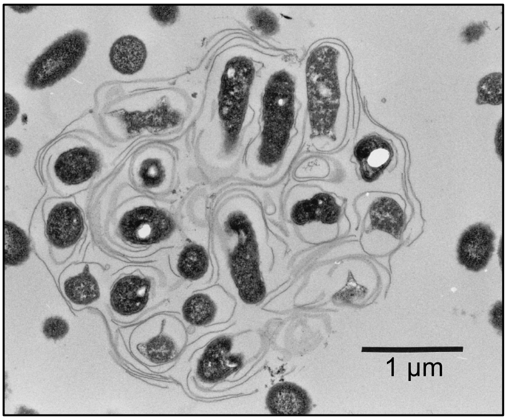


Figure 4 | Electron microscopy image of a vesicle filled with *L. pneumophila* cells expelled from *A. castellanii* (adapted from Berk *et al*, 1998)[46].

The presence of replicating pathogens attached to biofilms in artificial ecosystems that produce significant amount of aerosols, such as industrial evaporative condensers or cooling towers, creates an inherent risk for public health. The detachment of individual cells or biofilm aggregates that reach susceptible people can result in human infection. It is important to remark that these community-living bacteria have a higher chance of exchanging genetic material and this could lead to the spread of pathogens with increased resistance to antibiotics and even to host immune system clearance[31].

1.3.2. Life cycle and pathogenesis: an accidental infection

The life cycle of *L. pneumophila* is determined by its ability and need for intracellular replication, which influences its pathogenicity[47]. It has a two-phase life cycle: an infectious, stationary growth phase in which bacterial cells are flagellated and motile, and an exponential phase in which the bacteria are non-flagellated and have the ability to replicate, among other several differences[48].

Rowbotham (1980)[25] was the first author to show the ability of this bacterium to infect amoebae and reside, move and multiply within large vacuoles. Several experiments have confirmed that *L. pneumophila* can multiply within human monocytes[49,50] and, specifically, alveolar macrophages[51], thus considering the bacterium a facultative intracellular parasite. Human infection has been frequently considered as accidental[10,52,53] because of the multiple similarities in the infection route compared to that in protozoa[54].

The phagocytosis pathway followed by *Legionella* was discovered to consist of a novel mechanism, termed ‘coiling phagocytosis’ (Fig. 5)[55]. However, conventional entry has also been observed in a low proportion of cases[56]. During the coiling phagocytosis process, the bacterium is engulfed within a pseudopod coil and remains surrounded by a membrane layer that protects it to be degraded by lysosomes.

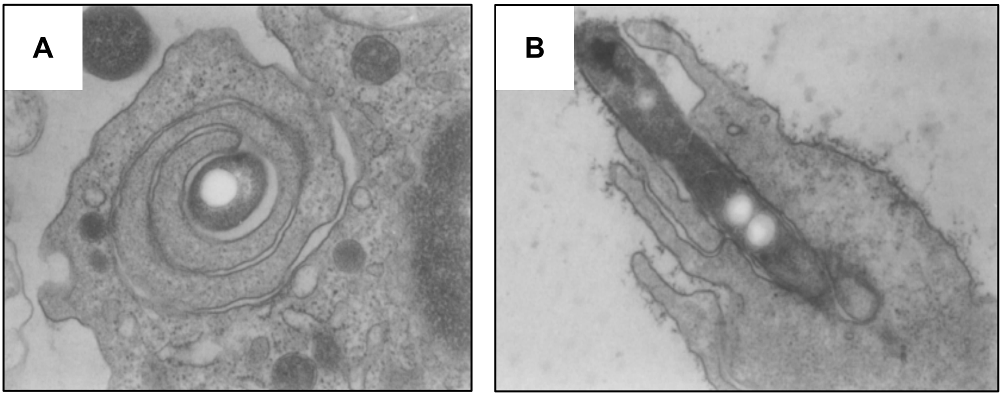


Figure 5 | Coiling phagocytosis of *L. pneumophila* by an amoeba. Both images show the bacterium in the centre of a coiled pseudopod in front (A) and cross-sectioned views (B) (adapted from Horwitz, 1984)[55].

Phagocytes can engulf both live and dead cells. However, living cells, through a series of cytoplasmic events, induce the formation of a phagosome studded with ribosomes that avoids the fusion and degradation by host cell lysosomes. Dead cells can enter the vacuole and are rapidly degraded[49,55]. Later works confirmed that, already during phagocytosis, the *Legionella*-containing vacuoles (LCV) start to create and recruit host cell organelles to their surface, such as mitochondria, ribosomes or small vesicles (Fig. 6)[27,57]. LCVs intercept the vesicle trafficking from the endoplasmic reticulum (ER) to the Golgi apparatus and get coated from the ER, as well as polyubiquitinated proteins. As a consequence, the bacteria are able to evade the normal trafficking pathway to the lysosomes[58,59]. It is at this point that a rapid intracellular proliferation of *L. pneumophila* starts[27,57] aided by host cell proteins transported from the ER to the LCV[60].

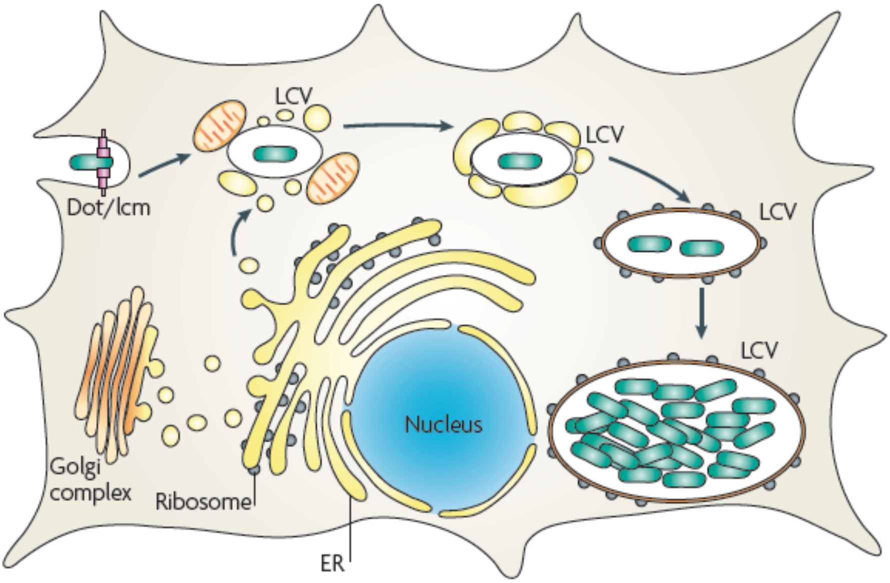


Figure 6 | Representation of the formation of the replicating vacuole within amoebae or macrophages. After uptake, the *Legionella*-containing vacuole (LCV) is able to evade fusion and lysosome-mediated lysis. ER: endoplasmic reticulum (adapted from Isberg *et al*, 2009)[57].

L. pneumophila uses the Dot/Icm (defect in organelle trafficking/intracellular multiplication) system, part of the bacterial type IV secretion network, to translocate and recruit host proteins[61–64]. The Dot/Icm system can also be involved in DNA conjugal transfer[65,66]. Several proteins of the Dot/Icm system perform different roles[57]. These include substrate recognition (IcmS, IcmW, LvgA), the formation of a coupling ATPase (DotL/IcmO, DotM/IcmP, DotN/IcmJ), components of the system core (DotC, DotD, DotF/IcmB, DotG/IcmE, DotH/IcmK), inner-membrane proteins that determine core stability (DotU/IcmH, IcmF), different cytoplasmic components, such as a pore-forming molecule (IcmQ) and its chaperone (IcmR), an ATPase (DotB) or an inner-membrane protein (DotO/IcmB), as well as other inner-membrane or periplasmic components whose function are still unknown (DotA, DotE/IcmC, DotI/IcmL, DotJ/IcmM, DotK/IcmN, DotP/IcmD, DotV, IcmT, IcmV, IcmX). Over 300 different substrates translocated by the Dot/Icm system have been identified in *L. pneumophila*[57,67–69], although only 24 of them have been detected as core in other *Legionella* species[68]. Functional redundancy between those effectors allow the bacteria intracellular replication to be highly resistant to perturbation of single bacterial or host components[59]. They also provide the required repertoire for optimal growth in different hosts[70,71]. In fact, studies using different *L. pneumophila* have shown many of the substrates to be present in just part of the analysed strains, thus forming part of the auxiliary genome[52,68,72]. Besides, the Mip (Macrophage Infectivity Potentiator) protein is needed for an efficient infection of host cells, as studies introducing null mutations in the encoding gene have shown[73–75]. In order to prevent premature apoptosis of infected macrophages caused by toxic microbial products or the immune system, the bacterium is able to interfere with caspase activation in the host[76–78]. The creation of LCV for other *Legionella* species has not been described in such detail with the exception of *L. micdadei* and *L. longbeachae*[53].

During the last stages of intracellular replication, the LCV is disrupted and bacterial cells are released into the host cytosol, where they can complete the last 1-2 rounds of proliferation including some phenotypic modulations in response to nutrient depletion[27,79–81]. This egression allows the infection of new host cells, which can be other environmental host or a susceptible mammal. As already speculated, the infectious particle could be free *L. pneumophila* cells, vesicles filled with the bacteria and also intact parasitized amoeba. Man-made devices and installations colonized by amoeba-containing biofilms, such as cooling towers, whirlpools or showerheads can produce aerosols loaded with these infectious particles and cause infection[27].

2. Legionnaires' Disease: epidemiology and surveillance

2.1. First descriptions

The investigation of an explosive outbreak of severe pneumonia affecting 182 attendants to the American Legion convention in Philadelphia (Pennsylvania, United States) in July 1976 was the first epidemiological alert that led to the identification of Legionnaires' Disease (LD) as a new respiratory illness. This first detected LD outbreak caused 29 fatalities[82]. An intense epidemiological investigation discarded person-to-person transmission and pointed to a potential air-borne agent that explained both patients staying in the same hotel and those just walking near the facility. A Gram-negative bacillus isolated from the lung tissues of Guinea pigs inoculated with patients' sera from this outbreak was described as the etiological agent of LD[83].

However, this was not the first time in which this pathogen caused disease. After its initial description, retrospective analyses of unclassified agents producing severe pneumonia were confirmed to have the same characteristics. This was the case of an organism isolated from a febrile respiratory illness in 1947[84], an outbreak in a meat packing plant in 1957[85], a psychiatric hospital in 1965[86] and even an outbreak affecting 11 persons staying in a convention held in 1974 in the same hotel in which the American Legion met two years later[87]. This agent was also found to cause an outbreak characterized by different symptoms, mainly acute fever, headache and myalgia, affecting 144 people from a Health Department facility in Pontiac (Michigan, United States). This symptomatology was designated as

Pontiac Fever[88]. Brenner *et al* (1979)[1] classified the agent of these severe pneumonia and acute fever illnesses as *Legionella pneumophila*.

2.2. Clinical features, risk factors and treatment

There are two clinical presentations of *Legionella* infections: Legionnaires' Disease, a severe multisystem disease with pneumonia as main symptom[82] or Pontiac Fever, a flu-like illness[88]. A wide range of clinical manifestations and symptoms apart from pneumonia have been described, including fever, organ-specific symptoms and other signs, such as diarrhoea, confusion, multisystem disease, etc., as recently reviewed in Phin *et al* (2014)[89]. Clinically, *Legionella*-caused pneumonia is very difficult to differentiate from other types of pneumonia, although alveolar infiltrates seem to be more common in LD[5]. So, in order to detect a potential LD case when a patient is diagnosed with pneumonia, it is essential to apply microbiological or molecular methods for *Legionella* detection, which will be reviewed in Section 3. The typical incubation period for developing LD after infection is from 2 to 10 days, with a median of 6-7 days[89], although point deviations have been described[90]. Legionellosis is considered an opportunistic disease because it only develops after *Legionella* infection in a susceptible host. Risk factors that make a person susceptible to this disease are gender (males are more frequently affected), smoking, age, chronic respiratory or cardiovascular disease, diabetes, alcohol misuse, cancer and immunosuppression[91]. No person-to-person transmission cases of legionellosis have been described up to now.

Legionella infections are usually treated with antibiotics such as macrolides or fluoroquinolones. They are applied in the initial stages of the disease and result, in most cases, in the clearance of the infection. Other therapies are applied in the case of complications or additional comorbidities[89]. Up to date, there is no evidence of antibiotics creating high-level resistance in *Legionella*[92–95]*.* Onody *et al* (1997)[95] hypothesized that *Legionella* does not develop antibiotic resistance in the clinical setting because there is not a carrier step before the infection or person-to-person transmission. Because infection occurs directly from the environment, the potential exposure to antibiotics before infection is very low.

2.3. Environmental factors and implications for public health

The opportunistic nature of LD is not only due to clinical factors. Different environmental factors have also been associated to this disease: water temperatures in the optimal growth range of 20ºC to 45ºC in man-made installations[5], pH, levels of copper or iron, the presence of dead-end loops, stagnant water, periods of non-use and plastic-made pipes instead of steel, as reviewed in Mekkour *et al* (2013)[10], but also atmospheric phenomena, such as rainfall[96]. Epidemiological studies have shown other factors that are dependent on patients' activities, such as travelling abroad, spending one or more nights away from home without leaving the country, using the shower at a swimming pool or a sports facility[91] and even being a professional driver[91,97]. However, the ultimate risk factor for human exposure is always the formation of aerosols loaded with an infective dose of bacteria. Humidity is crucial in determining their persistence and suspension in the environment. Different *L. pneumophila* strains can have different survival capacity in aerosols, being serogroup 1 monoclonal subtype Pontiac the most resistant strain to long exposures and high humidity conditions[98].

LD cases are mostly sporadic, acquired in a residential environment and associated with domestic potable water and disruptions in plumbing systems[99–101]. However, infections can affect more than one case in a community, forming a cluster of cases that can turn into an outbreak if three or more cases accrue in the same location and period. These clusters can be travel-associated, when people staying in the same hotel, cruise ship, etc., are infected[102–109] or affect people living in the same residential blocks[110,111], hospitals (nosocomial infections) and other large buildings supplied by public water[112–114]. Unsuspected sources such as car washers[115], dental units[116], asphalt paving machines[117] or baby delivery tubs[118] have also been reported.

2.4. Incidence and surveillance schemes

The distribution of LD cases by age and sex are homogenous among countries, being children the less affected age-class. Most cases correspond to men older than 50 years (74-91%)[89]. The mortality rate is normally between 8-12% and the case-fatality rate in Europe is 10% (range 0-27% in countries reporting at least 30 cases) and 8% in the USA[6,89,119–121]. Nosocomial cases have a higher case-fatality rate, in the 15-34% range[89].

Comparisons of the incidence of legionellosis cases in different countries are highly dependent on the number of laboratory confirmations but also on the notification rates. LD is considered as underreported because clinicians prescribe broad-spectrum antibiotics that cover for *Legionella* spp. and rarely ask for laboratory confirmation. Besides, although being tested, positive results are not always notified to health authorities[122]. Currently, Europe[6,122], the USA[123], Canada[124], New Zealand[125], Australia[126,127], Japan[128] and Singapore[129] have surveillance schemes for LD. In the last surveillance report of the European Centre for Disease Prevention and Control (ECDC 2012)[122], 5,852 LD cases were reported by the 27 EU member states of the European Legionnaires' Disease Surveillance Network (ELDSNet) (Fig. 7) plus Iceland and Norway, which is within the range of the notifications between years 2005 and 2011.

Just six countries (France, Italy, Spain, Germany, Netherlands and United Kingdom) accounted for 84% of the notified occurrences. Most of the cases (69%) were acquired in a community, 20% were travel-associated and 8% linked to healthcare facilities. *L. pneumophila* was found as the causal agent in 98% of the cases confirmed by culture, 85% of them corresponding to serogroup 1. These notified cases included 99 clusters, the largest one involving 42 cases in a hotel in Spain[107], which is further investigated in Chapter 4 of this thesis. The incidence of LD has increased in the USA in the period 2000-2009, with the north-eastern states reporting most of the cases[123]. Australia and New Zealand have an additional peak of LD cases due to *L. longbeachae* potentially linked to potting soil and gardening activities[17–19].

2.5. Preventive regulations

LD surveillance in Europe is mainly coordinated by ELDSNet, which focuses on the detection of clusters and monitoring of epidemiological trends by collecting, analysing and reporting LD cases on an annual basis[122]. Cases are classified as travel-associated if they have stayed at any accommodation site away from home during the incubation period. Clusters are defined as at least two or more cases that have been exposed to the same source and with dates of onset within a plausible period of time.

The ELDSNet developed a European case definition in which a LD case is defined as a patient who has developed pneumonia and there is laboratory confirmation of at least one of the following three methods: isolation of *Legionella* spp. from respiratory secretions, detection of *L. pneumophila* antigen in urine or a significant increase in specific antibody levels to *L. pneumophila* serogroup 1 in paired serum samples. If the bacterium is detected using monoclonal-antibody-derived reagents, molecular amplification of respiratory secretions, a rise in specific antibody level to non-serogroup 1 *L. pneumophila* or other species, the case is defined as probable.

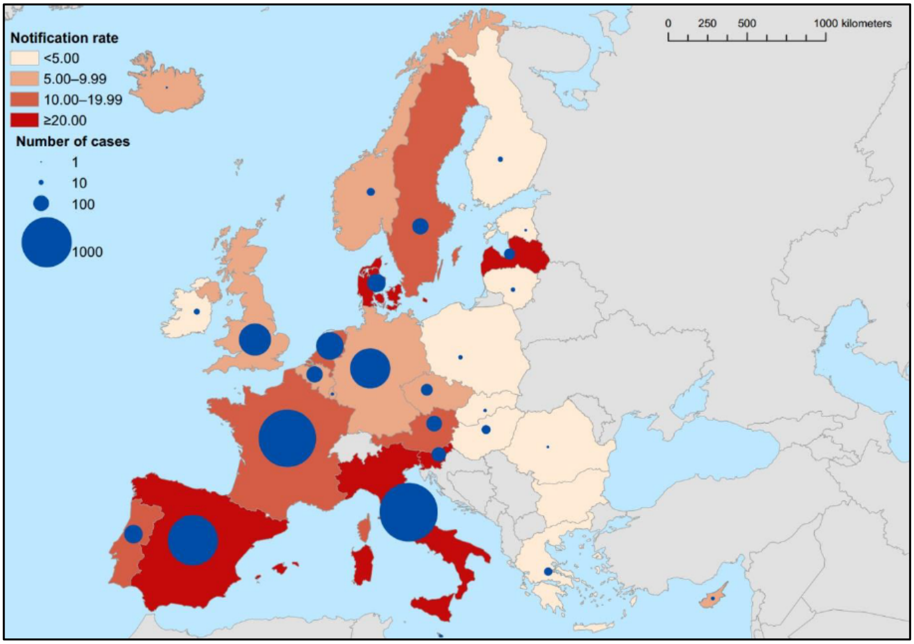


Figure 7 | Representation of the reported cases of Legionnaires' disease and notification rates per million in Europe, 2012 (adapted from De Jong *et al*, 2012)[122].

Apart from the notification programs, the World Health Organization (WHO) established the Water Safety Plan (WSP), which contains specific guidelines for *Legionella* control in risk installations, such as cooling towers, hot tubs or water distribution systems where susceptible people might be exposed. These are valid for outbreak investigations and validation of the effectiveness of decontamination and control measures. Besides, they encourage surveillance agencies to undertake audits and inspections of high-risk facilities. Control methods have both advantages and disadvantages, such as keeping temperature below 20ºC, which is easily monitored but only really applicable to drinking water systems, or above 50ºC, which is difficult to reach in old systems and does not eliminate *Legionellae*. Dosing with different chemicals such as sodium hypochlorite or monochloramine has been proved effective but can produce dangerous by-products. An in-depth list of advantages and disadvantages of different control measures can be found in the WHO report for Legionellosis prevention[130].

In Spain, legionellosis was included in 1996 in the list of compulsory notifiable diseases (RD 2210/95 from the 28th December) when the National Network of Epidemiologic Surveillance (*Red Nacional de Vigilancia Epidemiológica*) was created. In 2003 (RD 865/2003 from the 4th July), national health authorities established the hygienic-sanitary criteria for the prevention and control of legionellosis. This regulation classified the different risk installations as those with high probability of *Legionella* proliferation and spreading (refrigeration towers, evaporative condensers, hot water systems or industrial humidifiers), those with low probability (cold water systems, humidifiers, ornamental fountains, urban sprinkler systems, etc.), and those related to respiratory therapies, such as breathing systems, nebulizers and other medical devices in contact with the airways. Installations in private homes are excluded unless they affect the external environment of the buildings. Specifically, temperature follows WHO guidelines, with cold water intended to be below 20ºC and hot water between 50-70ºC. Pipe or tower materials that favour the colonization of bacteria or fungi such as wood, leather, concrete or cellulose sub-products should be avoided. Chlorine or other disinfectants should be automatically dosed in refrigeration towers or analogues, routine disinfection programs or after an outbreak. The addition of the biocide must be accompanied from a thorough cleaning of the system. Chlorine levels should not be above 5 mg/L and pH between 7 and 8. After a period of inactivity, the system should recirculate during at least 3 hours without ventilation to avoid the spreading of aerosols.

3. *Legionella* detection and typing methods

3.1. Methods for *Legionella* identification

Laboratory confirmation of any potential LD case is essential for public health authorities to start an environmental investigation. The main aim is to identify the *Legionella* reservoir that could pose a risk for other susceptible people and originate an outbreak. Also, even in the absence of any clinical occurrence, environmental testing for *Legionella* is used within hazard analyses and control risk management plans as well as to verify the effectiveness of decontamination procedures[131]. It is important to remark that *Legionella* is ubiquitous in aquatic environments, so it may be present during these surveillance programs but not causing disease. In fact, the identification of the bacteria without clinical cases combined with a risk assessment of the system can be of important use in order to test its likelihood as a source[89].

The gold standard for *Legionella* detection remains microbiological culture[5,131,132]. Since the first culture media that contained L-cysteine as essential amino acid and starch[133], many improvements have been made[47,134,135] until the currently used Buffered Charcoal-Yeast Extract (BCYE) agar enriched with α-ketoglutarate. Two modifications of this basal medium with and without selective agents are essential for the confident identification of the bacterium and have led to the many commercial media currently available[136]. However, culture-independent methods based on serology, such as direct fluorescent antibody analysis, urinary antigen tests and nucleic acid targeting are in increasing use[131,132].

3.1.1. Identification in environmental samples

During the environmental investigation of outbreaks, sporadic cases or routine screenings, three main types of samples can be studied: water, biofilms and aerosols[131]. Standard regulations for *Legionella* spp. detection from water samples are mainly based on ultra-filtering of 1 or 2 L of water through a series of sterile polycarbonate membranes of decreasing pore size (AFNOR NF T 90-431 2 (2003) and UNE-EN ISO 11731-2:2008). Centrifugation has also been used as a concentration strategy but it provides lower recovery rate[137,138]. Survival to specific thermal and acid treatments as well as confirmation of no growth in the absence of L-cysteine are the microbiological markers that lead to the identification of *Legionella* spp. Biofilms are usually sampled using sterile Dacron, cotton or polyester swabs by gently scraping the surface of pipes, faucets or any other surface with signals of biofilm formation, and can be cultured directly on selective media[131]. Further details on water and biofilm processing can be found in Chapter 3. Sampling of aerosols potentially loaded with pathogenic bacteria has been performed using different methods, such a bioerosol collector[139], impactors or impingers[140].

One of the most frequently used methods for routine identification of isolates is polyclonal antiserum coupled to latex beads, the so-called latex agglutination tests[141]. These tests are used to distinguish between *L. pneumophila* and other species, as well as the 15 different serogroups in the *L. pneumophila* species. Other serological methods, applicable directly on uncultured samples, are the fluorescein-conjugated polyclonal antibody (Direct Fluorescent Antibody, DFA) identification[142,143] and the Dresden monoclonal antibodies[7,8,144]. Mass spectrometry MALDI-TOF is another method that can be used to generate a molecular fingerprint of the isolated strains, which can be distinguished because of specific spectra dependent on the bacterial molecular content[145–147].

Legionella is known to be fastidious, with a slow growth *in vitro* and can enter into a viable but non-culturable state (VBNC) under appropriate conditions[11,148–151] as many other bacteria[152,153]. Hence, culture-independent methods are being increasingly applied for detection and typing, as reviewed in Section 3.1.3. Co-culturing with amoeba has been used to recover the cultivability of *L. pneumophila* strains[154].

3.1.2. Identification in clinical samples

After a clinical diagnosis of pneumonia, the confirmation of *Legionella* as the causal agent is primarily performed by detecting soluble antigens in urine, specifically, a component of the lipopolysacharide (LPS) cell wall[155]. Once the urinary antigen is detected in a sample, a respiratory secretion is taken from the patient, usually sputum or brochoalveolar lavage. However, tracheal aspirate, tracheobronchial aspiration, pleural fluid, abscesses or lung biopsies can also be used[156]. These secretions are fluidised and used for culturing using selective media as described for environmental samples in Section 3.1.1. As the urine test is biased towards *L. pneumophila* serogroup 1, negative urines can result in positive isolation by culture from the respiratory secretions of the patients[155,156]. Moreover, clinical specimens with very low bacterial load or VBNC forms of *Legionella* have been shown to produce virulence proteins[157]. Detectable levels can be reached after co-culturing with amoeba[158,159].

The species, serogroup and subgroups of the isolates from culture-positive samples can be identified using latex agglutination and monoclonal antibody tests as described previously for environmental samples. However, background fluorescence and dilution effects of DFA limit the number of different conjugates to be included in a single screening pool[160,161]. To overcome that problem in clinical samples, indirect fluorescent antibody (IFA) assays targeting IgG, IgM and IgA antibodies[160–162] generated during the immune response to *Legionella* infection have also been applied[124,125,163].

3.1.3. Sequence-based identification methods

All the alternative methods mentioned in the above sections for classification, serogrouping and subtyping of *Legionella* species are based on phenotypic features and do not have enough resolution to accommodate new species or detect intra-species variability. Only DNA-targeted methods have this potential. The *mip* gene[147,164–168] is the most widely used target for polymerase chain reaction (PCR) although others, such as 16S rRNA[169–171], *rpoB*[172], *rnpB*[173] or *proA*[174] have also been used. There are also some works using the 23S-5S ribosomal intergenic spacer region (ISR)[175] and the *gyrA* gene[176] but the published primers are not universal or the presence of multiple copies make sequence interpretation difficult. Also, as any of these regions can be involved in recombination events, it is recommendable to use several targets for confident species determination[174].

Legionella detection by genetic amplification targeting a fraction of its genome has been widely performed using mainly quantitative real-time PCR approaches in both clinical[177–184] and environmental samples[179,181,182,185–190]. This procedure can be applied directly to total DNA extracted from specimens and it also allows a quantification of the bacterial load. As an example of the comparison between the efficiency of isolation by culture and direct PCR, Mentasti and colleagues[183] found an overall sensitivity of quantitative PCR (qPCR) above 30% greater than that of culture (~65%) in respiratory samples. In the case of environmental samples, it is common to have more incongruence between results from culturing and PCR from the same sampling points. Although there are studies detecting similar results between qPCR and culture from water samples[190], there are works reporting differences, especially for cold water[189,191]. In general, culturing efficiency is reported to be around 70-80% and PCR above 95% in water from warm environments[186,189,192]. The principal inconvenient of PCR is that it cannot distinguish between viable and dead cells, although recent works show new methods that selectively remove DNA from dead cells prior to amplification using interchelating agents such as propidium monoazide or ethidium monoazide[193,194]. DNA extracted from biofilm swabs can also be directly used for molecular detection of *Legionella*, with different advantages and drawbacks that are further discussed in Chapters 3 and 4.

3.2. Typing methods in outbreak investigations

The investigation of LD outbreaks, clusters of cases or even a single case starts with the search for a potential *Legionella* reservoir. To this end, different water-related risk points are sampled, starting from the patient's home or a specific area where the epidemiological investigation points that the involved patients have stayed or even walked through[195]. Both respiratory secretions and environmental samples (mainly water and/or biofilms) from the potential sources are analysed simultaneously in the laboratory. In the case of isolation from both sources, typing methods that provide further discrimination between strains than the traditional 1-15 serogrouping need to be applied[196]. As mentioned in previous sections, a standardized subtyping scheme using monoclonal antibodies (mAb)[7,8] allows higher discrimination within *L. pneumophila* (mainly serogroup 1), with the possibility of further classification in 23 phenoms. Besides, a pan-European study identified a significant association between mAb 3/1-positive or negative strains, the category of infection and the region where the infection was acquired, thus encouraging its use as a rapid and reproducible tool[144]. Further works studying the distribution of *L. pneumophila* serogroups and subtypes used the detection of mAb3/1 as a virulence marker[197,198]. Other methodologies coupled with serogrouping and mAb subtyping, such as Multi-Locus Enzyme Electrophoresis (MLEE), have been applied to obtain higher discrimination in outbreak investigations and to study the genetic variability of *L. pneumophila* populations[115,199].

Nevertheless, the resolution attained by mAb subtyping and MLEE is still lower than that obtained with DNA-based methods. Molecular techniques such as Pulsed-Field Gel Electrophoresis (PFGE)[200–204], Arbitrarily Primed PCR (AP-PCR)[205–207], Restriction Fragment Length Polymorphism (RFLP)[208–210], ribotyping[205,206], Amplified Fragment Length Polymorphism (AFLP)[204,211,212], the detection of repetitive elements[213,214] and the current gold standard, Sequence-Based Typing (SBT)[215–218], have been found to discriminate within mAb subtypes[219]. AFLP became the gold standard for *L. pneumophila* typing for a few years in the last decade[5] and it has the advantage that it can be performed without sequencing. However, fragment sizing is imprecise and reproducibility and comparison between laboratories are problematic[196].

SBT has currently established as the most discriminatory genotyping tool without considering genome sequencing[196]. The method follows the basis of Multi-Locus Sequence Typing (MLST)[220] and consists on the amplification and sequencing fragments of seven gene targets (*fliC/flaA, pilE, asd, mip, mompS, proA* and *neuA*) [215–218]. The retrieved sequences are then compared to those in the European Study Group for *Legionella* Infections (ESGLI) database[221] and the allelic pattern is reported as a

Sequence Type (ST). More than 1,800 different STs have been reported until January 2015. SBT has been applied to outbreak investigations in order to compare the ST(s) of the clinical cases with the ST(s) of the environmental isolates[117,222–224], but also in studies of genetic variability and distribution of *L. pneumophila* populations[224–226]*,* which will be further addressed in Chapters 1 and 2. However, infections caused by unculturable forms for *Legionella* that do not produce a positive isolation of the bacteria make the finding of the infectious strain in every patient difficult, thus complicating the epidemiological search of one or more potential sources. To overcome this problem, semi-nested PCR and qPCR approaches accommodated to the SBT scheme have been developed in order to get the ST pattern from the infecting strains directly from respiratory samples[183,227,228]. In fact, the application of SBT over uncultured specimens has shown cases of mixed infection[229] that have also been observed using culturing strategies[230,231]. The application of SBT on total DNA extracted from environmental samples such as water or biofilms without a culturing step is still not widely accepted in the scientific community because of the detection of both live and dead cells and the difficulty in obtaining clean sequences to get the ST. However, in cases of VBNC forms of *Legionella*, molecular detection is necessary to detect colonization of risk facilities. Besides, complex samples such as biofilm swabs can be difficult to culture, so direct PCR has been proven to give sensitive and rapid results, as further discussed in Chapter 2.

Specific markers to improve discrimination between strains from the same group or sequence type have been developed. As an example, a spoligotyping tool based on the diversity of the CRISPR (Clustered Regularly Interspaced Palindromic Repeats) locus has been described for discriminating within *L. pneumophila* serogroup 1 ST1 strains[232].

4. Lessons from *L. pneumophila* genomics

4.1. Comparative genomics and diversity

Twelve complete genome sequences of *L. pneumophila* clinical and environmental strains are currently (January 2015) available in the databases. The first genome (Philadelphia 1) was published in year 2004[52] from a serogroup 1 strain isolated from the 1976 outbreak[83] and was described to have a single circular chromosome of nearly 3.4 Mb with 38% G+C content and a plasmid-like element of 45 kb. This first genome was enriched in genes encoding efflux transporters for heavy metals and other toxics, eukaryotic-like proteins predicted to modulate host cell functions and molecular markers of horizontal gene transfer (HGT) or recombination such as tRNA, phages, transposases, etc. The transfer of genetic material in *Legionella* had already been inferred by topological incongruences between the *rpoB* and *dotA* phylogenetic analyses of isolates of *L. pneumophila* subspecies *pneumophila* and *fraseri*[233]. A high genomic plasticity was also found for strains Paris and Lens[72] (Table 1), which were reported to contain a plasmid of 131 kb and 59 kb respectively. Many proteins and motifs of eukaryotic origin were also detected and were discussed to have been acquired by HGT from their hosts or by convergent evolution during its coevolution within free-living amoebae[72,234]. After the sequencing of Corby[235] and Alcoy[236] strains (Table 1), a detailed structural analysis of the five first genomes provided a better overview of the genomic content of *L. pneumophila* as a species[236].

Table 1 | List of reference *L. pneumophila* genomes and associated metadata.

Reference genome	Source	Origin	Sg.*	mAb subgroup	ST**	Accession number	Ref.
Philadelphia1	Clinical	USA	1	Philadelphia	36	AE017354	52
Paris	Clinical	France	1	Philadelphia	1	CR628336	72
Lens	Clinical	France	1	Benidorm	15	CR628337	72
Corby	Clinical	UK	1	Knoxville	51	CP000675	235
Alcoy	Clinical	Spain	1	NA	578	CP001828	236
Lorraine	Clinical	France	1	NA	47	FQ958210	♯
HL06041035	Environmental	France	1	NA	734	FQ958211	♯
ATCC 43290	Clinical	USA	12	NA	187	CP003192	237
Thunder Bay	Clinical	Canada	6	NA	187	CP003730	238
LP0509	Environmental	China	NA	NA	NA	CP003885	239
Hextuple_2q	Clinical	Mutant	1	Philadelphia	36	CP003023	70
Hextuple_3a	Clinical	Mutant	1	Philadelphia	36	CP003024	70
130b	Clinical	USA	1	Benidorm	42	FR687201	240

* Sg: Serogroup.

** ST: Sequence Type. Extracted from the corresponding reference and Reuter *et al*, 2013[241].

♯ Unpublished, direct submission (see http://www.ncbi.nlm.nih.gov/genome/genomes/416).

NA: Non-available information.

G+C content remained around 38% for the five genomes, which had a total length within the range 3.39-3.57 Mb, with a coding portion of 86-88% and three ribosomal operons. The work also reported the presence of several genomic islands with different factors that could give additional virulence to each strain (see Section 4.2). These five genomes were all retrieved from clinical serogroup 1 isolates. Other currently closed genomes correspond to strains LPE509[239], Lorraine, HL06041035, ATCC 43290[237], Thunder Bay[238] and two targeted mutants of the Philadelphia 1 strain[70] (Table 1). The genome of *L. pneumophila* strain 130b[240] has not been completely closed but has already been incorporated as reference sequence in some works[241,242].

The core genome for *L. pneumophila* isolates, defined as the number of orthologous genes shared by all of the analysed strains[243], has been described to encompass 2,405 orthologous genes among eight genomes [244]. The accessory genes accounted for approximately 10% of the ~3,000 present in each of the strains (Fig. 8). Oppositely to the core genome, the total number of genes in all the strains is known as the pangenome[245].

L. pneumophila shares approximately 42% of its genome with its closest relative with a complete genome sequence, *Coxiella burnettii,* despite their difference in genome size (3.5 and 1.9 Mb, respectively). Interestingly, 60% of *L. pneumophila* genes are orthologous to other phylogenetically distant but intracellular bacteria, such as *Salmonella*, *Chlamydia, Rickettsia, Brucella* or *Mycobacterium* species[52]. Genome size and redundancy in the *L. pneumophila* genome have been discussed being subjected to host variation[70]. No evidence of significant differences between clinical and environmental strains of this bacterium has been found[246], contrary to previous reports using RFLP[167] or SBT data[197,225].

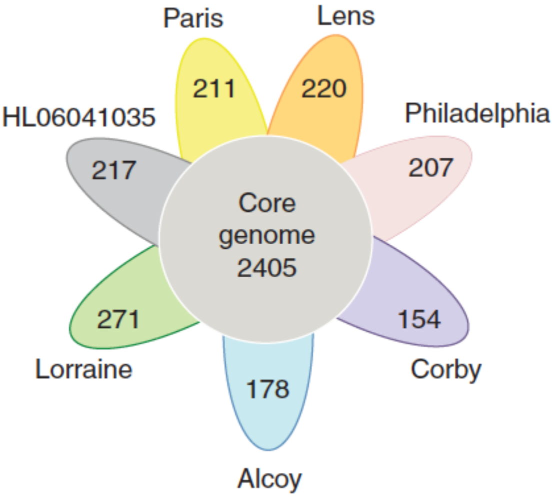


Figure 8 | Pangenome of seven *L. pneumophila* strains. The central circle represents the core genome (number of orthologous genes in all strains) and each color represents the number of accessory genes in each specific strain (adapted from Gómez-Valero and Buchrieser, 2010)[244].

Genome sequencing provides the maximum discrimination power possible, and that is why its use in bacterial diversity studies is increasing[242,247–251]. Underwood and colleagues[242] performed an exhaustive analysis to compare genetic diversity between *L. pneumophila* strains using SBT and genomic data and found STs that were not clustered within the genetic lineage described by whole-genome sequencing (WGS), indicating signals of admixture.

Specific WGS projects have been addressed to obtain higher discrimination between different strains of the same ST, clonal complex (CC) or lineage in many bacteria, such as *Staphylococcus aureus*[252–254]*,* *Chlamydia trachomatis*[255], *Salmonella typhimurium*[256] or *Clostridium difficile*[257]. This thesis reports the first in-depth analysis of intra-ST *L. pneumophila* variability (see Chapter 5).

4.2. Recombination and mobile genetic elements

The use of low-resolution molecular markers for bacterial typing and species delineation, such as MLEE or PFGE but also sequence-based typing schemes, led to the first extended idea of clonal population structure in bacteria[199]. This was the case, for instance, of different species of *Salmonella*[258], *Neisseria meningitidis*[259,260] or *L. pneumophila*[261,262]. However, clonality started to be questioned, with the discussion that pathogenic bacteria might need a certain level of localized recombination to maintain useful variation within the population[263]. Besides, in many bacterial species, evolutionary changes with long-term consequences, such as a major shift in pathogenicity or ecological niche, or short-term, such as the acquisition of antibiotic resistance, are more likely to occur by recombination rather than point mutation[264].

L. pneumophila has been proven to be naturally competent for external DNA acquisition[22], to contain a chromosomal conjugation system with an origin of transfer (*oriT*)[265], and to be able to mobilize and conjugate plasmids and genomic islands mediated by the Dot/Icm system[65,66,235,265]. However, it is important to emphasize that although there are integrative and conjugative elements that encode for their own excision, transfer and integration, others require the recombination machinery of the recipient to integrate[266]. Genome analyses have confirmed the results from the first studies that revealed possible intergenic recombination events using SBT data[267,268], although intragenic events have only been detected very recently for the virulence effector gene *sidJ*[269].

Coscollá and colleagues (2011)[268] reported an estimation for the ratio between the population recombination rate (ρ) and the population mutation rate (θ) of approximately $\rho/\theta = 0.44\text{-}0.47$ (95% credibility region) and a ratio of nucleotide changes introduced by recombination events (r) relative to point mutations (m) of r/m = 2.6-5.7 (95% credibility region) for SBT data. This data showed that it is around four times more likely that a site has changed by recombination than by mutation. Similar recombination rates were found using Multi-Locus Sequence Typing (MLST) data for example in *Campylobacter jejuni*, *Haemophilus parasuis* or *Pseudomonas syringae*[270]*.*

The use of genomic data to calculate these recombination rates has shown differences with respect to the rates obtained using MLST data. In the study by Coscollá and colleagues (2011)[268], an r/m between 0.280 and 0.313 was found for an alignment of four *L. pneumophila* reference genomes. They also inferred that approximately 34-57% of the genome could be involved in recombination events, including two hotspots containing proteins of types II and IV secretion systems. Furthermore, HGT was found to occur with other phyla normally present within the cytoplasm of amoebas in more than 40% of the genes studied[268]. In the work presented in Chapter 5 of this thesis, genomic recombination rates were inferred from a considerable collection of clinical and environmental strains of this bacterium, and its comparison with other species is discussed.

Cazalet and colleagues (2008)[246] used DNA arrays to characterize the genomic variability in a large collection of *Legionella* strains from different species. They concluded that, despite the high variability of the genus, *L. pneumophila* has many known virulence and eukaryotic-like genes that are highly conserved within the species but absent or highly divergent in others. Besides, they found the operon coding for the core and the O side-chain synthesis of the lipopolysaccharide (LPS cluster), determining of serogroup 1 *L. pneumophila*, in other strains with very different backgrounds. This result suggested the involvement of the LPS cluster in HGT. Further works using genomic sequences confirmed the high frequency of HGT and recombination of large chromosomal fragments of different origins[271]. Siefert (2009)[272] first introduced the concept of 'mobilome', which in the case of *Legionella* is formed by plasmids, integrative conjugative elements, insertion sequences, genomic islands and also the accessory genes[244].

Glöckner and colleagues (2008)[265] were the first authors to describe two genomic islands (Trb-1 and Trb-2, of 42 and 34 kb respectively) in the Corby strain of *L. pneumophila*. They also demonstrated that Trb-1, which contains a complete conjugation system, had the capability of excising from the chromosome, being transferred horizontally by conjugation and site-specifically integrated into the acceptor genome. Another work[236] provided a detailed list of all genomic islands found in the first five sequenced strains (Philadelphia 1, Paris, Lens, Corby and Alcoy) and classified them as resistance-related islands, transport/secretion systems, DNA transfer, integrated and phage-related, and others without a well-defined role or composition (Fig. 9).

These islands were discussed to be associated to the virulence of the different strains, being as an example, Corby and Alcoy the ones with the highest number of islands related to virulence and DNA transfer. Some of these islands provide *L. pneumophila* systems for protecting themselves against foreign DNA, such as CRISPRs or toxin-antitoxin systems[236,271]. As an example of the potential benefit of incorporating one of these mobile elements, Flynn and colleagues (2014)[273] reported an integrative conjugative element, denoted as ICE-βox, that provides resistance to β-lactam antibiotics and oxidative stresses. This element was discussed to increase survival within macrophages as well as *L. pneumophila* fitness in natural and engineered water systems.

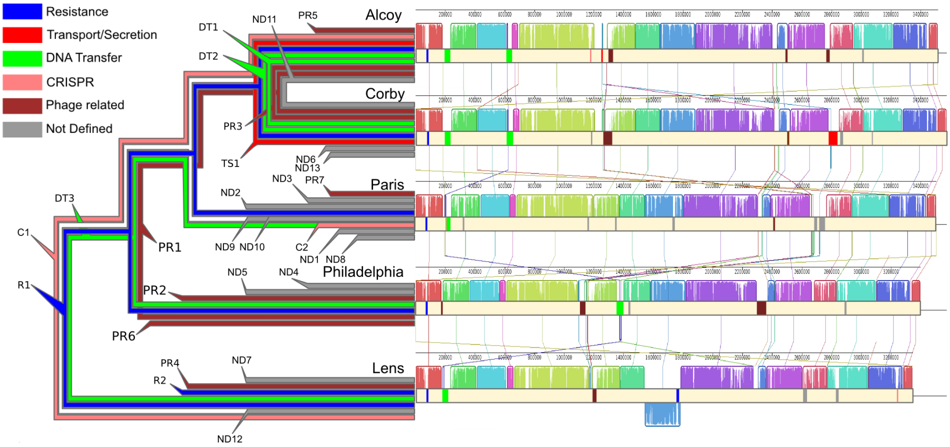


Figure 9 | Distribution of genomic islands over the Sequence-Based Typing tree topology of five *L. pneumophila* reference genomes (Alcoy, Corby, Paris, Philadelphia and Lens). Branch colours represent different island types (see legend). R: Resistance; TS: Transport/Secretion; DT: DNA transfer; C: CRISPR; PR: Phage related; ND: Not defined. Next to the tree, the multiple genome alignment of the five strains is shown (MAUVE software[274]), with the different islands marked along the genomes in their corresponding colours (adapted from D'Auria *et al*, 2010)[236].

4.3. Implications for outbreak investigations

Whole-genome sequencing is being increasingly applied to other fields apart from the determination of a complete genome from a new species or comparative genomics. Public health has been shown to benefit from these high-throughput approaches as they can complement the epidemiological investigation during outbreaks. Specifically, when cases of infection are suspected to be part of a cluster, one of the key questions to solve is if they have been caused by same or different strains[275]. Molecular epidemiology typing tools have been widely used to establish the relatedness between the strains causing the different cases. This has been achieved either by discrimination through electrophoretic bands in a gel or the assignment of bacterial isolates to sequence types, which can be further grouped into clonal complexes[220,275,276]. However, even sequence-based approaches, such as MLST, only include a small portion of the genome under study. This means that they can have low resolution when there is further variability outside the analysed loci determining sublineages or clusters.

From Sanger sequencing[277], which pioneered the massive study of complete genomes especially in organisms of small genome size such as viruses[278], the biggest explosion in genome sequencing of microorganisms came with the arrival of bench-top sequencers[279]. These have allowed an increasing throughput of WGS at a rapidly decreasing cost, thus already making it the gold standard molecular typing method for surveillance and outbreak investigation[280]. This approach has allowed not only the sequencing of new bacterial species, but also the study of outbreaks in real time, such as during the 2011 European outbreak of *Escherichia coli* O104:H4[281,282], or the development of massive re-sequencing projects. As an example, several clusters within a single sequence type causing a methicillin-resistant *Staphylococcus aureus* outbreak were detected using WGS[283]. In this thesis we have used this approach to characterize two sublineages of *L. pneumophila* ST578 colonizing the endemic area of Alcoy (Alicante, Spain) (see Chapter 5).

The higher discrimination provided by WGS leads to the discovery of SNPs between strains that can help to determine the transmission routes, a strategy that has been widely used for pathogens with patient-to-patient transmission such as *M. tuberculosis*[284,285], *M. abscessus*[286] or *K. pneumoniae*[287]. Interestingly, WGS has already unveiled the historical patterns of global dissemination of pathogens[251] such as *Yersinia pestis*[288], responsible for the Black Death, *Mycobacterium leprae*[289], the causal agent of leprosy, or *Vibrio cholerae*[290], currently causing the seventh cholera pandemic. Furthermore, apart from additional discrimination between outbreak strains, WGS links epidemiology to the biology of the pathogen, its genome structure, evolution and gene content, providing also information on important markers such as those involved in resistance and virulence[276].

WGS can contribute to different steps of an outbreak investigation, from the initial confirmation of the outbreak, to the generation of hypotheses using geographical or temporal data or the institution and verification of control measures, as reviewed in Robinson *et al* (2013)[275]. However, all the new opportunities provided by WGS also imply many challenges, mainly the processing and storage of huge amounts of data[275,291]. For genomic epidemiology to be applied routinely in outbreak investigations and surveillance programs, there is still a further need of standardization of methods, centralization and expansion of genomic databases. Besides, having a good genomic representation of the pathogen population in order to have a deeper overview of the global strain diversity, as well as a centralized amount of accessible reference genomes, would enhance the epidemiological investigation. However, this is still not available for most species[292]. Although large genome sequencing centres and some public health departments can currently afford to use WGS and have the expertise to process the resulting data, this is still a medium-term goal for most public health and hospital laboratories that cannot replace the standard typing methods yet. Furthermore, although providing loads of data, the genomic investigation of an outbreak can complement but not replace the epidemiological investigation, which is indispensable for linking possible sources of transmission, contact tracing and environmental reservoirs of the outbreak.

In the case of legionellosis, the use of standard typing tools has led to the general belief that outbreaks are caused by a single source contaminated by a clone that overgrows due to environmental factors. Then, bacterial clones are released in aerosols from which they infect and produce disease in susceptible people. Thus, a match between the sequence type of clinical and environmental isolates, although infrequent, has been considered sufficiently decisive to declare the source as identified. Chapter 5 shows the first work published up to date that uses WGS to characterize 13 different outbreaks caused by *L. pneumophila* ST578 during 10 years in a locality in which it is considered to be endemic. Results revealed the underlying diversity within this sequence type, with different strains contributing to the same outbreak, and the high impact of recombination on outbreak-causing strains. Only two other works published up to date have applied WGS to study legionellosis outbreaks[241,293].

5. Life within a community: lessons from metagenomics.

Legionella has a biological cycle in which intracellular life is essential[5]. Other bacterial pathogens also replicate inside protozoa, such as *Mycobacteria*, *Chlamydia, Listeria, Burkholderia* or *Francisella*, as reviewed in Harb et al (2000)[294]. Other non-pathogenic microorganisms have also been found in the cytoplasm of amoebae[268], thus meaning that *Legionella* and other intracellular organisms are not alone inside their hosts or biofilms[21]. In order to study bacterial communities, the standard procedure until the early 90's relied on microbial identification of isolates from cultures followed by further microbiological and biochemical characterization[295]. However, it has been known for many years that some viable strains of different species are not culturable, either because the culturing conditions are not suitable for them and they enter in a non-culturable state or because they are not culturable[295,296]. It is interesting to point out the results of a recent report by Tanaka and colleagues (2014)[297] in which they discuss the high proportion of unculturable species to result from a pitfall in the preparation of agar media. This non-culturability has been estimated to affect up to 99.8%[298] of the microbial species, which makes their characterization very difficult. This has led to the development of culture-independent strategies and the application of genomic sequencing to complex microbial communities, in what is currently known as metagenomics.

Genomic procedures using cloning of DNA fragments and Sanger sequencing started to be applied to complex samples in the early 2000s[299]. This further evolved to high throughput approaches with larger studies such as the Global Ocean Sampling (GOS)[300] or TARA Oceans[301]. Other large projects are aimed at characterizing the global microbial diversity of our planet (Earth Microbiome Project[302]; www.earthmicrobiome.org) as well as the bacterial populations colonizing indoor environments such as the Home Microbiome Study (www.homemicrobiome.com) or the Hospital Microbiome Project (www.hospitalmicrobiome.com). Human microbiota analyses are also leading the field with different consortia[303] studying the bacterial communities of different parts of the human body, from the oral cavity[304] to the gastrointestinal tract[305].

Many of the works applying WGS to the study of microbial communities have used a direct approach to analyse the population composition by targeting the 16S ribosomal RNA (rRNA) subunit gene[306]. Specifically, DNA is extracted from the samples and a targeted PCR is applied to amplify this segment using universal primers. 16S rRNA is used because it is known to be conserved among species and thus, to be representative of the taxonomic classification[307]. This methodology overcomes the problem of the high amount of material needed for constructing the sequencing libraries, because the product comes from PCR amplification. However, the method is targeted only to bacteria and has numerous pitfalls in the PCR amplification step, such as the formation of chimeras, the introduction of contaminants in the process that are amplified during PCR or the amplification bias toward certain species[308]. In cases in which total DNA yields sufficient material, direct shotgun sequencing of the extracted nucleic acid material is recommended in order to overcome these problems. However, both techniques have been shown to be extremely sensitive to contamination, especially in highly diluted or low biomass samples[309].

Regardless the sequencing approach, the resulting composition of the population represents just a fraction of the community. A large portion of the reads normally remains taxonomically unassigned due to lack of representation of all the variability in the databases. Woyke and Rubin (2014) hypothesized that these unassigned reads could come from an undiscovered domain of life[310]. In the case of metagenomics, data handling also remains a challenge, because the methodology needs to be adjusted from the study of single species in genomics to dealing with a mixture of sequences from a complex population. In addition, processing huge amounts of data requires a higher expertise in bioinformatics than other simpler works in which fully automated pipelines can be sufficient[311–313].

Metagenomic efforts to characterize environmental communities have been mainly focused on different water sources and to which the GOS project has made a huge contribution. For instance, they have characterized different surface ocean, marine, freshwater and hypersaline samples[300,314,315], following the pioneering study of the Sargasso Sea by Craig Venter and colleagues (2004)[316]. In fact, although using standard cloning approaches, the firsts studies analysing marine communities date from the early 90's[317]. Although the bacterial composition of surface waters seem to be homogeneous, salinity gradients have an important effect with different taxonomic groups dominating distinct salinity regimes[318]. Other natural water sources such as deep-sea hydrothermal vent chimneys[319], coastal lagoons[320], hot springs[321] or salty ponds[322,323] have also been studied with high throughput sequencing approaches. Besides, metagenomics has also been used to study the effect of disinfectants on the bacterial communities in water distribution systems[324–326], where the genus *Legionella* was found in a fraction below 1% of the total bacterial abundance[325,326]. Yooseph and colleagues (2013) also reported *Legionella* as one of the most abundant genera in an air sample taken from an outdoor environment in New York City[327].

Analyses on biofilms created in the surfaces of drinking water networks started to be reported in 2003 using 16S-targeted and random sequence cloning, with *Rhizobium*, *Pseudomonas* and *Escherichia* being the most abundant genera[328]. A more limited diversity was found in a study of the bacterial composition of a corrosive biofilm associated to steel pipelines[329]. Recently, 16S rRNA pyrosequencing has been applied to the study of differences in abundance and composition of biofilm communities in drinking water distribution pipes[330,331], mainly dominated by *Methylomonas*, *Acinetobacter*, *Mycobacterium* and *Xanthomonadaceae*[331]*.* The impact of water treatment procedures as well as water conditions and pipe materials has also been reported targeting 16S rRNA and even 18S rRNA to evaluate microbial and eukaryotic diversities. These works also evaluate the impact of introducing *L. pneumophila* and one of its amoebal hosts, *Acanthamoeba polyphaga*, in the ecosystem[332–334]. To our knowledge, there is no work that assesses both bacterial and eukaryotic diversity in biofilms obtained from natural springs using direct shotgun sequencing. Further discussion on this topic will be addressed in Chapter 6.

Objectives

This thesis aims at getting a deep insight into the molecular epidemiology of *L. pneumophila* populations through standard genetic and high-throughput genomic approaches and its applicability to public health investigation of legionellosis outbreaks. The specific goals are the following:

- To study the genetic variability and structure of *L. pneumophila* sequence types populations in Comunidad Valenciana, second region in Spain with the highest incidence of legionellosis.
- To develop a sensitive method for *L. pneumophila* detection from environmental samples independent from the microbiological culture with an immediate application to real-time outbreak investigations.
- To apply high throughput sequencing approaches to study how variability is acquired in *L. pneumophila* populations at a local scale and find genomic traces of recombination.
- To evaluate the phylogenetic and phylodynamic relationships between outbreak-related strains, specifically of ST578, recurrent in the locality of Alcoy (Alicante, Spain), at the genomic level.
- As *Legionella* is a frequent inhabitant of the freshwater microbiota and, specially, the associated biofilms, to explore the composition of the microbial, eukaryotic and viral community of biofilms from natural springs in the endemic area of Alcoy.

Chapter 1

Geographical and temporal structure of *Legionella pneumophila* sequence types in Comunidad Valenciana (Spain), 1998-2013

Leonor Sánchez-Busó[1,2], Fernando González-Candelas[1,2].

1. Unidad Mixta Infección y Salud Pública FISABIO/Universitat de València. Avenida Cataluña, 21; 46020, Valencia, Spain.

2. CIBER en Epidemiología y Salud Pública, Valencia, Spain.

This work has been submitted to *Applied and Environmental Microbiology*.

Abstract

Legionella pneumophila is an accidental human pathogen associated to aerosol formation in water-related sources. High recombination rates make *Legionella* populations genetically diverse, and near two thousand different sequence types (STs) have been described for this environmental pathogen up to date. The spatial distribution of STs is extremely heterogeneous, with some variants being present worldwide while others are detected only at a local scale. Similarly, some STs have been associated to disease outbreaks, such as ST578 or ST23.

Spain is among the European countries with highest incidence rates of reported legionellosis cases and, specifically, Comunidad Valenciana (CV) is the second most affected area in the country. In this work, we aimed at studying the overall diversity of *L. pneumophila* populations found in the period 1998-2013 in 79 localities encompassing 23 regions within CV. To do so, we performed Sequence-Based Typed (SBT) on 1,088 *L. pneumophila* strains detected in the area from both environmental and clinical sources. A comparison with the genetic structuring detected in a global dataset that included 28 mainly European countries was performed. Our results reveal a level of diversity in CV that can be considered representative of the diversity found in other countries worldwide.

Introduction

The main reservoirs of *Legionella pneumophila* in natural habitats are water-related environments, where it is known to form biofilms in surface interphases[31]. *L. pneumophila* can multiply actively within those complex structures[21], disperse through the water flow and colonize different urban water distribution systems and risk facilities, such as cooling towers[335]. Aerosols loaded with sufficient amount of this bacterium contribute to its dispersal[5] and, when inhaled by humans, it can produce an opportunistic pneumonia known as Legionnaires' Disease (LD)[82]. No person-to-person transmission has been reported for *Legionella* so that it is commonly considered as a strictly environmental pathogen[5].

Different molecular markers have been used to characterize *Legionella* populations in previous genetic studies[196], but Sequence-Based Typing (SBT) has become the current gold standard typing tool[216,217]. Apart from its utility in the epidemiological investigation of LD outbreaks[117,336,337], SBT provides nucleotide sequence data that can be used for further analysis on genetic variability and population structure[267,338]. A recent study has shown that, even though undetected genomic variability within sequence types (ST) could mislead the identification of *Legionella* reservoirs during outbreak investigations, genetic distances using SBT data correlate significantly with genome-wide estimates[339].

Characterizing the distribution of different STs across space and time is essential to better understand dispersal patterns of environment-associated bacteria. This is particularly true for *L. pneumophila*, which is generally considered as an accidental pathogen[10]. Our team has been analysing most of the detected strains during outbreak investigations and surveillance programs in the Comunidad Valenciana (CV) region for more than 15 years. During the period 1999-2011, CV and Cataluña summed over three-quarters of the total legionellosis outbreaks reported in Spain, with 124 and 331 registered episodes, respectively[340]. Despite being the second most important Spanish region in number of reported LD cases, *Legionella* populations in the CV region are still poorly characterized. An initial report using only three loci from the SBT scheme to genotype a few isolates from the Alicante province (within the CV region) already revealed a large variability within this area[338], comparable to that found in Europe.

In this work, we present an extensive survey of the genetic variability and population structure of *L. pneumophila* strains, including over 1,000 isolates sampled throughout the whole CV during the 1998-2013 period. In addition, we compare the genetic diversity detected in the CV with the *L. pneumophila* Sequence-Based Typing database[221], which includes strains from 28 different countries, mostly from Europe. Our objective is to provide a more thorough view of the genetic diversity of this opportunistic pathogen at different geographic scales, which can be used for epidemiological analyses in the area in subsequent years.

Materials and Methods

Strain collection and processing

A total of 1,088 clinical and environmental samples were collected across the CV region during the period 1998-2013 (Table S1). Specifically, 398 samples were from clinical origin (sputums, broncho-alveolar aspirates or *L. pneumophila* isolates) and 690 were from environmental sources (Table S1), mainly isolates from water. Samples were obtained during routine surveillance programs for *Legionella* control as well as outbreak investigations in 97 different localities spanning 24 regions (Fig. S1 and Table S2). All the samples were referred to our laboratory for *Legionella* detection and genetic characterization.

DNA was extracted from pure cultures by suspending bacterial biomass in water and applying a thermal-shock protocol as described previously[341]. Sputums and similar samples were treated with UltraClean® BloodSpin® DNA Isolation Kit following the manufacturer's instructions to extract total DNA. Quantity and purity of the nucleic acids were measured using NanoDrop™ 1000 (Thermo Scientific) and DNA was stored at -20ºC until used.

Sequence-Based Typing (SBT)

The seven loci of the SBT scheme for *L. pneumophila* (*fliC, pilE, asd, mip, mompS, proA* and *neuA*)[215–217] were amplified using standard PCR. Primers, mixture, amplification and sequencing conditions were performed as described previously[341]. A semi-nested PCR was subsequently applied on the samples coming from sputums and similar samples as described elsewhere[228]. Consensus sequences were retrieved for the amplified loci in all the samples by using forward and reverse chromatograms with the Staden package[342]. The corresponding allele for each sequence was assigned comparing with the *Legionella* SBT database[221].

SBT clustering and phylogenetic reconstruction

Isolates with complete allelic profiles (n=643) were used for clustering the different patterns into groups of Single Locus Variants (SLV) using the goeBURST full MST (Minimum Spanning Tree) algorithm in PHYLOViZ v1.1[343]. Strains from the CV with at least four successfully genotyped loci (n=778) were used for phylogenetic inference. Sequences for the SBT loci were concatenated for each isolate using R[344]. Missing loci were replaced by gaps. Maximum likelihood (ML) phylogenetic reconstruction was performed over the concatenated alignment with RAxML v7.2.8[345] using the

GTRGAMMA model of nucleotide substitution and 1,000 bootstrap replicates. The Interactive Tree of Life (iTOL)[346] utility was used to visualize and collapse the resulting phylogenetic tree.

We have previously shown that genetic variability in the genome of *L. pneumophila* strains is derived from recombination rather than substitution processes[339]. This frequent exchange or acquisition of genetic material makes the actual relationships among strains to be better represented by a network than by a bifurcating tree. Therefore, the concatenated SBT haplotype sequences of the CV strains were used to build a Median Joining Network using Network v4.613 with an epsilon value of 10[347].

The allelic profile of the SBT scheme was downloaded for the 1,709 different sequence types available in the ESGLI database[221] at the time of starting this work (February 2014; 6,478 entries). An alignment of the seven loci for each ST was retrieved by extracting and concatenating the different alleles with R[344]. The ESGLI-dataset multiple sequence alignment was also used for ML phylogenetic reconstruction as described above for CV data.

Analysis of diversity and population structure

Estimates of the nucleotide divergence between *L. pneumophila* populations in the two different datasets (CV and ESGLI) was obtained using DnaSP v5.10.01[348]. The Ewens-Watterson test was performed using Arlequin v3.5.1.2[349] over the different localities and regions in the CV as well as the different years of isolation. This test was applied to test whether the haplotype frequencies significantly deviated from the expected values under neutrality. Arlequin was also used for testing the partitioning of the genetic variation within and among localities and areas in the CV dataset (79 localities over 23 areas) as well as different regions within countries in the ESGLI dataset (258 regions over 28 countries, mainly European but also some American and Asian locations) using AMOVA. Travel-associated cases were not considered in the AMOVA test. Bayesian modeling (BAPS software)[350] was applied in order to cluster the different STs into 15 groups[242]. The resulting tree and BAPS clusters were visualized with iTOL[346].

Two factors were used as metadata to analyze the population structure of *L. pneumophila* isolates in Comunidad Valenciana: year of isolation and geographic region. The xAMOVA test (Daniel Wilson, unpublished; http://www.danielwilson.me.uk/xAMOVA.html) was performed in order to estimate the component of variance explained by each factor (two-way AMOVA test). This is an extension of the original AMOVA test (Analysis of Molecular Variance)[351] that considers the simultaneous effect of two variables over the overall variance. Statistical significance was estimated using 10,000 permutations. Waves of the most common sequence types along the different years were studied using R[344].

Results

L. pneumophila diversity in the Comunidad Valenciana region

A total of 1,088 samples were recovered from 24 different areas across the CV and used for *Legionella* detection and additional SBT analysis in the period between 1998-2013 (Tables S1 and S2). At least one locus could be amplified from 888 (74.5%) samples, 220 from clinical origin and 668 from environmental sources. From the 490 isolates with a known serogroup (3 clinical and 487 environmental), 73.8% (n=362) were of serogroup 1 (3 clinical and 359 environmental), known to be responsible for most of the clinical cases over the world[5]. From the 1,088 isolates, 643 could be successfully assigned to a specific ST (Tables S1 and S2) and up to 778 had at least four loci sequenced and were used for further phylogenetic analyses.

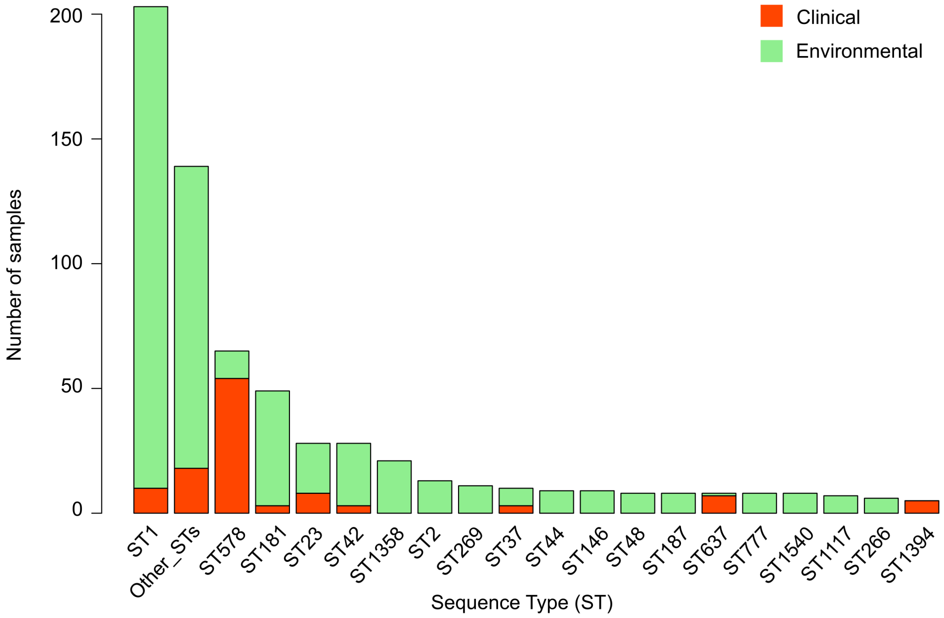


Figure 1 | Bar plot representing the number of clinical and environmental strains of each Sequence Type (ST) detected in CV during the period 1998-2013. STs with less than five strains are grouped as "Other_STs".

From a total of 102 different STs found in the completely characterized CV dataset (n=643), 54 (52.9%) were found at least twice, and the rest were singletons (Fig. 1). About 55% of the strains belonged to only 9 STs. Specifically, the most frequently isolated sequence type was ST1 (n=203) followed by ST578 (n=65), ST181 (n=49), ST42 (n=28), ST23 (n=28), ST1358 (n=21), ST2 (n=13), ST269 (n=11) and ST37 (n=10). These STs were found spread all over the global phylogenetic tree, showing the high diversity of the *L. pneumophila* populations in CV (Fig. 2). Twenty-two different STs were found in clinical samples, representing only a subgroup of the overall environmental variability. Twelve of those STs were also found in the environment (ST1, ST20, ST23, ST37, ST42, ST75, ST181, ST367, ST448, ST578, ST637 and ST1012). For the ten STs only found in patients, we only had one or two isolates, except for ST1394 (n=5).

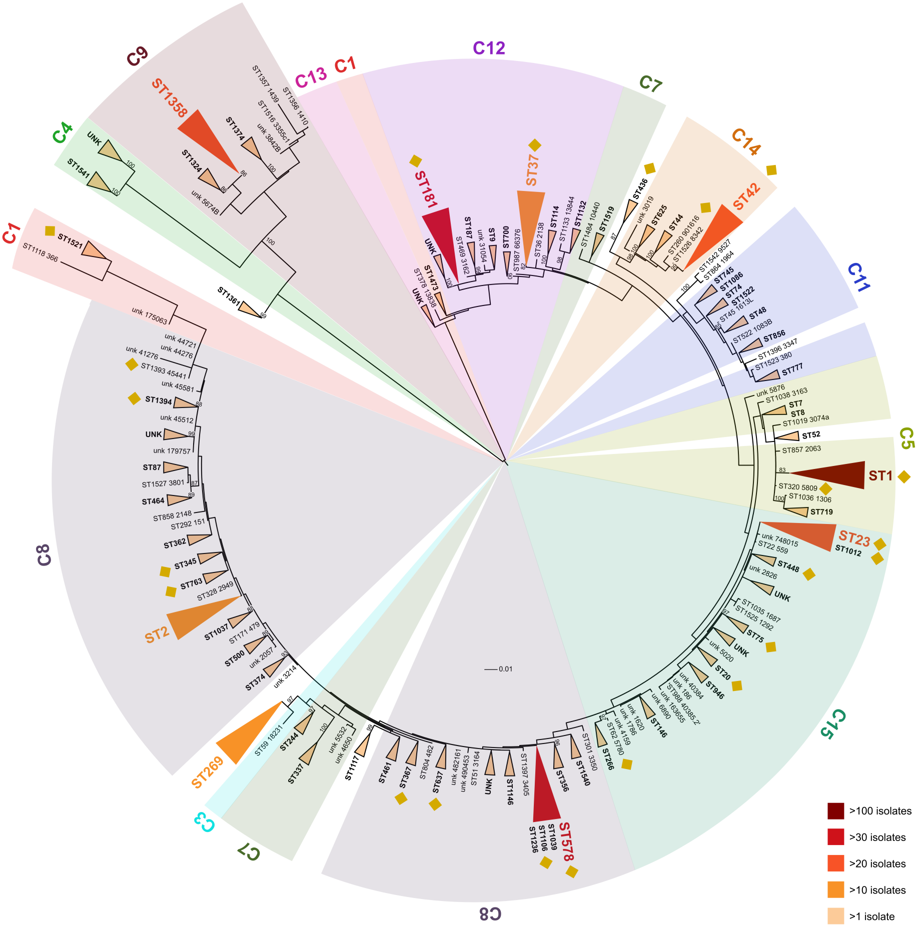


Figure 2 | Maximum likelihood phylogenetic tree of the 778 *L. pneumophila* strains from CV with at least 4 SBT loci sequenced. Clades with more than one strain of the same sequence type (ST) have been collapsed (see legend for colour differentiation of collapsed clades according to the number of samples in each ST). Shadowed clades (C1, C2, etc.) represent the different BAPS clusters as estimated from the ESGLI dataset. Yellow diamonds mark STs that have been detected in clinical cases of Legionnaires’ disease. Bootstrap support values higher than 80% are shown.

The Median Joining Network shown in Fig. 3 depicts the complex relationships between the different STs and suggests the existence of much more environmental variability than the one represented in our dataset. Using the SBT allelic profiles we detected 15 complexes of Single Locus Variants (SLVs) derived from a specific ST that could have acted as a founder for each of the complexes (Fig. 4). The different SLVs could have been formed either by accumulation of polymorphisms in one of the locus involved in the SBT scheme or by their involvement in recombination events.

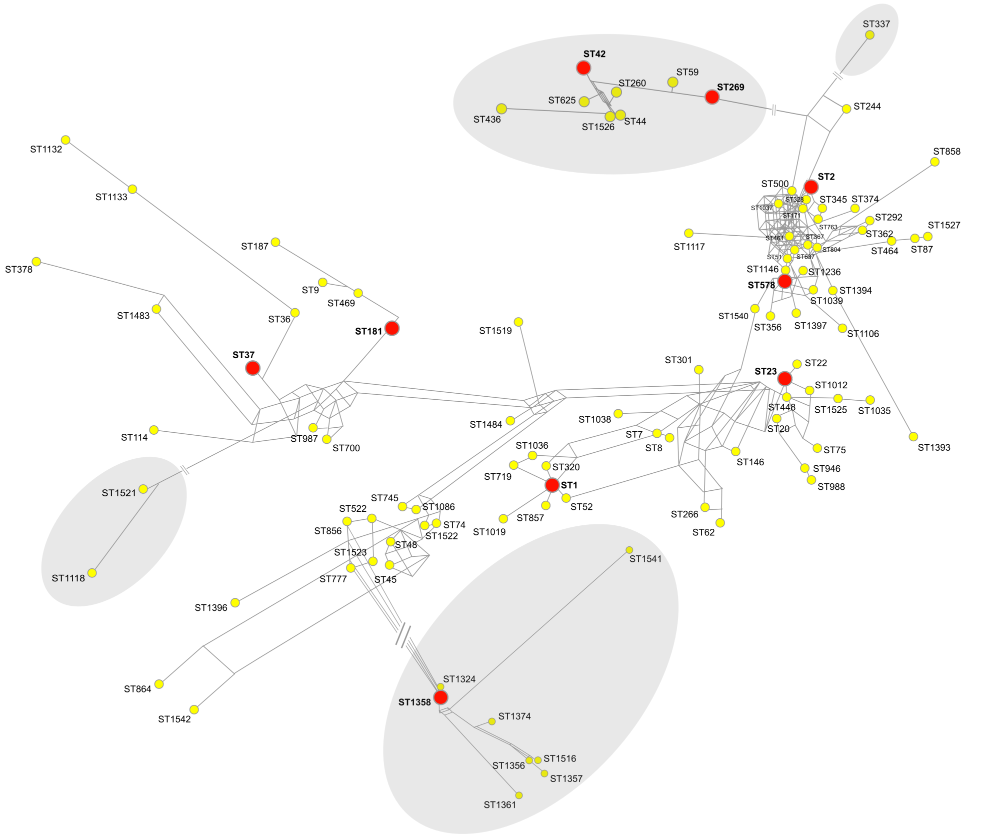


Figure 3 | Median-joining network built using Network for the haplotypes of the 643 *L. pneumophila* strains with an ST assigned. Coloured dots correspond to STs found in the CV dataset: yellow dots represent STs with less than 10 cases and reds those with more. Nodes without dots represent undetected intermediate states necessary to explain the observed variability. Branch lengths are proportional to the number of nucleotide differences. Shadowed clusters represent distant groups of STs whose links to the main network structure have been shortened for illustrative purposes.

Geographical and temporal structure

Samples with a complete ST were obtained from 23 different regions in the CV. Several STs showed a wide distribution across those sampling regions (Fig. 4), with ST1 being the most widespread profile (18/23 regions), followed by ST181 (9/23 regions) and ST42 (8/23 regions). Nevertheless, ST578 was mainly found in the locality of Alcoy, where it is known to be endemic[352] and other SLVs such as ST51 or ST637 were also exclusively found in the same area. ST1 was found in 41 different localities from 18 regions whereas 73 STs were exclusively found in different localities spread over 36 regions. Forty-eight of these STs were detected only once (singletons). The Ewens-Watterson test was performed over the 23 different regions and 79 localities and revealed that the observed haplotype frequencies in the different locations did not deviate significantly from the expected values under neutrality (p-value > 0.05).

A temporal overview of the frequency of each ST per year revealed waves of STs arising in the area (Fig. 5), with ST1 being dominant in most years. The number of ST1 strains detected in the 16-year period analysed correlated with the diversity of haplotypes found per year (Pearson's r = 0.8644; p-value = 1.57E-05; Fig. S2). Nevertheless, years 2006 and 2009 were found to deviate from this linear relationship, mainly because of the outburst of ST181 in 2006, affecting 4 different localities, and ST578 and ST42 in 2009, affecting one locality each (Fig. S2). The ST578 peak in 2009 corresponded to an outbreak in the locality of Alcoy in 2009[117]. No further conclusion can be made from these results because of the implicit bias in the dataset under study, which is dependent on the different sampling efforts made on specific points due to outbreaks or sporadic cases. To study the contribution of year and geographical region of isolation in the distribution of the genetic variability of *L. pneumophila* in the CV, we performed a two-way AMOVA test (Table 1). The results showed that 33.2% of the total variance could be explained by the geographic distribution (p-value < 0.0001) and that a 24.6% of the variance could be explained by differences in the population structure by year (p-value < 0.0001).

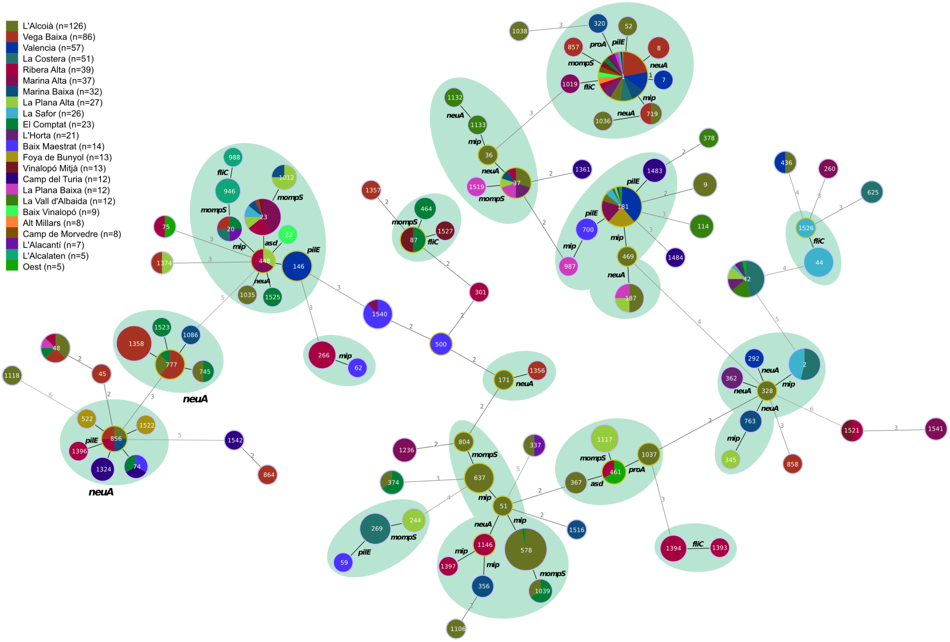


Figure 4 | Minimum-spanning tree obtained with PhyloViz on the SBT profiles of 643 *L. pneumophila* strains. Clusters of Single-Locus Variants (SLVs) have been shadowed. Pie charts represent the 102 different STs in the CV with size proportional to haplotype frequencies and the different colours refer to the 23 different areas under study in the CV region (see legend). Numbers on branches represent the number of loci different from the founder ST. For SLVs the name of the changing locus is represented.

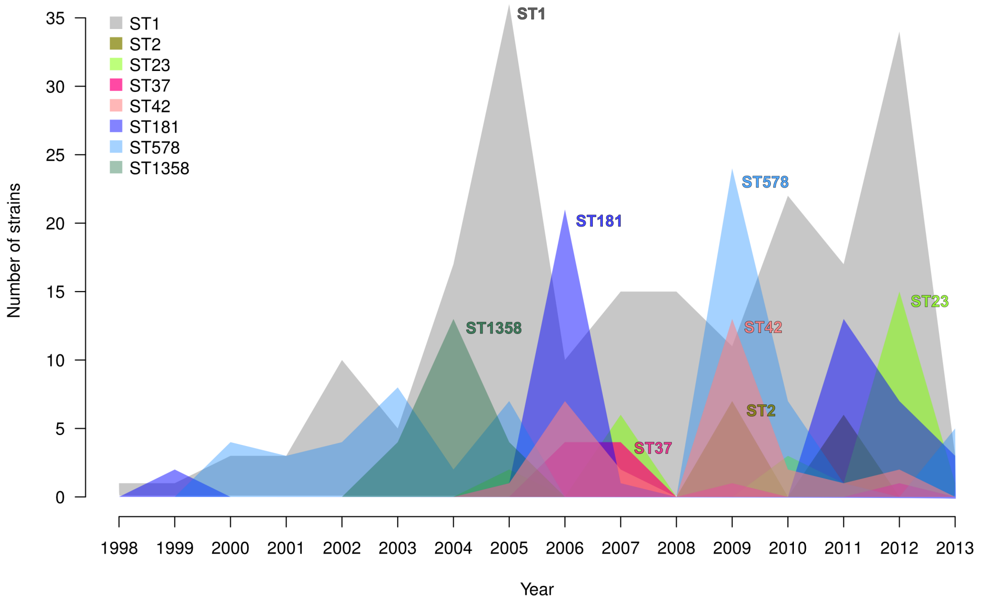


Figure 5 | *L. pneumophila* STs detected in the period 1998-2013 in the CV represented as number of strains per year. Only the eight most abundant STs that appear more than one year are shown.

Table 1 | Two-way AMOVA test. The effect of geographical and temporal structuring on the genetic variability of *L. pneumophila* samples from the Comunidad Valenciana in the 1998-2013 period is evaluated.

Factor	df	Seq SS	Adj SS	MS	F	Variance	%Variance	P-value
Region	22	3835.2	5434.1	247.00	17.615	10.99	33.15	<0.001
Year	16	3644.5	3644.5	227.78	16.244	8.15	24.58	<0.001
Others	604	8469.6	8469.6	14.02	-	14.02	42.27	

df = degrees of freedom.
Seq SS = Sequential sum of squares deviations for each factor.
Adj SS = Adjusted sum of squares deviations for each factor.
MS = Mean square deviation (Adj SS/df).
F = F-statistic representing the ratio between each factor (region/year) and other factors.

Analysis of local versus global variability

To better evaluate the levels of genetic diversity of *L. pneumophila* in CV, we compared results from our dataset with the global variability reported to the ESGLI database. Sequences from the seven loci of one representative of each of the 1,709 STs included in the ESGLI database were concatenated and a ML tree was inferred (Fig. S3). Concatenated sequences were clustered in 15 groups using BAPS and the results were incorporated into the phylogenetic tree. Approximately 20% of the profiles (n=339) could not be assigned to specific clusters and were considered as resulting from admixture. Cluster 8 was the group including more profiles (n=413, 24% of the total number of STs in the ESGLI dataset).

Isolates from the CV were present in 12 of the 15 BAPS clusters, although clinical strains only spanned six of them (Fig. S3; Table S3). Cluster 8 was also the most frequent in the CV (n=30, 29%), containing two of the most often detected STs (ST2 and ST578), followed by clusters 11 and 12, with 24% and 13% of the CV STs respectively (Fig. S3). The other six most frequently reported STs clustered in five different groups and ST269 was found as a mixed profile (Table S3). Interestingly, one of the largest clusters in the ESGLI dataset (Cluster 3; 110 different STs) is scarcely represented in the Comunidad Valenciana, with only one ST detected in the entire region (Table S3). In general, the number of STs by cluster in the ESGLI dataset correlated significantly with the number of STs by cluster in the local CV dataset (Pearson’s r = 0.80; p-value < 0.001).

An AMOVA test using the nucleotide sequences from the SBT typing scheme was performed with the two datasets. In the ESGLI dataset, different regions (n=258) were grouped by country (n=28), while in the CV dataset different localities (n=79) were grouped by area (n=23). In both cases, the highest proportion of variation was found within populations rather than among populations within groups or among groups (Table S4). However, the distribution of variation differed between the two datasets: the ESGLI dataset showed a lower proportion of variation among regions within countries (8.67% of variance) compared to among localities within areas in the CV dataset (43.79% of variance). In addition, a higher proportion of the total variation was found within populations in the global dataset (84.77%) than in CV (58.06%), reflecting the higher diversity within the first dataset (Table S4). In both cases, the highest proportion of the total variation was found within the smaller areas considered (regions within countries and localities within areas), meaning that the diversity of *L. pneumophila* found in CV is comparable to the diversity of other locations.

Genetic differentiation was assessed through the estimation of DNA divergence corrected using the Jukes-Cantor model of nucleotide substitution. The average number of nucleotide substitutions per site between both the ESGLI and CV datasets was estimated as Dxy(JC) = 0.0214 ± 0.0005 and the number of net nucleotide substitutions per site between populations as Da(JC) = 0.0004 ± 0.0005. The intra-population nucleotide diversity estimates were comparable for both datasets (Pi(JC, ESGLI) = 0.0217 ± 0.0003; Pi(JC, CV) = 0.0204 ± 0.0009). The average number of nucleotide differences between ESGLI and CV was estimated to 52.24 changes, also similar to the intra-population estimates (k(ESGLI) = 53.37; k(CV)=49.69). These results further confirm that the high variability found in the Comunidad Valenciana is comparable to the overall variability found in other areas worldwide.

Discussion

The genetic variability of *L. pneumophila* has been assessed for countries such as the United Kingdom, Belgium, Portugal and Italy[210,223,226,353] but there is no similar information available from Spain, despite it being significantly affected by this opportunistic pathogen. Spain is known to be one of the countries with the highest number of reported cases of Legionnaires' Disease[6], and the Comunidad Valenciana is the second region with the most cases reported in the country.

In a previous work we described the variability of *L. pneumophila* in the Alicante province (South of CV) during three years, although only three of the current seven loci included in the SBT scheme were available at that time[338]. Here, we have analysed all the clinical and environmental strains that have been typed in our laboratory between 1998 and 2013. It is important to take into account that, as the samples were mainly obtained during the investigation of outbreaks and sporadic cases, there might be some bias to specific locations where these cases had occurred and where control measures were applied in a more intense manner. However, yearly data from routine environmental surveillance programs applied over risk installations in the whole area were also included, thus balancing the possible sampling bias.

L. pneumophila is a strictly environmental pathogen and human infection is considered as accidental[10]. As such, controlling this bacterium in the environment is the key point for controlling associated human infections. The CV has a high rate of reported cases of legionellosis and little is known about population structure of this bacterium in the environment. Our results show that, from the 643 samples fully typed in CV during 16 years, there is a high diversity of STs, with 102 different profiles detected in the region. ST1 was the most frequently found in environmental samples, with 203/643 (31.6%) of the strains having this genetic profile, in agreement with previous reports in other areas[197,226,354]. ST578 was found as the second most frequently reported type (65/643, 10.1%), mainly due to the recurrent outbreaks that this ST has caused over the years in the locality of Alcoy[339]. Only a subset of all CV STs was found in the clinical cases, a result that has also been reported in other studies. For example, in a surveillance work that used 443 clinical and environmental isolates from England and Wales[197], 82 different STs were detected among the environmental strains and only 42 different STs in the clinical isolates.

A Median Joining Network created using the SBT sequence data on the CV isolates showed a complex relationship between different STs. This complexity depicts the high rate of genetic exchange among *L. pneumophila* strains[267,339]. Many nodes are predicted that represent undetected genetic profiles and whose existence is postulated in order to explain the observed relationships between the STs in the network. These unobserved profiles could be STs that have rarely caused a clinical case, that are present in very low abundance in the environment or that are uncultivable. In fact, the environmental screening of water and biofilm samples using direct molecular methods has revealed the presence of mixed genetic profiles in samples resulting from the high underlying diversity present in the area[352]. Besides, looking at the combination of alleles of the SBT loci we can identify 15 complexes among the 102 STs, most of them formed by an ST that acts as a founder and SLVs that could result from intergenic recombination events from the former one.

Other surveillance works have reported that, apart from the widely-distributed ST1, particular STs might be more common in specific locations, such as ST47 in the United Kingdom and Belgium[197,353] or ST23 in Italy[226].

However, these STs can also be found sporadically in other places, showing the important spreading potential of this environmental pathogen. For instance, ST23 was found to cause an important outbreak in a hotel in the coastal city of Calpe (Alicante, Spain) in 2012[107]. Our analyses show that approximately 50% of the observed variability can result from geographical (33.2%) or temporal (24.6%) structuring (Table 1). Although 29 STs were found in more than one locality, the majority of profiles (73 STs considering the 48 singletons) were found exclusively in specific localities spanning 23 different regions of the CV. However, no significant deviation from the expected haplotype diversity was found in the different localities and regions. The same result was obtained for samples taken in different years. Temporal structuring was found in the shape of waves of STs bursting at specific years, such as ST181 in 2006 or ST578 in 2009. Nevertheless, this result can be affected by the aforementioned sampling bias, because the high numbers of these STs are dependent on the sampling effort in those particular years due to outbreaks. ST181 was found in four different localities in 2006, so we can say more confidently that there was an outburst of this ST in the area. All the ST578 cases in 2009 corresponded to a specific outbreak occurred at that time in the locality of Alcoy[117].

The comparison between the local (CV) and global (ESGLI) diversity of *L. pneumophila* revealed large similarities and little genetic differentiation between the two datasets. For instance, we mapped the different STs found in CV onto the global tree (Fig. S3) and found environmental representatives from the CV of 12 out of 15 of the clusters. STs found in clinical cases were included in only 6 of those clusters. In general, a good correlation was found between the numbers of STs from each dataset in each cluster. An analysis of variance over the global dataset, encompassing 258 regions from 28 different countries, showed that most of the variability was found within the different populations, meaning that there is a high diversity of *L. pneumophila* in most of the locations included in the test. A comparable result was obtained with the 79 localities spanning 23 areas in the CV. These results support the results of a previous work that already detected that about 80% of the total genetic variability in a specific area within the CV could attributed to intra-population differences[338]. The level of nucleotide differentiation between the two datasets is similar to those of genetic diversity within each dataset. These results indicate that two strains from the same dataset are almost as likely to be as different as any other two strains from different datasets each. Furthermore, these high levels of variability are mostly found at the smallest geographic scales tested in each dataset and in many different places. Interestingly, this variability is mainly due to variants isolated only once, which might imply that even a higher diversity from undetected or uncultivable environmental strains remains yet undiscovered.

In conclusion, we have characterized the genetic diversity of clinical and environmental samples of *Legionella pneumophila* from the Comunidad Valenciana region in Spain from 1998 to 2013 and we have compared it with the overall variability as reported to the ESGLI database[221] (mainly including samples from European countries). Although some STs have been found in specific localities, the most abundant STs can be found repeatedly in different countries. The high genetic variability detected can be created through the exchange of genetic material between strains that can spread to other locations. *L. pneumophila* is thought to be a recent pathogen for humans, that can spread to cause infections through aerosols and for which no person-to-person transmission has been reported, facts that contrast with the wide distribution of this pathogen. Genomic and metagenomic analyses will be crucial to evaluate the fitness of different lineages under distinct environmental conditions and more intense sampling projects could potentially find additional reservoirs that explain the yet undetected genetic diversity of this bacterium. Furthermore, this information will be crucial to elucidate the factors responsible for the observed global distribution of some STs and the adaptive or accidental nature of the minority variants found at local scales.

Acknowledgements

We thank the valuable contribution of other professionals to this work, most notably those working at the Clinical Microbiology services of public hospitals and at the Environmental Health units at the Public Health Centers of the Comunitat Valenciana. This work was funded by Conselleria de Sanitat de la Generalitat Valenciana and project BFU2011-24112 from Ministerio de Economía y Competitividad (Spanish Government). Leonor Sánchez-Busó has been recipient of a FPU fellowship from Ministerio de Educación y Ciencia (Spanish Government).

Supplementary material

Table S1 | Number of clinical and environmental samples analysed per year in the period 1998-2013. The same breakdown is also shown for the samples with a complete Sequence Type (ST). Two clinical samples with a complete ST do not have information about sampling year.

Year	All samples (1,088)		Complete ST (643)	
	Clinical	Environmental	Clinical	Environmental
1998	0	1	0	1
1999	3	4	0	4
2000	13	14	7	12
2001	4	16	3	15
2002	5	21	5	21
2003	10	18	8	18
2004	2	39	1	39
2005	78	91	16	72
2006	19	69	6	62
2007	17	66	1	43
2008	70	35	7	25
2009	48	61	21	45
2010	29	36	13	33
2011	22	48	7	41
2012	34	122	10	89
2013	44	49	4	12
Total	398	690	109	532

Table S2 | Number of clinical and environmental samples analysed per region in the Comunidad Valenciana (CV) in the period 1998-2013. The same breakdown is also shown for the samples with a complete Sequence Type (ST).

CV region	**All samples (1,088)**		**Complete ST (643)**	
	Clinical	**Environmental**	**Clinical**	**Environmental**
Alt Millars	0	8	0	8
Baix Maestrat	15	33	1	13
Baix Vinalopó	4	9	0	9
Camp de Morvedre	0	12	0	8
Camp del Túria	0	25	0	12
El Comptat	5	24	1	22
Els Ports	0	2	0	0
La Foia de Bunyol	0	14	0	13
L' Alacantí	1	7	1	6
L' Alcalaten	0	9	0	5
L' Alcoià	182	73	65	61
L' Horta	4	41	0	21
La Costera	6	60	0	51
La Plana Alta	8	32	5	22
La Plana Baixa	4	16	0	12
La Safor	0	28	0	26
La Vall d'Albaida	1	14	0	12
Marina Alta	23	36	7	30
Marina Baixa	59	36	17	15
Oest	3	4	1	4
Ribera Alta	43	39	8	31
Valencia	12	68	1	56
Vega Baixa	11	86	2	84
Vinalopó Mitjà	17	14	2	11
Total	398	690	111	532

Table S3 | Distribution of STs in the 15 BAPS clusters of the ESGLI (global) and CV (local) datasets. The nine STs most frequently found in CV are assigned to their corresponding BAPS cluster in the last column.

BAPS clusters	STs in the global dataset (ESGLI)	STs in the local dataset (CV)			% STs in CV	Most abundant STs in CV
		Total	Clinical	Environmental		
Cluster 1	69	2	1	1	2.90	-
Cluster 2	18	0	0	0	0.00	ST2, ST578
Cluster 3	110	1	0	1	0.91	-
Cluster 4	12	1	0	1	8.33	-
Cluster 5	134	9	2	8	6.72	ST1
Cluster 6	11	0	0	0	0.00	-
Cluster 7	79	4	0	4	5.06	-
Cluster 8	413	30	8	25	7.26	-
Cluster 9	72	6	0	6	8.33	ST1358
Cluster 10	12	0	0	0	0.00	-
Cluster 11	42	10	0	10	23.81	-
Cluster 12	85	11	2	11	12.94	ST37, ST181
Cluster 13	38	2	0	2	5.26	-
Cluster 14	106	5	2	4	4.72	ST42
Cluster 15	169	13	6	12	7.69	ST23
Admixture	339	8	1	7	2.36	ST269

Table S4 | Summary results from AMOVA tests performed on the global (ESGLI) and local (CV) datasets. Statistical significance (p-values) was obtained using 10,000 random permutations.

Global (ESGLI) dataset						
Source of variation	**df**	**Sum of squares**	**Variance components**	**% Variation**	**F-statistics**	**P-value**
Among countries	27	9485.3	1.78	6.56	0.066	< 0.001
Among regions within countries	230	14697.2	2.35	8.67	0.093	< 0.001
Within regions	4567	105032.5	22.99	84.77	0.152	< 0.001
Total	4824	129215.1	27.13	100		
Local (CV) dataset						
Source of variation	**df**	**Sum of squares**	**Variance components**	**% Variation**	**F-statistics**	**P-value**
Among areas	22	3855.1	-0.47	-1.85	-0.018	0.193
Among localities within areas	57	3832.8	11.19	43.79	0.430	< 0.001
Within localities	563	8352.6	14.84	58.06	0.419	< 0.001
Total	642	16040.5	25.55	100		

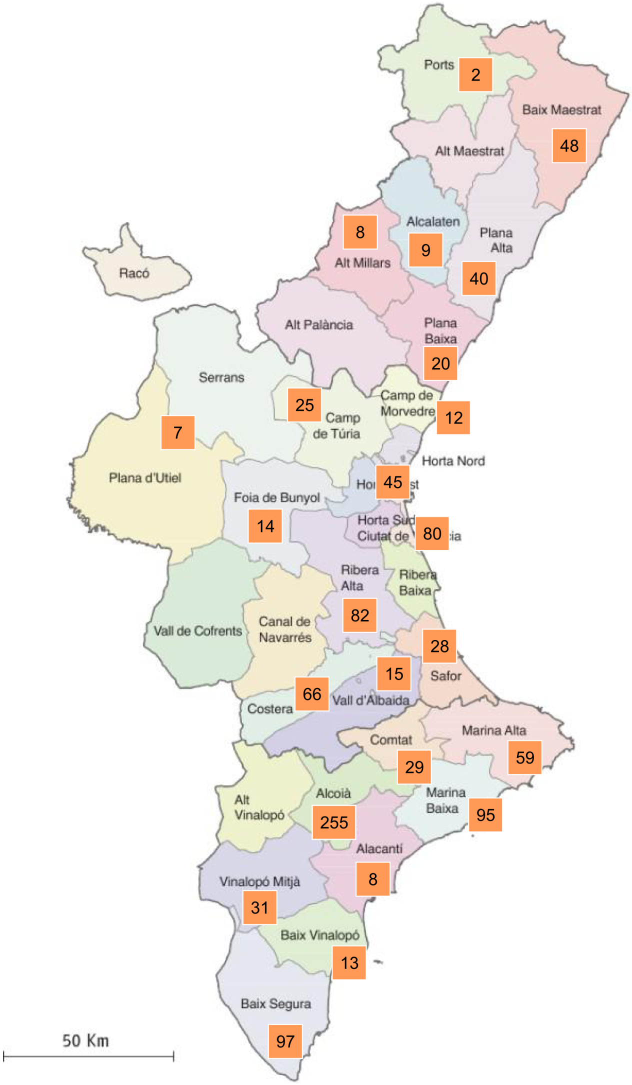


Figure S1 | Map of the CV area (East of Spain). The number of samples analysed in the period 1998-2013 in our laboratory are shown in the corresponding sampling region.

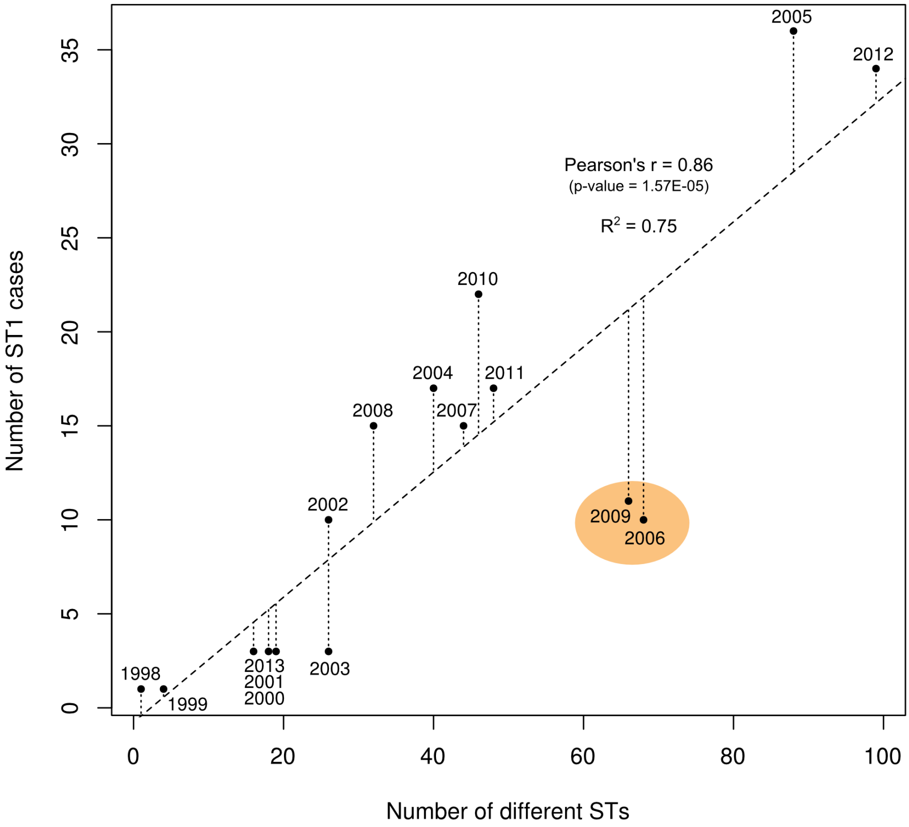


Figure S2 | Correlation between the number of ST1 strains and the number of different profiles detected in the CV per year. Years 2006 and 2009 are shadowed to highlight their evident deviation from the linear regression.

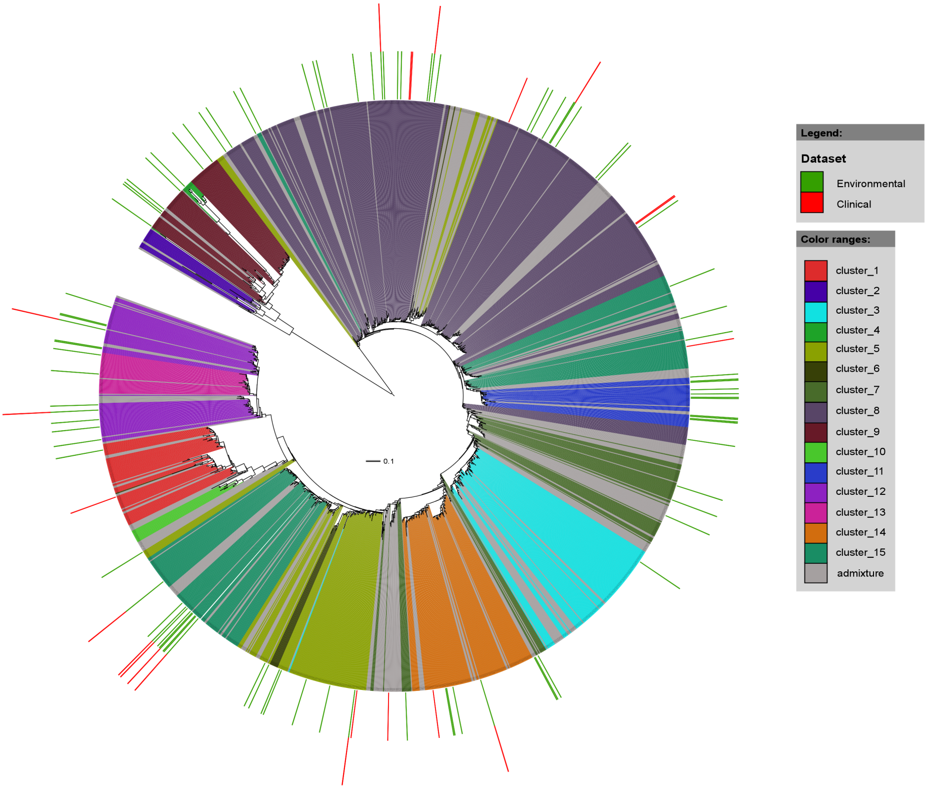


Figure S3 | Maximum likelihood phylogenetic tree performed over 1,709 different STs from the ESGLI database. Clade colours represent BAPS clusters (see legend). Green (environmental samples) and red (clinical samples) lines in the outer circle mark STs that have been detected in the CV.

Chapter 2

Genetic characterization of *Legionella pneumophila* isolated from a common watershed in Comunidad Valenciana, Spain

Leonor Sánchez-Busó[1,2]; Mireia Coscollá[3], Marta Pinto-Carbó[1], Vicente Catalán[4], Fernando Gonzalez-Candelas[1,2].

1. Genomics and Health Joint Unit CSISP (FISABIO)-University of Valencia/Cavanilles Institute. Valencia, Spain.
2. CIBER Epidemiology and Public Health, Valencia, Spain.
3. Tuberculosis Research Unit, Swiss Tropical and Public Health Institute, Basel, Switzerland.
4. Labaqua, S.A. Alicante, Spain.

Published in:

Sánchez-Busó L, Coscollá M, Pinto-Carbó M, Catalán V, González-Candelas F. Genetic Characterization of *Legionella pneumophila* Isolated from a Common Watershed in Comunidad Valenciana, Spain. PLoS One. 2013;8(4):e61564.

Abstract

Legionella pneumophila infects humans to produce legionellosis and Pontiac fever only from environmental sources. In order to establish control measures and study the sources of outbreaks it is essential to know the extent and distribution of strain variants of this bacterium in the environment. Sporadic and outbreak-related cases of legionellosis have been historically frequent in the Comunidad Valenciana region (CV, Spain), with a high prevalence in its Southeastern-most part (BV). Environmental investigations for the detection of *Legionella pneumophila* are performed in this area routinely. We present a population genetics study of 87 *L. pneumophila* strains isolated in 13 different localities of the BV area irrigated from the same watershed and compare them to a dataset of 46 strains isolated in different points of the whole CV.

Our goal was to compare environmental genetic variation at two different geographic scales, at county and regional levels. Genetic diversity, recombination and population structure were analyzed with Sequence-Based Typing data and three intergenic regions. The results obtained reveal a low, but detectable, level of genetic differentiation between both datasets, mainly, but not only, attributed to the occurrence of unusual variants of the *neuA* locus present in the BV populations. This differentiation is still detectable when the 10 loci considered are analyzed independently, despite the relatively high incidence of the most common genetic variant in this species, sequence type 1 (ST1). However, when the genetic data are considered without their associated geographic information, four major groups could be inferred at the genetic level, which did not show any correlation with sampling locations. The overall results indicate that the population structure of these environmental samples results from the joint

action of a global, widespread ST1 along with genetic differentiation at shorter geographic distances, which in this case are related to the common watershed for the BV localities.

Introduction

Legionella pneumophila is a Gram-negative bacterium commonly found in superficial-water ecosystems and in association with microbial biofilms[5]. From there, it is capable of colonizing urban and industrial water-supply systems, spreading into the environment through aerosols and causing infection when inhaled by susceptible persons.

L. pneumophila was first reported as a pathogen after an acute pneumonia outbreak in Philadelphia (USA) during a convention of the American Legion in 1976[82]. Since then, numerous community- and travel-associated outbreaks of *Legionella*-associated cases have been reported. The most severe form of pneumonia caused by *Legionella* infection is known as Legionnaires‘ disease but there is also a milder, flu-like form known as Pontiac fever[88].

Although other species of *Legionella* are capable of producing infection, *L. pneumophila* serogroup 1 is responsible for about 84% of sporadic cases and outbreaks of legionellosis in the world and 95% in Europe[355]. These bacteria can multiply intracellularly in amoebas and other ciliate hosts[25]. Moreover, the pathogenesis of *Legionella* is comparable between amoebas and human macrophages[5]; in fact, these bacteria are able to enter their hosts by both traditional and coiling phagocytosis[356].

The currently accepted typing scheme for *L. pneumophila* is known as Sequence-Based Typing (SBT)[215,217], a variant of the Multilocus Sequence Typing (MLST) method[220]. In the case of *L. pneumophila*, the SBT scheme is based on PCR amplification and sequencing of 7 loci, including two housekeeping genes (*asd*, *neuA*) and five genes associated with virulence (*fliC*, *pilE*, *mip*, *proA*, *mompS*). But these loci only provide genetic information on approximately 3 kb of the analyzed strains, in comparison to the whole *L. pneumophila* genome (approximately 3.5 Mb)[52,72,235,236,240]. In order to increase the level of resolution for epidemiological studies, Coscollá and González-Candelas[267] studied 13 intergenic regions of the bacterial genome. In fact, the combination of only three of these markers provided an index of discrimination (ID) of 0.88, exactly the same as the discriminating ability of the six genes established by ESGLI[216], which increased when combined with these intergenic regions.

Our research group has reported previously about the genetic variability of *L. pneumophila* strains distributed all around the Comunidad Valenciana (CV) region (Spain) isolated from different environments[225,267]. This region is located along the East of Spain, in the Mediterranean coast, with a surface of 23,255 km^2 and a population of about 5,000,000 inhabitants. Legionellosis outbreaks and cases have been frequent in this area, with the first documented outbreak having occurred in a hotel in 1973[357,358]. The main interest of this work resides on the characterization of the genetic variability, recombination and population structure of *L. pneumophila* strains isolated from a specific water distribution network, fed from the same watershed, in a localized area within the Comunidad Valenciana. This is a relatively small territory, with 1,000 km^2 and 400,000 inhabitants located at the South of the CV (Alicante province), hereafter denoted as BV. We were interested in gaining insight on the distribution of genetic variation in this species at local geographic scales and its relation to those in related localities, in this case connected by the same major watershed.

Materials and methods

Samples

A total of 133 environmental samples were included in this study. Of those, 87 were collected from 1998 to 2006 during regular surveillance for *Legionella* in different points of the water distribution network that supplies 13 populations in the Alicante province (Comunidad Valenciana, Spain) from the same watershed. These locations are denoted as BV. Sampling points were always pre-defined places of the water system where routine water quality control monitoring was usually performed and consisted in small and closed installations in the street for collecting representative water samples, avoiding cross-contamination. Data from 46 additional isolates were retrieved from a previous work by our group[225], and consisted of *L. pneumophila* environmental strains sampled around other localities of the Comunidad Valenciana (denoted as CV). Samples were obtained with permission from the Environmental Health Service, Conselleria de Sanidad, the authority in charge of environmental surveillance for water-borne pathogens across the Comunidad Valenciana, the Spanish region studied in this work. No clinical or human origin samples were used in this study and no animals were used in it.

Bacterial colonies from pure cultures were suspended in 200 µL of 20% Chelex 100 resin (Bio-Rad Laboratories, Richmond, CA). DNA was then extracted by three freeze-thaw cycles (4ºC for 5 min and 99ºC for 5 min), and cellular debris was removed by pelleting at maximum speed for 1 min. The quantity of genomic DNA and its purity were measured by spectrophotometry at 260 nm in triplicates using the A260/A280 ratio with NanoDrop^{TM} 1000 (Thermo Scientific). Purified DNA was stored at -20ºC until used.

PCR amplification and product purification

The seven regions of the *L. pneumophila* genome used for typing (*fliC*, *pilE*, *asd*, *mip*, *mompS*, *proA* and *neuA*)[216,217] and three intergenic regions (L2, L6 and L14)[267] were amplified by standard PCR.

Amplification mixtures contained 1X standard reaction buffer with 2 mM $MgCl_2$ (Biotools), 1 U DNA polymerase (Biotools), 100 µM of each primer, 50 ng of sample DNA and ultrapure water until a final volume of 50 µl. The oligonucleotides used for the SBT scheme gene amplification were described by Gaia *et al* (2005), except those for *neuA*, which were designed by our group to improve amplification results (neuAB_F: ACCGATAGTAAACAAATAGC, neuAB_R: TTCTGTTAGAGCCCAATCGA, optimal melting temperature 56ºC; Coscollá *et al*, unpublished), although other combination of primers were published in Farhat *et al*[359].

The amplification program consisted of a 2 min denaturation step at 94ºC, 35 cycles of denaturation (30 s at 94ºC), annealing (30 s at the optimal annealing temperature for each pair of primers[216,267] and extension (30 s at 72ºC). Finally, the reaction was subjected to a final step at 72ºC for 8 min. PCR products were then purified using NucleoFast® 96 PCR Plates (Macherey-Nagel) following the centrifugation protocol, eluted in 50 µl of ultrapure water and finally stored at -20ºC until sequencing.

DNA sequencing

Purified PCR products were subjected to Sanger sequencing using BigDye™ Terminator v3.0 Ready Reaction Cycle Sequencing Kit (Applied Biosystems) and analyzed in an ABI PRISM 3700 sequencer (Applied Biosystems, Foster City, CA). The program consisted of 60 cycles of 10 min at 94ºC, 5 s at 50ºC and 4 min at 60ºC. The oligonucleotides used for sequencing were the same used in the amplification reaction except for *mompS*. In this case, an inner reverse primer was applied, as previously described[216]. Chromatograms were processed by *gap4* and *pregap4*, from the Staden package[342], to obtain a consensus sequence for each locus.

Newly determined sequences are publicly available in GenBank with accession numbers KC409659-KC410574.

Sequence analysis

Two concatenates of sequences were prepared, one with the six loci of the initial SBT scheme (without *neuA* and the three intergenic regions)[216] and the other with all ten loci. The concatenation was made using BioEdit (available from http://www.mbio.ncsu.edu/BioEdit/bioedit.html) according to their relative position in the *L. pneumophila* str. Philadelphia genome. Sequences were aligned with Muscle[360,361], implemented in MEGA v5.0[362].

The best substitution model for both concatenates was assessed with jModelTest[363]. The application of the Akaike Information Criterion (AIC)[364] resulted in the selection of GTR+Γ as the best model for these data, and this was used for phylogenetic reconstruction following the maximum likelihood (ML) method implemented in RAxML v7.2.8[345].

Genetic variability

Genetic variability was analyzed with DnaSP v5.10.01[348]. The studied parameters were the number of polymorphic sites (S), haplotype diversity (Hd)[365], nucleotide diversity (π)[365], average number of pairwise differences (k) [366], and population mutation rate per site (θ)[367].

Topological congruence

We investigated the topological congruence between the ML trees of each DNA fragment analyzed independently, and also with the trees resulting from two concatenates: 9 (without *neuA*) and 10 loci. As different methods for testing topology congruence have distinct drawbacks, potentially resulting in biased results, three different tests were performed. The Shimodaira-Hasegawa (SH) test[368] is dependent on the best tree to be included among those considered in the test and behaves conservatively as the number of input trees increases. The Expected-Likelihood Weight (ELW) test[369] is independent of the true best tree, but it needs long sequence datasets to correct for possible model miss-specification. Finally, the Approximately Unbiased (AU) test[370] is based on a multiscale bootstrap to correct for selection bias, although, if many of the candidate trees are almost equally well supported, the best true tree might be missed due to an over-confidence in the wrong ones. In order to assess the potential rejection of any of the 12 tested topologies by each of the corresponding datasets, the first two tests were performed using TREE-PUZZLE v5.2[371] and the third one with CONSEL[372].

Recombination

RDP3[373] was used to test for possible intergenic and intragenic recombination events in this dataset by applying seven detection methods: RDP[374], GENECONV[375], Bootscan/Recscan[376], MaxChi[377], Chimaera[378], SiScan[379] and 3Seq[380]. The circularity of the *Legionella* genome was taken into account for these tests. The significance level was established at $p<0.05$ and Bonferroni's correction for multiple comparisons was applied. Only recombination events detected by at least two of the methods were considered.

As RDP3 detects potential breakpoints in the alignment, in order to confirm the distinct phylogenetic history of the regions involved in recombination events and their flanking fragments, the 10-loci alignment was split in two parts and ML trees were inferred for each one. One portion of the alignment included the putative recombinant region detected by RDP3 while the other included the concatenate of the corresponding flanking regions. Subsequently, these topologies were used for testing their reciprocal congruence with the corresponding alignments using the SH[368] and ELW tests[369] with TREE-PUZZLE v5.2[371].

Population structure

Structure v2.3[381] was applied for inferring population structure using haplotype data with no prior information. The potential number of populations (K) was assessed 10 times from K = 2 to K = 10 using the linkage model[382] with a burn-in of 20,000 generations and 100,000 MCMC iterations. The results were processed with Structure Harvester online (http://taylor0.biology.ucla.edu/struct_harvest/) and subsequently with CLUMPP v1.1.2[383] to obtain a consensus result for the 10 replicates for each K value, which was graphically represented with Distruct[384]. The optimal value of K was assessed using Evanno's method, which is based on the second-order rate of change of the log probability of the data between consecutive values of K[385].

Moreover, population structure was further studied by estimating variance components and F-statistics through an Analysis of Molecular Variance (AMOVA)[351], as implemented in Arlequin v3.5 (available from http://cmpg.unibe.ch/software/arlequin3). This analysis provides the percentage of genetic variation within and among populations (fixation index, F_{ST}), which represents the correlation of random haplotypes within populations, relative to that of random pairs drawn from the whole species[351]. The statistical significance was non-parametrically tested using 10,000 random permutations.

Neutrality tests

We tested for deviations of neutrality in each of the 10 loci using Tajima's D, Fu & Li's D* and F* and Fu's Fs tests as implemented in DnaSP v.5.10.01[348]. Statistical significance was evaluated from 1,000 coalescent simulations and the false discovery rate (FDR)[386,387] method was applied to correct for multiple comparisons ($\alpha = 0.025$).

Results

Sequence typing

The allelic profile of the 87 isolates from BV was obtained by sequencing the seven loci in the ESGLI typing scheme. For the 46 strains from CV, the *neuA* locus was sequenced, thus allowing the assignment of the corresponding sequence type (ST) from the information in the ESGLI database. Four *neuA* alleles in BV samples were widely divergent from those traditionally included in the ESGLI database, but have recently been described as a different group of *neuA* variants[359] (Table S1) denoted as *neuAh*.

The 133 isolates corresponded to 30 different STs (Table S1). Most of these (22 STs from 28 isolates) were exclusive to one single locality, and only three (ST1, ST777, and ST1356) were present in more than 3 locations (Table S2). The most frequent variant was ST1, which corresponds to strain Paris[72], present in 63 isolates from 16 different localities. In fact, this ST was found in all but two sampling locations (CV-5 and BV-14) where only 4 and 1 samples had been taken, respectively. ST1356 was present in 6 locations from BV, with a total of 21 isolates, whereas ST777 was found in 4 localities, one from CV and three from BV, for a total of 6 isolates. Three additional STs (ST48, ST719, and ST856) were shared by locations from the CV and BV.

Genetic variability

The main genetic variability parameters estimated from our data are shown in Table 1. These data allowed the comparison between the genetic variability of 87 *L. pneumophila* strains from BV and the 46 isolates from the rest of the CV. The 67 bp non-coding fragment at the beginning of the *pilE* region was analyzed separately. Also, the presence in BV of four divergent alleles in the *neuA* gene (see above) led us to split this locus into two groups for comparison purposes.

Haplotype diversity (Hd) correlates positively with the number of sequences in a sample. However, the estimates of Hd for the CV sample (n=46) were higher than those in BV (n=87) in all loci considered. A similar result was obtained for most comparisons between the two samples for the remaining genetic variability parameters in which the CV samples presented higher values than those from BV (Table 1). The only exceptions corresponded to the number of pairwise differences (k) and nucleotide diversity (π) in *proA* and the number of polymorphic sites (S) in *mip* and *mompS*. In *neuA*, the comparison between the two groups considering only the non-*neuAh* alleles yielded similar results to those previously reported, with only a few more polymorphic sites and mutations in the BV than in the CV samples.

Table 1. Genetic variability by locus of the 133 isolates included in the study.

	n		m		h		Hd		π		S		θ (no recomb)		k (no recomb)		Syn		Non-Syn		dN/dS	
	BV	CV	BV	CV	BV	CV	BV	CV	BV	CV	BV	CV	BV	CV	BV	CV	BV	CV	BV	CV	BV	CV
L14	87	46	453	453	12	12	0.5850 (0.0430)	0.7950 (0.0500)	0.0421 (0.0025)	0.0478 (0.0030)	53	54	0.0239 (0.0067)	0.0278 (0.0086)	18.5540 (8.3037)	21.1070 (9.4859)	-	-	-	-	-	-
proA	87	46	440	440	4	7	0.5410 (0.0420)	0.6920 (0.0630)	0.0127 (0.0008)	0.0116 (0.0019)	13	19	0.0059 (0.0021)	0.0098 (0.0035)	5.5960 (2.6941)	5.1010 (2.5197)	13	18	0	1	0.0000	0.0050
***pilE*NC**	87	46	67	67	3	4	0.1110 (0.0460)	0.5610 (0.0410)	0.0027 (0.0012)	0.0165 (0.0013)	3	4	0.0089 (0.0054)	0.0136 (0.0075)	0.1800 (0.2366)	1.1020 (0.7355)	-	-	-	-	-	-
pilE	87	46	346	346	6	6	0.5570 (0.0390)	0.6930 (0.0410)	0.0097 (0.0013)	0.0224 (0.0009)	20	21	0.0115 (0.0038)	0.0138 (0.0048)	3.3560 (1.7378)	7.7650 (3.6831)	18	18	2	3	0.0580	0.0170
L2	87	46	481	481	7	10	0.5430 (0.0360)	0.8190 (0.0360)	0.0248 (0.0016)	0.0285 (0.0030)	46	49	0.0196 (0.0056)	0.0240 (0.0074)	11.5430 (5.2822)	13.2520 (6.0716)	-	-	-	-	-	-
neuA+neuAh	87	46	476	476	11	8	0.5600 (0.0500)	0.7480 (0.0480)	0.1433 (0.0150)	0.0105 (0.0004)	180	14	0.0756 (0.0194)	0.0067 (0.0025)	67.7820 (29.5004)	4.9810 (2.4668)	n.a.	10	n.a.	4	0.3030	0.1370
neuAh	24	46	476	-	4	-	0.2390 (0.1130)	-	0.0051 (0.0030)	-	22	-	0.0125 (0.0047)	-	2.4200 (1.3601)	-	19	-	3	-	0.0450	-
neuA	63	46	476	-	7	-	0.2650 (0.0730)	-	0.0045 (0.0013)	-	24	-	0.0107 (0.0035)	-	2.1230 (1.1979)	-	18	-	6	-	0.1320	-
mip	87	46	498	498	12	11	0.5850 (0.0430)	0.8440 (0.0400)	0.0047 (0.0009)	0.0099 (0.0011)	25	20	0.0100 (0.0031)	0.0091 (0.0032)	2.3340 (1.2876)	4.9250 (2.4423)	23	18	2	2	0.0110	0.0280
fliC	87	46	200	200	5	5	0.5420 (0.0360)	0.7430 (0.0380)	0.0174 (0.0010)	0.0180 (0.0014)	10	12	0.0099 (0.0039)	0.0137 (0.0054)	3.4860 (1.7944)	3.5970 (1.8590)	10	9	0	3	0.0000	0.0260
L6	87	46	407	407	7	8	0.5440 (0.0400)	0.7890 (0.0440)	0.0085 (0.0008)	0.0186 (0.0018)	22	29	0.0109 (0.0035)	0.0162 (0.0054)	3.4110 (1.7621)	7.5680 (3.5972)	-	-	-	-	-	-
asd	87	46	501	501	7	8	0.5290 (0.0360)	0.7490 (0.0530)	0.0070 (0.0004)	0.0089 (0.0005)	11	12	0.0044 (0.0017)	0.0055 (0.0021)	3.5070 (1.8036)	4.4680 (2.2419)	11	11	0	1	0.0000	0.0090
mompS	87	46	509	509	9	8	0.5740 (0.0350)	0.7760 (0.0450)	0.0078 (0.0013)	0.0130 (0.0027)	41	37	0.0161 (0.0047)	0.0166 (0.0053)	3.9490 (1.996)7	6.5520 (3.1537)	n.a.	n.a.	n.a.	n.a.	0.0650	0.0940

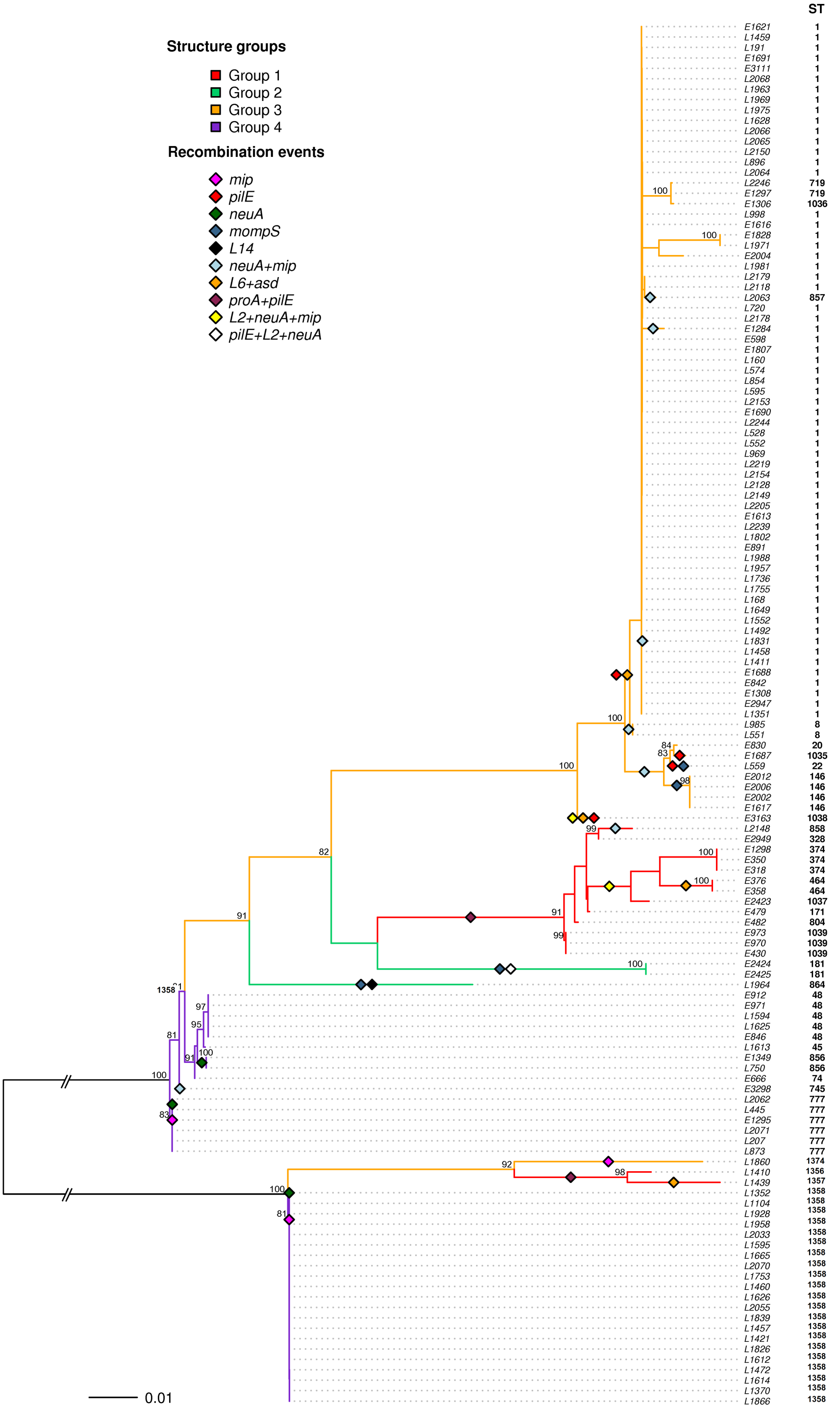
Both groups of *neuA* alleles found in BV and the 67 bp non-coding fragment of *pilE* (*pilE*NC) are analyzed separately. BV and CV account for the two datasets under study. Standard deviations are given in parentheses. **n:** Number of sequences. **m:** Sequence length. **h:** Number of haplotypes. **Hd:** Haplotype diversity. **π:** Nucleotide diversity. **S**: number of polymorphic sites. **k:** Number of pairwise differences. **θ:** Expected heterozygosity per site from S. **Syn:** Number of synonymous changes. **Non-Syn:** Number of non-synonymous changes. **ω:** dN/dS ratio. **n.a.:** Not available.

In agreement with previous results[267], intergenic regions L2 and L14 were more diverse than the protein-coding loci. However, the intergenic region L6 presented a low diversity, especially in BV, comparable to that of coding fragments such as *fliC* or *mompS*, and even lower than the coding region of *pilE*. The genetic variability estimates in the two types of *neuA* alleles were similar and were among the lowest for all the loci considered. Only the non-coding portion of *pilE* presented lower estimates of nucleotide diversity, likely resulting from its small size (67 positions).

Topological congruence and recombination testing

Phylogenetic trees were constructed separately for each locus and for the concatenated alignment of all the loci. ELW and AU phylogenetic congruence tests between the 12 alignments and all the topologies resulted in the complete rejection of the null hypothesis for the topologies not directly derived from each alignment (Table S3). In consequence, these results pointed to the independence of every DNA fragment analyzed from each other, an indication of the potential participation of these regions in intergenic recombination events.

The possibility that recombination might explain the observed lack of congruence was tested using RDP3 with the 36 haplotypes resulting from the concatenate alignment of the 10 loci from the 133 *L. pneumophila* strains (Fig. 1; Table S4). A total of 31 recombination events were detected, most of them by six or seven of the recombination detection methods implemented in the program, and 33 haplotypes showed from one to three recombination events. A single event involving the *pilE* region was found in the clade of the phylogram containing the isolates of ST1, ST8, ST719, ST857, ST1036 and ST1038 (Fig. 1).



Structure groups
🟥 Group 1
🟩 Group 2
🟧 Group 3
🟪 Group 4
Recombination events
◆ *mip*
◆ *pilE*
◆ *neuA*
◆ *mompS*
◆ *L14*
◇ *neuA+mip*
◆ *L6+asd*
◆ *proA+pilE*
◆ *L2+neuA+mip*
◇ *pilE+L2+neuA*
ST
E1621 1
L1459 1
L191 1
E1691 1
E3111 1
L2068 1
L1963 1
L1969 1
L1975 1
L1628 1
L2066 1
L2065 1
L2150 1
L896 1
L2064 1
L2246 719
E1297 719
E1306 1036
L998 1
E1616 1
E1828 1
L1971 1
E2004 1
L1981 1
L2179 1
L2118 1
L2063 857
L720 1
L2178 1
E1284 1
E598 1
E1807 1
L160 1
L574 1
L854 1
L595 1
L2153 1
E1690 1
L2244 1
L528 1
L552 1
L969 1
L2219 1
L2154 1
L2128 1
L2149 1
L2205 1
E1613 1
L2239 1
L1802 1
E891 1
L1988 1
L1957 1
L1736 1
L1755 1
L168 1
L1649 1
L1552 1
L1492 1
L1831 1
L1458 1
L1411 1
E1688 1
E842 1
E1308 1
E2947 1
L1351 1
L985 8
L551 8
E830 20
E1687 1035
L559 22
E2012 146
E2006 146
E2002 146
E1617 146
E3163 1038
L2148 858
E2949 328
E1298 374
E350 374
E318 374
E376 464
E358 464
E2423 1037
E479 171
E482 804
E973 1039
E970 1039
E430 1039
E2424 181
E2425 181
L1964 864
E912 48
E971 48
L1594 48
L1625 48
E846 48
L1613 45
E1349 856
L750 856
E666 74
E3298 745
L2062 777
L445 777
E1295 777
L2071 777
L207 777
L873 777
L1860 1374
L1410 1356
L1439 1357
L1352 1358
L1104 1358
L1928 1358
L1958 1358
L2033 1358
L1595 1358
L1665 1358
L2070 1358
L1753 1358
L1460 1358
L1626 1358
L2055 1358
L1839 1358
L1457 1358
L1421 1358
L1826 1358
L1612 1358
L1472 1358
L1614 1358
L1370 1358
L1866 1358
100
100
100
100
84
83
98
99
100
100
91
99
100
82
91
1358
97
95
81
100
91
100
83
92
98
100
81
0.01

Figure 1 | Maximum likelihood phylogenetic reconstruction of the 10-loci concatenate using partitioned data with RAxML. Colored clades represent the four groups detected by Structure (G1 in red, G2 in green, G3 in orange and G4 in purple). Sequence types (ST) of each sample are represented next to the tips of the tree. Colored diamonds on the branches represent recombination events detected by RDP3. Bootstrap support values higher than 80% are shown.

A joint event including *proA* and *pilE* was also found in ten of the haplotypes and another one involving *neuA* and *mip* was detected in 13 haplotypes. The mapping of these events onto the ML phylogenetic tree (Fig. 1) indicates that several *neuA+mip* events might have happened independently in the genealogical history of these isolates. However, recombination events in *neuA* and *mip* were also detected independently in three haplotypes (ST1358, ST777 and ST856) and some of the *neuA+mip* events also involved locus L2, as shown in Fig. 1. Another frequent recombination event included L6+*asd*, which might have happened in the ancestral node of the clade containing isolates ST1 and ST8, and also independently in the isolates E3163 and L1439. Locus *mompS* was also detected as being involved in several independent intergenic recombination events but locus L14 was detected as recombinant by four methods only in isolate L1964.

Apart from the statistical significance of each event given by the different methods implemented in RDP3, the alignment and the ML topology of the fragment within the detected breakpoints were compared with the alignments and ML topologies of the flanking regions for each recombination event. All reciprocal comparisons resulted in each alignment significantly rejecting the topology given by the other alignment, thus confirming the real existence of incongruent genealogical histories for the genomic stretches involved in all the inferred recombination events.

Population structure

In order to investigate the extent and distribution of genetic differentiation among the samples studied, we considered two datasets, one including the *L. pneumophila* strains isolated from the same watershed (BV), and the other with isolates from different locations in the Comunidad Valenciana (CV). The phylogenetic reconstructions for each locus and the concatenated alignment failed to group variants by sampling location except for the *neuA* locus, for which one well-defined cluster contained all newly described *neuA* alleles from BV (Fig. 1). To gain further insight on the geographic structure of these 133 isolates at the genetic level, we used the Bayesian clustering method implemented in Structure, and applied it to both the 9-loci (Fig. S1) and 10-loci concatenates (Fig. 1) separately using the linkage model. This method assumes genome admixture and also the possibility of linked loci coming from the same population. The objective of this double approach was to take into account the potential effect of the newly described *neuA* alleles in the estimation of the global genetic structure.

From the different number of populations which were assumed *a priori* (from K = 2 to K = 10), both concatenates resulted in 4 as being the most likely number of genetically distinct groups in our data, estimated from ΔK (Fig. 2). However, the 9-loci concatenate also showed a high support for K = 8, although a bit lower than K = 4 (Fig. S2). In these cases, Evanno *et al.*[385] recommend using the smallest value of K because it represents the major structure in the data. Fig. 1 also shows the 4 clusters detected by Structure mapped onto the phylogenetic tree.

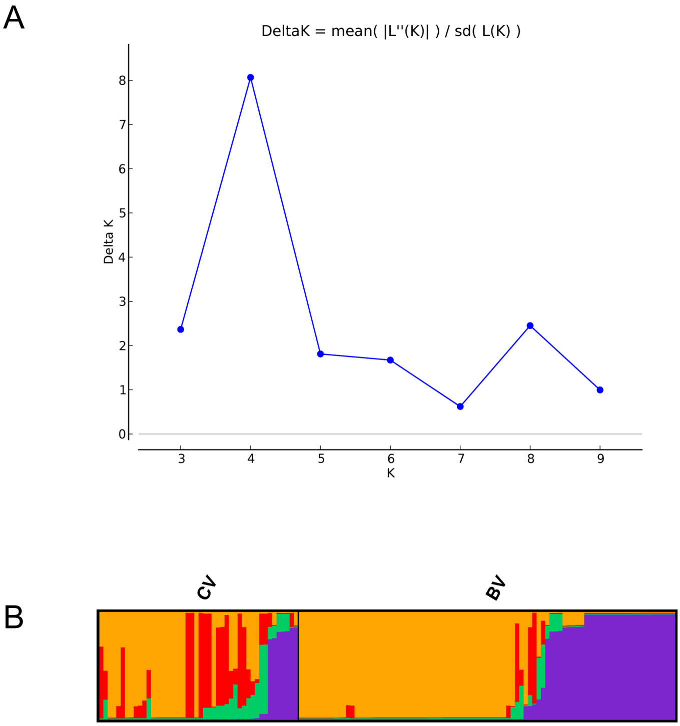


Figure 2 | Summary of population assignment analyses using Structure. (A) Delta K values calculated by Evanno's method detecting K = 4 groups as the most genetically probable within the 10-loci data by Structure Harvester Online. (B) Bar plot representing the probabilities of assignation of the isolates included in the study to each genetic group detected by Structure within CV and BV.

To further characterize the genetic divergence of the four groups detected by Structure, we computed population pairwise F_{ST}'s from haplotype frequencies with Arlequin (Table 2). Using the nomenclature defined in Table 2, the test resulted in group 4 (containing two of the most abundant STs in our data, ST777 and ST1358, that differ only in the *neuA* allele) being the most different with respect to the other three. Also, group 2 was more than 30% genetically different from groups 1 and 3, leaving these two groups being the most similar, as can also be seen in the phylogenetic reconstruction.

Table 2 | Pairwise comparison between populations defined by Structure calculated with Arlequin. Average numbers of pairwise differences within populations are shown in diagonal. Upper matrix represents population pairwise F_{ST} and lower matrix shows the corrected average pairwise differences.

	G1	**G2**	**G3**	**G4**
G1	0.933	0.346***	0.153*	0.529***
G2	0.286	0.495	0.473***	0.629***
G3	0.200	0.419	0.667	0.827***
G4	0.486	0.705	0.619	0.095

p-value<0.05; **p-value<0.01; ***p-value<0.001

A hierarchical analysis of molecular variance (AMOVA) was performed using Arlequin for each of the 10 loci independently considering two populations, the one from BV and the other one from the rest of CV. Results (Table S5) showed significant differentiation between the two groups considered (p-value < 0.05) despite most F_{ST} values were below 0.10. The only exception corresponded to locus *pilE* in which 13.73% of the total genetic variability was found between the two populations.

Neutrality tests

A summary of the neutrality tests for each locus is shown in Table S6 in File S1. Tajima's D values suggest an excess of polymorphisms at low frequencies in *mip*, *mompS* and L6, which is confirmed by D* and F* for the two coding regions, directing these excess of single polymorphisms to the external branches of the phylogeny. Fu's Fs gives evidence for an excess in the number of alleles in *mip*, as would be expected from a recent population expansion or genetic hitch-hiking. However, after FDR correction, only the intergenic L14 region was detected as significantly departed from neutralism by two of the tests (D and F*), and Fu's Fs also rejected the null hypothesis of neutral evolution in the *neuA* genic fragment. Fu & Li's D* was not able to reject neutralism in any of the 10 loci. So, although signs of selection or demographic effects were found especially in *mip*, no significant evidence of deviation from neutralism was finally found, neither truly demographic effects, in which case all loci would be affected similarly.

Discussion

We have previously studied in detail the genetic variability and population structure of clinical and environmental isolates of *Legionella pneumophila* from different points of Comunidad Valenciana[225,267,338]. In this work, our main objective has been to analyze the extent and distribution of genetic variability in the environment of this species at a smaller geographic scale. For this, we have analyzed samples in a small area within the Comunidad Valenciana in which the water distribution systems of the sampled localities are supplied from the same watershed.

Although it has been known for a long time that the main natural habitats of *L. pneumophila* are freshwater environments such as rivers, lakes, ponds, and springs[12,388], there have been very few studies addressed at characterizing the genetic variation of this species in these systems. Instead, most similar efforts have been devoted to characterize the diverse *Legionella* spp. present in different environments[389–392]. One recent study by Parthuisot *et al*[393] analyzed the spatial and temporal dynamics of *Legionella* spp. in a French river watershed subject to seasonal and anthropogenic changes along the year. These authors found a higher prevalence of *L. pneumophila* over other species from the same genus, a result consistent with similar observations from other natural water environments[390,394]. *Legionella* spp. colonize urban distribution systems were they can survive despite disinfection and control measures undertaken. Several studies have revealed a high prevalence of *L. pneumophila* also in these artificial environments[395] both before and after treatment. Despite our efforts, we have not been able to find any publication reporting on the genetic variation of *L. pneumophila* in different locations of the same watershed. Nevertheless, some studies in other water-borne human pathogens such as *E. coli*[396–398] and *Listeria monocytogenes*[399] have been performed with a similar approach.

Genetic variability in the *L. pneumophila* loci analyzed here was higher in the whole Comunidad Valenciana region than in the reduced BV area. These results are expected, because the CV dataset includes more distant localities and with different water supply sources than the BV group. However, these general results do not apply to all the loci considered in the analysis, due to the unusual diversity pattern in locus *neuA* which results from the presence in some samples from the BV population of a particular group of alleles that corresponds to an alternative *neuA* gene from the one found in most serogroup 1 isolates[359].

The study of recombination in the haplotypes derived from the concatenate alignment of the 10-loci considered revealed four main events that were detected as statistically significant by six or even seven of the detection methods used. These events involved mainly loci *proA*, *pilE*, *neuA, mip, asd* and L6, either as separate loci or as a combination of at least two of them, as in *proA*+*pilE*, *neuA*+*mip* and *asd*+L6. These results are congruent with those obtained by Coscollá *et al*[267] on a similar population from the Comunidad Valenciana, and also with the role that homologous recombination and horizontal gene transfer have had on the evolution of *L. pneumophila*[268,271].

Recombination must be taken into account to understand the population structure of bacterial species, as recently shown in populations of

Salmonella enterica[400]. In the present work, we have used isolates from 24 different geographical locations, but the phylogenetic reconstruction and analysis of the 30 resulting STs showed that their genetic composition was largely independent of their sampling location, except for group 4 (which only includes samples with the new *neuA* alleles), which was found only in the BV dataset. The four genetically distinct groups detected using Structure could be easily mapped onto the phylogenetic tree of all the loci, both when locus *neuA* was included (Fig. 1) or not (Fig. S1) in the concatenated alignment. Average population pairwise differences between the four genetic groups resulted in levels of diversity within groups comparable to those of some pairwise comparisons (Table 2).

Given that groups 1, 2 and 3 include isolates from both datasets, this result points to lack of genetic differentiation between BV and CV populations. However, we have found clear evidence of genetic differentiation between these two groups when the 10 loci were compared independently (Table S5). In this case, the highest percentage of variability was found within populations, which held more than 90% of the total genetic variation in all loci except for *pilE*. Additionally, except for the ubiquitous ST1 (strain Paris)[246], only 4 of the remaining 29 STs found were present in at least one population of each datasets. Therefore, there is some evidence of genetic differentiation between the BV and CV datasets which is somewhat disguised by the highly frequent presence of ST1 in both groups.

The results derived from the present study give further insight into the population genetics of *L. pneumophila*, both at the macro and micro-environmental level. Despite the ubiquity of the most common genetic variant, ST1, in the two datasets considered, we have found some evidence of genetic differentiation between them which cannot be attributed only to the presence in BV of some divergent alleles in locus *neuA*. We have confirmed that recombination likely plays an important role in shaping the genetic variability of this bacterium, and also in the independent evolution of different genes within the whole genome[268,271]. However, further studies with complete genomes and more detailed samplings at different geographic scales are needed to draw more conclusions about the effect of selection and demographic events on the distribution of *L. pneumophila* in the environment.

Acknowledgements

We thank the staff from Centros de Salud Pública and Servicio de Vigilancia Epidemiológica (Dirección General de Salud Pública, Conselleria de Sanidad, Comunitat Valenciana) for their help in obtaining environmental samples of *Legionella pneumophila*. LSB is recipient of a FPU scholarship from MEC (Spain).

Supplementary material

Table S1 | List of sequence types assigned to the 133 samples included in the study. The first letter in each sample name denotes the population group of origin: BV (L), rest of Comunidad Valenciana (E).

Sample	*fliC*	*pilE*	*asd*	*mip*	*mompS*	*proA*	*neuA*	ST
L1351	1	4	3	1	1	1	1	**1**
L1411	1	4	3	1	1	1	1	**1**
L1458	1	4	3	1	1	1	1	**1**
L1459	1	4	3	1	1	1	1	**1**
L1492	1	4	3	1	1	1	1	**1**
L1552	1	4	3	1	1	1	1	**1**
L160	1	4	3	1	1	1	1	**1**
L1628	1	4	3	1	1	1	1	**1**
L1649	1	4	3	1	1	1	1	**1**
L168	1	4	3	1	1	1	1	**1**
L1736	1	4	3	1	1	1	1	**1**
L1755	1	4	3	1	1	1	1	**1**
L1802	1	4	3	1	1	1	1	**1**
L1831	1	4	3	1	1	1	1	**1**
L191	1	4	3	1	1	1	1	**1**
L1957	1	4	3	1	1	1	1	**1**
L1963	1	4	3	1	1	1	1	**1**
L1969	1	4	3	1	1	1	1	**1**
L1971	1	4	3	1	1	1	1	**1**
L1975	1	4	3	1	1	1	1	**1**
L1981	1	4	3	1	1	1	1	**1**
L1988	1	4	3	1	1	1	1	**1**
L2064	1	4	3	1	1	1	1	**1**
L2065	1	4	3	1	1	1	1	**1**
L2066	1	4	3	1	1	1	1	**1**
L2068	1	4	3	1	1	1	1	**1**
L2118	1	4	3	1	1	1	1	**1**
L2128	1	4	3	1	1	1	1	**1**
L2149	1	4	3	1	1	1	1	**1**
L2150	1	4	3	1	1	1	1	**1**
L2153	1	4	3	1	1	1	1	**1**
L2154	1	4	3	1	1	1	1	**1**
L2178	1	4	3	1	1	1	1	**1**
L2179	1	4	3	1	1	1	1	**1**
L2205	1	4	3	1	1	1	1	**1**
L2219	1	4	3	1	1	1	1	**1**
L2239	1	4	3	1	1	1	1	**1**
L2244	1	4	3	1	1	1	1	**1**
L528	1	4	3	1	1	1	1	**1**
L552	1	4	3	1	1	1	1	**1**
L574	1	4	3	1	1	1	1	**1**
L595	1	4	3	1	1	1	1	**1**

L720	1	4	3	1	1	1	1	**1**
L854	1	4	3	1	1	1	1	**1**
L896	1	4	3	1	1	1	1	**1**
L969	1	4	3	1	1	1	1	**1**
L998	1	4	3	1	1	1	1	**1**
L551	1	4	3	1	1	1	9	**8**
L985	1	4	3	1	1	1	9	**8**
L559	2	3	6	10	2	1	6	**22**
L1613	5	1	22	26	6	10	12	**45**
L1594	5	2	22	27	6	10	12	**48**
L1625	5	2	22	27	6	10	12	**48**
L2246	1	4	3	19	1	1	1	**719**
L2062	5	2	22	10	6	25	1	**777**
L207	5	2	22	10	6	25	1	**777**
L2071	5	2	22	10	6	25	1	**777**
L445	5	2	22	10	6	25	1	**777**
L873	5	2	22	10	6	25	1	**777**
L750	5	1	22	30	6	10	1	**856**
L2063	1	4	3	1	6	1	1	**857**
L2148	2	10	19	44	19	4	36	**858**
L1964	6	41	45	51	55	10	31	**864**
L1410	6	10	15	3	21	4	207	**1356**
L1439	2	10	3	16	9	4	208	**1357**
L1860	4	3	18	10	5	1	218	**1374**
L1104	5	2	22	10	6	25	203	**1358**
L1352	5	2	22	10	6	25	203	**1358**
L1370	5	2	22	10	6	25	203	**1358**
L1421	5	2	22	10	6	25	203	**1358**
L1457	5	2	22	10	6	25	203	**1358**
L1460	5	2	22	10	6	25	203	**1358**
L1472	5	2	22	10	6	25	203	**1358**
L1595	5	2	22	10	6	25	203	**1358**
L1612	5	2	22	10	6	25	203	**1358**
L1614	5	2	22	10	6	25	203	**1358**
L1626	5	2	22	10	6	25	203	**1358**
L1665	5	2	22	10	6	25	203	**1358**
L1753	5	2	22	10	6	25	203	**1358**
L1826	5	2	22	10	6	25	203	**1358**
L1839	5	2	22	10	6	25	203	**1358**
L1866	5	2	22	10	6	25	203	**1358**
L1928	5	2	22	10	6	25	203	**1358**
L1958	5	2	22	10	6	25	203	**1358**
L2033	5	2	22	10	6	25	203	**1358**
L2055	5	2	22	10	6	25	203	**1358**
L2070	5	2	22	10	6	25	203	**1358**
E1284	1	4	3	1	1	1	1	**1**
E1308	1	4	3	1	1	1	1	**1**

E1613	1	4	3	1	1	1	1	**1**
E1616	1	4	3	1	1	1	1	**1**
E1621	1	4	3	1	1	1	1	**1**
E1688	1	4	3	1	1	1	1	**1**
E1690	1	4	3	1	1	1	1	**1**
E1691	1	4	3	1	1	1	1	**1**
E1807	1	4	3	1	1	1	1	**1**
E1828	1	4	3	1	1	1	1	**1**
E2004	1	4	3	1	1	1	1	**1**
E2947	1	4	3	1	1	1	1	**1**
E3111	1	4	3	1	1	1	1	**1**
E598	1	4	3	1	1	1	1	**1**
E842	1	4	3	1	1	1	1	**1**
E891	1	4	3	1	1	1	1	**1**
E830	2	3	18	15	2	1	6	**20**
E846	5	2	22	27	6	10	12	**48**
E912	5	2	22	27	6	10	12	**48**
E971	5	2	22	27	6	10	12	**48**
E666	5	1	22	30	6	10	6	**74**
E1617	2	10	18	10	2	1	6	**146**
E2002	2	10	18	10	2	1	6	**146**
E2006	2	10	18	10	2	1	6	**146**
E2012	2	10	18	10	2	1	6	**146**
E479	6	10	15	3	21	4	9	**171**
E2424	3	10	1	12	14	9	9	**181**
E2425	3	10	1	12	14	9	9	**181**
E2949	6	10	19	28	19	4	9	**328**
E1298	6	10	17	3	4	14	9	**374**
E318	6	10	17	3	4	14	9	**374**
E350	6	10	17	3	4	14	9	**374**
E358	2	10	3	28	21	4	13	**464**
E376	2	10	3	28	21	4	13	**464**
E1297	1	4	3	19	1	1	1	**719**
E3298	5	2	22	10	6	25	9	**745**
E1295	5	2	22	10	6	25	1	**777**
E482	6	10	15	3	21	14	6	**804**
E1349	5	1	22	30	6	10	1	**856**
E1687	2	3	18	10	2	1	1	**1035**
E1306	1	4	3	19	1	1	3	**1036**
E2423	6	10	14	28	21	4	9	**1037**
E3163	1	6	3	10	1	5	11	**1038**
E430	6	10	15	13	21	14	6	**1039**
E970	6	10	15	13	21	14	6	**1039**
E973	6	10	15	13	21	14	6	**1039**

Table S2 | Distribution of *L. pneumophila* sequence types (STs) found in localities of the CV (Spain) and the BV area.

	Sequence Type																												
Location	**1**	**8**	**20**	**22**	**45**	**48**	**74**	**146**	**171**	**181**	**328**	**374**	**464**	**719**	**745**	**777**	**804**	**856**	**857**	**858**	**864**	**1035**	**1036**	**1037**	**1038**	**1039**	**1356**	**1357**	**1374**	**1358**
CV-1	6					2			1	2	1				1		1					1	1	1	1	1				
CV-2	1																													
CV-3	2		1			1	1					2	2													1				
CV-4	1																													
CV-5												1		1		1		1												
CV-6	2																									1				
CV-7	4							4																						
BV-1	11				1									1		3		1	1	1							9			
BV-2	7					2																					6		1	1
BV-3	4	1																									2	1		
BV-4	3																										2			
BV-5	3															1														
BV-6	6			1												1											1			
BV-7	2																													
BV-8	3																													
BV-9	4																										1			
BV-10	1																													
BV-11	2																				1									
BV-13	1																													
BV-14		1																												
Total	63	2	1	1	1	5	1	4	1	2	1	3	2	2	1	6	1	2	1	1	1	1	1	1	1	3	21	1	1	1

Table S3 | Summary of topological congruence tests for each locus tree and alignment. Summary of the p-values given by Shimodaira-Hasegawa (SH), Expected Likelihood Weight (ELW) and Approximately Unbiased (AU) tests using TREE-PUZZLE and CONSEL. Non-shadowed cells represent topological incongruence by rejection of the null hypothesis of the likelihood of the topology and the corresponding alignment being significantly different (p-value < 0.05). cat9 and cat10 account for the 9-loci and 10-loci concatenates respectively.

Tree	L14			*proA*			*pilE*			L2			*neuA*			*mip*		
	SH	ELW	AU	SH	ELW	AU	SH	ELW	AU	SH	ELW	AU	SH	ELW	AU	SH	ELW	AU
L14	1.0000	1.0000	1.0000	0.2760	0.0000	2.00E-05	0.1580	0.0000	5.00E-109	0.0080	0.0000	9.00E-48	0.0000	0.0000	9.00E-05	0.0000	0.0000	6.00E-63
proA	0.1980	0.0000	3.00E-65	1.0000	0.9897	0.9960	0.0000	0.0000	5.00E-07	0.0000	0.0000	4.00E-05	0.0000	0.0000	4.00E-26	0.0030	0.0000	8.00E-60
pilE	0.0020	0.0000	4.00E-48	0.0010	0.0000	2.00E-04	1.0000	1.0000	0.9990	0.0000	0.0000	5.00E-08	0.0000	0.0000	1.00E-05	0.0000	0.0000	4.00E-70
L2	0.0030	0.0000	3.00E-06	0.0900	0.0000	9.00E-05	0.0020	0.0000	8.00E-05	1.0000	1.0000	1.0000	0.0000	0.0000	1.00E-05	0.0000	0.0000	6.00E-59
neuA	0.0000	0.0000	2.00E-04	0.0020	0.0000	6.00E-16	0.0000	0.0000	5.00E-40	0.0000	0.0000	2.00E-10	1.0000	1.0000	1.0000	0.0000	0.0000	9.00E-89
mip	0.0040	0.0000	2.00E-05	0.2060	0.0000	4.00E-67	0.0230	0.0000	1.00E-81	0.0000	0.0000	6.00E-67	0.0000	0.0000	8.00E-05	1.0000	1.0000	1.0000
fliC	0.0030	0.0000	1.00E-39	0.0930	0.0000	1.00E-19	0.0440	0.0000	2.00E-51	0.0000	0.0000	4.00E-82	0.0000	0.0000	9.00E-07	0.0000	0.0000	2.00E-54
L6	0.0270	0.0000	1.00E-53	0.1660	0.0000	4.00E-09	0.0130	0.0000	2.00E-99	0.0000	0.0000	2.00E-219	0.0000	0.0000	4.00E-05	0.0000	0.0000	5.00E-82
asd	0.0060	0.0000	6.00E-52	0.2940	0.0000	3.00E-04	0.0000	0.0000	3.00E-50	0.0030	0.0000	4.00E-05	0.0000	0.0000	1.00E-64	0.0000	0.0000	9.00E-49
mompS	0.0230	0.0000	0.0010	0.1650	0.0000	8.00E-05	0.0480	0.0000	2.00E-06	0.0000	0.0000	1.00E-10	0.0000	0.0000	4.00E-61	0.0000	0.0000	2.00E-69
cat10	0.1630	0.0000	3.00E-37	0.3170	0.0000	2.00E-04	0.2070	0.0000	1.00E-04	0.0760	0.0000	5.00E-09	0.3020	0.0000	3.00E-72	0.1010	0.0000	4.00E-112
cat9	0.4570	0.0000	1.00E-87	0.6960	0.0103	0.0040	0.4170	0.0000	0.0010	0.2610	0.0000	5.00E-35	0.0000	0.0000	1.00E-04	0.2700	0.0000	3.00E-52

Tree	*fliC*			L6			*asd*			*mompS*			cat10			cat9		
	SH	ELW	AU	SH	ELW	AU	SH	ELW	AU	SH	ELW	AU	SH	ELW	AU	SH	ELW	AU
L14	0.5400	0.0045	0.0530	0.1330	0.0000	4.00E-06	0.2630	0.0000	4.00E-05	0.2460	0.0000	1.00E-05	0.0000	0.0000	1.00E-155	0.0860	0.0000	3.00E-04
proA	0.0110	0.0000	2.00E-06	0.0020	0.0000	1.00E-14	0.0190	0.0000	2.00E-07	0.0440	0.0000	9.00E-06	0.0000	0.0000	1.00E-95	0.0000	0.0000	2.00E-47
pilE	0.2970	0.0000	4.00E-53	0.0000	0.0000	0.0010	0.0260	0.0000	2.00E-09	0.0050	0.0000	1.00E-05	0.0000	0.0000	1.00E-59	0.0000	0.0000	2.00E-46
L2	0.0460	0.0000	9.00E-05	0.0090	0.0000	8.00E-08	0.1410	0.0000	9.00E-05	0.0080	0.0000	3.00E-41	0.0000	0.0000	1.00E-69	0.0000	0.0000	5.00E-05
neuA	0.0010	0.0000	3.00E-05	0.0000	0.0000	2.00E-07	0.0030	0.0000	9.00E-09	0.0010	0.0000	2.00E-07	0.0000	0.0000	3.00E-45	0.0000	0.0000	2.00E-29
mip	0.2100	0.0000	9.00E-05	0.0000	0.0000	3.00E-09	0.0530	0.0000	2.00E-10	0.0480	0.0000	9.00E-81	0.0000	0.0000	2.00E-05	0.0000	0.0000	3.00E-06
fliC	1.0000	0.9832	1.0000	0.0000	0.0000	1.00E-05	0.1520	0.0000	1.00E-76	0.0530	0.0000	8.00E-36	0.0000	0.0000	5.00E-53	0.0000	0.0000	3.00E-09
L6	0.2300	0.0000	4.00E-04	1.0000	1.0000	1.0000	0.1580	0.0000	8.00E-10	0.1850	0.0000	3.00E-05	0.0000	0.0000	2.00E-07	0.0020	0.0000	8.00E-77
asd	0.1460	0.0000	4.00E-14	0.0570	0.0000	1.00E-09	1.0000	1.0000	1.0000	0.0830	0.0000	1.00E-08	0.0000	0.0000	9.00E-10	0.0000	0.0000	4.00E-09
mompS	0.3960	0.0000	3.00E-42	0.1730	0.0000	2.00E-08	0.1300	0.0000	1.00E-11	1.0000	1.0000	1.0000	0.0000	0.0000	3.00E-05	0.0000	0.0000	1.00E-48
cat10	0.4330	0.0000	1.00E-07	0.2730	0.0000	3.00E-44	0.2530	0.0000	2.00E-11	0.2970	0.0000	3.00E-56	1.0000	1.0000	1.0000	0.2120	0.0000	3.00E-50
cat9	0.5780	0.0123	3.00E-04	0.1410	0.0000	1.00E-06	0.5670	0.0000	1.00E-06	0.2760	0.0000	6.00E-95	0.0000	0.0000	1.00E-79	1.0000	1.0000	1.0000

Table S4 | Recombination events detected by RDP3. Colors represent the number of methods that significantly detect each of the events. Haplotypes in blank/grey distinguish between different clades on the phylogenetic tree.

Haplotypes	**L14**	***proA***	***pilE***	**L2**	***neuA***	***mip***	***fliC***	**L6**	***asd***	***mompS***
cE430_cE970_cE973		🟥	🟥							
cE479		🟥	🟥							
cE482		🟥	🟥							
cE2949		🟥	🟥							
cE2423		🟥	🟥	🟥	🟥	🟥				
cE1298_cE318_cE350		🟥	🟥	🟥	🟥	🟥				
cE358_cE376		🟥	🟥	🟥	🟥	🟥		🟩	🟩	
cL2148		🟥	🟥		🟩	🟩				
cE2424_cE2425			🟪	🟪	🟪					🟩
cE1308_cE1613_cE1616…			🟥		🟩	🟩		🟩	🟩	🟥
cE1284			🟥		🟩	🟩		🟩	🟩	
cL2063			🟥		🟩	🟩		🟩	🟩	
cL2118_cL2179			🟥					🟩	🟩	
cE2004			🟥					🟩	🟩	
cE1828_cL1971			🟥					🟩	🟩	
cE1306			🟥					🟩	🟩	
cE1297			🟥					🟩	🟩	
cL2246			🟥					🟩	🟩	
cL551_cL985			🟥		🟩	🟩		🟩	🟩	
cL1831			🟥		🟩	🟩		🟩	🟩	
cL559			🟦		🟦	🟦				🟥
cE1687			🟦		🟩	🟩				
cE1617_cE2002_cE2006_cE2012					🟦	🟦				🟥
cE830					🟦 🟥	🟦 🟥				
cE3163			🟥 	🟪 🟥	🟪 🟥	 🟥		🟩	🟩	
cL1964	🟨									🟥
cL1410		🟥	🟥		🟥					
cL1439		🟥	🟥		🟥	🟥		🟩	🟩	
cL1860					 🟥	🟦 🟥				
cL1104_cL1352_cL1370_cL1421…					🟥	🟦				
cE1295_cL2062_cL207…					🟦	🟦				
cE3298					🟦	🟦				
cE1349_cL750					🟦					
cE666										
cL1613										
cE846_cE912_cE971_cL1594_cL1625										

Table S5 | Analysis of Molecular Variance for each locus. AMOVAs were performed with Arlequin. Levels of diversity explained by the variation among and within populations comparing the BV and CV datasets are shown. (d.f.: degrees of freedom).

	Source of variation	d.f.	Sum of squares	Variance components	Percentage of variation	Fixation Index (F_{ST})	Significance test (10000 perm.)
L14	Among populations	1	1.96600	0.02721	7.65000	0.07645	0.0099+- 0.00030
	Within populations	131	43.06400	0.32873	92.35000		
	Total	132	45.03000	0.35594			
proA	Among populations	1	1.60300	0.02170	6.82000	0.06820	0.00168+- 0.00042
	Within populations	131	38.84100	0.29650	93.18000		
	Total	132	40.44400				
pilE	Among populations	1	3.19200	0.04803	13.73000	0.13732	0.00000+- 0.00000
	Within populations	131	39.52900	0.30175	86.27000		
	Total	132	42.72200	0.34978			
L2	Among populations	1	2.08200	0.02929	8.41000	0.08413	0.00129+- 0.00033
	Within populations	131	41.76800	0.31884	91.59000		
	Total	132	43.85000	0.34813			
neuA	Among populations	1	2.38300	0.03440	9.92000	0.09921	0.00010+- 0.00010
	Within populations	131	40.91800	0.31235	90.08000		
	Total	132	43.30100	0.34675			
mip	Among populations	1	2.39100	0.03414	9.19000	0.09193	0.00000+- 0.00000
	Within populations	131	44.17200	0.33719	90.81000		
	Total	132	46.56400	0.37133			
fliC	Among populations	1	1.90100	0.02651	7.99000	0.07986	0.00218+- 0.0046
	Within populations	131	40.01600	0.30547	92.01000		
	Total	132	41.91700	0.33198			
L6	Among populations	1	2.29200	0.03286	9.47000	0.09471	0.00010+- 0.00010
	Within populations	131	41.15200	0.31413	90.53000		
	Total	132	43.44400	0.34700			
asd	Among populations	1	1.31800	0.01688	5.29000	0.05289	0.01050+- 0.00085
	Within populations	131	39.60600	0.30234	94.71000		
	Total	132	40.92500	0.31922			
mompS	Among populations	1	1.54200	0.02028	5.93000	0.05931	0.00703+- 0.00089
	Within populations	131	42.13500	0.32164	94.07000		
	Total	132	43.67700	0.34192			

Table S6 | Summary of neutrality tests. Tests performed with DnaSP for the 10 loci of all samples included in the study. Shadowed cells indicate significant deviation from neutrality after multiple-testing correction using FDR (α = 0.025).

Locus	**Tajima's D**	**Fu & Li's D***	**Fu & Li's F***	**Fu's Fs**
L14	2.821 ***	1.408 *	2.412 ***	20.748 ***
proA	1.658 *	1.717 **	2.028 *	9.507 *
pilE	0.868	0.314	0.639	10.936 **
L2	1.103	0.784	1.103	16.468 **
neuA	1.851 *	1.005	1.664 *	58.988 ***
mip	-0.735	-0.158	-0.462	-1.328
fliC	1.124	0.348	0.758	6.508 *
L6	-0.301	0.905	0.505	4.592
asd	1.362	0.348	0.863	2.828
mompS	-1.392	-1.649	-1.851	3.086

*p-value<0.05
**p-value<0.01
***p-value<0.001

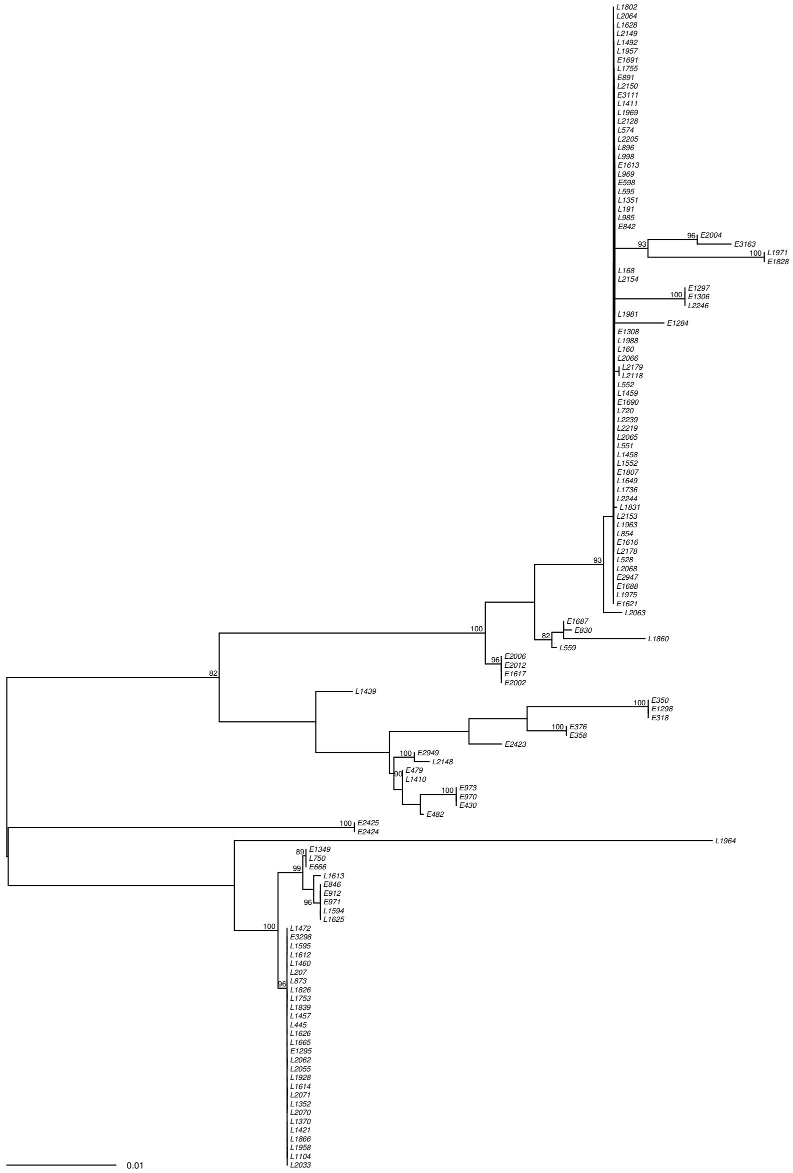


Figure S1 | ML phylogenetic reconstruction of the 9-loci alignment from data of the 133 environmental isolates from BV (L) and the rest of Comunidad Valenciana (E) using RAxML. Bootstrap support values higher than 80% are shown.

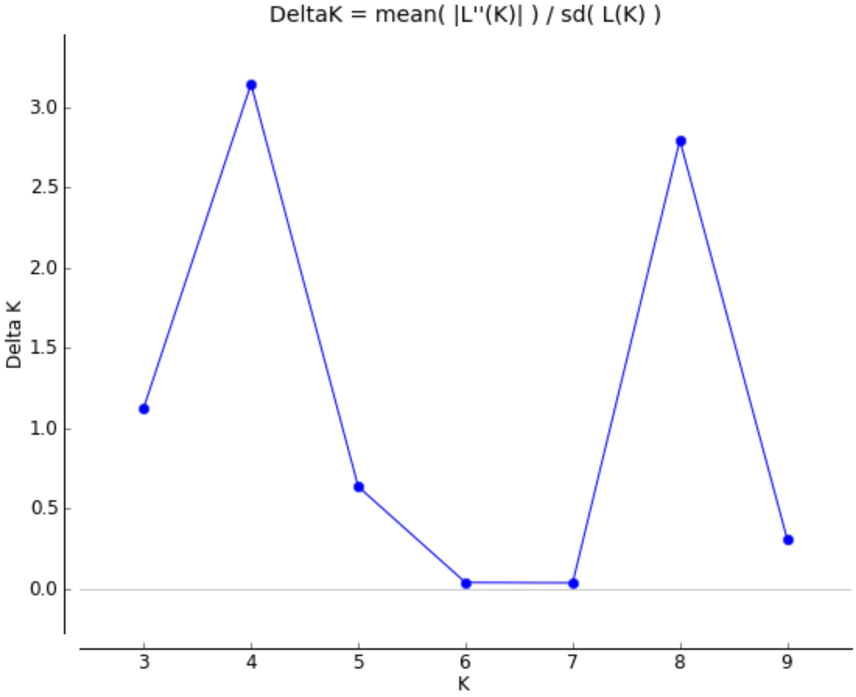


Figure S2 | Delta K values calculated by Evanno's method using the 9-loci data by Structure Harvester Online.

Chapter 3

Phylogenetic analysis of environmental *Legionella pneumophila* isolates from an endemic area (Alcoy, Spain)

Leonor Sánchez-Busó[1,2], María Piedad Olmos[3], María Luisa Camaró[3], Francisco Adrián[4], Juan Miguel Calafat[4], Fernando González-Candelas[1,2].

1. Unidad Mixta Infección y Salud Pública FISABIO/Universitat de València. Avenida Cataluña, 21; 46020, Valencia, Spain.
2. CIBER en Epidemiología y Salud Pública, Valencia, Spain.
3. Laboratorio de Salud Pública de Valencia. Conselleria de Sanitat. Valencia, Spain. Avenida Cataluña, 21; 46020, Valencia, Spain.
4. Centro de Salud Pública de Alcoy. DGSP. Alcoy, Spain.

Published in:

Sánchez-Busó L, Olmos MP, Camaró ML, Adrián F, Calafat JM, González-Candelas F. Phylogenetic analysis of environmental *Legionella pneumophila* isolates from an endemic area (Alcoy, Spain). Infection, Genetics and Evolution. 2015;30:45–54.

Abstract

Environmental surveillance of *Legionella pneumophila* is a key component of the control measures established in urban settlements to ensure water safety and quality, with the aim of minimizing and limiting opportunistic infections in humans. In this work, we present results on the detection and genetic characterization of these bacteria in the outbreak-recurrent region of Alcoy (Comunidad Valenciana, Spain) using water and biofilm samples. We were particularly interested in studying the presence and distribution of *L. pneumophila* in the absence of outbreak or sporadic cases of legionellosis and in comparing the efficacy of culturing from water samples with a biofilm-based detection procedure using molecular amplification. To this end, water samples were taken from 120 sites distributed all around the city and its surroundings, as well as 60 biofilm swabs from half of the sampling sites. *L. pneumophila* could be isolated from water in just 4 of the locations. Touchdown PCR was applied to DNA extracted from water and also biofilm swabs, as a rapid method for both routine and outbreak investigations. *L. pneumophila* was detected by this method in 14 of the sites in which both water and biofilms were taken, although 13 of them tested positive using only the biofilm samples. These results show a ten-fold increase in the success rate of *Legionella* detection over water samples. The application of this method to study the presence of *L. pneumophila* in the water-supply system and risk facilities of Alcoy revealed different strains distributed in different areas of the city. Sequence Type ST578, endemic in the area and responsible for most clinical cases, was detected in one of the sampling sites. The number of positive samples correlated with water temperature but not with chlorine levels. The direct analysis of biofilm swabs improves the detection rate and genetic characterization of *L. pneumophila* and can complement analyses based on bacterial culture.

Introduction

Potable drinking water carries a diverse but poorly identified microbiota that often forms very persistent structures such as biofilms that may harbor human pathogens[401]. *Legionella pneumophila* is a waterborne Gram-negative bacterium, an opportunistic pathogen, and the main causative agent of Legionnaires' disease (LD) and Pontiac fever[53,88]. Infection of humans occurs by inhalation of aerosols loaded with bacteria. *L. pneumophila* can live as free cells and in vacuoles inside the cytoplasm of amoebas in biofilms[21] of different human-made and natural aquatic environments, which can eventually become sources of legionellosis outbreaks.

Environmental investigations of outbreaks and preventive routine analyses aimed at the detection and characterization of *L. pneumophila* have focused traditionally on microbiological and biochemical classifications of isolates. These can be classified into 15 different serogroups and more than ten subtypes[355] mainly within serogroup 1[8]. However, the recovery of *L. pneumophila* requires selective media and prolonged incubation periods, which have favored the application of alternative molecular biology techniques[182,183,402] to detect, identify and characterize these bacteria. The use of PCR-based methods allows also the detection of viable but non-culturable *L. pneumophila* strains (VBNC), which are potentially infectious, and also the development of a highly discriminant Sequence-Based Typing (SBT) scheme[215–217,359], which is currently the most frequently used method for *L. pneumophila* typing[196]. Amplification and sequencing of the seven loci included in this scheme has been the starting point for assessing the actual genetic variability of clinical and environmental strains of *L. pneumophila*[225,267].

Although the role of *L. pneumophila* in biofilms developed in water distribution networks has been described extensively[21,335,403–406], most detection efforts have focused on applying microbiological and molecular biology techniques directly to water samples[40,192,407,408]. Only a few biofilm-based studies on the ecology and dynamics of these bacteria[406,409–411] and the effect of disinfectants on their eradication from human-made facilities have been published up to date[412–414].

The city of Alcoy (Alicante province, Comunidad Valenciana, east of Spain; 61,000 inhabitants) has been one of the main areas of legionellosis outbreaks and sporadic cases in Spain, affecting a total of 407 people between 1999 and 2011[117,415,416]. As in most *L. pneumophila* outbreaks, the match between clinical and environmental samples by using SBT methods has been difficult to establish, with only a few exceptions[117]. The high incidence of legionellosis in this city suggests that *L. pneumophila* is a frequent inhabitant of suitable environments in the area. However, there is little information about its presence and distribution in this locality apart from outbreak investigations.

In this work, we present an analysis of the occurrence of *L. pneumophila* in the water distribution system and other risk facilities of Alcoy using two different types of samples: water and biofilms. This allowed us to compare the traditional method of detection of *L. pneumophila* from water samples and a PCR-based specific approach from biofilm swabs. The latter method may improve the chances of identifying the sources of community-related outbreaks and also be useful in routine epidemiological investigations by complementing culture-based analyses.

Materials and methods

Sample collection

A total of 120 2-liter water samples were taken weekly from February to July (with a break in April) during 2011 in different points of the water network and other risk facilities of Alcoy, including fountains using non-drinking water, hydrants, pipes dead-ends, nozzles and deposits (Fig. 1, Table S1). Water temperature was measured at each sampling point and residual chlorine levels were detected by the DPD (diethyl paraphenilene diamine) indicator test (Table S1). Chlorine-positive samples were neutralized with sodium thiosulphate. Additionally, 60 biofilm samples were also taken at the same sampling points and at the same time that water samples during the months of June and July by gently scratching with cotton swabs the inner surfaces of pipes or nozzles. Permits for sampling were obtained from Alcoy City Hall and Public Health General Directorate. Sampling was performed under direct supervision of representatives of these bodies.

Water processing and Legionella culturing

Water filtering and *Legionella* culturing were performed following the UNE-EN ISO 11731-2:2008 and AFNOR NF T 90-431 2 (2003), which describe the standard method for the detection and counting of *Legionella* spp. Briefly, one liter of each water sample was filtered using a Millipore vacuum filtering system through sterile polycarbonate membranes with 47 mm diameter and 0.45 μm pore size. Membranes were suspended in 10 mL of deionized sterile water and subjected to ultrasounds at intermediate power for 5 min to remove the attached microorganisms.

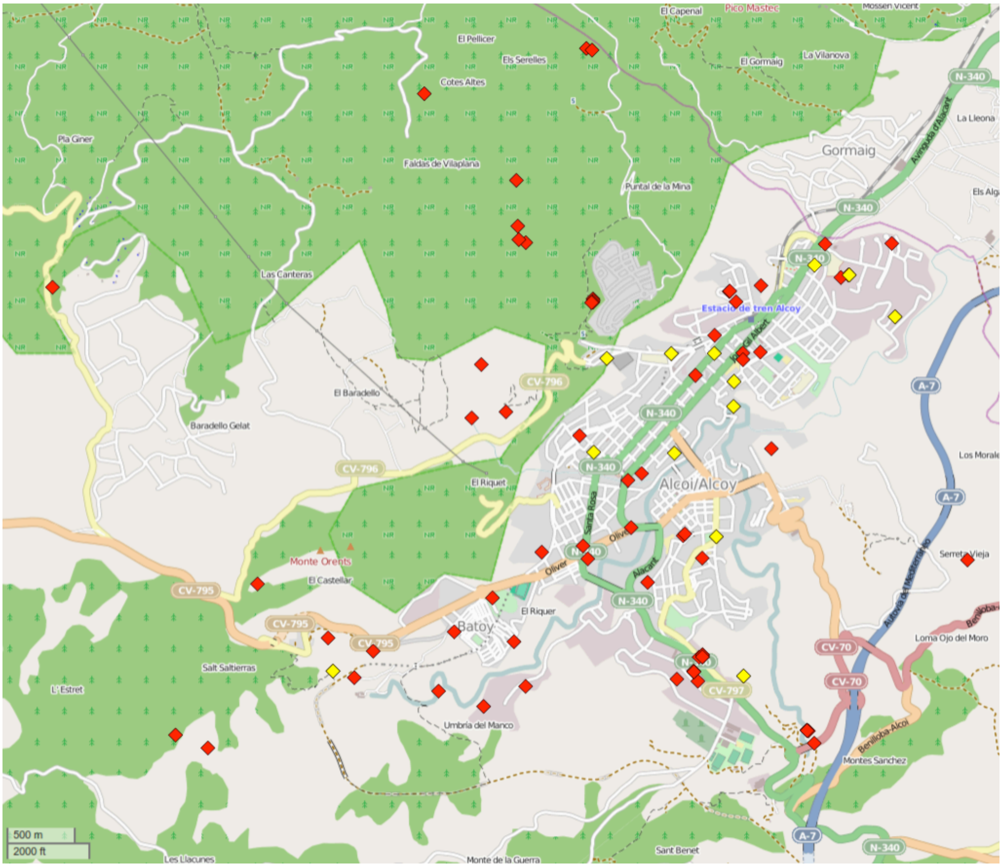


Figure 1 | Water and biofilm sampling points of the water distribution network and other risk facilities in Alcoy, Spain. *L. pneumophila*-positive sites by any detection method are shown in yellow, negatives are shown in red. Two positive points were sampled twice and two street cleaners are not shown.

Aliquots of this bacterial suspension were used for different treatments: 1 mL was cultured on GVPC medium plates, 0.5 mL were subjected to thermal treatment and 0.5 mL to treatment with an acidic agent, and subsequently cultured on GVPC plates. After 10 days of culturing at 37±1ºC, colonies with a typical morphology of *Legionella* spp. were re-cultured on BCYE medium with and without L-cysteine. Bacteria surviving the thermal and acid treatments and only able to grow in the BCYE medium with L-cysteine were serogrouped using Oxoid *Legionella* Latex Text (Thermo Scientific, Waltham, MA). Duplicated biofilm samples were also plated in the same media to test for *L. pneumophila* isolation.

DNA extraction

Single colonies of *L. pneumophila* were suspended in 50 µL of autoclaved and filtered ultrapure water. DNA was extracted by thermal shock using two cycles of a 5-min incubation step at 99ºC followed by 5 min at 4ºC. Cell debris was pelleted by centrifuging during 3 min at 12,000 rpm. The supernatant was stored at -20ºC for further analysis.

The remaining filtrated water from the culturing step was concentrated by centrifugation using Amicon® Ultra-15 100 kDa Ultracel tubes (Millipore Corporation, Billerica, MA) at 4,000 rpm during 8 min. A final volume of 200 µL was collected per sample and centrifuged for 20 minutes at maximum speed. After removing the supernatant, 200 µL of the chelex-based InstaGene™ Matrix (Bio-Rad, Richmond, CA) were added for mechanical and thermal lysis by incubation at 56ºC during 30 min, vortexing, and heating at 99ºC for 8 min. DNA concentration and purity (A260/A280 ratio) were checked by triplicate with NanoDrop™ 1000 (Thermo Scientific, Waltham, MA).

Biofilm cotton swabs were separated from the stick using a scalpel and subjected to total DNA extraction using the UltraClean® BloodSpin® DNA Isolation Kit (MoBio Laboratories Inc., Carlsbad, CA). Briefly, samples were incubated with proteinase K and lysis buffer during 10 minutes at 65ºC. Cotton swabs were introduced in 0.7 mL microtubes with an opening at the bottom placed in a 2 mL-microcentrifuge tube[417] and centrifuged at maximum speed for 3 min to retrieve the maximum possible amount of embedded lysate. Final volumes were used for DNA purification using the spin filter protocol of the kit, and were stored at -20ºC until further processing. DNA concentration and purity were tested as described above.

TD-PCR for detection of L. pneumophila in water and biofilm samples

Purified DNA extracted from water and biofilms samples was used for *L. pneumophila* detection using a molecular amplification technique known to be highly sensitive and specific, the touchdown PCR (TD-PCR)[418–420], targeting a region of the *pilE* gene, originally set up as specific for *L. pneumophila* by the SBT scheme. *pilE* was chosen for being the locus producing the best results (cleaner and more intense bands) in the SBT scheme under our previous experience. Amplification reactions were performed with 20-50 ng of total DNA using 2.5 U FastStart High Fidelity PCR System (Roche Applied Science, Hoffmann-La Roche AG, Basel, SWTZ), 200 µM of each dNTP, 1x buffer with 1.8 mM $MgCl_2$, 4% DMSO, 0.2 µM of the corresponding primer[216] and ultrapure water until a final volume of 25 µL. For the *pilE* region, amplification conditions included an initial denaturation step of 5 min at 95ºC followed by 10 cycles of denaturation (30 s at 95ºC), annealing (30 s at 65-55ºC with a ramp rate of -1 ºC/cycle) (Table S2) and extension (30 s at 72ºC). Immediately after the touchdown step, 35 cycles of standard PCR were applied using the same conditions but fixing the annealing temperature at 55ºC; a final extension step of 8 min at 72ºC was also used to ensure complete amplification. Positive amplifications were confirmed using Gel-Red (Biotium Inc, Hayward, CA) after 1.4 % agarose gel electrophoresis.

Sequence-Based Typing from isolates and direct samples

Amplified DNA from isolates was used for typing following the SBT scheme. The seven loci were amplified using previously described primers[216] except for *neuA* (neuAB_F: ACCGATAGTAAACAAATAGC, neuAB_R: TTCTGTTAGAGCCCAATCGA; Coscollá *et al*, unpublished). PCR reactions contained 10 ng of genomic DNA, 1 U Taq polymerase (Biotools, B&M Labs, S.A, Madrid, SPN), 200 µM of each dNTP, 1x buffer with 2 mM $MgCl_2$, 0.2 µM of the corresponding primer and ultrapure water until a final volume of 25 µL. Amplification conditions included an initial denaturation step at 95ºC during 2 min, 35 cycles of denaturation (30 s at 95ºC), annealing (30 s at the corresponding temperature for each locus[215,216] and extension (30 s at 72ºC), followed by a final extension step of 8 min at 72ºC. The six remaining loci for the *L. pneumophila*-positive biofilms were amplified using TD-PCR as described above (Table S2).

PCR products were purified using NucleoFast® 96 PCR plates (Macherey-Nagel GmbH & Co, Düren, Germany) and suspended in 50 µL of ultrapure water. Sanger sequencing reactions were performed using BigDye™ Terminator v3.0 Ready Reaction Cycle Sequencing Kit and read in an ABI 3730 DNA Analyzer (Applied Biosystems, Foster City, CA). Chromatograms were processed using the Staden package[342] and allele numbers were assigned with the ESGLI database[221]. PCR products from non-cultured samples that showed signs of presence of multiple *L. pneumophila* strains were cloned to further study intra-sample variability.

Cloning of non-typeable PCR products and M13 PCR reactions

For four biofilm and one water sample (Table 1), which could not be typed because of the detection of multiple signals in the corresponding chromatogram, PCR products from *pilE*, *asd* and *mip* loci were cloned into *Escherichia coli* JM109 Competent Cells (Promega Corporation, Madison, WI) using pGEM®-T Easy vector System II (Promega Corporation, Madison, WI). Ligation reactions and transformation into JM109 competent cells were performed following the instructions of the manufacturer.

From 10 to 15 colonies were selected from each plate and suspended in 20 µL of ultrapure water. 5 µL of the suspension was used for amplifying the insert using universal forward and reverse M13 primers and the same reaction mix used with pure cultures. The amplification program was set as described above for isolates but with an annealing temperature of 55ºC. PCR products were purified and Sanger-sequenced as described previously.

Phylogenetic reconstruction

Sequences obtained from clones of the *pilE*, *asd* and *mip* PCR products and the respective available alleles of the ESGLI database were aligned using Muscle[360,361] in Mega v5.10[362]. Phylogenetic reconstruction was performed with maximum likelihood using RAxML v7.2.8[345] with the GTRGAMMA model for nucleotides and 1000 bootstrap re-samplings.

Concatenated sequences from the seven loci of the SBT scheme for all the *L. pneumophila* samples obtained in this analysis were aligned with sequences from the 28 different sequence types found in Alcoy by our group in samples collected since 1999 (Fig. 4), 31 sequence types retrieved from the ESGLI database containing *neuAh* alleles[359], and those STs whose sequences matched any of the previous strains in six loci but differed in *neuA*, totaling 63 strains. The alignment and subsequent phylogenetic reconstruction by maximum likelihood were performed with and without *neuA/neuAh* as described above.

Results

L. pneumophila isolation and characterization from culture

A total of 120 water samples as well as 60 biofilm swabs were taken in the same sampling and time points per duplicate between February and July 2011 at different points of the water distribution system and other risk facilities of Alcoy (Table S1). *L. pneumophila* could be isolated from only four water samples (4/120, 3.3%). One biofilm swab from each sampling point was plated for *Legionella* culturing; however, no *Legionella* isolates were retrieved because of the overgrowth of a considerable amount of different microbiota, even after diluting the samples, or no growth was observed.

The four isolates, all of them retrieved from street dead ends, were serogrouped and typed using the SBT scheme, resulting in four different STs (Table 1), ST578 serogroup (sg) 1 (340 Colony Forming Units per Liter – CFU/L), ST1086 sg 2-15 (450 CFU/L), ST1324 sg 2-15 (<40 CFU/L) and ST1358 sg 2-15 (290 CFU/L), the two latter presenting *neuAh* alleles[359]. ST578 is known to be endemic in the area; the other genetic profiles had not previously been described to have either caused any clinical case or being found in the environment in the routine surveillance programs.

PCR-based L. pneumophila detection and characterization

Apart from culturing, water samples and duplicates of each biofilm swab were also tested by TD-PCR for the presence and subsequent genetic characterization of *L. pneumophila*. The bacterium was found in 5 water (5/120, 4.2%) and 13 (13/60, 21.7%) biofilm samples (Table 1). Of the 60 sampling points from which both water and biofilm samples were taken, only one water specimen (1/60, 1.7%) was positive by PCR, but the associated biofilm (sample 9733, Table S1) was negative.

Table 1 | Allelic profiles of isolates and water and biofilm samples that were detected as positive for *L. pneumophila* and for which sequencing data could be retrieved.

Sample ID	Type of sample	Source	*fliC*	*pilE*	*asd*	*mip*	*mompS*	*proA*	*neuA*	ST
8523	Water	Industry	-	3[a]	-[b]	-	-	-	-	-
8524	Water	Industry	-	+[c]	15	-	-	+	-	-
8527	Water	Industry	-	10	-	-	-	-	-	-
8787	Culture	Dead-end	5	2	22	10	6	25	6	1086
9185	Culture	Dead-end	6	10	15	13	9	14	6	578
9729	Biofilm	Dead-end	-	**3**[d]	**18, 28**	**10**	-	+	-	-
9733	Culture	Dead-end	5	2	22	10	6	25	203	1358
	Water		+	**2**	**22**	**10**	+	+	-	-
10015	Biofilm	Industry	+	**4**	**1, 3**	**1,10**	+	+	+	-
10402	Culture	Dead-end	5	1	22	30	6	10	203	1324
10403	Biofilm	Dead-end	-	4	-	-	-	+	-	-
10405	Biofilm	Dead-end	+	4	**1, 3, 9**	**1**	+	+	+	-
10406	Biofilm	Dead-end	-	4	-	-	-	+	-	-
10407	Biofilm	Dead-end	-	**4**	**1**	+	+	+	-	-

[a] Numbers indicate ESGLI alleles.
[b] Negative symbols ('-') indicate negative amplification.
[c] Positive symbols ('+') indicate positive amplification but non-assignable allele by direct sequencing.
[d] Shadowed results indicate PCR products used for cloning. Those not in bold (*pilE* in sample 10405 and *mip* in sample 10407) could not be cloned.

In summary, eighteen of the sampling points were positive (18/120, 15%) when analyzed by TD-PCR of the *pilE* region and were used for subsequent amplification of the remaining loci of the SBT scheme (Table 1). In most cases, it was not possible to amplify all the 7 loci, probably because of the low concentration and high complexity of the DNA in these samples. After sequencing the PCR products, only 18.2% (10/55, Table 1) could be identified with a specific allele at this point, with sequences showing a mixture of variants.

Analysis of cloned sequences from water and biofilms

In general, samples that were positive for the bacteria were not distributed uniformly throughout the sampling period as they appeared only from May onwards. Geographically, positive samples were mainly concentrated in the northern part of the city (Fig. 1). Dead-ends from different streets accounted for most positive samples (57.1%), while 33.3% were found in industries (including 2 street cleaners), and only 9.5% in untreated fountains (Table S1).

In order to better characterize the *L. pneumophila* found in water and biofilm samples for which no sequence type could be assigned, PCR products of the *pilE*, *asd* and *mip* loci for five of the samples were chosen for sub-cloning (Table 1). Between 3 and 15 clones containing the insert were retrieved and subsequently sequenced per fragment. The analysis of the cloned sequences with the ones deposited in the ESGLI database using phylogenetic methods showed intra-sample variability (Fig. 2 and Fig. S1), being the *asd* the locus the most heterogeneous one (Fig. 2).

Three different ESGLI alleles were detected in the *asd* locus of sample 10405, as well as two different alleles in the *asd* and/or *mip* loci of samples 9729 and 10015.

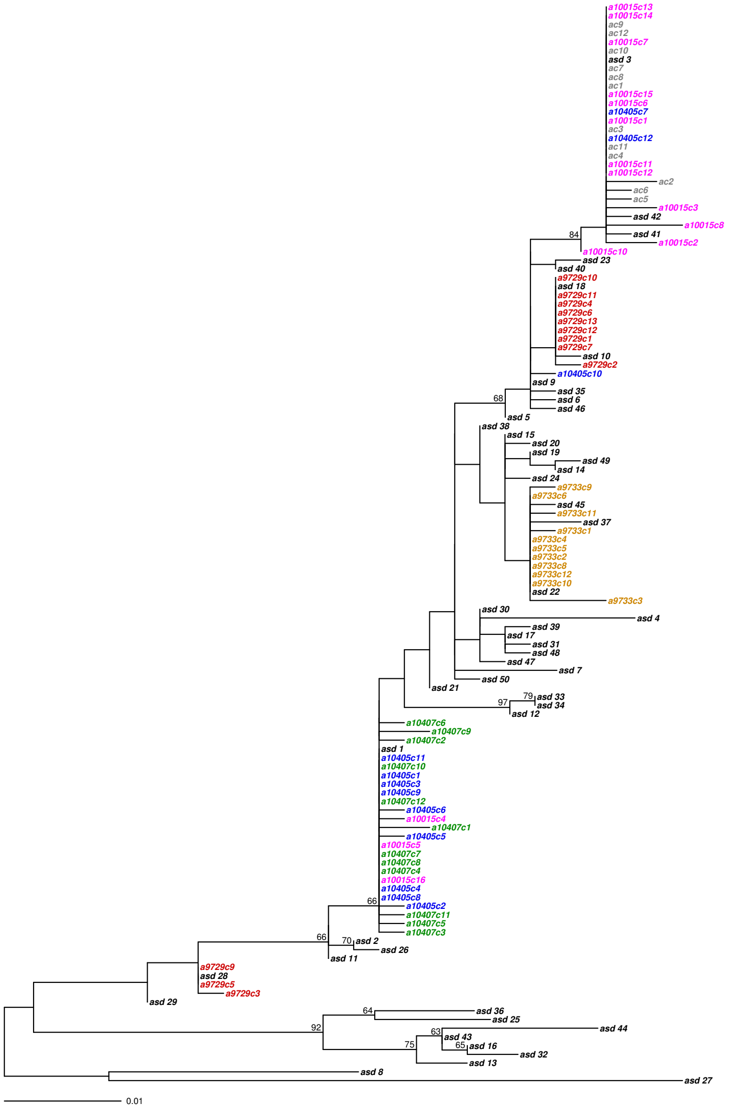


Figure 2 | Maximum likelihood phylogenetic reconstruction of the sequences obtained from clones of the *asd* fragment amplified by PCR in biofilm samples (9729, 9733, 10015, 10405 and 10407, marked in red, orange, pink, blue and green, respectively) performed using RAxML. Alleles from the experimental control (ac1-ac11) are colored in grey and the remaining tips represent alleles from ESGLI. Nodes with bootstrap support values higher than 60% are shown.

Polymorphic positions between different alleles in the same sample varied from 3 SNPs between alleles asd_3 and asd_9 in sample 10405 to 14 SNPs between alleles asd_18 and asd_28 in sample 9729 (Fig. 2). Some previously non-described alleles were also found among the sequenced clones that diverged in up to three nucleotides from alleles in the ESGLI database. To evaluate the possibility that these variants were the result of amplification and/or cloning errors, the *asd* region of one colony from a pure culture was also amplified, cloned and sequenced following the same protocol. The results from this control were used to calculate the experimental error rate and to compare it with the levels of polymorphism obtained previously. The experimental control presented an error rate of 7.05×10^{-4}, significantly lower than the polymorphism rates obtained for *asd* in the five cloned samples (Table 2).

Table 2 | Rate of nucleotide change (μ) in the amplified and cloned region of *asd* in five biofilm samples, separately and together (All).

	Samples						
Parameter	**9729**	**9733**	**10015**	**10405**	**10407**	**All**	**EC**[a]
N[b]	12	11	15	12	12	62	12
k[c]	16	6	16	12	9	44	4
μ[d]	2.82E-03	1.15E-03	2.26E-03	2.11E-03	1.59E-03	1.50E-03	**7.05E-04**[e]

[a] Experimental control (EC) performed from a single colony of *L. pneumophila*.
[b] Number of sequences (N).
[c] Number of nucleotide differences (k).
[d] $\mu = k/(N \cdot L)$ with L (length) = 473 bp.
[e] Experimental error rate**.**

Sequence type distribution and diversity in Alcoy

L. pneumophila strains isolated from environmental sources in Alcoy and with complete SBT patterns presented a higher diversity than those derived from clinical cases (Fig. 3). ST578 was the strain most frequently associated to outbreaks and sporadic cases (n=54; Fig. 3). Other clinically associated STs were ST637 (n=7), ST37 (n=1), ST181 (n=1) and ST1106 (n=1). However, 26 different STs were detected in environmental sources, being ST1 (n=14) the most frequently isolated. Most of the environmental ST578s (n=11) were found during the investigation of an outbreak occurred in Alcoy in 2009[117].

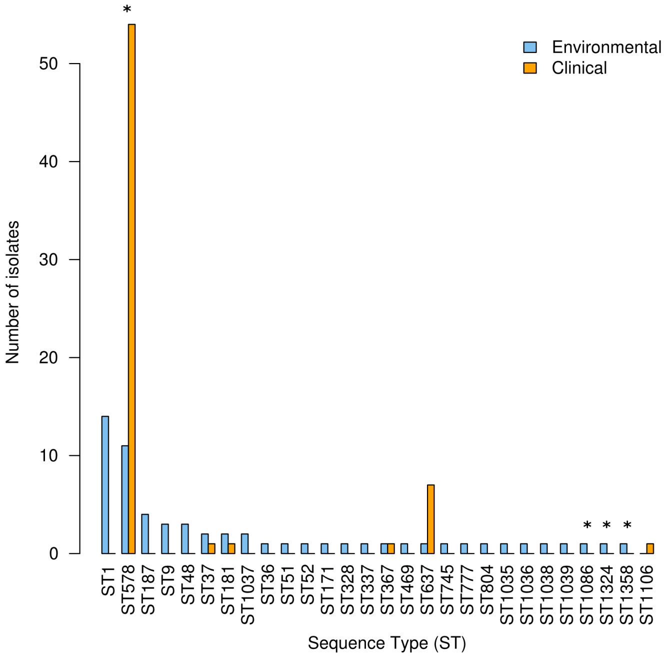


Figure 3 | Diversity of sequence types detected by our group in the locality of Alcoy since 1999. Environmental isolates are shown in blue and clinical isolates in orange. STs found in this study are marked with an asterisk.

A ML phylogenetic tree (Fig. 4) was obtained with the concatenated sequences of the 7 loci in the SBT scheme from the four strains isolated in this work along with other STs found in Alcoy and additional representatives from the ESGLI database. A similar tree without the *neuA* locus is shown in Fig. S2. The 7 loci tree presents two major groupings that correspond to the presence of a *neuA* or a *neuAh* allele in this locus[218,359]. Two of the isolates from this work clustered in the former group (8787c and 9185c, Fig. 4) and two in the later (9733c and 10402c). However, the exclusion of this locus from the reconstruction (Fig. S2) led sample 9733c to cluster with other STs detected previously in this area, such as ST777. The allele patterns included in this phylogenetic reconstruction are shown in Table S3.

Effect of temperature and chlorine levels

Water temperature levels might play a role in the proliferation of *L. pneumophila* in the distribution network of the city. No positive samples by culture or direct amplification were obtained from February to March, when water temperatures at the sampling points ranged from 7.7ºC to 19.5ºC (Table S1). Most sampling points with positive detection of *L. pneumophila* showed water temperatures from 20.1ºC to 30.9ºC (Table S1) with the peak of mean temperature levels on the fourth week of June, coinciding with the highest number of positive samples. The correlation between the per-week average temperature and number of *L. pneumophila* occurrences was positive and statistically significant (Pearson's r = 0.72, p-value < 0.05).

Chlorine levels had a wide range of variation between sampling points, from 0.00 mg/L in natural sources to a range of 0.00-6.00 mg/L in human-made environments, and they did not show any significant correlation with the presence or absence of *L. pneumophila* (Pearson's r = 0.12, p-value > 0.20). The bacterium was found in places with complete lack of chlorine, such as natural fountains, and also in chlorinated urban water pipes dead-ends and street cleaners.

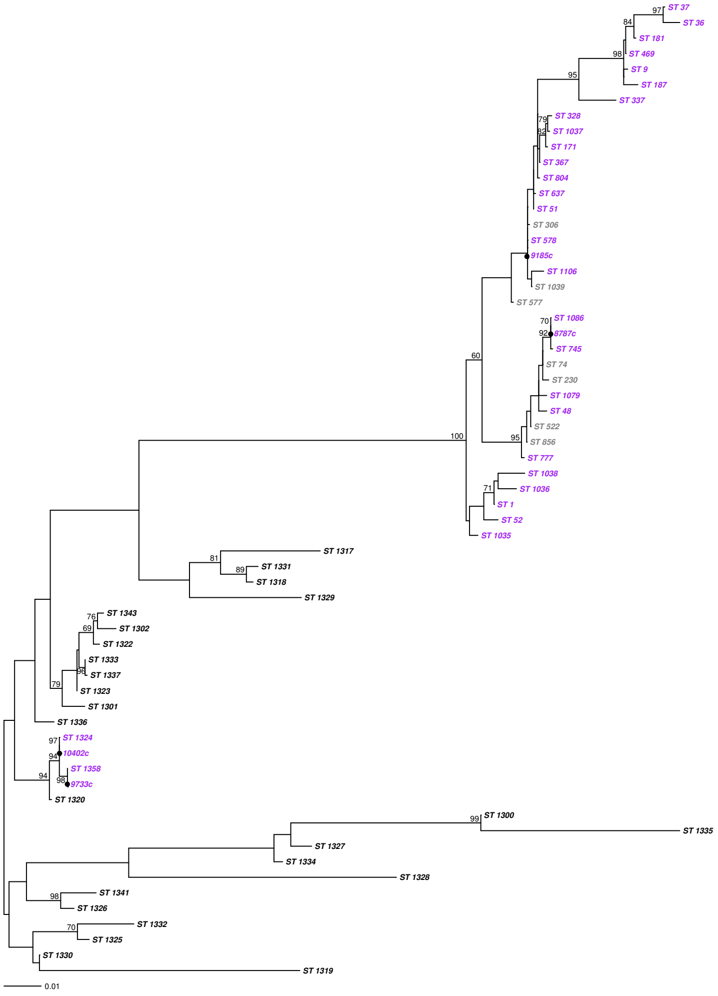


Figure 4 | Maximum likelihood phylogenetic reconstruction performed by RAxML of the seven SBT sequences of the 4 environmental isolates (marked with a black dot), 28 STs found previously in Alcoy (in purple), and 31 STs from ESGLI (in black), 7 of which differed from the 4 involved in this study only in *neuA* (in grey) and the rest containing *neuAh* alleles (lower clade of the tree). Nodes with bootstrap support values higher than 60% are shown**.**

Discussion

Molecular studies of *L. pneumophila* have been mainly derived from outbreak investigations. However, to better understand outbreaks it is important first to learn about the presence and distribution of the pathogen in the environment in non-outbreak conditions by studying potential sources of infection such as natural and artificial reservoirs and other risk facilities[117,402,406]. Besides, taking into account the role of amoebas and biofilms as *L. pneumophila* reservoirs[335,403,404,406,411,421], the usage of other types of samples, apart from water, in epidemiological studies should be further encouraged. Here, we report the results of an intensive and homogeneous sampling of water and biofilms throughout the city of Alcoy (Alicante, Spain), where recurrent outbreaks of legionellosis arise frequently[117,416,422]. These results allowed us to study the basal diversity of *L. pneumophila* in this endemic area in the absence of epidemiological alert and to test the suitability of the direct analysis of biofilms compared to water samples for the environmental detection of this pathogen.

L. pneumophila was detected by at least one of the methods in 21 samples (17.5%, Table S1), although isolation of *L. pneumophila* was only possible for 4 water samples (3.3%). Other works have shown a higher rate of isolation by culture[109,182,186,187] but these used mostly hot-water samples and *Legionella* is known to replicate actively in warm environments[5]. However, in previous reports in which direct PCR over water samples was performed the detection rate was higher than by culturing. The strains isolated in this study were all retrieved from street dead-ends, which accounted for the majority of positive sampling points (57.1%). In the 60 sampling points in which both water and biofilm samples were studied, only 1.7% of the water samples were *L. pneumophila*-positive by PCR, while the analysis of biofilms increased the percentage of detection up to 21.7%.

Besides, the presence of *L. pneumophila* was not found to correlate with the levels of chlorine detected in the corresponding sampling points, as shown previously by other studies[423,424]. Positive samples were found in different areas of the city, with a higher concentration in the northern part and a peak during the month of June with an average temperature at the sampling points in the range 25-30ºC, around the optimum temperature at which *Legionella* can multiply actively[5].

Although biofilm swabs have been already used for *L. pneumophila* detection in environmental samples[131,406,409–411], to our knowledge no specific comparison between the detection of *L. pneumophila* from water and biofilm samples has been performed yet. Analyzing one liter of water involves a more diluted bacterial sample that has to be processed through filtering and re-suspension steps, which might reduce the final yield of extracted DNA. However, biofilms can be scraped and transported easily using cotton or polystyrene swabs, which can be directly used for total DNA extraction and subsequent amplification. The biofilm-based detection method for *L. pneumophila* surveillance presented here can be readily applied to the study of outbreaks (Sánchez-Busó *et al*, submitted) because it has been proven to be rapid and sensitive enough to detect and even identify outbreak-causing *L. pneumophila* reservoirs and spreading devices in different human-made environments. However, further work needs to be performed in order to validate this method for routine analysis.

Changes in water flow velocity can produce biofilm detachment (Horn *et al*, 2003), although most *L. pneumophila*-positive samples were retrieved from sites with stagnant water. Thus, it is likely that the attached biofilms were not eroded enough by the water collection process at the time of sampling. This, along with the low bacterial load in water, can explain the low rate of detection in water samples compared to biofilms, which were mechanically scraped. Sites in which the biofilm sample tested negative for *L. pneumophila* but positive for the other methods can be explained by a reduced amount of biofilms collected by scraping with a cotton swab during sampling or by the previous detachment of the bacteria, which also shows some of the limitations of this method.

The STs found in this work (ST578, ST1086, ST1324 and ST1358) show a level of diversity that is congruent with that known to be present in the Alcoy area by the SBT culture-based routine surveillance programs performed in our laboratory (Fig. 3 and Fig. 4). Specifically, ST578 has caused many of the recent outbreaks in Alcoy[117], and has been found in this study in a similar frequency to other STs. Considering the high incidence of ST578-related legionellosis cases in this area, it is probable that, when conditions favor the multiplication of *Legionella*, ST578 presents a replicative and infective advantage. Further studies at the microbiological and genomic level are needed in order to confirm this hypothesis.

Despite that less than 20% of the sampling points tested in this study resulted positive for *L. pneumophila*, the results presented in this work show a level of environmental variability in an endemic area higher than expected. These results lead us to hypothesize that, in the absence of epidemiological risk, this pathogen is mostly attached to biofilms in a viable but non-culturable form in the water distribution system, thus explaining the low detection rate from water samples both by culturing or molecular amplification. Moreover, our results provide an applicable and rapid approach to increase the rate of *L. pneumophila* detection for routine analysis and outbreak investigations by direct amplification from biofilm samples. However, its combination with culture-based methods is essential in order to define precisely the sequence types involved in the investigation when the presence of more than one genetic pattern makes the assignment from direct amplification difficult. Finally, additional high-throughput studies are also of interest to better evaluate the actual heterogeneity of *L. pneumophila* in natural environments as well as the bacterial communities in which *Legionella* can be found, and also to gain insight on the mechanisms driving their evolution and adaptation to specific microenvironments.

Acknowledgements

This work has been supported by Conselleria de Sanidad, Generalitat Valencia and projects BFU2008-03000/BMC from MICINN, and BFU2011-24112 from MINECO (Spanish Government). LSB is recipient of a FPU predoctoral fellowship from Ministerio de Educación, Spanish Government. We thank the editor and three anonymous reviewers for their comments on previous versions of the manuscript.

Supplementary material

Table S1 | Sampling sites tested positive for *L. pneumophila*. C, W and B denote culture, water and biofilm, respectively (positive or negative results are shown as +/-). Sampling sources, water temperature and chlorine levels at each point are also shown.

Sample ID	Date	C	W	B	Source	Sampling point	Chlorine in water (mg/l)	Water temperature (ºC)
8523	19/05/2011	-	+		Industry	Water nozzle	1.07	20.90
8524	19/05/2011	-	+		Industry	Spray nozzle	NA	NA
8527	19/05/2011	-	+		Industry	Water nozzle	0.90	21.90
8787	26/05/2011	+	+		Street	Dead-end	1.15	22.40
8794	26/05/2011	+	-		Street	Dead-end	1.21	24.00
9185	02/06/2011	+	-		Street	Dead-end	1.53	20.10
9729	09/06/2011	-	-	+	Street	Dead-end	1.43	20.20
9733	09/06/2011	+	+	-	Street	Dead-end	0.77	21.70
10006	16/06/2011	-	-	+	Street cleaner	Water deposit	6.00	27.00
10007	16/06/2011	-	-	+	Street cleaner	Water nozzle	5.05	24.20
10015	16/06/2011	-	-	+	Industry	Spray nozzle	0.35	30.90
10401	23/06/2011	-	-	+	Street	Dead-end	1.24	25.40
10402	23/06/2011	+	-	-	Street	Dead-end	1.23	26.00
10403	23/06/2011	-	-	+	Street	Dead-end	1.05	17.50
10404	23/06/2011	-	-	+	Street	Dead-end	1.42	25.40
10405	23/06/2011	-	-	+	Street	Dead-end	0.91	24.30
10406	23/06/2011	-	-	+	Street	Dead-end	1.38	29.10
10407	23/06/2011	-	-	+	Street	Dead-end	1.31	24.70
10718	30/06/2011	-	-	+	Fountain	Natural non-drinking water	0.00	16.10
10720	30/06/2011	-	-	+	Fountain	Natural non-drinking water	0.00	12.80
11209	11/07/2011	-	-	+	Industry	Water nozzle	0.00	27.40

Table S2 | Optimal annealing temperatures for the touchdown PCR of the seven loci in the SBT scheme.

Region	Touchdown temperatures (ºC)		Reference
	10 cycles	35 cycles	
fliC	66-56	56	Gaia *et al*, 2003
pilE	65-55	55	Gaia *et al*, 2005
asd	66-56	56	Gaia *et al*, 2005
mip	66-56	56	Gaia *et al*, 2005
mompS	60-50	50	Gaia *et al*, 2003-2005
proA	64-54	54	Gaia *et al*, 2003
neuA	66-56	56	Coscollá *et al*, unpublished

Table S3 | Allelic profiles of the 63 sequences included in the phylogenetic reconstructions shown in Fig. 4 and Fig. S1. The samples from this study are shown in bold and the remaining sequences were retrieved from the ESGLI database.

Name	***fliC***	***pilE***	***asd***	***mip***	***mompS***	***proA***	***neuA***	**ST**
ST_1	1	4	3	1	1	1	1	1
ST_9	3	10	1	3	14	9	11	9
ST_36	3	4	1	1	14	9	1	36
ST_37	3	4	1	1	14	9	11	37
ST_48	5	2	22	27	6	10	12	48
ST_51	6	10	15	28	9	14	6	51
ST_52	1	10	3	1	1	1	1	52
ST_74	5	1	22	30	6	10	6	74
ST_171	6	10	15	3	21	4	9	171
ST_181	3	10	1	12	14	9	9	181
ST_187	3	10	1	28	14	9	3	187
ST_230	5	1	22	30	6	10	10	230
ST_306	6	10	15	13	9	14	11	306
ST_328	6	10	19	28	19	4	9	328
ST_337	10	22	7	28	16	18	6	337
ST_367	6	10	15	28	21	14	9	367
ST_469	3	10	1	28	14	9	9	469
ST_522	5	1	22	30	6	10	13	522
ST_577	6	10	15	13	9	14	16	577
ST_578	6	10	15	13	9	14	6	578
ST_637	6	10	15	3	9	14	6	637
ST_745	5	2	22	10	6	25	9	745
ST_777	5	2	22	10	6	25	1	777
ST_804	6	10	15	3	21	14	6	804
ST_856	5	1	22	30	6	10	1	856
ST_1035	2	3	18	10	2	1	1	1035
ST_1036	1	4	3	19	1	1	3	1036
ST_1037	6	10	14	28	21	4	9	1037
ST_1038	1	6	3	10	1	5	11	1038
ST_1039	6	10	15	13	21	14	6	1039
ST_1070	5	1	22	30	6	10	15	1079
ST_1086	5	2	22	10	6	25	6	1086
ST_1106	6	10	15	12	17	14	9	1106
ST_1300	11	14	16	18	15	13	201	1300
ST_1301	2	11	3	28	9	4	207	1301

ST_1302	2	10	4	28	4	4	207	1302
ST_1317	16	21	12	19	31	21	215	1317
ST_1318	6	10	5	10	9	1	209	1318
ST_1319	2	6	17	14	12	8	211	1319
ST_1320	8	1	22	30	6	10	203	1320
ST_1322	6	10	22	3	21	3	207	1322
ST_1323	6	10	3	28	9	4	207	1323
ST_1324	5	1	22	30	6	10	203	1324
ST_1325	7	6	3	20	13	11	205	1325
ST_1326	3	10	1	28	14	9	207	1326
ST_1327	11	14	16	31	15	13	210	1327
ST_1328	20	26	27	34	46	27	212	1328
ST_1329	1	4	3	5	50	1	213	1329
ST_1330	2	6	17	28	13	31	207	1330
ST_1331	6	10	14	10	33	1	209	1331
ST_1332	12	8	11	20	40	12	216	1332
ST_1333	2	10	3	28	9	14	207	1333
ST_1334	11	14	16	25	7	13	206	1334
ST_1335	14	18	8	18	28	19	201	1335
ST_1336	6	35	38	42	1	14	207	1336
ST_1337	2	10	3	3	9	14	207	1337
ST_1341	3	13	1	3	14	9	207	1341
ST_1343	6	10	21	28	4	14	207	1343
ST_1358	5	2	22	10	6	25	203	1358
8787c	**5**	**2**	**22**	**10**	**6**	**25**	**6**	**1086**
9185c	**6**	**10**	**15**	**13**	**9**	**14**	**6**	**578**
9733c	**5**	**2**	**22**	**10**	**6**	**25**	**203**	**1324**
10402c	**5**	**1**	**22**	**30**	**6**	**10**	**203**	**1358**

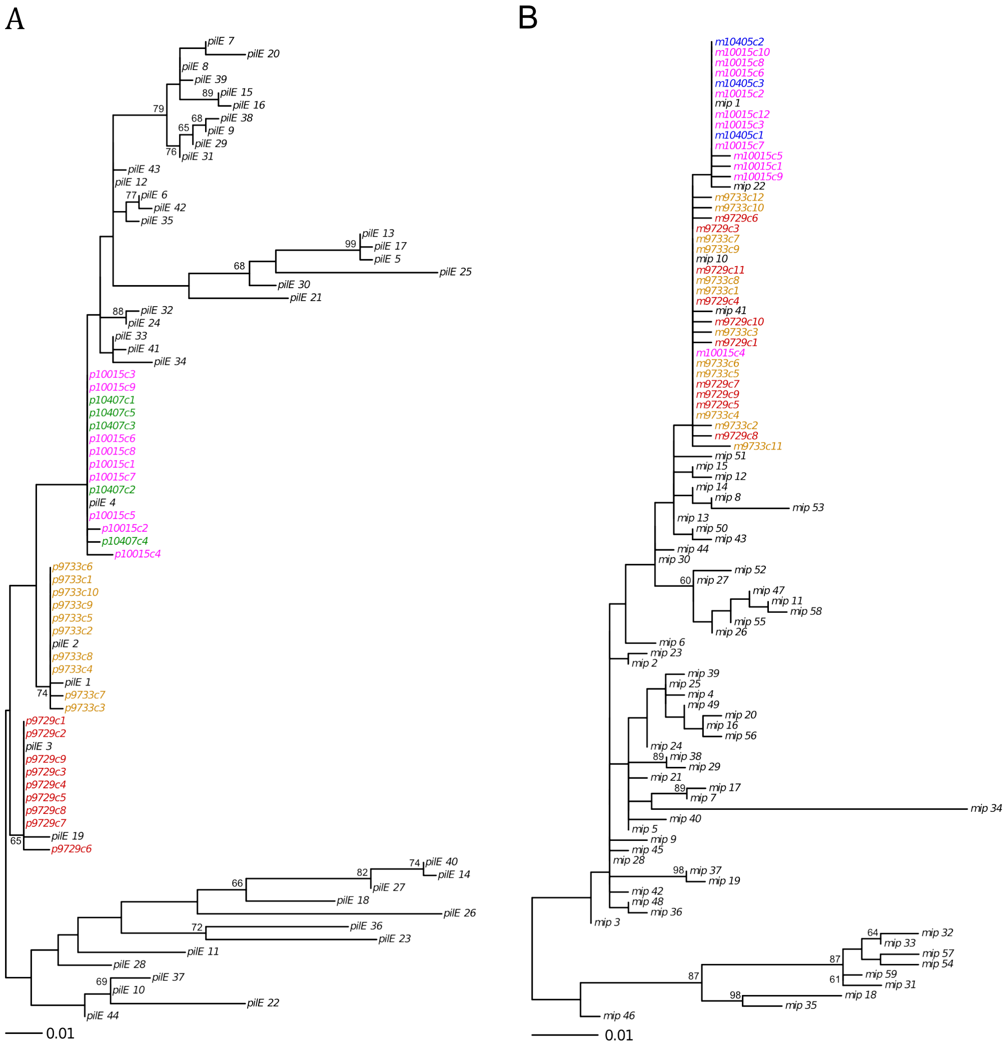


Figure S1 | Maximum likelihood phylogenetic reconstruction of the sequences obtained from clones of the *pilE* (A) and *mip* (B) fragments amplified by TD-PCR in biofilm samples (9729, 9733, 10015, 10405 and 10407, marked in red, orange, pink, blue and green, respectively) performed using RAxML. Note that clones from all samples could not be retrieved for these loci. Nodes with bootstrap support values higher than 60% are shown.

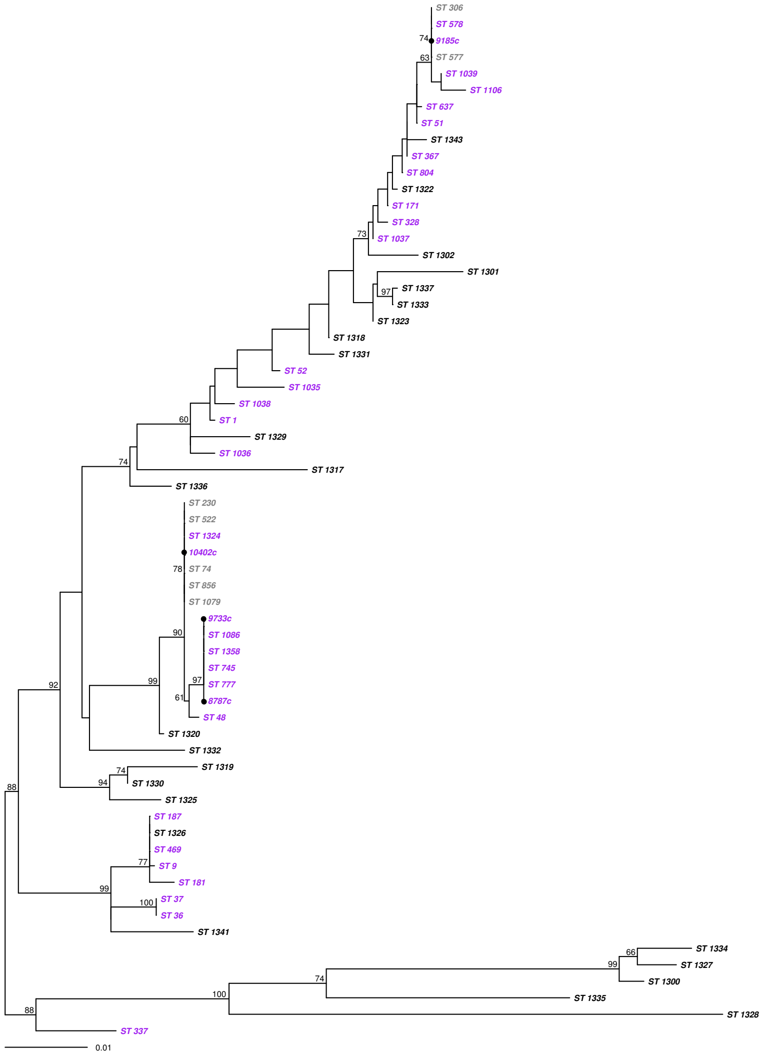


Figure S2 | Maximum likelihood phylogenetic reconstruction performed by RAxML of six fragments of the SBT scheme (all except *neuA*) of the 4 environmental isolates (marked with a black dot), 28 STs found previously in Alcoy (in purple), and 31 STs from ESGLI (in black), 7 of them differing from the 4 involved in this study only in *neuA* (in grey) and the rest containing *neuAh* alleles (lower clade of the tree). Nodes with bootstrap support values higher than 60% are shown.

Chapter 4

Genomic investigation of a legionellosis outbreak in a persistently colonized hotel

Leonor Sánchez-Busó[1,2], Silvia Guiral[1,3], Sebastián Crespi[4,5], Víctor Moya[6], María Luisa Camaró[7], María Piedad Olmos[7], Francisco Adrián[6], Vicente Morera[8], Francisco González Morán[1,3], Hermelinda Vanaclocha[1,3], Fernando González-Candelas[1,2].

1. Unidad Mixta "Infección y Salud Pública" FISABIO-Salud Pública, Universitat de València. Valencia, Spain.
2. CIBER Epidemiología y Salud Pública. Valencia, Spain.
3. DGSP. Subdirección de Epidemiología y Vigilancia de la Salud.
4. Biolinea Int. Palma de Mallorca, Spain.
5. Policlínica Miramar. Palma de Mallorca, Spain.
6. Legioprev. Valencia, Spain.
7. DGSP. Laboratorio de Salud Pública. Valencia, Spain.
8. Centro de Salud Pública Denia. Denia, Spain.

This work has been submitted to *Journal of Infection*.

Abstract

A long-lasting legionellosis outbreak was reported between November 2011 and July 2012 in a hotel in Calpe (Spain) affecting 44 patients including 6 deaths. Intensive epidemiological and microbiological investigations were performed in order to detect the reservoirs. Clinical and environmental samples were tested for the presence and genetic characterization of *Legionella pneumophila*. Six of the isolates were subjected to whole-genome sequencing.

Sequencing of all clinical isolates revealed sequence type (ST) 23 as persistently found in the spa pool, from where it spread to other hotel public spaces, explaining the ST23 clinical cases, including guests who had not visited the spa. Uncultured clinical specimens showed profiles compatible with ST23, ST578 and mixed patterns. Profiles compatible with ST578 were obtained by direct sequencing from biofilm samples collected from the domestic water system. Whole genome data from five ST23 strains provided evidence that these patients had been infected by distinct STs that likely colonized the hotel since its opening.

Both epidemiological and molecular data are essential in the investigation of legionellosis outbreaks. Whole-genome sequencing data revealed significant intra-ST variability and allowed to make further inference on the short-term evolution of a local colonization of *L. pneumophila*.

Introduction

Legionellosis infections are opportunistic and the inhalation of aerosols with enough bacterial loads can cause a severe form of pneumonia, known as Legionnaires' disease[82], or a milder flu-like condition, denoted as Pontiac fever[88]. After its first identification in 1976 in a Legionnaires' convention in a hotel in Philadelphia, many legionellosis outbreaks have involved travel-associated clusters[103,104,212].

Legionellosis outbreaks are usually studied by typing pure *Legionella* cultures from infected patients with different molecular techniques; assuming that only one strain is causing the disease. However, co-infection with different *Legionella* species has also been reported[231,425–428], as well as with different serogroups of *L. pneumophila*[230,231,425]. The introduction of new typing methods based on direct amplification and sequencing of *Legionella* from clinical samples[183,227,228] has also revealed that outbreak patients can be infected simultaneously by more than one *Legionella* strain[229], even from the same serogroup. Dual or multiple infections pose an additional difficulty for the identification and subsequent control of outbreak sources.

In November 2011, the first cases of a legionellosis outbreak that lasted 33 weeks were reported in a hotel in the locality of Calpe (Comunidad Valenciana, Spain). The outbreak comprised four different clusters with 2 suspected and 42 confirmed patients that matched the European case definition including 6 fatalities. The hotel had a previous history of six legionellosis cases in 2006, just three months after it was opened to the public, and one case in 2007. However, no additional cases had been reported until the outbreak analyzed in this work, in which both tourists and workers were affected.

Upon notification of the first cases, public health officials started an environmental investigation and the implementation of control measures[107]. These included the disinfection of the water distribution system of the hotel and its closure. Here, we present the molecular analyses performed during the investigation of this outbreak using both cultured and uncultured clinical and environmental samples that helped unveil the infection source and route for all the patients. Whole-genome sequencing (WGS) was performed for some of the isolates to get higher discrimination than the traditional DNA Sequence-Based Typing (SBT) and to infer when *L. pneumophila* might have colonized the facility.

Materials and methods

Case definition and epidemiological description of the outbreak

Confirmed cases were defined as patients who, having stayed or worked at the hotel between two and ten days before the onset of symptoms, showed a clinical diagnosis of pneumonia with laboratory findings which confirmed infection by *Legionella*, including a positive urine test for *L. pneumophila* antigen or a positive culture isolation of the bacteria from respiratory secretions.

The outbreak was epidemiologically divided into four temporal clusters that involved 44 cases, 38 tourists (average age 71.5) and 6 hotel workers (average age 49.5) (Fig. 1). Visitors had stayed at 28 different rooms in the hotel, distributed in 11 floors. Six of the rooms accommodated two cases and other three patients had also used the same room, two of them simultaneously. The outbreak involved 6 deaths (average age 77.2), and deceased patients had stayed at 5 different rooms: two of those patients had stayed in the same room at different times. Only 5 tourists and 2 workers had ever been in the spa facility of the local, which is located below the main hall from where it is easily visible through a glass dome.

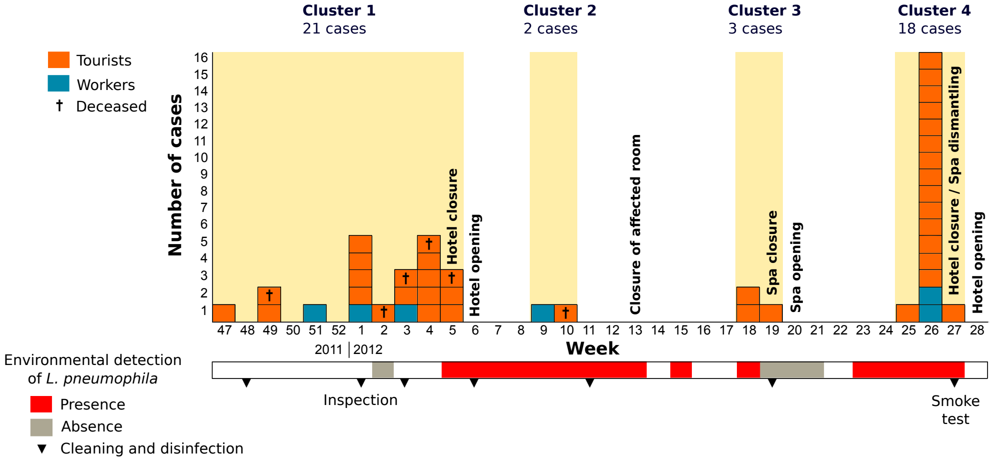


Figure 1 | Number of cases affected by legionellosis grouped per week of onset of symptoms. Detection of *L. pneumophila* during the environmental investigation is indicated by colors in the lower bar (see legend). Measures undergone by local health authorities are also marked in the corresponding weeks.

Clinical samples

Clinical samples from 14 outbreak patients were received for study in our local reference laboratory for *L. pneumophila* infections. Specifically, 8 sputum, 2 broncho-alveolar aspirates (BAS), 1 lung puncture aspirate and 3 cultures obtained in the Microbiology Service of the corresponding hospitals, two of them from autopsies, were collected and transported at 4ºC for genetic testing. Total DNA was extracted from uncultured clinical samples using UltraClean® BloodSpin® DNA Isolation Kit (MoBio). DNA from the 3 isolates was extracted using a thermal shock as described below for environmental isolates.

Environmental sampling

A total of 632 two-liter water samples and 164 biofilm swabs (per duplicate) were taken during the environmental investigation of the outbreak from the different water systems of the hotel including the spa pool. Water samples were tested for biocides, pH and temperature. Pieces of air filters of the air-conditioning system connecting the spa and the hall area were also collected for testing. All samples were transported immediately to the lab for *L. pneumophila* detection and genetic characterization.

Water filtering and culturing were performed following the AFNOR NF T 90-431 2 (2003) and UNE-EN ISO 11731-2:2008 regulations on water quality and standard culture-based detection of *Legionella* spp. isolates were serogrouped using Oxoid *Legionella* Latex Test (Thermo Scientific), subtyped using monoclonal antibodies[7,8], and subsequently characterized genetically. Different colonies from the culturing plates were analyzed, when available, to check for strains of different genetic profile.

Biofilm swabs and air filters were submerged into lysis buffer (100 mM Tris HCl, 100 mM NaCl, 1 mM EDTA, 10% SDS and ultrapure water, final pH 8.1) and incubated with 1 mg/mL proteinase K during 10 minutes at 65ºC for biofilm swabs and 30 minutes at the same temperature for the other samples. Cotton swabs were introduced into 0.7 mL microtubes with an opening at the bottom placed inside a 2-mL microtube[417] and centrifuged at maximum speed for 10 minutes to recover the maximum amount of lysate. Final retrieved volumes were used for direct DNA purification using the spin filter protocol of the UltraClean® BloodSpin® DNA Isolation Kit (MoBio).

DNA from *L. pneumophila* culture isolates was extracted by thermal shock. A small amount of bacteria was suspended in 200 µL of autoclaved, ultrapure water and subjected to two cycles of a 5-min incubation step at

99ºC followed by 5 min at 4ºC. Cell debris was pelleted by centrifuging during 3 min at 12,000 rpm, and the DNA-containing supernatant was used for molecular amplification.

L. pneumophila detection and Sequence-Based Typing (SBT)

Touchdown PCR (TD-PCR)[418] targeting two regions of the Sequence-Based Typing (SBT)[216,359] scheme (*pilE* and *asd*) was performed on the DNA samples extracted from biofilms for *L. pneumophila* detection. Positive samples were subsequently used for amplifying the 5 remaining loci (*fliC*, *mip*, *mompS*, *proA* and *neuA*) in the SBT scheme for this species. Amplification was performed using 2.5 U FastStart High Fidelity PCR System (Roche Applied Science), 1X Buffer with MgCl2 1.8 mM, 200 µM of each dNTP, 4% DMSO, 0.2 µM of the corresponding pair of primers, an amount of DNA suspension depending on its concentration (measured three times by NanoDrop 1000TM, Thermo Scientific) and autoclaved and filtered ultrapure water to a final volume of 25 µL. The thermal profile of the TD-PCR included an initial denaturation step of 5 minutes at 95ºC followed by 10 cycles of denaturation (30 s at 95ºC), annealing (30 s at a primer-dependent temperature[352] and at a ramp rate of -1 ºC/cycle) and extension (30 s at 72ºC). Immediately after the TD step, 35 cycles of standard PCR were applied using the same conditions but fixing the annealing temperature at the optimal for each region; a final extension step of 8 min at 72ºC was also used to ensure complete amplification.

DNA retrieved from *L. pneumophila* pure cultures and uncultured clinical samples from confirmed patients were used for amplification of the 7 loci in the SBT scheme by traditional[267] and seminested PCR amplification[228]. PCR products were purified and Sanger-sequenced as previously described in Coscollá *et al*[228,267].

Cloning of PCR products

PCR products from 5 positive biofilm samples obtained in hotel rooms associated to clinical cases and that showed evidence of multiple *L. pneumophila* variants in *fliC*, *pilE* and *asd* were cloned into *Escherichia coli* JM109 Competent Cells (Promega) using pGEM®-T Easy vector System II (Promega). In order to check whether the minor variants detected could be considered actual polymorphisms and not artifacts resulting from the amplification and/or cloning process, TD-PCR products from pure cultures of these three regions were cloned and sequenced. PCR products from the same regions obtained from 4 sputum/BAS samples were also cloned in order to check for intra-patient variability. Ligation and transformation reactions were performed following the instructions of the manufacturer.

Phylogenetic reconstruction of SBT data

The sequences of the alleles involved in the STs found in the analyzed samples were downloaded from the ESGLI SBT database[221] and concatenated in order to create a haplotype sequence for each strain. Complete and incomplete profiles (with at least two sequenced loci) were used for maximum likelihood phylogenetic reconstruction using RAxML v7.2.8[345] with the GTRGAMMA model of substitution and 1000 bootstrap replicates along with reference STs.

Whole-genome sequencing on outbreak isolates

Four clinical ST23 and 2 environmental isolates (ST23 and ST1236) were selected because of their DNA quality and quantity for whole-genome sequencing using the SOLiD 5500XL platform. Single-end reads of 75 bp with an average per nucleotide coverage of 90X were mapped against the assembly of one of the ST23 strains (VelvetOptimiser v2.2.5)[429], ID_125 (resulting scaffold of 3,307,11 bp), following the pipeline described in

Sánchez-Busó *et al* (2014)[339]. Filtered SNPs from the core genome of the six strains and 9 reference genomes from GenBank (Paris - NC_006368[72], Lens - NC_006369[72], Philadelphia1 - NC_002942[52], Alcoy - NC_014125[236], Corby - NC_009494[235], Lorraine - NC_018139[271], 130b - FR687201[240], ATCC43290 - NC_016811 and HL06041035 - NC_018140) were used for phylogenetic reconstruction under the minimum evolution criterion[430]. To do so, pairwise distances and 1,000 bootstrap replicates were applied with MEGA5.0[362]. SOLiD genomic data have been deposited in the European Nucleotide Archive (ENA) under project accession PRJEB5990.

Bayesian estimation of colonization time

In order to study the time at which ST23 could have started proliferating within the hotel, we used the ST23 non-homoplastic core alignment (3,267,949 bp) for Bayesian estimation with BEAST v1.7.5[431]. The HKY model of nucleotide substitution and an uncorrelated lognormal molecular clock were selected using the substitution rate previously estimated for ST578[339] and the isolation dates as priors. Four parallel Markov chains using the same parameters were run for 10,000,000 generations with samples taken every 10,000 steps. Convergence of the chains was tested with Tracer v1.5 when the ESS for all the parameters was >200.

Results

Clinical L. pneumophila SBT profiles

L. pneumophila could be cultured from only 4 (28.57%) of 14 the clinical samples studied in our laboratory, two from pulmonary tissue retrieved during autopsy and the other two from sputums. All of them were typed as ST23 (Table 1), in agreement with the results obtained by Public Health England (PHE) from the first travel-associated cases related to the hotel that were reported after returning to their country of origin[107]. Eleven uncultured clinical samples (one of them was also culture-positive) were analyzed by direct amplification and sequencing as described. Three of them provided a complete genetic profile (ST23) and an additional sample (163655) resulted in a new profile that differed from ST23 in two loci (*fliC* and *mompS*; Table 1). Other two samples (191712 and 192147) presented very divergent profiles to ST23. However, some of the sequencing reads from these and two other samples (192061 and 163655) presented unresolved nucleotides at several positions, which led us to seek determining their sequence with better precision.

Cloning and sequencing of PCR products from three loci (*fliC*, *pilE* and *asd*; Table 2) from the 4 uncultured sputums detailed above revealed the presence of two variants (*asd*-1 and *asd*-15) in patient 191712, providing evidence for the presence of an infecting strain that matched that found in patient 192147. It is worth mentioning that these two patients were a couple who had stayed in the same hotel room and that although the corresponding profiles missed one locus each, their combination corresponded to ST578 (Table 1).

Table 1 | Sequence-based Typing of *L. pneumophila* strains from clinical samples. Shadowed bold numbers correspond to PCR products that were cloned for further genetic testing.

Sample	Type of	*fliC*	*pilE*	*asd*	*mip*	*mompS*	*proA*	*neuA*	ST
125	Culture	2	3	9	10	2	1	6	23
191712	Sputum	**6**	**10**	**1,15**	13	-	14	6	(578)
192091	Culture	2	3	9	10	2	1	6	23
	Sputum	**-**	**3**	**9**	10	2	1	6	23
163655	Sputum	**6**	**3**	**9**	10	9	1	6	A[*]
192147	Sputum	**6**	**10**	**1**	13	2	-	6	-
50291	Culture	2	3	9	10	2	1	6	23
50726	Culture	2	3	9	10	2	1	6	23
191620	Lung	2	3	9	10	2	1	6	23
193975	Sputum	6	3	-	-	2	1	6	-
9138	Sputum	-	-	9	-	-	-	-	-
160613	Sputum	2	3	9	10	2	1	6	23
18484	Sputum	2	3	9	-	2	1	6	-
160717	BAS	2	3	9	10	2	1	6	23

[*] ST-A corresponds to a new profile that could not be reassigned because results came from an uncultured sample.

Table 2 | List of clinical and environmental biofilm samples submitted to cloning of PCR products. Underlined bold regions were used (*fliC, pilE, asd, mip, mompS, proA, neuA*). Selected biofilm samples were taken from different rooms.

Sample	C/E[a]	Allele pattern[b]	ST	Implication
191712	C	**6,10,1/15**,13,X,14,6	578[c]	Patient
192147	C	**6,10,1**,13,2,X,6	-	Patient
163655	C	**6,3,9**,10,9,1,6	B	Patient
192091	C	**X,3,9**,10,2,1,6	23[c]	Patient
cal12_18	E	**+,+,+**,+,+,+,+	+	Hotel room with 2 cases (1 death)
cal12_40	E	**6,10,15**,X,9,14,6	578[c]	Hotel room with 1 case (1 death)
cal12_46	E	**6,10,15**,13,9,14,6	578	Hotel room with 2 cases (1 death)
cal12_97	E	**6,10,+**,+,9,14,+	578[c]	Hotel room with 1 case
cal12_151	E	**6,10,15**,13,9,14,6	578	Hotel room with 2 cases (2 deaths)
F3	E	**2/6,3,9**,10,2,1,6	23	Air filter

[a] C: clinical, E: environmental.

[b] X: PCR amplification failed, +: positive amplification but allele assignation failed.

[c] Incomplete allele pattern but compatible with ST23 or ST578.

Environmental L. pneumophila SBT profiles

From the sampled biofilms, 106 (64.63%) resulted positive for at least one of the two SBT loci initially tested, *pilE* and *asd*. PCR and further sequencing of the remaining other loci revealed that a profile compatible with ST578 was frequently found in samples from most of the rooms studied and other common hotel facilities (Table 3). Only one biofilm swab was taken from the spa at the end of January 2012 and it tested negative.

Positive cultures were obtained from water samples taken at the spa pool (Table 3), revealing mostly *L. pneumophila* serogroup 1 MAb type Allentown/France ST23, and *L. pneumophila* serogroup 1 ST1236. However, *L. pneumophila* serogroup 10 (ST1358) and *L. micdadei* were also observed when several colonies from the same plate were analyzed*.* ST1236 isolates showed a SBT profile (6, 10, 15, 10, 21, 40, 6) that is genetically close to that of ST578 (6, 10, 15, 13, 9, 14, 6), found only in biofilm samples, with a difference in only 5 nucleotide sites (Fig. 2).

The link between the environmental strain ST23 found in the spa pool and most clinical cases was firmly established by the corresponding SBT sequencing results. However, only the direct SBT analysis of uncultured environmental and clinical samples provided evidence for the existence of a second infecting strain found only in the domestic water system, compatible with ST578, which can be linked reliably to at least one clinical case (Tables 1 and 2). In general, the phylogenetic tree derived from all these sequence data revealed links between all the clinical cases and environmental samples from the hotel (Fig. 2).

Table 3 | Results of the molecular investigation of environmental samples.

Date	Sample ID	Location	N	Type of sample	*fliC*	*pilE*	*asd*	*mip*	*mompS*	*proA*	*neuA*	ST/Species	Serogroup/ MAb type
Jan-March 2012	-	Hotel rooms, Hall, Restrooms	164	Biofilm	6	10	15	13	9	14	6	578	-
31/01/2012	Al47	Spa pool	3	Culture	6	10	15	10	21	40	6	1236	1
			3	Culture	-	-	-	-	-	-	-	*Legionella micdadei*	-
31/01/2012	918	Spa pool	2	Culture	6	10	15	10	21	40	6	1236	1
11/02/2012	12392	Spa pool	1	Culture	2	3	9	10	2	1	6	23	1
23/02/2012	2084	Spa pool	1	Culture	-	-	-	-	-	-	-	*Legionella micdadei*	-
12/04/2012	4029	Spa pool	4	Culture	2	3	9	10	2	1	6	23	1
			1	Culture	6	10	15	10	21	40	6	1236	1
			1	Culture	-	-	-	-	-	-	-	*Legionella micdadei*	-
02/05/2012	4855	Spa pool	9	Culture	-	-	-	-	-	-	-	*Legionella micdadei*	-
02/05/2012	4856	Spa pool	12	Culture	2	3	9	10	2	1	6	23	1
02/05/2012	4857	Spa pool	3	Culture	2	3	9	10	2	1	6	23	1 Allentown/France
			1	Culture	6	10	15	10	21	40	6	1236	1 Allentown/France
			1	Culture	5	2	22	10	6	25	203	1358	10
			2	Culture	-	-	-	-	-	-	-	*Legionella micdadei*	-
26/06/2012	8298	Room, cold wáter	9	Culture	-	-	-	-	-	-	-	-	2-15
04/07/2012	Al657	90 m3 water tank	10	Culture	2	3	9	10	2	1	6	23	1
07/07/2012	PF2	Air filter	-	Biofilm	6	3	-	-	-	-	-	-	-
07/07/2012	PF3	Air filter	-	Biofilm	-	3	9	-	-	1	6	-	-

07/07/2012	F1	Air filter	-	Biofilm	6	3	9	10	2	1	6	1164	-
07/07/2012	F2	Air filter	-	Biofilm	6	3	9	10	2	1	6	1164	-
07/07/2012	F3	Air filter	-	Biofilm	2	3	9	10	2	1	6	23	-
09/07/2012	8965	Osmosis (in)	7	Culture	2	3	9	10	2	1	6	23	1
09/07/2012	8966	Osmosis (out)	5	Culture	2	3	9	10	2	1	6	23	1
09/07/2012	8968	Spa (fountain)	10	Culture	2	3	9	10	2	1	6	23	1
18/07/2012	69135	Spa pool	1	Culture	-	-	-	-	-	-	-	*Legionella micdadei*	-
18/07/2012	69147	Spa pool	1	Culture	-	-	-	-	-	-	-	*Legionella micdadei*	-
18/07/2012	69148	Osmosis device	1	Culture	2	3	9	10	2	1	6	23	1 Allentown/France
18/07/2012	69153	Spa pool	1	Culture	2	3	9	10	2	1	6	23	1 Allentown/France
18/07/2012	69183	Spa pool drainer	1	Culture	2	3	9	10	2	1	6	23	1 Allentown/France
27/07/2012	Al763	Watering network (hotel entrance)	1	Culture	5	2	22	10	6	25	203	1358	8

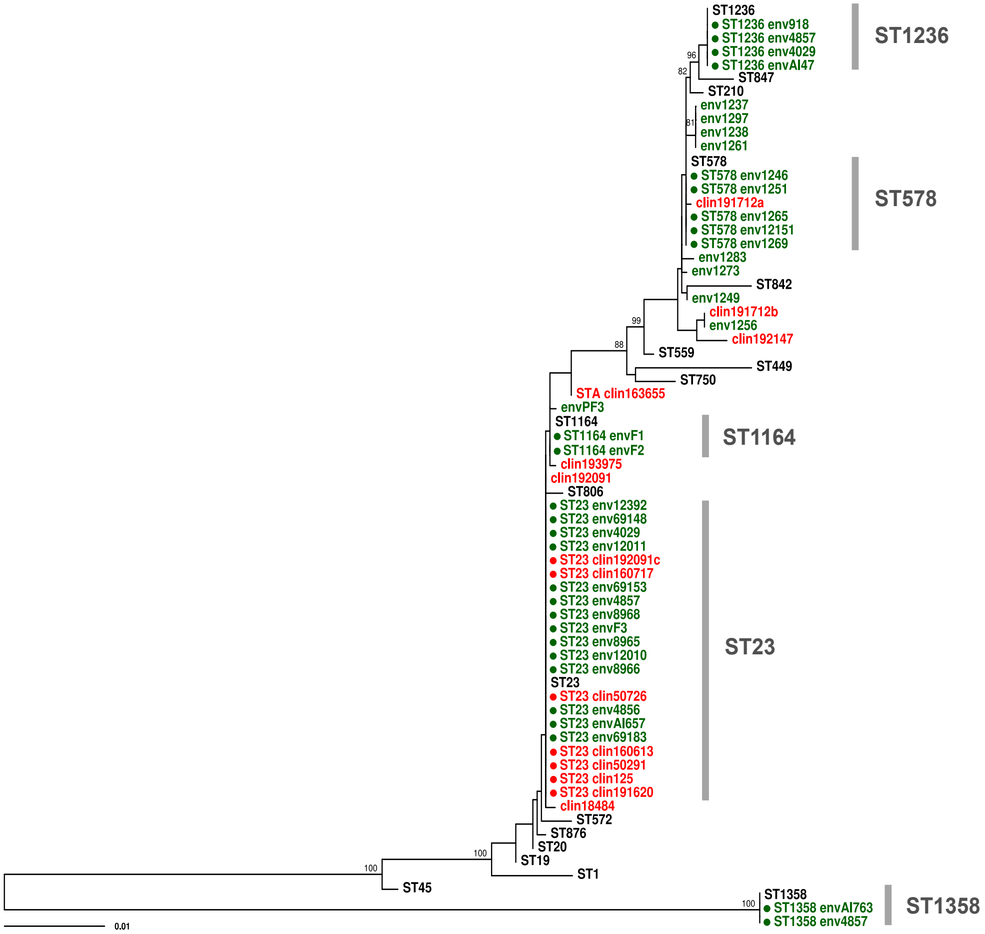


Figure 2 | Maximum likelihood tree constructed with RAxML using SBT data of all clinical and environmental samples included in the study. Red and green tips represent clinical and environmental samples, respectively. Reference STs are shown in black. Complete profiles are marked with a filled circle. Bootstrap support values higher than 70% are shown.

Smoke-tracing studies and airflow dynamics models showed that aerosols from the spa area could reach easily the hotel hall by different ways, including the air-conditioning system. Pieces of the air filters collected from the air-handling unit for the hall area were further analyzed by direct SBT.

Five of the six filter samples analyzed were positive for two loci of the bacteria (*pilE* and *asd*) by TD-PCR. A mixed pattern was detected for locus *fliC* in one of the samples, which prompted us to use the same procedure for cloning and sequencing of PCR products described previously. The analysis of sequences derived from filter samples revealed the presence of strain ST23, thus providing clear evidence of the presence of this ST in the air-conditioning system of the hotel, and also a compatible profile (ST1164) which differs from ST23 in only one polymorphic site (Tables 2 and 3). The sequences obtained from these samples were also included in the phylogenetic tree shown in Fig. 2.

ST23 colonization history analyzed by WGS

Five ST23 isolates (4 clinical and 1 obtained from the spa pool) were subjected to whole-genome sequencing (WGS) to assess the intra-ST23 variability and to gain insight about its colonization history. Additionally, one ST1236 isolate was included to check its relatedness to ST578, from which it is genetically close according to the SBT data.

Seventy-one polymorphic sites were detected in the core genome of the five ST23 strains (3,267,949 bp). Twenty-four were estimated as homoplastic and removed from further analyses to account for the effect of recombination, leaving 47 non-homoplastic SNPs. The average number of core SNPs between ST23 isolates was estimated to be 22.4 (range 9-34; 6.85E-06 substitutions per site, s/s) (Fig. 3). Strain ID_918 (ST1236) showed a high divergence from the ST23 strains, with approximately 38,810 SNPs between both STs (core genome size 2,844,379 bp; 0.0136 s/s). SBT data showed just 5 nucleotide differences from ST578 (Fig. 2). However, at the genome level, these two strains differ in more than 21,441 positions (0.0083 s/s).

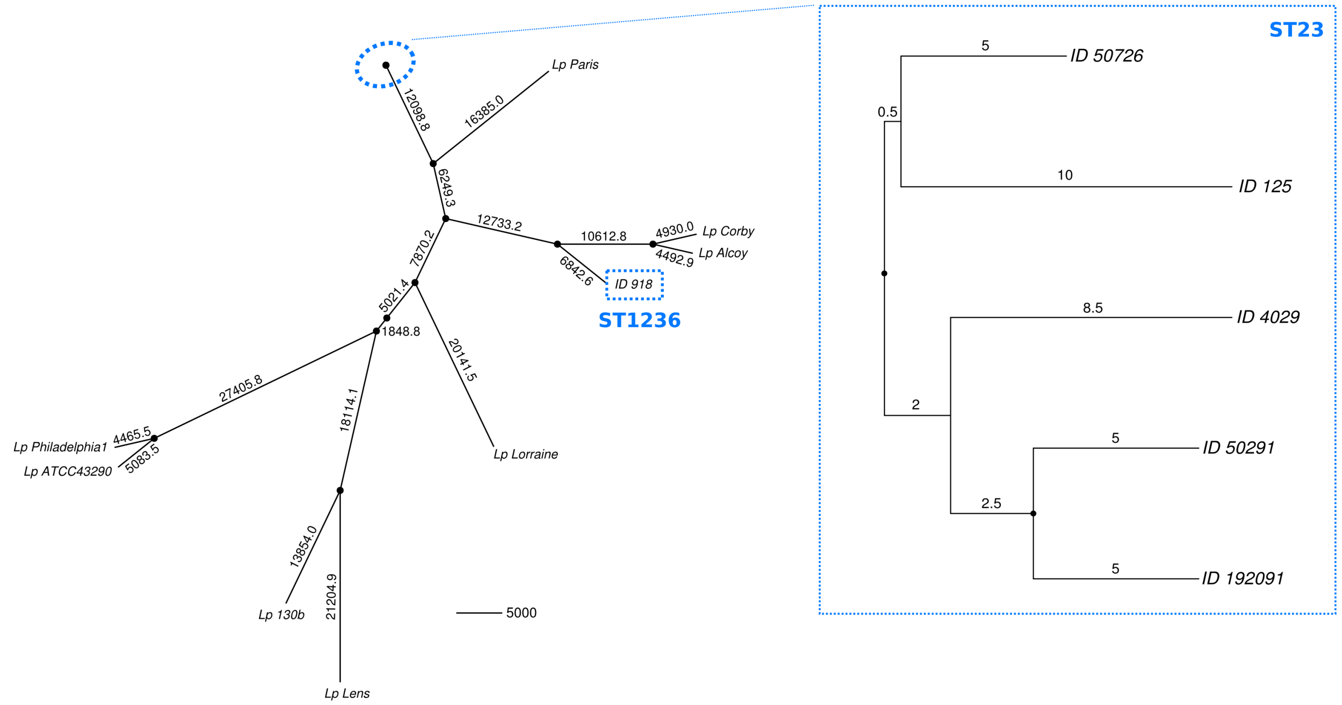


Figure 3 | Minimum-evolution tree constructed using pairwise SNP distances between the sequenced strains and the reference genomes from the databases. Black dots at nodes represent bootstrap supports higher than 90%. Branch labels indicate number of polymorphisms.

Using a Bayesian inference approach, the substitution rate for the five ST23 strains was estimated to be 7.76E-07 s/s/y (95% Highest Posterior Density (HPD) interval 2.32E-06 - 2.09E-08), corresponding to an accumulation of approximately 2.5 nucleotide changes in the core genome per year. Hence, these isolates started to diverge from their most recent common ancestor by approximately year 1994 (95% HPD, 1956-2011), much earlier than the hotel opening (2006).

Discussion

We have described the results of the epidemiological, environmental and genetic investigations of a complex and long-lasting legionellosis outbreak that occurred between November 2011 and July 2012 in a hotel (Calpe, Spain) with 44 cases and 6 fatal outcomes. During the outbreak, different control measures were adopted, which included repeated cleaning and disinfections of the different water systems of the hotel and two temporary hotel closures (Fig. 1). The outbreak was controlled only when the hotel spa pool was permanently closed and completely dismantled. A new spa pool was built with renewed air-conditioning ducts. No new case of legionellosis related to this hotel has appeared since then.

The initial epidemiological investigation revealed a clear link to the hotel that prompted the first control measures. The first signs of the presence of this bacterium in the hotel were obtained by the direct amplification of two loci from biofilm samples in several hotel rooms only ten days after the cleaning and disinfection and led to the first closure of the hotel[107]. Remarkably, the corresponding water samples were negative for *Legionella* by culture. Repeated samples from the same spots revealed that biofilm-containing *Legionella* recolonized most facilities only a few days after their apparently removal by chlorine treatment.

Most of the sequenced clinical isolates and direct samples corresponded to ST23 whereas a profile compatible with ST578 was repeatedly identified in biofilm samples of the domestic water system. The only exception corresponded to a couple (samples 192712 and 192147) who shared the same room and whose incompletely determined ST differed in 5 of the six sequenced loci from ST23 but were compatible with ST578 (Table 1). Additional cloning and sequencing of PCR products from these patients revealed the presence of common strains in both of them, thus confirming the existence of a second infecting strain along the outbreak which, in turn, suggested that the domestic water system was another source of infection. Several recent reports[229–231,293] have shown the presence of mixed *Legionella* infections in outbreak patients. Here, we have shown the presence of more than one strain co-existing in the same location and simultaneously producing infections that led to an outbreak.

After the first cluster of cases, ST23 was isolated for the first time in the spa pool, located at the hotel basement. However, the initial epidemiological investigation failed to find a link between many of the patients infected with *L. pneumophila* ST23 and the spa area. In July 2012, a thorough environmental investigation of the spa structure led to the discovery of multiple hidden cavities with stagnant water under the pool vessel that had connections to the bathing water. Subsequent environmental studies suggested the dissemination of ST23 from the spa pool to the hall area through different ways including the air-conditioning system. The identification of *Legionella* profiles compatible with ST23 in the air filters of the air-handling unit for the hall area confirmed the implication of the air-conditioning system in the dissemination of this strain. This fact contributed to explain the detection of this ST in samples from patients who had not visited the spa. Sequencing of isolates from water samples revealed mainly

ST23, ST1236 and *Legionella micdadei* as colonizers in different parts of the spa pool, a reverse osmosis plant, a drinking fountain at the spa and a pressure vessel of a non-potable water system.

Genomic data can be used to estimate the evolutionary rates of living organisms[286,339,432,433], and even the time at which these organisms could have colonized a specific location[339]. Only three works up to date have used high throughput sequencing to the retrospective study of legionellosis outbreaks[241,293,434]. In the latter case, *L. pneumophila* isolation was only possible from 15/91 patients. In our case, isolation was only retrieved from 4/14 of the clinical samples. The difficulty in obtaining *L. pneumophila* pure cultures from all the samples is reflected in the subsequent poor representation of the outbreak strains that we could use for WGS.

Using Bayesian inference, we estimated the substitution rate of the non-recombinant core genome of the ST23 strains within the hotel as 7.76E-07 s/s/y, higher but in the same order of magnitude than that estimated for the persistent ST578 in the locality of Alcoy (1.39E-07 s/s/y)[339]. The ST23 strains found in the spa were not identical, as would be expected from a single clone colonizing the spa and causing an outbreak. Besides, the time at which ST23 could have started proliferating was estimated to be approximately year 1994. Despite the long HPD intervals from the Bayesian analysis (1956-2011), the finding of 71 polymorphic sites (47 non-homoplastic) showed higher diversity than expected. These results, as well as the finding of other STs and *Legionella* species colonizing the hotel, provide evidence that there was a complex community of *L. pneumophila* inhabiting the area previous to the building of the hotel, including a population of ST23. Thus, the hotel could have been colonized by different STs during its construction and before its opening to the public in 2006. Moreover, there was a precedent of legionellosis cases in the period 2006-

2007 related to the hotel, but no genetic information about the causing strain was retrieved at that moment, so further comparisons between both clusters cannot be performed.

The results derived from the investigation of this outbreak revealed that information from both cultured and uncultured samples collected during the epidemiological investigation of a complex legionellosis outbreak can be essential for its clarification. Moreover, WGS data from an outbreak investigation can help, not only to the epidemiological investigation in order to detect the main reservoir, but it also allows the application of inference tools to trace back the local evolution of the pathogen. The information retrieved from this type of analysis can be important for public health officials to better understand the local behavior of a pathogen such as *L. pneumophila*, capable of colonizing rapidly several facilities in a complex building and remain silently evolving until conditions favor its multiplication and spread, resulting in the infection of exposed, susceptible persons.

Acknowledgements

The authors would like to acknowledge the contributions of all members of the investigation team: R Carbo, M Castellanos, C Marin, E Perez (Surveillance and Epidemiologic Control Unit. Public Health Directorate, Regional Government of Valencia, Spain); J Añó, C del Hierro (Environmental Health Unit. Public Health Directorate, Regional Government of Valencia, Spain); J Miralles, J Miravet, E Rubiales (Environmental Health Unit, Public Health Center, Denia, Spain); R Gonzalez (Epidemiological Unit, Public Health Center, Denia, Spain); J Calafat (Environmental Health Unit, Public Health Center, Alcoi, Spain); C Gonzalez (Microservices, Benidorm, Spain); J del Barrio, B Santandreu (Biolinea Int.); P Santa-Olalla Peralta (Coordinating Centre for Health Alerts and Emergencies, Ministry of Health, Social Services and Equity, Madrid, Spain) and V Drasar (National *Legionella* Reference Laboratory, Public Health Institute Ostrava, Czech Republic). We would also like to thank the members of the visiting ECDC team (B. de Jong, C.L.R. Bartlett, and J. Walker) for their suggestions and recommendations at the closing of the investigation as well as the *Legionella* labs of Public Health England (PHE) and Instituto de Salud Carlos III (Madrid, Spain) for sharing the results of their typing of clinical samples related to this outbreak.

Chapter 5

Recombination drives genome evolution in outbreak-related *Legionella pneumophila* isolates

Leonor Sánchez-Busó[1,2], Iñaki Comas[1,2], Guillermo Jorques[3] & Fernando González-Candelas[1,2]

1. Infection and Public Health Joint Unit FISABIO/University of Valencia-Cavanilles Institute for Biodiversity and Evolutionary Biology, Valencia, Spain.
2. CIBER in Epidemiology and Public Health, Valencia, Spain.
3. Public Health Centre, Alcoy, Spain.

Published in:

Sánchez-Busó L, Comas I, Jorques G, González-Candelas F. Recombination drives genome evolution in outbreak-related *Legionella pneumophila* isolates. Nature Genetics. 2014;46(11):1205–11.

Abstract

Legionella pneumophila is a strictly environmental pathogen and the etiological agent of legionellosis. It is known that non-vertical processes play a major role in the short-term evolution of pathogens but little is known about the relevance of these and other processes in environmental bacteria. We have used next generation sequencing to obtain nearly complete genome sequences of 69 *L. pneumophila* strains linked to recurrent outbreaks in a single location (Alcoy, Spain) along 11 years. Occasionally, genome sequences from isolates of the same sequence type (ST) and outbreak did not cluster together and were more closely related to sequences from different outbreaks. Our analyses reveal that 16 recombination events are responsible for almost 98% of the SNPs detected in their core genome and an apparent acceleration of the evolutionary rate. These results have profound implications for our understanding of microbial populations and for public health interventions in *Legionella* outbreak investigations.

Introduction

Mutations are the ultimate source of genetic variation. However, not all the variants found in an organism directly result from point or structural mutations appearing on it or its immediate ancestors. Bacteria have developed specific strategies to incorporate and exchange genomic and plasmid segments from external sources[435–437]. Processes of non-vertical transmission allow microorganisms to acquire new genes or variants that provide new adaptations such as resistance to environmental stresses and antibiotics or virulence factors at both long and short time scales[438–441]. In addition, they are major determinants of the rate of adaptive evolution, especially under rapidly changing environmental conditions[438]. Recent developments in sequencing and analysis methodologies are providing a wealth of information on the genome sequences of many organisms and the different processes providing variation at this level. However, most of these analyses at the microevolutionary scale have been performed on microbial pathogens[281,283,284,442–444] and no similar information is yet available on the processes occurring outside the human host(s).

Legionella pneumophila is an environmental opportunistic bacterium that inhabits aquatic and soil environments from where it can colonize human-made environments and spread in aerosols infecting susceptible persons. It causes Legionnaires' disease, a potentially fatal form of pneumonia, or Pontiac fever, a milder, flu-like disease[5] and it is not transmissible from person to person. Since its first identification as a potential human pathogen[82], legionellosis outbreaks have been recurrent in many countries[103,117,230,337,415,445,446]. Legionellosis outbreaks are usually studied using the Sequence-based Typing scheme (SBT)[215–218], which provides a reliable genealogical marker in outbreaks caused by clonal strains. However, the genome of *L. pneumophila* has been shown to be highly plastic[268], including recombinant regions, genomic islands, conjugative elements and insertion sequences[244]. A recent study with complete genomes[242] has shown mixed profiles that could result from recombination between strains, thus confirming previous results based on SBT patterns[242,244,267,268].

Until now, most studies applying whole-genome sequencing (WGS) to molecular epidemiology have focused on the investigation of specific outbreaks[241,281–284,433,444], the changes after the introduction of new vaccines[447], or the global spread and transmission of pathogens[257,448], but limited information is available about the population genomics of strictly environmental pathogens such as *Legionella*. The exchange of genomic segments between bacteria could alter the expected clonality in *L. pneumophila* outbreaks and, thus, complicate their investigation. Therefore, there is a need for further studies focused on the genome evolution of strains from different outbreaks in order to have a real measure of how the introduction of new variability can interfere in outbreak investigations.

Here, we present a genomic analysis of *L. pneumophila* strains isolated in 13 different outbreaks occurred in Alcoy (Alicante, Spain) from 1999 to 2010. The peculiar orography of this city, a deep valley almost completely enclosed by the surrounding mountains, is considered to facilitate the maintenance of aerosol emissions within the urban area. After the first cases notified in 1997, more than 200 cases related to legionellosis outbreaks were diagnosed during 1999 and 2000[415]. Since then, the public health authorities in this city, affecting a total of 343 patients, have declared 18 outbreaks. Clinical and environmental strains of *L. pneumophila* were isolated by microbiological methods in 13 of them. A particular *L. pneumophila* variant, identified as sequence type (ST) 578, has been found repeatedly as the causing strain of these outbreaks. However, only in 2009 was it possible to find this ST in an environmental source related to an outbreak[117].

The persistence of *L. pneumophila* in the study site offers a unique opportunity to analyze an environmental, opportunistic pathogen population at the genomic level and, simultaneously, validate WGS as an epidemiological tool. By sampling and sequencing the environmental and clinical diversity of this pathogen in a region where it can be considered endemic over almost 15 years, we are also able of analyzing the factors generating genetic variability in natural populations of this bacterium at a short time-scale.

Materials and Methods

Choice of samples

Public health authorities have identified 18 legionellosis outbreaks from 1999 to 2010 in the locality of Alcoy (Table S1). They affected from a minimum of 3 to a maximum of 97 patients and lasted between 9 and 160 days. In 13 of the outbreaks in which the pathogen could be isolated by microbiological methods, the strain most frequently associated to clinical cases was ST578. Apart from one ST578 strain isolated from an evaporative condenser in 2004, this profile has only been found in an environmental source linked to an outbreak in 2009 (Alcoy 16, Table S1), when it was isolated from an asphalt paving machine and a water-tank truck[117]. The remaining ST578 strains used in this study were isolated from sputums or bronchoalveolar aspirates from legionellosis patients both in outbreak and sporadic cases. Moreover, clinical and environmental ST637 isolates found in the outbreak occurred at the end of year 2000 (Alcoy 3) were also included in our study because of the close relationship between this sequence profile (6, 10, 15, 3, 9, 14, 6) and ST578 (6, 10, 15, 13, 9, 14, 6), differing in just 4 SNPs in the *mip* locus. Additional STs found in environmental sources during outbreak investigations in this locality, either genetically close to ST578 in the SBT scheme, such as ST51 or ST171, or more distant, such as ST1 – a world-wide distributed type of *L. pneumophila*, were also included for analysis (Table S2).

DNA extraction and sequencing

DNA from pure cultures was extracted using a thermal shock procedure by repeating twice a two-step incubation of the bacteria for 5 minutes at 99ºC and 5 minutes at 4ºC. Purification of all DNA samples was performed by ethanol precipitation using 3M sodium acetate (pH 5.2). After an evaporation step using Savant DNA SpeedVac (Thermo Scientific), DNA was resuspended in 1X LowTE (10 mM Tris-HCl and 0.1 mM EDTA, pH 8). Quantity and quality of DNA were tested by triplicate using NanoDrop 1000 (Thermo Scientific) and Quant-iT™ PicoGreen® (Invitrogen). From 1-3 µg of DNA were used for 75 bp single-end sequencing with the SOLiD 5500XL platform (LifeTechnologies, Applied Biosystems, Carlsbad, CA, USA).

Reference mapping, SNP calling and genome alignment

The 69 strains were mapped against the *L. pneumophila* 2300/99 Alcoy[236] reference genome (sequence type ST578) with BWA v0.6.2[449]. Unmapped reads were filtered with prinseq-lite v0.19.5 (see URLs) by removing all reads with a base quality lower or equal to 15 at base 50. Reads passing the filter were assembled using Velvet v1.2.07[450] and VelvetOptimiser.pl v2.2.5. Sequences in contigs were identified using BLAST v2.2.25+[451] versus the NCBI nt database.

Pileup files were generated for each strain using SAMtools v0.1.18 and BCFtools and VCFutils within the same package[452] were used for SNP filtering. The minimum root mean square (RMS) mapping quality for SNPs was set to 25 and the base quality to 20, with a read depth for SNP calling within 5 and 250. Consensus variants were assigned from heterozygous calls evaluated with LoFreq v0.5[453] and were considered for further analyses with the homozygous calls. Positions detected as low-frequency variants were assigned as indeterminations. A non-redundant list of SNPs was retrieved for the 69 strains and used for false-negatives recovery in each of them.

A draft genome was created for each strain by parsing the pileup files created with SAMtools v0.1.18[452] using Perl. Constant positions versus the reference sequence were called as such and the alternative base was called if the polymorphic site had passed the SNP filters described above. A non-mapped position was called as a gap. The resulting draft genomes were reconstructed as collinear to the 2300/99 Alcoy reference sequence. In order to create a proper alignment of the 69 new strains and the nine reference genomes, the latter were downloaded from GenBank (Paris[72] - NC_006368, Lens[72] - NC_006369, Philadelphia1[52] - NC_002942, Alcoy[236] - NC_014125, Corby[235] - NC_009494, Lorraine[271] - NC_018139, 130b[240] - FR687201, ATCC43290 - NC_016811 and HL06041035 - NC_018140) and aligned using progressiveMauve (MAUVE v2.3.1)[274,454]. Then a Perl script was used to convert the extended multifasta (xmfa) output of MAUVE to traditional fasta format by reordering the Locally Collinear Blocks (LCBs) following the coordinates of the 2300/99 Alcoy genome. All positions of the alignment that were not shared with the reference were removed to provide a collinear sequence for each genome to the reference. Finally, the fasta sequences from the 69 strains and the 9 references were joined in the same file to create the final multiple genome alignment.

Phylogenetic inference and genetic variability testing from core genome

The core genome of the multiple genome alignment of the 78 strains was retrieved by removing sites with gaps with trimAl v1.4[455]. The resulting alignment of 2,448,996 bp containing 140,257 SNPs was used for phylogenetic inference by maximum likelihood (ML) using RAxML v7.2.8[345] with the GTRGAMMA model and 1000 bootstrap replicates.

Genetic variability within and between groups of samples (by ST, outbreak and year of isolation) was assessed from the core genome using Variscan v2[456]. The number of polymorphic sites, number of mutations, nucleotide diversity (Pi) and population mutation rate (Theta) were inferred using a sliding window of 1,000 effective sites with an overlap of 25%. SNP distances between and within groups were computed using MEGA v5.0[362] and R[344].

Bayesian phylodynamic analysis of the ST578 coding core genome

Coordinates of the coding sequences of the Alcoy genome (lpa) (NC_014125.ptt file) were downloaded from the NCBI FTP web page (ftp://ftp.ncbi.nih.gov/genomes/Bacteria) and used with the “ape” R package[457] to check for the presence or absence of each lpa gene in each of the 45 ST578 *L. pneumophila* strains and sample ID_3164 (SBT-typed as ST51), which was also included because of its clustering with ST578 strains. Genes covered in at least 80% of their sequences were considered as present. 3,120 out of the 3,190 lpa genes were detected as ST578 core, with only 70 being auxiliary within this ST. R was also used to extract each of the 3120 core genes from the initial alignment using coordinates of the lpa genome and create a concatenate for further analyses.

The existence of temporal information in the data set was tested by regression analysis of the root-to-tip distance over time with Path-O-Gen v1.4 (see URLs). The demographic history of ST578 samples was studied using BEAST v1.8[431]. Different data sets were used, before and after removing recombinant genes (see below). Different demographic (constant population size, exponentially growing population, and Bayesian skyline plot) and molecular clock (strict, random, and uncorrelated lognormal) models were compared by AICM values. For the three demographic models, the uncorrelated molecular clock performed substantially better than the remaining models. The selection of demographic model was based on the comparison of Bayes factors after thermodynamic integration to compute the marginal likelihood of each model using path sampling and stepping stone methods.

The first analysis was performed with the 2,202 polymorphic sites at the 3,120-gene coding core genome of the 46 ST578 strains (including the lpa reference) and sample ID_3164 (ST51). The GTR substitution model, accounting for site heterogeneity with a gamma distribution (4 categories) was used for the inference. A Bayesian piecewise-constant skyline model of coalescence with 10 groups and an uncorrelated lognormal molecular clock were chosen as prior, and 6 parallel chains running for 100,000,000 states were used to check for convergence in this dataset sampling by 10,000 states. A burn-in of 10,000,000 states was applied before checking that all ESS for the model parameters in the combined runs were within 100-200. A second demographic and evolutionary analysis was performed after removal of the genes detected with non-vertical signal of evolution (recombinant genes). The concatenated alignment of the 45 SNPs and 2,607,219 non-polymorphic sites from the remaining 2,688 core genes was also used for Bayesian inference as described above but using the HKY model of nucleotide substitution instead of the GTR+G.

Recombination testing

From the alignment of the ST578 strains and the Ipa reference, all polymorphic sites were extracted using the *seg.sites* function from the “ape” R package[457]. Using the coordinates for all 3120-core genes, the total number of SNPs per gene was assigned. Mesquite v2.74 (see URLs) was used to map the 2,202 SNPs of the 3,120-gene coding core over the corresponding Bayesian tree reconstruction and to extract the ancestral sequence at each node. Perl and R scripts were applied to process the output and make the assignment of the number of SNPs per gene between each pair of nodes in the tree. The observed number of polymorphisms in each gene in the core genome was compared to a theoretical expectation to detect potential signals of recombination or horizontal transfer. Using the nomenclature in Cui *et al*[68], the probability of observing a polymorphism at a certain location was calculated as $m=n/N$, where n was the total number of SNPs, 2,202 bp, and N was the sum of the gene lengths of the 3,120 genes in the coding core genome, 3,012,426 bp. The expected number of SNPs in each gene i was computed as $m \cdot L_i$, being L the length of the gene.

A phylogenetic approach was used to detect signals of recombination. First, genes for which the observed number of SNPs exceeded in more than four times the expected number were used for a first round of recombination testing. The alignment of each of the 149 detected genes was extracted from the concatenate of all the strains and the 9 reference genomes using R and a ML tree was reconstructed for each of them using RAxML v7.2.8[345] with the GTRGAMMA model. The alignment and the tree for each gene were used for topology testing against the concatenate tree of the 3,120 genes in the core using TREE-PUZZLE v5.2[371] with the GTR model, 16 categories (maximum) for the gamma distribution of among-sites heterogeneity in substitution rates and the frequency of change between each pair of nucleotides extracted from the RAxML output. Although we were focused on ST578, external STs were also introduced in this test in order to check for transfer events between profiles. The alignment for each gene was checked using Perl before running TREE-PUZZLE to remove empty sequences (absence of the gene in specific strains) and, for comparison purposes, the concatenate tree was modified by removing the tips not present in the gene tree using the *drop.tip* function in the “ape” R package[457]. A command file was created for each gene and TREE-PUZZLE was run in batch mode. TREE-PUZZLE output files were parsed with Perl to retrieve the p-values given by the Shimodaira-Hasegawa (SH) and Expected Likelihood Weight (ELW) tests for both the gene and the concatenate trees. Genes that rejected the concatenate tree (p-value < 0.05 in both tests) were considered to have a different phylogenetic history than the core, and thus were considered as potentially recombinant. A second round of identification of genes with a non-vertical signal was performed by re-calculating the expected number of SNPs per gene considering the 2,971 genes left after the first search.

In order to detect recombinant segments, the genomic location and the phylogenetic trees of the 247 genes that rejected the concatenated tree after the two searching rounds were manually inspected. Isolated genes with <= 2 SNPs were not considered, whereas consecutive genes that were detected as significant by TREE-PUZZLE and whose trees were consistent with the genetic exchange with other STs within our dataset were considered a recombinant segment. 185 extra genes present within those segments but not detected as recombinant by the phylogenetic approach described above were also included. Results were compared to the recombination detection method described in Croucher *et al*[443].

Results

We obtained nearly complete genome sequences for 69 *L. pneumophila* strains using SOLiD 5500XL. These strains were obtained from clinical and environmental sources in the investigation of outbreak and sporadic cases in Alcoy between 1999 and 2010 (Tables S1 and S2). Most of the isolates (n=45) corresponded to ST578, the same profile as the reference strain (*L. pneumophila* str. 2300/99 Alcoy, NC_014125)[236]. The average number of reads retrieved per strain was 4,700,318 (range [1,872,888-42,874,666]) and the mapping step yielded an average per base coverage of 88.2X and an average mapping percentage of the positions of the reference genome of 96.2% (Table S2).

Overlapping diversity between outbreak and non-outbreak strains

Firstly, we explored the inter- and intra-ST SNP distances at the genomic level for the three most abundant STs (ST578, ST1, and ST637). For all cases, inter-subtype distances were substantially larger and practically did not overlap with intra-subtype differences (Fig. S1A). In general, the distance computed from the seven SBT loci is a good predictor of the inter-ST distance from complete genome sequences (Pearson's r = 0.90, p-value < 2.2E-16, Fig. S2).

At the intra-ST level we noticed that the average number of polymorphic sites in the whole genome was highly variable among STs. However, this was not related to sample size, with the least abundant of the three STs considered, ST1 and ST637, presenting an average of 2,294 and 1,626 SNPs in pairwise comparisons, whereas ST578, the most abundant in this dataset, presented an average of 424 SNPs between pairs of isolates (Fig. S1B and S3).

Next, we investigated whether this high intra-ST variability was also observed between strains assigned to the same outbreak, which are expected to be almost identical genetically. The analysis of ST578 revealed no significant differences in the genomic distance within and between outbreak strains (Wilcoxon rank-sum test, p-value = 0.54). The average number of SNPs in the core genome between ST578 isolates from different outbreaks was 594 (range 5-1,807) whereas the average number of SNPs between isolates from the same outbreak was 652.5 (range 6-1,760) (Fig. 1).

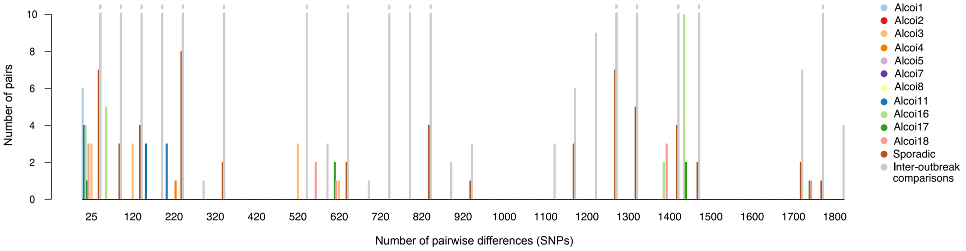


Figure 1 | Distribution of the number of pairwise SNP distances in the core genome between and within *L. pneumophila* strains from 13 different outbreaks occurred in Alcoy (Spain) between 1999 and 2010. Inter-outbreak distances are shown in grey and intra-outbreak comparisons are represented in different colours (see legend).

The overlap in diversity between outbreak and non-outbreak strains was also reflected in the phylogenetic relationships of these strains (Fig. 2). The evolutionary relationships of isolates within ST578 revealed two interesting observations. Firstly, isolates from the same outbreak do not define monophyletic groups and are even found quite distantly placed within the ST578 cluster, thus explaining the previously noted overlapping diversity between outbreak and non-outbreak strains. Secondly, clinical and environmental isolates of this ST clustered in two well-defined, highly supported clades, subsequently denoted as sublineages A and B (Fig. 2).

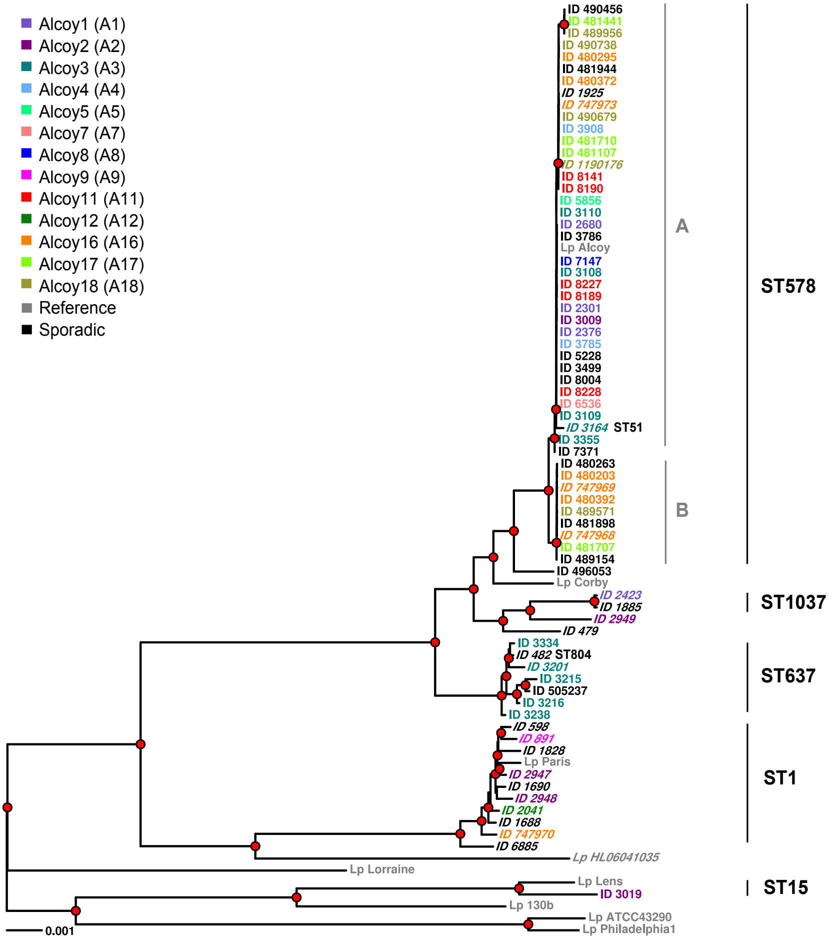


Figure 2 | Maximum likelihood phylogenetic reconstruction of the core genome (2,448,996 bp) of 69 strains of *L. pneumophila* and 9 reference genomes. Tip colours represent isolates either related epidemiologically to different outbreaks or sporadic cases (see legend). Environmental isolates are shown in italics and red dots at nodes indicate bootstrap support values higher than 90%.

Interestingly, strains isolated in outbreaks from 1999 to 2004 clustered only in sublineage A, whereas isolates from posterior outbreaks (2009 and 2010) clustered in both sublineages, which is reflected as an increase in the genetic variability of the strains in the later outbreaks (Fig. S4A) and, in consequence, over time (Fig. S4B). These results indicate that the population dynamics of this ST has been changing in this location over the time period considered. The fact that clinical strains from the same outbreak clustered in two different sublineages within ST578 suggests that the current SBT has not enough resolution to firmly establish the source of a *Legionella* outbreak when there is a substantial amount of underlying environmental diversity.

Recombination accounts for most of the observed diversity

To study the phylodynamics of *L. pneumophila* ST578 strains in Alcoy, we used a Bayesian analysis to fit a demographic model and calibrate a molecular clock (Fig. 3 and Table S3) using the isolation dates as prior information (Fig. S5A). For this, we firstly analyzed the core coding genome sequence (3,120 genes, 3,046,296 bp including 2,202 SNPs) of all the available ST578 strains (41 clinical, including the reference strain 2300/99 Alcoy, and 5 environmental isolates). We also included isolate ID_3164, previously typed as ST51 but included in the ST578 cluster (Fig. 2). The comparison of different combinations of demographic and molecular clock models led us to select a relaxed molecular clock (uncorrelated lognormal) and a Bayesian Skyline-plot demographic model as the best combination to analyze these data (Table S3). The maximum clade credibility trees shown in Fig. 3 were obtained after checking the convergence of six independent runs and achieving ESS > 200 for all the parameter estimates in the model. The mean evolutionary rate at the core genome of ST578 was estimated to be of 8.02E-6 (95% HPD interval 3.69E6 - 1.33E-5) nucleotide substitutions per site per year (s/s/y), accounting for 24.43 changes per year. Similar estimates of the evolutionary rate were obtained with the multiple alignment of the core genome including all the invariant sites (mean = 6.41E-6, 95% HPD interval 3.66E-6 - 1.03E-5 s/s/y).

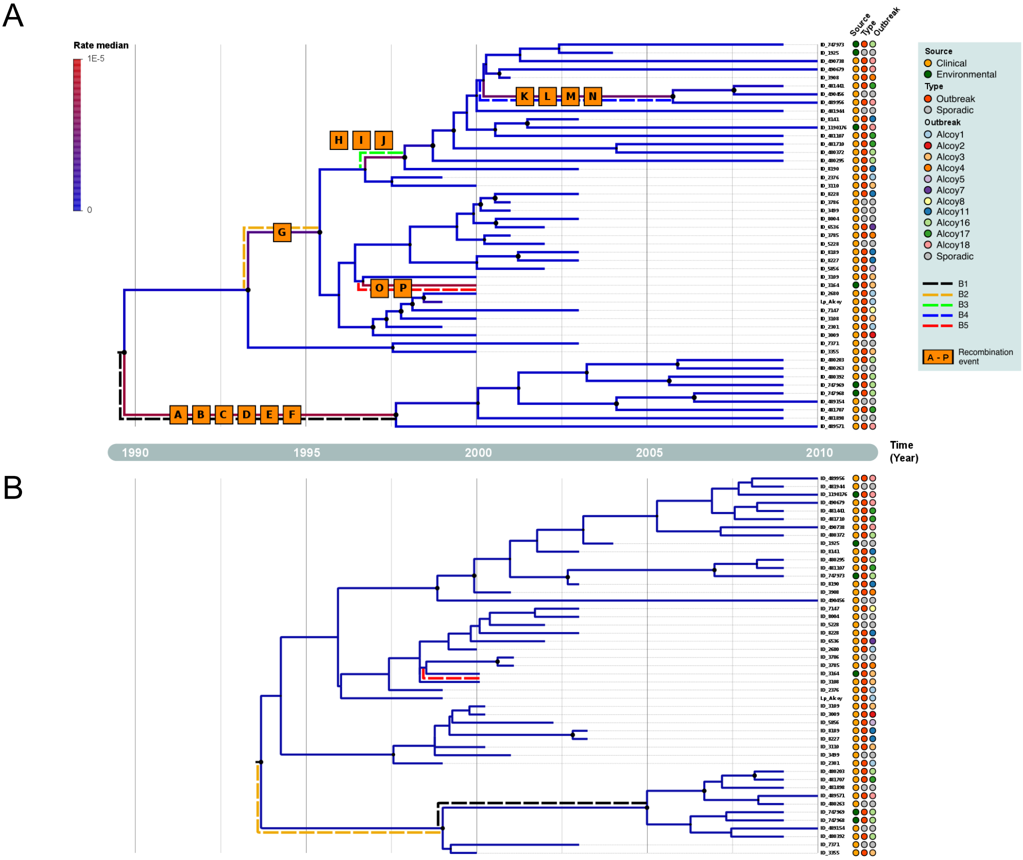


Figure 3 | (A) Bayesian Maximum Clade Credibility tree estimated using BEAST (Bayesian skyline demographic model and uncorrelated lognormal molecular clock) with prior temporal information (isolation dates) from the 2,202 SNPs in the coding core genome of ST578 strains (3120 genes). The colour gradient in the branches represents the median substitution rates, orange squares are recombination events and coloured dots represent the associated metadata (see legend). (B) The same tree obtained from the core genome (2,607,264 bp, 45 SNPs, 2,688 genes) after removal of SNPs and genes involved in the detected recombination events.

However, long terminal branches were found for most isolates as well as a wide range of substitution rates per branch (1.04E-7 – 5.46E-5 s/s/y), especially in five of them (Fig. 3 and Fig. S6). Polymorphisms at the core genes were mapped in these five branches revealing a non-random distribution of SNP densities (Fig. 4), with polymorphisms accruing in clusters of nearby genes, thus suggesting recombination or horizontal gene transfer. To test for non-vertical signals and to estimate the possible effect of this evolutionary force in *L. pneumophila* ST578 isolates, the expected number of SNPs for all core genes, assuming a random distribution along the genome, was calculated and compared to the observed ones (Fig. S7). 175 genes exceeded in more than four times the expected values (Table S4) and were subjected to a topology testing analysis comparing each gene tree with that obtained from the concatenated alignment of the 3120 core genes.

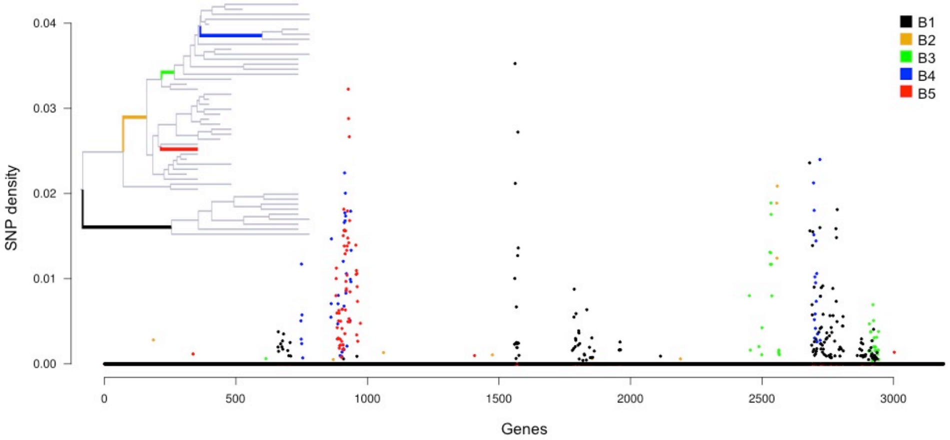


Figure 4 | SNP density per core gene in the five highlighted branches of the tree. Colours represent the different branches.

The 247 genes with evidence of recombination were removed from the genome alignment, as well as 185 extra *loci* that were interspersed among these genes and which contained 125 SNPs, 95 of them on branch B1 (Fig. 3), thus leaving a core of 2,688 non-recombining genes (2,607,264 bp, 45 SNPs). BEAST was re-run with the same prior parameters, as well as the tip dating information (Fig. S5B), and the mean evolutionary rate for the non-recombining core was estimated to be 1.39E-7 s/s/y (95% HPD interval = 5.41E-8 - 2.30E-7), with a per branch range of 4.29E-8 – 1.88E-7 s/s/y (Fig.

3 and S6). These estimates support an accumulation of 0.39 polymorphisms per year in the core genome of *L. pneumophila* ST578 merely due to nucleotide substitutions (2.04 % of the total variability found), with almost 98% of the total variability resulting from non-vertical inheritance. After removing the mentioned 432 genes involved in recombinant regions, the 45 SNPs left were located in 42 *loci,* two of which (lpa_02547 and lpa_03967) had 3 and 2 SNPs respectively, and accumulated on three of the previously studied branches (Fig. S9 and Table S5).

As the distribution of SNPs in the core genome of *L. pneumophila* ST578 was found to be highly concentrated at some points (Fig. 4), the number of SNPs per gene in the core was obtained on a per branch basis (Fig. S10). Results confirmed that many polymorphisms accumulated in consecutive genes in the five branches discussed above, most of them in short periods of time, and could be grouped in 16 recombination events (Fig. 3 and Table S6). These events include genes involved in housekeeping functions, such as cell membrane biogenesis (COG category M) or metabolic reactions (i.e. COG category C) but also in others that could be related to virulence, such as defense mechanisms (COG category V) (Fig. S11). Specifically, the branch that separates ST578A and ST578B (B1), presented 6 independent potential recombination events (A-F) with lengths ranging from 6.9 to 141 kb and which included 899 of the 939 SNPs (95.74%) found to have accumulated in this branch.

According to our previous dating for the colonization of ST578 in the Alcoy area, these events should have occurred during approximately 15 years, and included the incorporation into lineage ST578B of SNPs in genes with many diverse functions, including genes involved in virulence, DNA damage repair and mutagenesis (Fig. S11). Interestingly, the branches leading to the most recent outbreak strains (2009 and 2010) show convergence for genes involved in two recombination events (F and J, Fig. 3). Both events included a set of 14 genes encoding different subunits of the NADH dehydrogenase type I.

Phylogenetic analyses show that the donor strains for most of the genes involved in the recombination events clustered with different STs of the main clade of the ML tree containing profiles close but external to ST578 or ST637 (Fig. S8). The recombinant genes and regions inferred here were compared to those detected by the method described in Croucher *et al*[443]. With this method, the final number of SNPs retained for the non-recombining core genome was 84. We compared the coordinates of both sets of SNPs and 31 of them were coincident for both methods (Table S7). The 14 SNPs unique to our method correspond to genes that were not detected as statistically significant by the ELW and SH methods or were not present within the 16 recognized recombinant regions. The 53 SNPs unique to Croucher's method corresponded to mutations in genes with a non-significant change in topology and not detected as recombinant in all their length, so only a partial region was removed. The 16 recombination regions were also detected by this method along with other 5 fragments (Table S7) that were not removed by our pipeline because they were detected as isolated, with non-evident accumulation of SNPs over specific branches and with no phylogenetic incongruence of the corresponding gene trees versus the core genome tree. Thus, the method described in this paper is more conservative than that in Croucher *et al*[443], as it also incorporates a phylogenetic approach.

Genomic epidemiology of ST578 outbreaks in Alcoy

According to Bayesian analyses of the non-recombining core genome, ST578 colonized Alcoy around 1990 (95% highest posterior density (HPD) intervals ranging [1973.2-1998.6]) (Fig. 3 and Fig. S12). From the initial colonizers, ST578A started to expand a few years later, in agreement with the explosion of clinical cases in 1999. From the complete core genome, ST578B, which only includes cases of the two outbreaks occurred in 2009 and 2010, was estimated to have started diverging from its MRCA approximately in 2005 (95% HPD interval 2002.1-2007.5) (Fig. 3). The remaining clinical and environmental samples from outbreaks occurred in 2009 and 2010 clustered in clades derived from the ST578A lineage and their MRCA was estimated to have emerged around 1991 (95% HPD interval 1986.7 - 1998.3).

The diversity of *Legionella* in Alcoy also reflects the effectiveness of control measures implemented by public health authorities and roughly correlates with the number of cases, as shown by the results from the Bayesian skyline analysis (Fig. 5). Although stringent control measures were adopted since the first legionellosis outbreaks (Table S1), it was not until 2006-2008 that most high-risk facilities were completely removed from the urban area. This was reflected in a decrease in the number of cases during that period along with a decrease of diversity mainly in strains of sublineage A. However, almost simultaneously, a sharp increase in bacterial diversity was observed (Fig. 5). A previous study[117] revealed that an asphalt paving machine had been responsible for the 2009 outbreak (Alcoy16, Table S1). This machine used a water source colonized by ST578 located outside the city limits and not subjected to the usual control measures. Hence, we hypothesize that the original bacterial population (designated ST578A) was largely eliminated during the removal of urban risk facilities and that the increase observed in the skyline plot was due to the introduction of new colonizers from alternative, uncontrolled sources outside the city. The skyline plot clearly shows that new diversity was introduced in Alcoy around

2005 and, since then, the corresponding strains have further diversified probably outside the city limits and in unconventional facilities, thus escaping from the control measures adopted. This diversification is strongly correlated with an increase in the number of cases, from almost none in 2008 to about 50 in 2009, also coincident with the peak of maximum diversity observed in the skyline (Fig. 5).

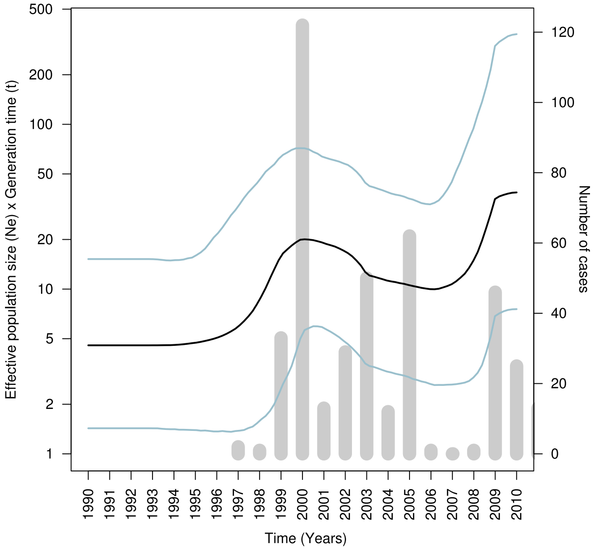


Figure 5 | Demographic dynamics of ST578 strains. The Bayesian skyline shows the demographic changes as the effective population size (Ne) per generation time (t). The grey bar plot represents the total number of clinical cases of legionellosis in 1997-2010 in the Alcoy area.

Discussion

We have performed a whole-genome sequence analysis of *L. pneumophila* strains isolated in a single locality throughout a period in which they have caused 18 outbreaks, with 343 infected patients, and several sporadic cases. From a public health perspective, our aim was to assess the utility of WGS in outbreak investigations in the case of a strictly environmental pathogen for which human infection is a dead-end. From an evolutionary perspective, we wanted to measure the impact of different evolutionary forces in shaping the genetic diversity in a population of this environmental pathogen during a short period of time.

Whole genome epidemiology of the 13 outbreaks and sporadic cases analyzed has revealed a non-clonal relationship among isolates epidemiologically linked to the same outbreak, with overlapping distances between and within outbreaks (Fig. 1) and the occurrence of at least 16 recombination events in five different branches during the short-term evolution of the ST578 strains sampled in Alcoy (Spain) from 1999 to 2010 (Fig. 3 and Table S6). These results differ markedly from other recent genomic epidemiology studies in which WGS has been applied to analyze specific outbreaks or the global tracking of person-to-person transmitted pathogens[283,284,444]. However, there are very few studies on environmental pathogens and they correspond to organisms that have different natural history characteristics than *L. pneumophila*, such as *Mycobacterium abscessus*[286] or *Vibrio cholerae*[442]. Most such analyses reveal an almost complete genomic identity among isolates from the same outbreak, whereas our analyses of legionellosis outbreaks indicate that they may be caused by different strains infecting susceptible people simultaneously when environmental conditions are appropriate.

The genetic diversity of *L. pneumophila* ST578 in Alcoy has been increasing since the first strains were isolated in 1999. From the initial colonizers, ST578 started diverging approximately in 1990 in two sublineages, A and B. The persistence and evolution of *Legionella* in Alcoy could be explained by the permanent colonization of difficult-to-reach spots from where the bacteria cannot be removed as well as by migration from external sources[117]. This indicates that an outbreak can be originated from a single physical source and not necessarily involving recently diverged (i.e., clonal) strains of *L. pneumophila*. At the same time, the analysis of bacterial diversity over time seems to be a good marker of the effectiveness of control measures (Fig. 5), as it can reveal the existence of uncontrolled diversity that can lead to new clinical cases[117].

Several studies[242,244,268] have described the plasticity of the *Legionella* genome, in which recombination, horizontal gene transfer and convergent evolution play a highly relevant role[244]. These events may have an impact on the SBT approach[242,267], with potential consequences in outbreak investigations because they may create paraphyletic groups. One example is described in this work, with a ST578 isolate assigned to ST51 (strain ID_3164) because of the transfer of a *mip* allele from a different group (Fig. 2).

Rapid adaptation of bacterial populations can result from an increased availability of genetic variation. The presence of hyper-recombination in bacterial populations has been described sporadically, entailing important consequences[458–460]. For instance, *Streptococcus pneumoniae* has adapted rapidly to changing environmental conditions through genomic changes brought about by recombination events[443]. Here, we present evidence that during a short period of time, about 20 years, recombination has played a major role in introducing genetic variation in *L. pneumophila* populations, leading to a higher than 10-fold difference in the substitution rates compared to those obtained only from mutation. The dead-end nature of *Legionella* infection implies that these recombination events have occurred in the environment and are not linked to increased selective pressures in the infected patients.

The magnitude of recombination in the *L. pneumophila* ST578 population of Alcoy can be highlighted by comparing the observed number of SNPs introduced in the core genome by 16 recombination events (2,157/2,202, 97.96%) during approximately 20 years with the equivalent figures in a world-wide sample of 240 *S. pneumoniae* isolates (50,720/57,736, 88% from 702 events)[443]. Previous analyses using MLST data show a wide range of estimates for the ratio of probabilities that a base change occurs by recombination or mutation (per site r/m rate) over different bacteria and archaea, from 0.02 in *Leptospira interrogans*[461] to 63.6 in *Flavobacterium psychrophilum*[270,462]. In the case of *L. pneumophila*, a previous study using SBT data showed a recombination rate of 0.9[267] and a recent work showed an estimate of 16.8 for the *L. pneumophila* SBT database[242]. However, the genomic data from the present study, which is the first one up to our knowledge showing a genomic estimate of the recombination rate for a single population of *L. pneumophila*, suggest a higher per site r/m = 47.93, calculated using the number of SNPs introduced by recombination and mutation (2,157/45). Croucher *et al*[443] also detected an important difference in the per site r/m estimate of *S. pneumoniae* using genomic data (r/m = 7.2) from a previous study focused on MLST data (r/m = 23.1)[463].

The observed recombination events in the *L. pneumophila* ST578 population spanned an average of 35.7 kb, with a maximum size of 141 kb. Large recombination events have been described in other bacteria, such as *Streptococcus agalactie*, with exchanged regions detected between 21 to >300 kb[464], as well as in *Clostridium difficile*, with imports ranging from 9 to 170 kb[465]. The authors of these studies point at chromosomal mobilization as a consequence of integrated mobile elements as the mechanism for these large replacements. We have found several regions usually associated to genome mobility in these recombining regions (Table S6) but we cannot postulate any specific mechanism. It is not possible yet to evaluate how common is recombination in other *L. pneumophila* populations but this phenomenon deserves further attention given its potential public health importance and impact on the biology of the bacteria. Within these recombinant regions, we have observed convergent events in different branches of the ST578 phylogeny affecting *nuoG*, a host cell apoptosis inhibitor in *M. tuberculosis*[466–468]. Experimental data will be needed to show whether these acquisitions provide a selective advantage during replication in macrophages. If this is the case, strains harboring these variants should be actively screened to prevent future outbreaks in the locality.

The direct inference of recent recombination events overcomes many of the difficulties usually encountered when analyzing bacterial populations in search for genetic exchanges. Other evolutionary processes, such as selection and genetic drift, might act after such events obscuring or completely removing their presence in a given population. In the case of *L. pneumophila*, we have shown that accelerated evolution in the environment is linked to recombination events. If the relative contribution of recombination in the generation of genetic variation in a population revealed in these analyses is confirmed for other bacterial populations and species, we must conclude that the adaptive and evolutionary relevance of recombination for bacteria is even higher than previously appreciated[469]. This can have profound implications for our general understanding of bacterial populations and public health investigations.

URLs

Software can be found at the following Web pages:
Path-O-Gen, http://tree.bio.ed.ac.uk/software/pathogen/
VelvetOptimiser, http://bioinformatics.net.au/software.velvetoptimiser.shtml
prinseq, http://prinseq.sourceforge.net/
Mesquite, http://mesquiteproject.org/mesquite

Accession codes

SOLiD genetic data have been deposited in the European Nucleotide Archive (ENA) under the project accession PRJEB5990 (http://www.ebi.ac.uk/ena/data/view/PRJEB5990).

Acknowledgements

The authors acknowledge the Dirección General de Salud Pública (DGSP, Conselleria de Sanidad, Generalitat Valenciana) and the Centro de Salud Pública de Alcoy for sharing information and comments on the *Legionella* outbreaks in Alcoy and facilitating access to epidemiological data. We also thank the Microbiology Service at the Hospital Virgen de los Lírios (Alcoy) and the Laboratorio de Salud Pública for sharing the *Legionella* samples used in this study, as well as the Genomics section of the SCSIE (University of Valencia) for their technical help in the deep sequencing of strains. We would also like to thank J. Parkhill and S. Harris (Wellcome Trust Sanger Institute, Hinxton, UK) for their comments as well as facilitating their computational resources during the revision of the manuscript. The work has been funded by project BFU2011-24112 from Ministerio de Economía y Competitividad (MINECO), Spanish Government. LSB is recipient of a FPU predoctoral fellowship from Ministerio de Educación, Spanish Government. IC was supported by an IEF (272086) from FP7 and Ramón y Cajal from the Spanish Government (RYC-2012-10627).

Supplementary material

Table S1 | Alcoy outbreaks. Detailed clinical and environmental information about the outbreaks taken place in Alcoy in the period 1999-2010.

Outbreak information			**Clinical information**									**Environmental information**
ID	**Dates (from first to last case)**	**Days**	**Number of cases**	**Male/ female ratio**	**Average age**	**Age range**	**Risk factors**	**Hospitali-zation rate (%)**	**Lethality (%)**	**Positive urine antigen**	**Number of isolates**	
ALCOY 1	September 1999 - February 2000	160	36	2.27	61.8	18 - 92	44% smokers; 50% chronic diseases	100	8.30	28	6	**Source:** Unknown. *L. pneumophila* isolated in cooling towers in the same municipality, but of different strain to those in clinical cases. **Control measures:** Disinfection of cooling towers positive for *L. pneumophila* **Chlorine level in distal points:** 0.5 ppm.
ALCOY 2	April - July 2000	113	11	2.67	63.3	38 - 86	45% smokers; 27.3% chronic diseases	63.60	0	11	2	**Source:** Unknown. *L. pneumophila* was isolated in three refrigeration towers but the strain was different to the clinical isolates. **Control measures:** Cleaning and disinfection of *L. pneumophila*-positive refrigeration towers. **Chlorine level in distal points:** 0.5 ppm.

ALCOY 3	September - December 2000	76	97	1.8	66.3	22 - 95	27.8% smokers; 50.5% chronic diseases	86.46	7.20	97	10	**Source:** Unknown. Four refrigeration towers were positive by culture. **Control measures:** Searching and control of risk facilities: 11 refrigeration towers were closed, 9 sealed and 2 under protocoled working. **Chlorine level in distal points:** 1.5-2 ppm.
ALCOY 4	May - June 2001	12	5	4	57.2	35 - 85	80% chronic diseases	100	20	5	2	**Source:** Unknown. All 194 environmental samples were negative. **Control measures:** Active searching of risk facilities: 63 found. **Chlorine level in distal points:** 1.5 ppm.
ALCOY 5	July - August 2002	13	9	8	56.2	32 - 83	66.6% smokers; 77.7% chronic diseases	100	0	9	1	**Source:** Unknown. Suspected saturation central and evaporative condenser in the urban area that were positive by PCR. **Control measures:** Cleaning and disinfection of the risk facilities positive for *L. pneumophila*. **Chlorine level in distal points:** 1.5 ppm.
ALCOY 6	October - November 2002	12	5	All males	62.8	44 - 82	20% smokers; 40% chronic diseases	80	0	5	0	**Source:** Unknown. Suspected humidifier central in the urban area that was positive by PCR. **Control measures:** Cleaning and disinfection of the risk facility positive for *L. pneumophila*. **Chlorine level in distal points:** 1.5 ppm.

ALCOY 7	November - December 2002	24	12	5	66.0	30 - 88	25% smokers; 25% chronic diseases	91.70	0	12	0	**Source:** Unknown. Four suspected evaporative condensers in the urban area positive by PCR. **Control measures:** Cleaning and disinfection of the risk facilities positive for *L. pneumophila*. **Chlorine level in distal points:** 1.5 ppm.
ALCOY 8	April - May 2003	32	11	2	66.0	38 - 84	36.36% smokers; 81% chronic diseases	100	0	11	1	**Source:** Unknown. Suspected evaporative condenser in the urban area that was positive by PCR. **Control measures:** Cleaning and disinfection of the risk facility positive for *L. pneumophila*. **Chlorine level in distal points:** 1.5 ppm.
ALCOY 9	July 2003	14	4	All males	72.2	47 - 86	50% smokers; 75% chronic diseases	100	0	4	0	**Source:** Unknown. Negative results by culture and PCR from the 50 risk facilities sampled during the environmental investigation. **Control measures:** Cleaning and disinfection of one risk facility with poor maintenance. Sealing of another inaccessible risk installation. **Chlorine level in distal points:** 1.5 ppm.

ALCOY 10	September - October 2003	9	3	All males	65.3	54 - 77	33.3% smokers; 66.6% chronic diseases	100	0	3	1	**Source:** Unknown. Suspected evaporative condenser positive by PCR. One culture was retrieved from a refrigeration tower and the serogroup was different to that of the clinical isolates. **Control measures:** Cleaning and disinfection of the risk facilities positive for *L. pneumophila*. **Chlorine level in distal points:** 1.5 ppm.
ALCOY 11	October - November 2003	30	31	2.1	67.5	41 - 89	31% smokers; 59% chronic diseases	100	9.70	31	6	**Source:** Unknown. Two suspected evaporative condensers in the urban area were positive by PCR. **Control measures:** Cleaning and disinfection of the risk facilities positive for *L. pneumophila*. **Chlorine level in distal points:** 1.5 ppm.
ALCOY 12	October - November 2004	50	12	1.4	76.2	61 - 90	17% smokers; 91% chronic diseases	100	25	12	2	**Source:** Unknown. Two saturation centrals and 3 evaporative condensers were positive in the urban area by PCR. **Control measures:** Cleaning and disinfection of the risk facilities positive for *L. pneumophila*. **Chlorine level in distal points:** 1.5 ppm.

ALCOY 13	June 2005	10	12	1	69.0	36 - 89	33% smokers; 65% chronic diseases	92	0	12	3	**Source:** Unknown. One suspected saturation central and 5 evaporative condensers in the urban area by PCR. **Control measures:** Cleaning and disinfection of the risk facilities positive for *L. pneumophila*. **Chlorine level in distal points:** 1.5 ppm.
ALCOY 14	June - July 2005	26	27	2	66.0	39 - 93	59.26% smokers; 44.4% chronic diseases	92	0	27	6	**Source:** Unknown. Suspected illegal saturation central in the urban area positive by PCR. **Control measures:** Closing and sealing of the illegal installation. **Chlorine level in distal points:** 1.5 ppm.
ALCOY 15	September - October 2005	25	19	1.1	66.0	35 - 93	37% smokers; 52.6% chronic diseases	86	0	19	2	**Source:** Unknown. Two suspected evaporative condensers positive by PCR. **Control measures:** Cleaning and disinfection of the risk facilities positive for *L. pneumophila*. **Chlorine level in distal points:** 1.5 ppm.

ALCOY 16	July - August 2009	38	11	4.5	70.0	49 - 88	45.5% smokers; 81.8% chronic diseases	100	9	11	2	**Source:** Confirmed source. Isolation *of L. pneumophila* from an auxiliary paving machine in the urban area. **Control measures:** Paving procedures in the urban area of Alcoy were protocoled. **Chlorine level in distal points:** 1-1.5 ppm.
ALCOY 17	September - October 2009	18	22	2.14	70.1	46 - 86	13.6% smokers; 59.1% chronic diseases	86	0	22	4	**Source:** Unknown. An auxiliary paving machine in the urban area was positive by PCR. **Control measures:** Paving procedures in the urban area of Alcoy were protocoled. **Chlorine level in distal points:** 1-1.5 ppm.
ALCOY 18	May - July 2010	38	16	2.2	64.6	38 - 85	68.7% smokers; 50% chronic diseases	75	0	16	6	**Source:** Unknown. Two car washing machines in the urban area were positive by PCR. **Control measures:** Protocol and control of car washing machines in the urban area. **Chlorine level in distal points:** 1-1.5 ppm.

Table S2 | Strains used in the study with geographical and temporal data. Sequencing and mapping statistics are also shown.

	Strain information					Sequencing statistics		Mapping statistics	
Strain	**ST**	**C/E**	**Year**	**Outbreak**	**Sampling point/Patient**	**Total number reads**	**Sequencing depth***	**Mapped reads**	**Per-base reference* coverage (%)**
ID_2301	578	C	1999	Alcoy1	Patient	6,820,701	145.48	3,486,056	99.85
ID_2376	578	C	1999	Alcoy1	Patient	2,912,596	62.12	1,479,667	99.70
ID_2423	1037	E	1999	Alcoy1	Food industry	3,799,441	81.04	1,778,657	90.66
ID_2680	578	C	2000	Alcoy1	Patient	3,290,621	70.19	1,776,035	99.78
ID_2947	1	E	2000	Alcoy2	Textile industry	5,339,318	113.88	2,083,691	85.04
ID_2948	1	E	2000	Alcoy2	Rubber factory	4,960,979	105.81	1,887,559	85.94
ID_2949	328	E	2000	Alcoy2	Food industry	4,995,195	106.54	2,390,175	92.13
ID_3019	15	C	2000	Alcoy2	Patient	3,630,880	77.44	1,180,121	79.31
ID_3009	578	C	2000	Alcoy2	Patient	4,156,412	88.65	2,272,409	99.82
ID_3164	51	E	2000	Alcoy3	Textile factory	3,753,666	80.06	1,992,072	99.77
ID_3108	578	C	2000	Alcoy3	Patient	3,370,906	71.90	1,745,193	99.77
ID_3109	578	C	2000	Alcoy3	Patient	3,731,005	79.58	1,859,651	99.77
ID_3110	578	C	2000	Alcoy3	Patient	3,326,900	70.96	1,708,208	99.71
ID_3355	578	C	2000	Alcoy3	Patient	3,502,645	74.71	1,750,520	99.72
ID_3201	637	E	2000	Alcoy3	Coffee-shop	5,885,756	125.54	2,793,448	94.25
ID_3215	637	C	2000	Alcoy3	Patient	4,377,547	93.37	1,998,822	94.24
ID_3216	637	C	2000	Alcoy3	Patient	4,373,652	93.29	1,937,986	94.16
ID_3238	637	C	2000	Alcoy3	Patient	4,668,595	99.58	2,075,840	94.25
ID_3334	637	C	2000	Alcoy3	Patient	3,170,486	67.62	1,425,610	94.35
ID_3785	578	C	2001	Alcoy4	Patient	3,045,636	64.96	1,642,859	99.68
ID_3908	578	C	2001	Alcoy4	Patient	9,812,615	209.29	5,257,572	99.20
ID_479	171	E	2001	Sporadic	Evaporative condenser	9,084,712	193.77	3,776,087	92.76
ID_3499	578	C	2001	Sporadic	Patient	3,913,063	83.46	1,948,733	99.75
ID_3786	578	C	2001	Sporadic	Patient	3,145,379	67.09	1,586,204	99.65
ID_482	804	E	2001	Sporadic	Evaporative condenser	5,481,713	116.92	2,131,507	94.32
ID_5856	578	C	2002	Alcoy5	Patient	4,428,053	94.45	2,439,109	99.77
ID_6536	578	C	2002	Alcoy7	Patient	3,981,736	84.93	1,996,436	99.73
ID_891	1	E	2002	Alcoy9	Evaporative condenser	3,879,116	82.74	1,355,236	87.96
ID_598	1	E	2002	Sporadic	Common industry	3,853,598	82.19	1,374,001	85.04
ID_5228	578	C	2002	Sporadic	Patient	4,194,459	89.46	2,113,916	99.72
ID_8141	578	C	2003	Alcoy11	Patient	3,071,569	65.51	1,474,690	99.02
ID_8189	578	C	2003	Alcoy11	Patient	3,609,695	76.99	1,863,426	99.72
ID_8190	578	C	2003	Alcoy11	Patient	4,445,909	94.83	2,122,688	99.73
ID_8227	578	C	2003	Alcoy11	Patient	4,984,637	106.32	2,415,038	99.78
ID_8228	578	C	2003	Alcoy11	Patient	2,339,945	49.91	1,254,560	99.56
ID_7147	578	C	2003	Alcoy8	Patient	2,412,556	51.46	1,176,194	99.49
ID_7371	578	C	2003	Sporadic	Patient	4,333,690	92.43	2,140,499	99.73
ID_8004	578	C	2003	Sporadic	Patient	3,588,458	76.54	1,794,573	99.66
ID_1688	1	E	2004	Sporadic	Evaporative condenser	4,735,698	101.01	1,699,719	85.03
ID_1690	1	E	2004	Sporadic	Evaporative condenser	4,491,561	95.80	1,681,617	84.68

ID_1828	1	E	2004	Sporadic	Evaporative condenser	5,488,079	117.06	1,875,822	85.96
ID_1925	578	E	2004	Sporadic	Evaporative condenser	6,963,114	148.52	2,965,178	99.17
ID_1885	1037	E	2004	Sporadic	Evaporative condenser	4,178,489	89.12	1,897,431	91.82
ID_2041	1	E	2005	Alcoy12	Water-treatment plant	7,145,529	152.41	1,868,438	89.19
ID_747970	1	E	2009	Alcoy16	Milling machine	3,341,834	71.28	1,204,865	85.58
ID_480203	578	C	2009	Alcoy16	Patient	3,016,217	64.33	1,392,000	99.37
ID_480295	578	C	2009	Alcoy16	Patient	3,763,446	80.27	1,916,349	99.00
ID_480372	578	C	2009	Alcoy16	Patient	4,010,993	85.55	1,875,862	99.05
ID_480392	578	C	2009	Alcoy16	Patient	3,800,877	81.07	1,914,090	99.47
ID_747968	578	E	2009	Alcoy16	Water tanker truck hose	4,241,659	90.47	1,900,993	99.52
ID_747969	578	E	2009	Alcoy16	Milling machine	3,370,204	71.88	1,654,536	99.46
ID_747973	578	E	2009	Alcoy16	Water tanker truck	3,521,351	75.11	1,662,718	98.98
ID_481107	578	C	2009	Alcoy17	Patient	4,411,214	94.09	1,971,516	99.04
ID_481441	578	C	2009	Alcoy17	Patient	3,412,628	72.79	1,470,833	98.78
ID_481707	578	C	2009	Alcoy17	Patient	4,377,335	93.36	1,971,991	99.53
ID_481710	578	C	2009	Alcoy17	Patient	3,767,135	80.35	1,596,371	99.01
ID_480263	578	C	2009	Sporadic	Patient	4,163,980	88.81	1,756,725	99.47
ID_481898	578	C	2009	Sporadic	Patient	4,160,472	88.74	1,796,481	99.45
ID_481944	578	C	2009	Sporadic	Patient	2,351,921	50.16	1,043,015	98.68
ID_1190176	578	E	2010	Alcoy18	Industry	1,872,888	39.95	930,577	98.71
ID_489571	578	C	2010	Alcoy18	Patient	3,423,481	73.02	1,490,117	99.37
ID_489956	578	C	2010	Alcoy18	Patient	2,605,865	55.58	1,038,359	98.61
ID_490679	578	C	2010	Alcoy18	Patient	42,874,666	914.48	17,951,594	99.39
ID_490738	578	C	2010	Alcoy18	Patient	2,301,664	49.09	899,430	98.47
ID_489154	578	C	2010	Sporadic	Patient	3,474,178	74.10	1,423,550	99.38
ID_490456	578	C	2010	Sporadic	Patient	3,274,422	69.84	1,329,070	98.69
ID_6885	1	E	2011	Sporadic	Hospital	4,643,468	99.04	1,107,769	84.94
ID_505237	637	C	2011	Sporadic	Patient	3,787,466	80.78	1,282,246	94.44
ID_496053	1106	C	2011	Sporadic	Patient	3,676,290	78.41	1,509,852	96.24

Table S3 | Comparison of parameter estimates and AIC values among combinations of demographic (constant population size, exponential growth and Bayesian skyline-plot) and molecular clock (strict, random and uncorrelated log-normal, ULN) models for the reconstruction of the evolutionary and demographic history of ST578 isolates from Alcoy using BEAST.

Demographic model	**Constant**			**Exponential**			**BSP (5 populations)**		
Clock	**Strict**	**Random**	**ULN**	**Strict**	**Random**	**ULN**	**Strict**	**Random**	**ULN**
Substitution rate (s/s/y) **95% HPD**	7.60E-08 [3.72E-8, 1.17E-7]	8.43E-08 [3.76E-8, 1.42E-7]	9.73E-08 [3.59E-8, 1.77E-7]	7.82E-08 [3.97E-8, 1.21E-7]	8.28E-08 [4.08E-8, 1.30E-7]	1.15E-07 [4.4423E-8, 1.9858E-7]	8.09E-08 [3.91E-8, 1.35E-7]	1.05E-07 [4.52E-8, 1.71E-7]	1.42E-07 [4.82E-8, 2.40E-7]
TMRCA **95% HPD**	48.98 [22.46, 81.64]	43.75 [16.01, 76.34]	41.61 [14.67, 80.34]	41.87 [20.67, 69.59]	38.93 [17.17, 65.37]	27.9329 [13.5857, 50.9384]	43.32 [13.99, 72.85]	28.82 [12.29, 58.86]	20.98 [11.35, 43.16]
TMRCA (578A) **95% HPD**	31.88 [17.85, 50.36]	30.86 [16.26, 49.50]	33.50 [15.10, 58.57]	29.16 [17.05, 45.24]	28.28 [16.43, 43.89]	25.1558 [13.6634, 40.8885]	27.93 [14.46, 44.68]	22.95 [12.51, 38.68]	19.45 [11.35, 33.89]
TMRCA (578B) **95% HPD**	7.94 [3.56, 12.86]	6.94 [1.99, 11.89]	7.03 [2.88, 11.79]	8.52 [4.25, 13.83]	7.82 [2.75, 12.67]	6.8453 [3.1327, 10.7893]	7.96 [3.31, 13.50]	5.17 [1.79, 10.37]	5.50 [2.56, 8.73]
AICM1	7015624.46 +/- 0.16	7015625.36 +/- 0.18	7015638.72 +/- 0.07	7015623.12 +/- 0.09	7015626.76 +/- 0.11	7015638.2392 +/- 0.1270	7015626.42 +/- 0.08	7015629.71+/- 0.06	7015635.85 +/- 0.32
AICM2	7015624.26 +/- 0.06	7015627.57 +/- 0.17	7015638.51 +/- 0.15	7015620.78 +/- 0.18	7015623.76 +/- 0.08	7015635.4060 +/- 0.1543	7015627.10 +/- 0.12	7015631.14 +/- 0.13	7015634.02 +/- 0.09

Table S4 | Summary of phylogenetic incongruence tests. For each gene in the *L. pneumophila* ST578 core genome with 4X (round 1) or 3X (round 2) more SNPs than expected we compared the phylogenetic tree obtained from the corresponding multiple alignment and that obtained from the concatenate alignment of the 3,120 genes using the Shimodaira-Hasegawa (SH) and Expected Likelihood Weight (ELW) tests implemented in TreePuzzle. P-values for these tests are shown per gene (*L. pneumophila* 2300/99 Alcoy) in the table, and lack of congruence was assumed when both tests were significant at the p<0.05 level.

Gene	Round	SH_gene	SH_concat	ELW_gene	ELW_concat	Incongruence?
lpa_00090	1	1.0000	0.0110	1.0000	0.0000	YES
lpa_00150	2	1.0000	0.0000	1.0000	0.0000	YES
lpa_00191	2	1.0000	0.2150	0.7160	0.2840	NO
lpa_00263	2	1.0000	0.0840	0.8993	0.1007	NO
lpa_00274	2	1.0000	0.0500	0.9500	0.0500	NO
lpa_00411	2	1.0000	0.2480	0.6326	0.3674	NO
lpa_00441	2	1.0000	0.0240	0.9866	0.0134	YES
lpa_00476	2	1.0000	0.0270	0.9915	0.0085	YES
lpa_00673	2	1.0000	0.0000	1.0000	0.0000	YES
lpa_00835	2	1.0000	0.0450	0.9913	0.0087	YES
lpa_00882	2	1.0000	0.0140	0.9994	0.0006	YES
lpa_00900	2	1.0000	0.0110	0.9915	0.0085	YES
lpa_00907	2	1.0000	0.1990	0.7577	0.2423	NO
lpa_00908	2	1.0000	0.0700	0.9733	0.0267	NO
lpa_00908	1	1.0000	0.0590	0.9697	0.0303	NO
lpa_00909	2	1.0000	0.0880	0.9238	0.0762	NO
lpa_00913	2	1.0000	0.0230	0.9923	0.0077	YES
lpa_00918	2	1.0000	0.0660	0.9881	0.0119	NO
lpa_00918	1	1.0000	0.0610	0.9881	0.0119	NO
lpa_00924	2	1.0000	0.0000	1.0000	0.0000	YES
lpa_00927	1	1.0000	0.0080	1.0000	0.0000	YES
lpa_00928	1	1.0000	0.0180	0.9994	0.0006	YES
lpa_00932	1	1.0000	0.0280	0.9901	0.0099	YES
lpa_00934	2	1.0000	0.0720	0.9030	0.0970	NO
lpa_00938	1	1.0000	0.0130	0.9997	0.0003	YES
lpa_00939	1	1.0000	0.0380	0.9990	0.0010	YES
lpa_00941	2	1.0000	0.0230	0.9993	0.0007	YES
lpa_00943	2	1.0000	0.0800	0.9664	0.0336	NO
lpa_00945	2	1.0000	0.3450	0.6001	0.3999	NO
lpa_00946	2	1.0000	0.0440	0.9795	0.0205	YES
lpa_00948	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_00949	1	1.0000	0.0110	0.9999	0.0001	YES
lpa_00951	1	1.0000	0.2420	0.7164	0.2836	NO
lpa_00951	2	1.0000	0.2240	0.7428	0.2572	NO
lpa_00952	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_00953	1	1.0000	0.0140	0.9992	0.0008	YES
lpa_00954	1	1.0000	0.0240	0.9821	0.0179	YES
lpa_00955	1	1.0000	0.0080	0.9973	0.0027	YES
lpa_00958	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_00960	2	1.0000	0.0030	0.9995	0.0005	YES
lpa_00968	2	1.0000	0.0600	0.9922	0.0078	NO
lpa_00970	1	1.0000	0.0040	1.0000	0.0000	YES

lpa_00971	1	1.0000	0.0070	0.9999	0.0001	YES
lpa_00972	1	1.0000	0.0100	0.9998	0.0002	YES
lpa_00973	2	1.0000	0.2090	0.7475	0.2525	NO
lpa_00974	2	1.0000	0.0000	1.0000	0.0000	YES
lpa_00975	2	1.0000	0.0000	1.0000	0.0000	YES
lpa_01005	2	1.0000	0.0010	1.0000	0.0000	YES
lpa_01025	1	1.0000	0.0120	0.9990	0.0010	YES
lpa_01027	2	1.0000	0.1760	0.8280	0.1720	NO
lpa_01027	1	1.0000	0.1740	0.8032	0.1968	NO
lpa_01031	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01032	2	1.0000	0.0150	0.9991	0.0009	YES
lpa_01034	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01040	2	1.0000	0.0200	0.9940	0.0060	YES
lpa_01129	2	1.0000	0.0720	0.9736	0.0264	NO
lpa_01146	1	1.0000	0.0960	0.9224	0.0776	NO
lpa_01146	2	1.0000	0.0940	0.9235	0.0765	NO
lpa_01183	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01184	1	1.0000	0.0060	1.0000	0.0000	YES
lpa_01186	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01196	2	1.0000	0.0010	1.0000	0.0000	YES
lpa_01207	1	1.0000	0.0040	1.0000	0.0000	YES
lpa_01210	1	1.0000	0.0030	1.0000	0.0000	YES
lpa_01212	1	1.0000	0.0010	1.0000	0.0000	YES
lpa_01216	1	1.0000	0.0020	1.0000	0.0000	YES
lpa_01217	1	1.0000	0.0010	1.0000	0.0000	YES
lpa_01218	1	1.0000	0.0240	1.0000	0.0000	YES
lpa_01219	1	1.0000	0.0010	1.0000	0.0000	YES
lpa_01220	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01222	1	1.0000	0.0010	1.0000	0.0000	YES
lpa_01225	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01231	1	1.0000	0.0040	1.0000	0.0000	YES
lpa_01232	1	1.0000	0.0170	0.9997	0.0003	YES
lpa_01233	2	1.0000	0.0450	0.9987	0.0013	YES
lpa_01234	2	1.0000	0.0120	1.0000	0.0000	YES
lpa_01236	1	1.0000	0.0140	0.9991	0.0009	YES
lpa_01237	2	1.0000	0.0090	0.9997	0.0003	YES
lpa_01241	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01245	2	1.0000	0.0010	1.0000	0.0000	YES
lpa_01248	2	1.0000	0.0000	1.0000	0.0000	YES
lpa_01249	1	1.0000	0.0060	1.0000	0.0000	YES
lpa_01251	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01253	1	1.0000	0.0100	1.0000	0.0000	YES
lpa_01254	2	1.0000	0.0090	1.0000	0.0000	YES
lpa_01255	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01256	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01258	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01261	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01262	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01264	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01265	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01266	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01267	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01268	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01269	1	1.0000	0.0020	1.0000	0.0000	YES

lpa_01270	1	1.0000	0.0030	1.0000	0.0000	YES
lpa_01271	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01272	1	1.0000	0.0040	0.9999	0.0001	YES
lpa_01273	1	1.0000	0.0010	1.0000	0.0000	YES
lpa_01275	1	1.0000	0.0010	1.0000	0.0000	YES
lpa_01276	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01277	1	1.0000	0.0410	0.9796	0.0204	YES
lpa_01278	1	1.0000	0.0070	1.0000	0.0000	YES
lpa_01279	1	1.0000	0.0020	1.0000	0.0000	YES
lpa_01281	1	1.0000	0.0100	0.9989	0.0011	YES
lpa_01282	1	1.0000	0.0010	1.0000	0.0000	YES
lpa_01283	1	1.0000	0.0020	0.9996	0.0004	YES
lpa_01284	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01286	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01287	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01289	1	1.0000	0.0010	1.0000	0.0000	YES
lpa_01290	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01292	1	1.0000	0.0120	0.9987	0.0013	YES
lpa_01294	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01295	1	1.0000	0.0020	1.0000	0.0000	YES
lpa_01319	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01320	1	1.0000	0.0090	1.0000	0.0000	YES
lpa_01322	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01323	1	1.0000	0.4340	0.5093	0.4907	NO
lpa_01323	2	1.0000	0.4220	0.5007	0.4993	NO
lpa_01324	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01325	1	1.0000	0.0010	1.0000	0.0000	YES
lpa_01326	1	1.0000	0.0270	0.9979	0.0021	YES
lpa_01327	1	1.0000	0.0010	0.9990	0.0010	YES
lpa_01346	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_01391	2	1.0000	0.0010	1.0000	0.0000	YES
lpa_01465	2	1.0000	0.0000	1.0000	0.0000	YES
lpa_01514	2	1.0000	0.0380	0.9787	0.0213	YES
lpa_01540	2	1.0000	0.0760	0.8966	0.1034	NO
lpa_01588	2	1.0000	0.0000	0.5926	0.4074	YES
lpa_01670	2	1.0000	0.0420	0.9998	0.0002	YES
lpa_01685	2	1.0000	0.0500	0.9500	0.0500	NO
lpa_01805	2	1.0000	0.0280	0.9947	0.0053	YES
lpa_01815	2	1.0000	0.3600	0.5446	0.4554	NO
lpa_01849	2	1.0000	0.0000	1.0000	0.0000	YES
lpa_01956	2	1.0000	0.0580	0.9559	0.0441	NO
lpa_02050	2	1.0000	0.1050	0.9149	0.0851	NO
lpa_02058	2	1.0000	0.0680	0.9274	0.0726	NO
lpa_02151	2	1.0000	0.0000	1.0000	0.0000	YES
lpa_02154	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_02155	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_02156	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_02160	1	1.0000	0.0220	0.9939	0.0061	YES
lpa_02161	2	1.0000	0.0000	1.0000	0.0000	YES
lpa_02162	1	1.0000	0.0010	0.9973	0.0027	YES
lpa_02164	2	1.0000	0.4040	0.5460	0.4540	NO
lpa_02166	2	1.0000	0.0450	0.9719	0.0281	YES
lpa_02167	2	1.0000	0.0470	0.9548	0.0452	YES
lpa_02168	2	1.0000	0.0330	0.9765	0.0235	YES

lpa_02172	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_02173	1	1.0000	0.0260	0.9999	0.0001	YES
lpa_02174	2	1.0000	0.2550	0.6731	0.3269	NO
lpa_02176	2	1.0000	0.0210	0.9830	0.0170	YES
lpa_02187	2	1.0000	0.0730	0.9223	0.0777	NO
lpa_02245	2	1.0000	0.1210	0.8580	0.1420	NO
lpa_02305	2	1.0000	0.0010	0.9999	0.0001	YES
lpa_02333	2	1.0000	0.0000	1.0000	0.0000	YES
lpa_02334	2	1.0000	0.0130	0.9952	0.0048	YES
lpa_02384	2	1.0000	0.3590	0.5571	0.4429	NO
lpa_02440	2	1.0000	0.1970	0.8033	0.1967	NO
lpa_02447	2	1.0000	0.0600	0.9353	0.0647	NO
lpa_02455	2	1.0000	0.0000	1.0000	0.0000	YES
lpa_02456	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_02457	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_02461	2	1.0000	0.0280	0.9777	0.0223	YES
lpa_02463	2	1.0000	0.1390	0.8625	0.1375	NO
lpa_02463	1	1.0000	0.1330	0.8846	0.1154	NO
lpa_02464	2	1.0000	0.0080	0.9985	0.0015	YES
lpa_02465	1	1.0000	0.1460	0.9018	0.0982	NO
lpa_02465	2	1.0000	0.1310	0.8755	0.1245	NO
lpa_02474	2	1.0000	0.0880	0.9612	0.0388	NO
lpa_02478	1	1.0000	0.0620	0.9849	0.0151	NO
lpa_02478	2	1.0000	0.0490	0.9850	0.0150	YES
lpa_02480	1	1.0000	0.0450	0.9992	0.0008	YES
lpa_02481	2	1.0000	0.0830	0.8937	0.1063	NO
lpa_02484	2	1.0000	0.0000	1.0000	0.0000	YES
lpa_02501	2	1.0000	0.0000	1.0000	0.0000	YES
lpa_02509	2	1.0000	0.0020	1.0000	0.0000	YES
lpa_02512	2	1.0000	0.0040	1.0000	0.0000	YES
lpa_02522	2	1.0000	0.0000	1.0000	0.0000	YES
lpa_02537	2	1.0000	0.0000	1.0000	0.0000	YES
lpa_02540	2	1.0000	0.1840	0.8107	0.1893	NO
lpa_02545	2	1.0000	0.0090	0.9999	0.0001	YES
lpa_02547	1	1.0000	0.3490	0.6213	0.3787	NO
lpa_02547	2	1.0000	0.2730	0.7108	0.2892	NO
lpa_02549	2	1.0000	0.1380	0.8777	0.1223	NO
lpa_02551	2	1.0000	0.1530	0.8342	0.1658	NO
lpa_02560	2	1.0000	0.0010	0.9999	0.0001	YES
lpa_02569	2	1.0000	0.0000	0.9978	0.0022	YES
lpa_02612	2	1.0000	0.0000	1.0000	0.0000	YES
lpa_02626	2	1.0000	0.0290	0.9915	0.0085	YES
lpa_02662	2	1.0000	0.0240	0.9832	0.0168	YES
lpa_02690	2	1.0000	0.0740	0.9807	0.0193	NO
lpa_02691	2	1.0000	0.0200	0.9999	0.0001	YES
lpa_02693	2	1.0000	0.0620	0.9009	0.0991	NO
lpa_02696	2	1.0000	0.0250	0.9962	0.0038	YES
lpa_02808	2	1.0000	0.1400	0.8190	0.1810	NO
lpa_02915	2	1.0000	0.3120	0.5679	0.4321	NO
lpa_02950	1	1.0000	0.0130	0.9986	0.0014	YES
lpa_02952	2	1.0000	0.0240	0.9998	0.0002	YES
lpa_02976	2	1.0000	0.0200	0.9876	0.0124	YES
lpa_03019	2	1.0000	0.0860	0.9024	0.0976	NO
lpa_03037	1	1.0000	0.0020	1.0000	0.0000	YES

lpa_03058	2	1.0000	0.1080	0.9086	0.0914	NO
lpa_03097	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_03174	2	1.0000	0.1020	0.8937	0.1063	NO
lpa_03250	2	1.0000	0.0000	1.0000	0.0000	YES
lpa_03374	1	1.0000	0.0090	0.9998	0.0002	YES
lpa_03378	2	1.0000	0.1020	0.9111	0.0889	NO
lpa_03424	2	1.0000	0.2200	0.7600	0.2400	NO
lpa_03437	2	1.0000	0.0840	0.9540	0.0460	NO
lpa_03439	1	1.0000	0.0140	0.9977	0.0023	YES
lpa_03447	2	1.0000	0.0140	0.9976	0.0024	YES
lpa_03491	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_03494	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_03495	1	1.0000	0.0030	0.9999	0.0001	YES
lpa_03496	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_03497	1	1.0000	0.0630	0.9784	0.0216	NO
lpa_03497	2	1.0000	0.0620	0.9783	0.0217	NO
lpa_03498	1	1.0000	0.0110	1.0000	0.0000	YES
lpa_03499	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_03517	2	1.0000	0.0110	0.9969	0.0031	YES
lpa_03527	2	1.0000	0.0370	0.9936	0.0064	YES
lpa_03530	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_03531	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_03532	1	1.0000	0.0010	1.0000	0.0000	YES
lpa_03541	2	1.0000	0.0080	1.0000	0.0000	YES
lpa_03542	2	1.0000	0.0730	0.9020	0.0980	NO
lpa_03543	2	1.0000	0.0680	0.9924	0.0076	NO
lpa_03544	2	1.0000	0.0540	0.9909	0.0091	NO
lpa_03619	2	1.0000	0.0060	0.9990	0.0010	YES
lpa_03701	2	1.0000	0.0920	0.8956	0.1044	NO
lpa_03704	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_03706	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_03716	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_03720	1	1.0000	0.0080	0.9999	0.0001	YES
lpa_03721	1	1.0000	0.0390	0.9959	0.0041	YES
lpa_03723	1	1.0000	0.0040	1.0000	0.0000	YES
lpa_03727	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_03729	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_03732	1	1.0000	0.0020	1.0000	0.0000	YES
lpa_03733	1	1.0000	0.0540	0.9920	0.0080	NO
lpa_03733	2	1.0000	0.0510	0.9985	0.0015	NO
lpa_03734	1	1.0000	0.0240	0.9993	0.0007	YES
lpa_03736	1	1.0000	0.0310	0.9979	0.0021	YES
lpa_03738	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_03739	1	1.0000	0.0010	1.0000	0.0000	YES
lpa_03741	1	1.0000	0.0620	0.9469	0.0531	NO
lpa_03741	2	1.0000	0.0460	0.9579	0.0421	YES
lpa_03743	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_03746	1	1.0000	0.0010	1.0000	0.0000	YES
lpa_03749	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_03750	2	1.0000	0.0780	0.8953	0.1047	NO
lpa_03751	1	1.0000	0.0030	0.9952	0.0048	YES
lpa_03752	1	1.0000	0.0060	1.0000	0.0000	YES
lpa_03766	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_03768	2	1.0000	0.0590	0.9708	0.0292	NO

lpa_03768	1	1.0000	0.0510	0.9781	0.0219	NO
lpa_03770	1	1.0000	0.0620	0.9956	0.0044	NO
lpa_03770	2	1.0000	0.0470	0.9962	0.0038	YES
lpa_03771	2	1.0000	0.0780	0.9932	0.0068	NO
lpa_03772	2	1.0000	0.0760	0.9501	0.0499	NO
lpa_03775	1	1.0000	0.0030	1.0000	0.0000	YES
lpa_03776	1	1.0000	0.0400	0.9960	0.0040	YES
lpa_03779	2	1.0000	0.0210	0.9969	0.0031	YES
lpa_03780	2	1.0000	0.2660	0.6663	0.3337	NO
lpa_03782	1	1.0000	0.0120	0.9958	0.0042	YES
lpa_03783	1	1.0000	0.0040	1.0000	0.0000	YES
lpa_03784	2	1.0000	0.1980	0.7513	0.2487	NO
lpa_03787	2	1.0000	0.0810	0.9623	0.0377	NO
lpa_03788	2	1.0000	0.1130	0.8434	0.1566	NO
lpa_03799	2	1.0000	0.2440	0.7403	0.2597	NO
lpa_03799	1	1.0000	0.2370	0.7510	0.2490	NO
lpa_03800	2	1.0000	0.1420	0.8527	0.1473	NO
lpa_03803	1	1.0000	0.0070	1.0000	0.0000	YES
lpa_03804	1	1.0000	0.0010	1.0000	0.0000	YES
lpa_03813	2	1.0000	0.0640	0.9508	0.0492	NO
lpa_03814	1	1.0000	0.0670	0.9900	0.0100	NO
lpa_03814	2	1.0000	0.0540	0.9875	0.0125	NO
lpa_03817	2	1.0000	0.0910	0.9406	0.0594	NO
lpa_03819	1	1.0000	0.0110	0.9999	0.0001	YES
lpa_03823	2	1.0000	0.1030	0.8997	0.1003	NO
lpa_03824	2	1.0000	0.4580	0.4847	0.5153	NO
lpa_03825	2	1.0000	0.1010	0.9500	0.0500	NO
lpa_03825	1	1.0000	0.0790	0.9657	0.0343	NO
lpa_03827	2	1.0000	0.0500	0.9500	0.0500	NO
lpa_03828	1	1.0000	0.0260	1.0000	0.0000	YES
lpa_03831	2	1.0000	0.0270	0.9969	0.0031	YES
lpa_03832	1	1.0000	0.0490	0.9818	0.0182	YES
lpa_03833	2	1.0000	0.0030	1.0000	0.0000	YES
lpa_03834	1	1.0000	0.0150	0.9994	0.0006	YES
lpa_03836	1	1.0000	0.0140	0.9995	0.0005	YES
lpa_03846	2	1.0000	0.0430	0.9864	0.0136	YES
lpa_03847	2	1.0000	0.3660	0.5629	0.4371	NO
lpa_03853	1	1.0000	0.0150	0.9999	0.0001	YES
lpa_03854	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_03855	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_03859	1	1.0000	0.0150	0.9992	0.0008	YES
lpa_03861	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_03863	1	1.0000	0.0080	1.0000	0.0000	YES
lpa_03867	1	1.0000	0.0010	1.0000	0.0000	YES
lpa_03869	1	1.0000	0.0130	0.9987	0.0013	YES
lpa_03873	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_03874	1	1.0000	0.0420	0.9857	0.0143	YES
lpa_03876	2	1.0000	0.0420	0.9950	0.0050	YES
lpa_03877	1	1.0000	0.0880	0.9686	0.0314	NO
lpa_03877	2	1.0000	0.0780	0.9642	0.0358	NO
lpa_03881	1	1.0000	0.0470	0.9819	0.0181	YES
lpa_03883	2	1.0000	0.0190	0.9987	0.0013	YES
lpa_03884	2	1.0000	0.0560	0.9544	0.0456	NO
lpa_03885	2	1.0000	0.0140	0.9963	0.0037	YES

lpa_03889	1	1.0000	0.3600	0.5856	0.4144	NO
lpa_03889	2	1.0000	0.0090	0.9973	0.0027	YES
lpa_03890	1	0.1120	1.0000	0.1853	0.8147	NO
lpa_03890	2	1.0000	0.3780	0.5886	0.4114	NO
lpa_03891	1	1.0000	0.0500	0.9500	0.0500	NO
lpa_03891	2	1.0000	0.2800	0.6218	0.3782	NO
lpa_03892	2	1.0000	0.3680	0.6408	0.3592	NO
lpa_03954	2	1.0000	0.0080	1.0000	0.0000	YES
lpa_03967	2	1.0000	0.1360	0.8647	0.1353	NO
lpa_03970	2	1.0000	0.0090	0.9969	0.0031	YES
lpa_03971	2	1.0000	0.0500	0.9500	0.0500	NO
lpa_03973	2	1.0000	0.0900	0.8998	0.1002	NO
lpa_03974	1	1.0000	0.0900	0.9068	0.0932	NO
lpa_03974	2	1.0000	0.0300	0.9799	0.0201	YES
lpa_03975	2	1.0000	0.0930	0.9332	0.0668	NO
lpa_03980	2	1.0000	0.4200	0.5329	0.4671	NO
lpa_03984	2	1.0000	0.0030	1.0000	0.0000	YES
lpa_03985	2	1.0000	0.0020	1.0000	0.0000	YES
lpa_03987	2	1.0000	0.0450	0.9685	0.0315	YES
lpa_03991	2	1.0000	0.0420	0.9885	0.0115	YES
lpa_03994	2	1.0000	0.0340	0.9877	0.0123	YES
lpa_03997	2	1.0000	0.0090	0.9999	0.0001	YES
lpa_04005	2	1.0000	0.0140	0.9970	0.0030	YES
lpa_04018	2	1.0000	0.0000	1.0000	0.0000	YES
lpa_04026	1	1.0000	0.0530	0.9943	0.0057	NO
lpa_04026	2	1.0000	0.0260	0.9970	0.0030	YES
lpa_04029	2	1.0000	0.0600	0.9950	0.0050	NO
lpa_04034	1	1.0000	0.0260	0.9997	0.0003	YES
lpa_04035	1	1.0000	0.0050	1.0000	0.0000	YES
lpa_04036	2	1.0000	0.0500	0.9797	0.0203	NO
lpa_04038	2	0.4710	1.0000	0.4754	0.5246	NO
lpa_04039	2	1.0000	0.0800	0.9579	0.0421	NO
lpa_04041	2	1.0000	0.0330	0.9931	0.0069	YES
lpa_04042	2	1.0000	0.0390	0.9864	0.0136	YES
lpa_04044	1	1.0000	0.0150	0.9979	0.0021	YES
lpa_04046	2	1.0000	0.1450	0.8344	0.1656	NO
lpa_04047	1	1.0000	0.0000	1.0000	0.0000	YES
lpa_04049	1	1.0000	0.0640	0.9947	0.0053	NO
lpa_04049	2	1.0000	0.0620	0.9922	0.0078	NO
lpa_04051	2	1.0000	0.0450	0.9882	0.0118	YES
lpa_04052	2	1.0000	0.3140	0.5727	0.4273	NO
lpa_04053	2	1.0000	0.1470	0.8351	0.1649	NO
lpa_04055	2	1.0000	0.0010	0.5014	0.4986	YES
lpa_04056	2	1.0000	0.1170	0.8787	0.1213	NO
lpa_04060	1	1.0000	0.0310	0.9998	0.0002	YES
lpa_04061	1	1.0000	0.0370	0.9982	0.0018	YES
lpa_04062	2	1.0000	0.0410	0.9640	0.0360	YES
lpa_04063	2	1.0000	0.0510	0.9838	0.0162	NO
lpa_04065	2	1.0000	0.0780	0.9552	0.0448	NO
lpa_04066	2	1.0000	0.0310	0.9922	0.0078	YES
lpa_04067	2	1.0000	0.0350	0.9862	0.0138	YES
lpa_04068	1	1.0000	0.0810	0.9364	0.0636	NO
lpa_04068	2	1.0000	0.0440	0.9989	0.0011	YES
lpa_04069	2	1.0000	0.0900	0.9258	0.0742	NO

lpa_04153	2	1.0000	0.0760	0.9991	0.0009	NO
lpa_04203	2	1.0000	0.0010	1.0000	0.0000	YES
lpa_04271	2	1.0000	0.1940	0.7844	0.2156	NO
lpa_04341	2	1.0000	0.0500	0.9500	0.0500	NO
lpa_04384	2	1.0000	0.0010	1.0000	0.0000	YES

Table S5 | Annotation of the 45 SNPs detected as being introduced by substitution events in the core genome of the ST578 strains. '→' refers to the branch of the non-recombining core tree where each SNP is mapped. B1-B5 denote the recombining branches described in the main text. The annotation of each gene corresponds to that from the *L. pneumophila* 2300/99 Alcoy reference genome.

SNP position	Change	Gene	Branch	Gene annotation
110128	G>T	lpa_00132	→ID_747969	Hypothetical protein
159071	T>A	lpa_00191	→Lp_Alcoy	Methylmalonate-semialdehyde dehydrogenase
208809	A>G	lpa_00263	B2	Hypothetical protein
215988	A>C	lpa_00274	→ID_490738	RNA polymerase Rpb8 subunit
310721	G>T	lpa_00411	→ID_747968	Multidrug efflux protein, outer membrane component
324977	G>A	lpa_00419	→ID_480372	SidE protein
893579	A>C	lpa_01129	→ID_7147	Phosphatidylglycerophosphatase A
909344	A>G	lpa_01146	→ID_6536	Iron sulphur-containing domain protein
1212503	A>G	lpa_01540	→(ID_3785,ID_3786)	Cadmium translocating P-type ATPase CadA
1240970	A>G	lpa_01577	→ID_5228	Cation efflux system protein cusA
1310005	A>G	lpa_01685	→ID_8141	Putative membrane protein, putative virulence factor
1409542	A>G	lpa_01815	→ID_3110	Acetylornithine deacetylase
1506009	T>A	lpa_01936	Homoplasy (ID_3908, ID_480203, ID_6356)	Motility protein FimV
1530075	C>T	lpa_01956	B5 (→ID_3164)	Hypothetical protein
1603285	A>G	lpa_02050	B2	3-oxoacyl-(acyl-carrier-protein)
1608925	G>A	lpa_02058	→ID_8228	DNA polymerase III, delta prime subunit
1716906	C>T	lpa_02187	→ID_480392	Aminopeptidase N
1770314	C>T	lpa_02245	→ID_490738	Histidyl-tRNA synthetase
1887297	G>C	lpa_02384	→(ID_8227,ID_8189)	Myo-inositol catabolism protein IolE
1949018	C>G	lpa_02440	→ID_481898	Aconitate hydratase 1

2030317	C>T	lpa_02540	B1	UDP-N-acetylmuramate-L-alanyl-gamma-D-glutamyl- meso-diaminopimelate ligase
2038992	G>A	lpa_02547	B1	Flagellar motor switch protein FliG
2038998	G>T	lpa_02547	B1	Flagellar motor switch protein FliG
2039115	A>C	lpa_02547	B1	Flagellar motor switch protein FliG
2041111	C>T	lpa_02549	B2	Flagellar M-ring protein FliF
2041810	C>T	lpa_02551	B1	Protein involved in the phosphorelay signal transduction system and the regulation of transcription
2104665	G>A	lpa_02619	→ID_3499	ATP-dependent DNA helicase (UvrD/Rep helicase)
2256069	C>G	lpa_02808	→ID_490738	Hypothetical protein
2341099	T>C	lpa_02915	→(ID_480203,ID_481707, ID_481898,ID_489571,ID_480263)	Histidinol-phosphate aminotransferase
2419305	T>C	lpa_03019	B2	Hypothetical protein
2453141	A>T	lpa_03058	→ID_489956	Chemiosmotic efflux system protein C-like protein
2586901	C>T	lpa_03174	→(ID_8227,ID_8189)	Hypothetical protein
2863518	A>G	lpa_03542	→(ID_489956,ID_481944, ID_1190176,ID_490679,ID_481441, ID_481710,ID_490738,ID_480372, ID_1925,ID_8141,ID_480295, ID_481107,ID_747973,ID_8190, ID_3908,ID_490456)	Nucleotidyltransferase
2864283	C>G	lpa_03543	→(ID_489956,ID_481944, ID_1190176,ID_490679,ID_481441, ID_481710,ID_490738,ID_480372, ID_1925,ID_8141,ID_480295, ID_481107,ID_747973,ID_8190, ID_3908,ID_490456)	Triphosphoribosyl-dephospho-CoA synthase
2865156	C>T	lpa_03544	→(ID_489956,ID_481944, ID_1190176,ID_490679,ID_481441, ID_481710,ID_490738,ID_480372, ID_1925,ID_8141,ID_480295, ID_481107,ID_747973,ID_8190, ID_3908)	Putative malonyl-CoA acyl-carrier-protein transacylase
2977091	C>A	lpa_03701	→ID_480203	Myoglobin-like protein
3184000	G>T	lpa_03967	B1	Protein with ligase activity within the cellular modified amino acid biosynthetic process
3184072	C>T	lpa_03967	B1	Protein with ligase activity within the cellular modified amino acid biosynthetic process
3187643	T>C	lpa_03971	B1	N-ethylmaleimide reductase

3188899	G>A	lpa_03973	B1	Hypothetical protein
3189713	G>A	lpa_03975	B1	Inner membrane protein
3191590	A>T	lpa_03980	B1	Queuine/archaeosine tRNA-ribosyltransferase
3337908	C>T	lpa_04153	Homoplasy (ID_2301,ID_3009,ID_3109,ID_3110, ID_3499,ID_480295,ID_481107, ID_481898,ID_5856,ID_7371, ID_747973,ID_8189,ID_8190,ID_8227)	Protein-binding domain
3415735	C>T	lpa_04271	→ID_489154	Protein with iron-sulphur oxidoreductase activity
3467887	T>C	lpa_04341	→Lp_Alcoy	Major outer membrane protein

Table S6 | Relevant features of the 16 recombination events detected in the core genome of *L. pneumophila* ST578 strains. Note that events marked with the same symbol (*, §, ‡) are completely or partially overlapping.

Branch	Total number of SNPs	Event	Gene interval	Positions	Length (bp)	Number of core loci	Adjacent or included recombining features	Number of core SNPs
B1	939	A	lpa_00900 - lpa_00975	687448 - 743613	56165	54	rRNA (5S, 16S, 23S), Ile tRNA	131
		B	lpa_02151 - lpa_02176	1691412 - 1711393	19981	17	-	149
		C	lpa_02447 - lpa_02484	1953606 - 1985949	32343	29	Adjacent transposase	50
		D	lpa_02690 - lpa_02696	2161769 - 2168758	6989	6	Leu tRNA, His tRNA, Arg, tRNA, Pro tRNA Adjacent Glu tRNA, Ala tRNA Recombinase	9
		E*	lpa_03704 - lpa_03892	2978293 - 3119309	141016	116	Pro tRNA	523
		F§	lpa_04035 - lpa_04065	3230002 - 3262184	32182	27	Met tRNA, Leu tRNA	37
B2	110	G	lpa_03530 - lpa_03532	2851732 - 2857447	5715	3	Ser tRNA	101
B3	172	H	lpa_03491 - lpa_03499	2825989 - 2833187	7198	7	-	79
		I	lpa_03374 - lpa_03439	2750696 - 2791809	41113	19	Ile tRNA Transposase, Integrase	17
		J§	lpa_04018 - lpa_04069	3220303 - 3267266	46963	40	Met tRNA, Leu tRNA Adjacents rRNA (5S, 16S, 23S), Ala tRNA	70
B4	491	K	lpa_01025 - lpa_01040	795968 - 809994	14026	8	-	86
		L	lpa_01183 - lpa_01186	940264 - 944157	3893	3	-	38
		M‡	lpa_01218 - lpa_01295	969679 - 1025828	56149	53	Adjacent Leu tRNA	248
		N*	lpa_03729 - lpa_03771	2990558 - 3018309	27751	19	Pro tRNA	117
B5	478	O‡	lpa_01207 - lpa_01289	961873 - 1021984	60111	55	Leu tRNA	395
		P	lpa_01319 - lpa_01346	1040861 - 1060781	19920	18	Ser tRNA	78

Table S7 | Recombination detection using the method described in Croucher *et al* (2011). Events per branch have been simplified as recombination hotspots longer than 1 kb and containing genes located less than 5 kb distant. Each region has been assigned to the 16 recombination events (A-P) described in the present study.

Hotspot	Start	End	Length	Genes involved	Recombination event
1	606661	648999	42339	lpa_00900 - lpa_00975	A
2	698008	702579	4572	lpa_01025 - lpa_01031	K
3	830787	833106	2320	lpa_01183 - lpa_01186	L
4	849214	861058	11845	lpa_01207 - lpa_01222	M-O
5	875244	888961	13718	lpa_01248 - lpa_01266	M-O
6	902499	905700	3202	lpa_01289 - lpa_01295	M-O
7	919617	934671	15055	lpa_01319 - lpa_01346	P
8	1490790	1505502	14713	lpa_02151 - lpa_02172	B
9	1719169	1746493	27325	lpa_02447 - lpa_02484	C
10	1781121	1782811	1691	lpa_02537	-
11	1791573	1796431	4859	lpa_02545 - lpa_02551	-
12	1903722	1908130	4409	lpa_02690 - lpa_02696	D
13	2073496	2074495	1000	lpa_02950 - lpa_02952	-
14	2130349	2131708	1360	lpa_03037	-
15	2415774	2428687	12914	lpa_03374 - lpa_03439	I
16	2458812	2462659	3848	lpa_03491 - lpa_03498	H
17	2477682	2491244	13563	lpa_03527 - lpa_03544	G
18	2590989	2592925	1937	lpa_03704 - lpa_03706	E
19	2599425	2608292	8868	lpa_03729 - lpa_03749	E-N
20	2632848	2663407	30560	lpa_03799 - lpa_03836	E-N
21	2674583	2690011	15429	lpa_03853 - lpa_03873	E
22	2761038	2778358	17321	lpa_03967 - lpa_03997	-
23	2789596	2814934	25339	lpa_04018 - lpa_04052	F-J
24	2816557	2824060	7504	lpa_04053 - lpa_04063	F-J

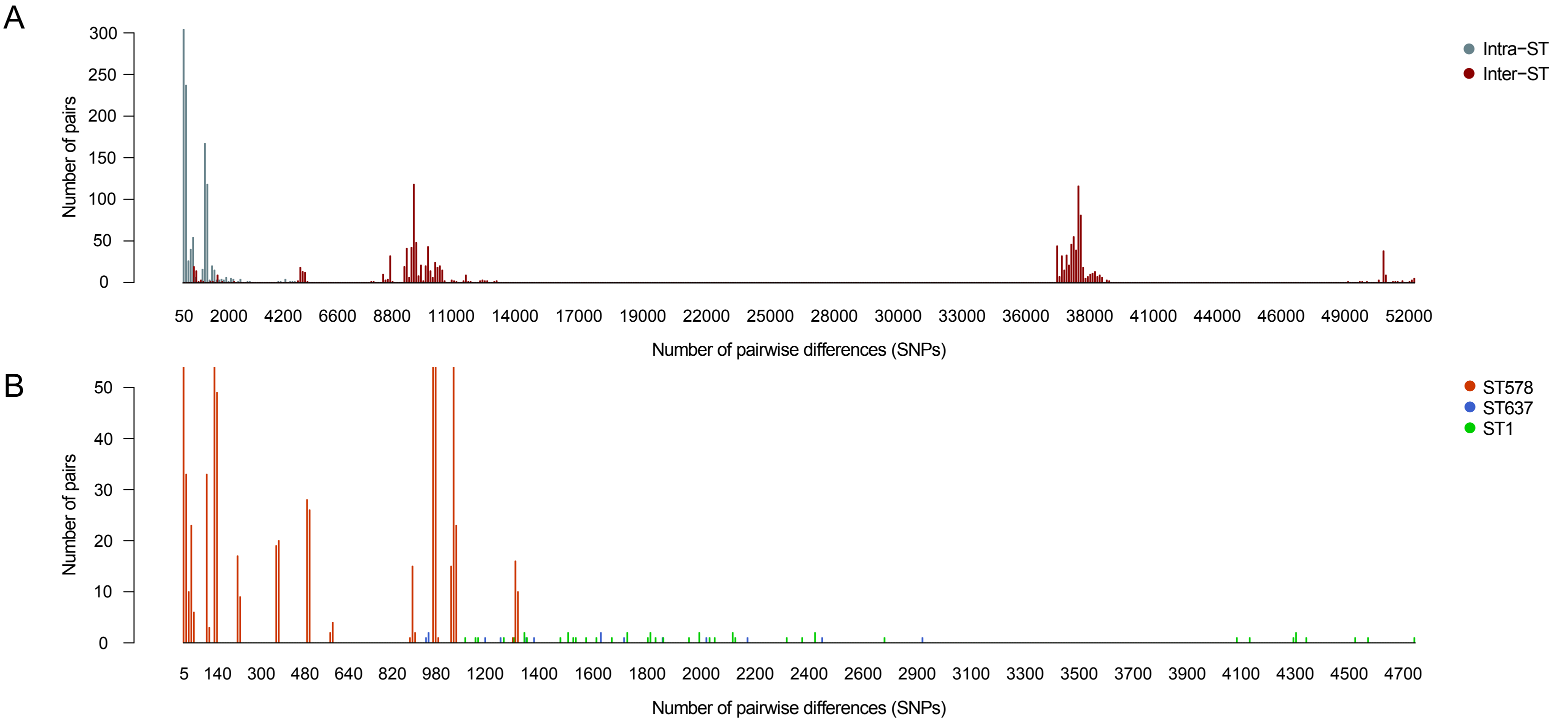


Figure S1 | Distribution of the number of core genome pairwise SNP distances between and within the *L. pneumophila* strains included in the study. (A) Intra– and inter–sequence type (ST) pairwise distances between *L. pneumophila* isolates are shown in gray and red, respectively. (B) Within-ST pairwise polymorphisms for the three most common profiles in the data set: ST578, ST637 and ST1.

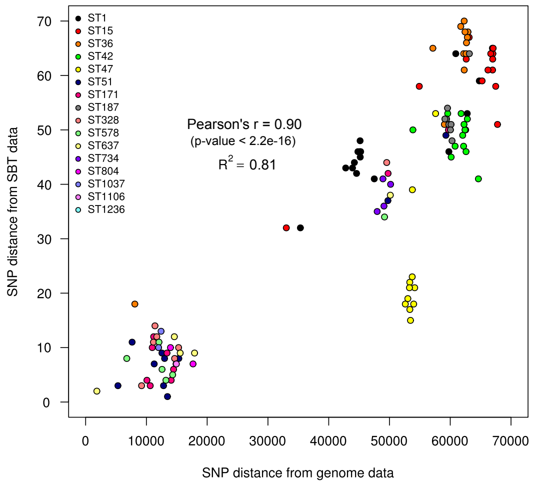


Figure S2 | Sequence-based typing (SBT) and genome-wide SNP distances. Correlation between the SNP distances found from SBT and core genome data for the 15 STs included in this study. Dots of different color show comparisons between each ST and the remainder.

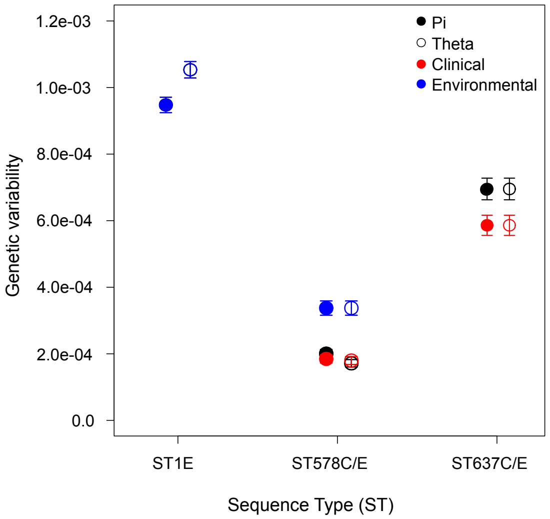


Figure S3 | Genetic variability measured by the nucleotide diversity (Pi; filled dots) and the population mutation rate (Theta; unfilled dots) between the core genome of the three most abundant STs in our data set (ST1, ST578 and ST637; with 10, 45 and 6 isolates, respectively) measured using Variscan. Red represents the estimates from clinical samples (C), blue represent the estimates from environmental samples (E) and black represents the estimates from both groups together.

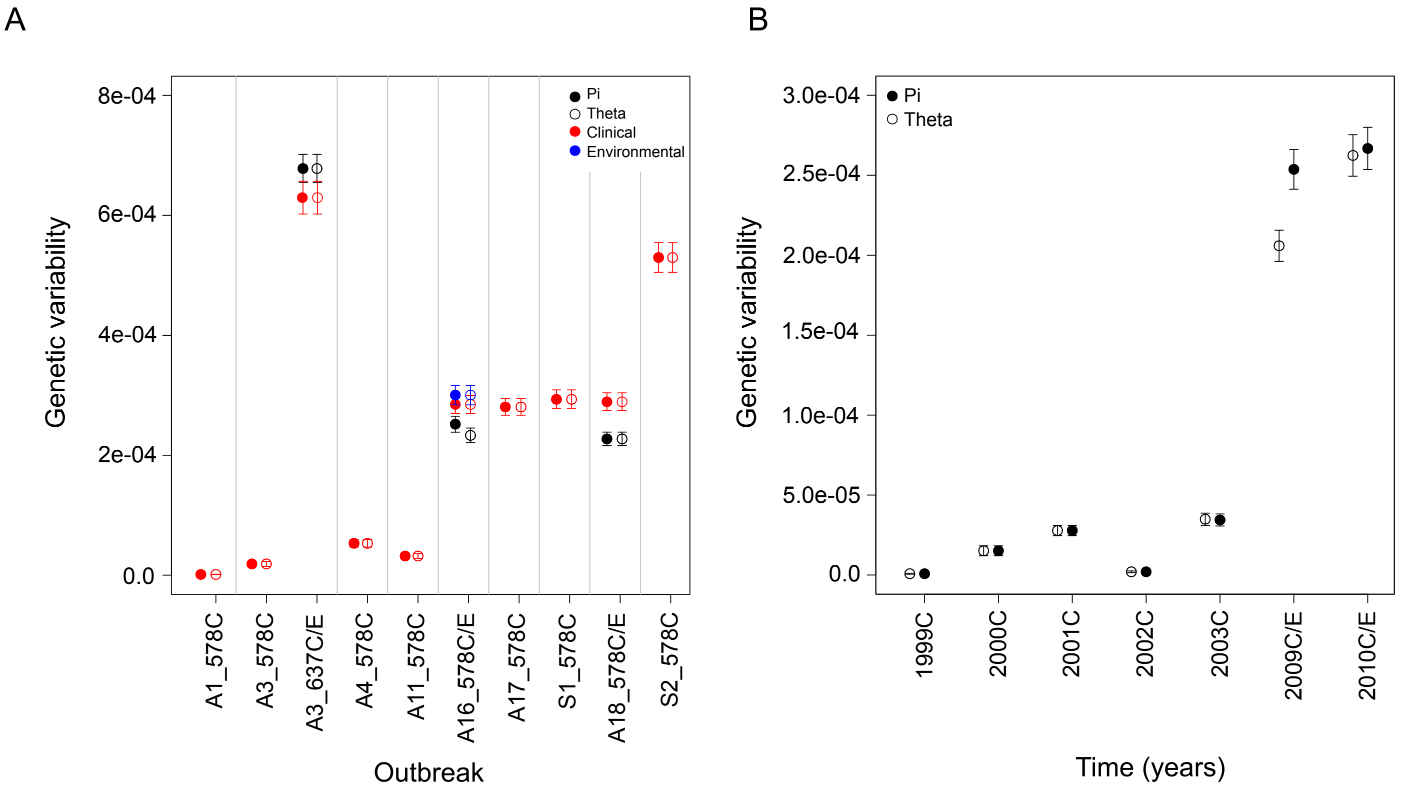


Figure S4 | Temporal variation in the genomic variability of *L. pneumophila* strains. Nucleotide diversity (Pi; filled dots) and population mutation rate (Theta; unfilled dots) calculated for the core genome using Variscan (A) within and between the ST578 and ST637 strains grouped by outbreak or sporadic cases (sorted by the time at which each outbreak occurred) and (B) between the ST578 strains isolated in different years. Red represents the estimates in clinical samples (C), blue represents the estimates in environmental samples (E) and black shows the grouping of both together.

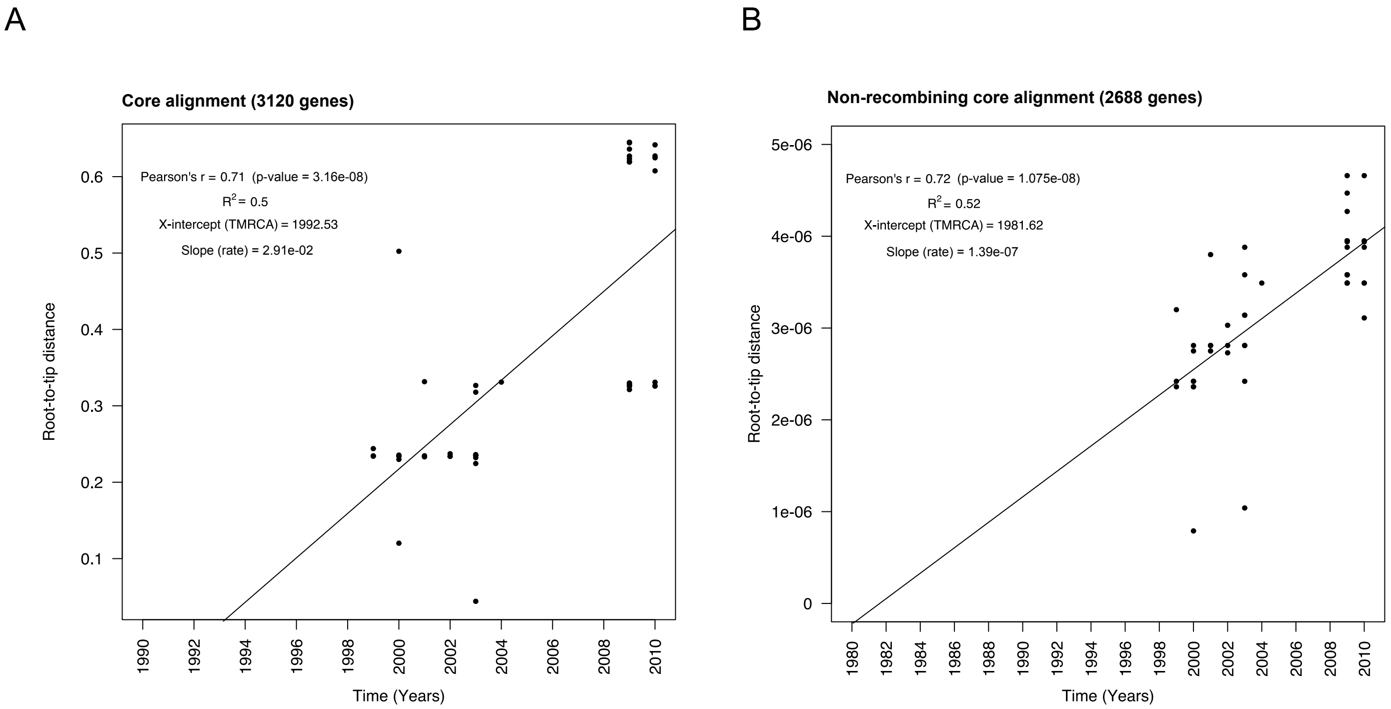
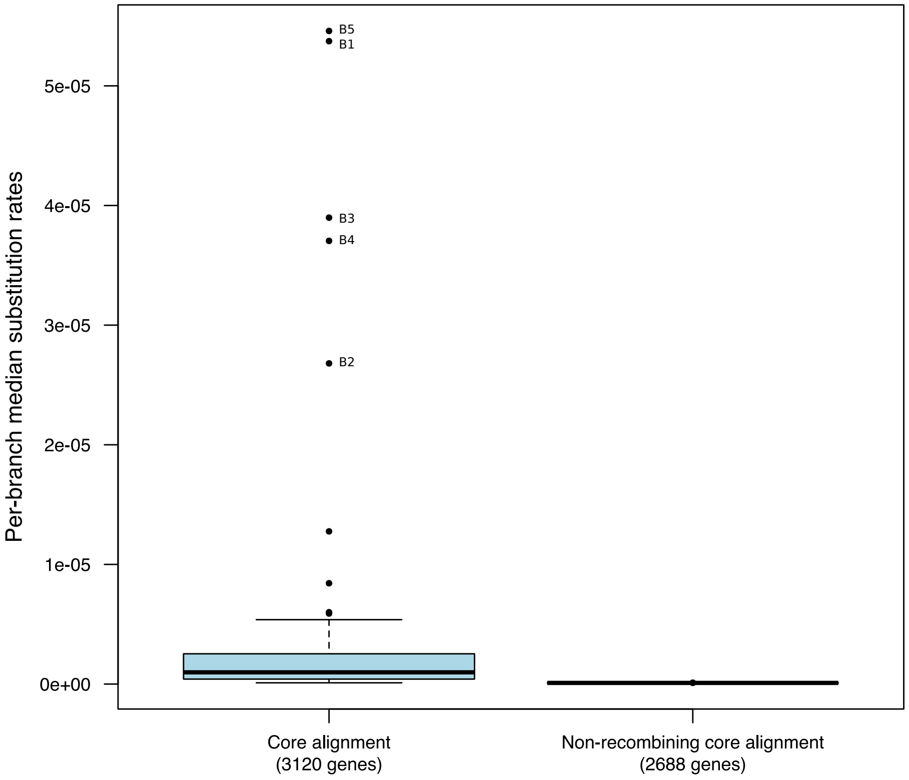


Figure S5 | Temporal signal in the data set used in the study. The correlation between the root-to-tip distances for the strains included in the study and the isolation time estimated using Path-O-Gen in the initial core alignment (a) and in the non-recombining core alignment (b). The two outlier points in b correspond to strains ID_3355 and ID_7371, which have probably acquired mutations at a slightly higher rate than the other strains present in the data set. Removing these two outliers, the correlation and R^2 parameters do not change significantly.



Mode:	3.837e-06	9.727e-08
Median:	9.810e-07	8.513e-08
Mean:	3.837e-06	9.757e-08
Min:	1.037e-07	4.298e-08
Max:	5.459e-05	1.875e-07

Figure S6 | Distribution of substitution rates. The distribution of the substitution rates per branch (evaluated from the MCC reconstruction with BEAST using the BSP and uncorrelated lognormal clock) of the ST578 strains from Alcoy is represented as boxplots for the complete core genome and the nonrecombining core genome, respectively. The B1–B5 labels for the outlier dots mark the five mainly recombinant branches found in the study. Summary statistics for the corresponding distributions are shown in the box below each plot.

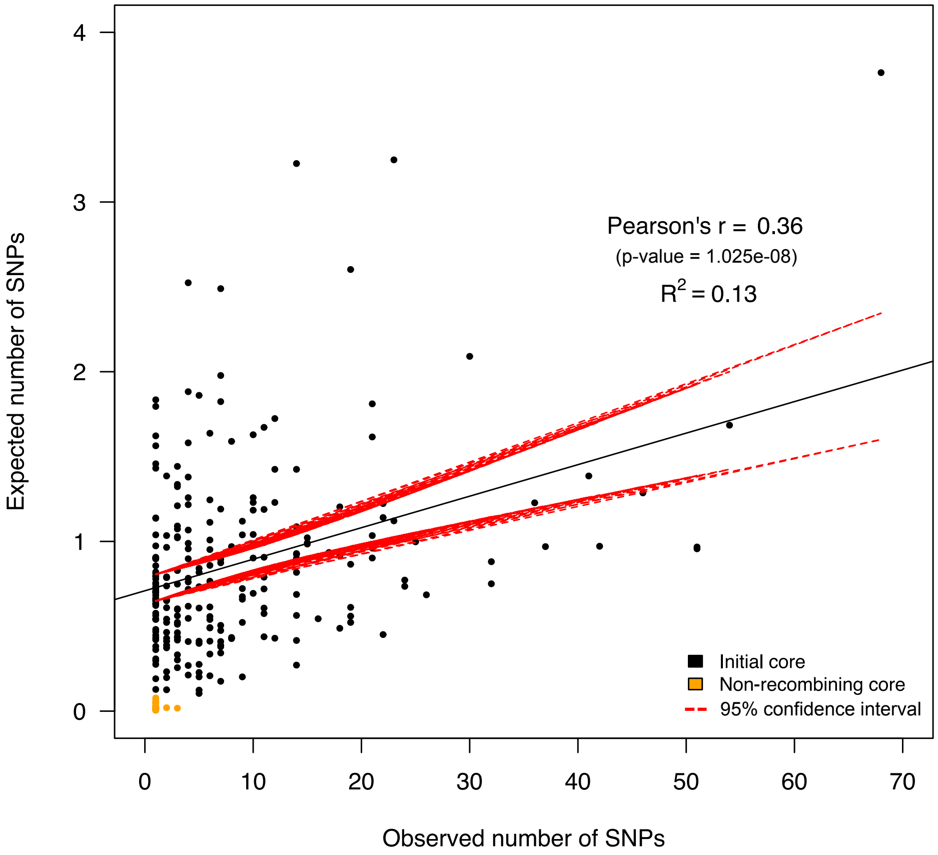


Figure S7 | Expected and observed number of SNPs per gene. Scatterplot showing the observed number of SNPs per gene against the expected number considering a random normal distribution. Red lines represent the 95% confidence interval for the regression analysis. The 3,120 genes of the initial core genome are represented as black dots, and the 2,688 genes of the non-recombining core genome are shown in orange.

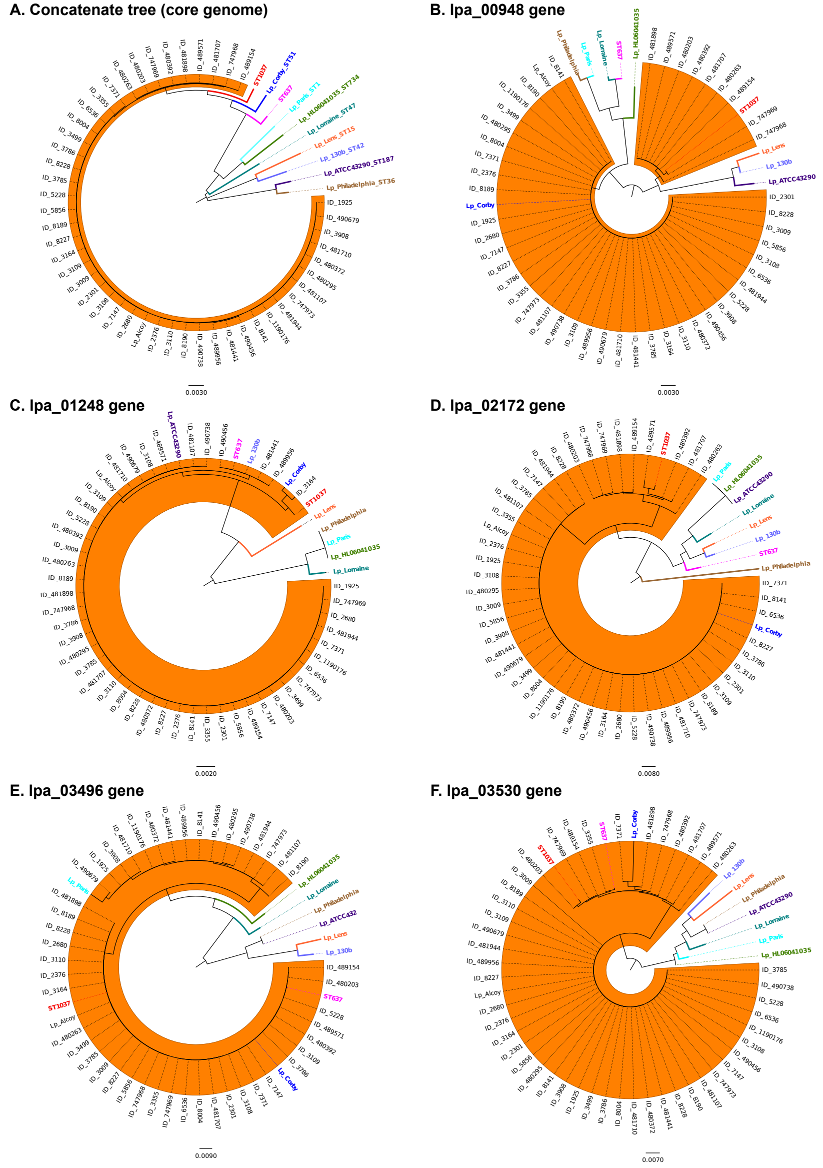


Figure S8 | Examples of topological incongruence in recombinant genes. Gene trees were compared topologically to the concatenate of the whole-core alignment. Sequences from other STs were included to account for recombination with external strains and are shown in different colors. (A) The topology of the concatenate. (B–F) The topologies of five different genes detected as recombinant. Note the external STs interspersed within the ST578 population.

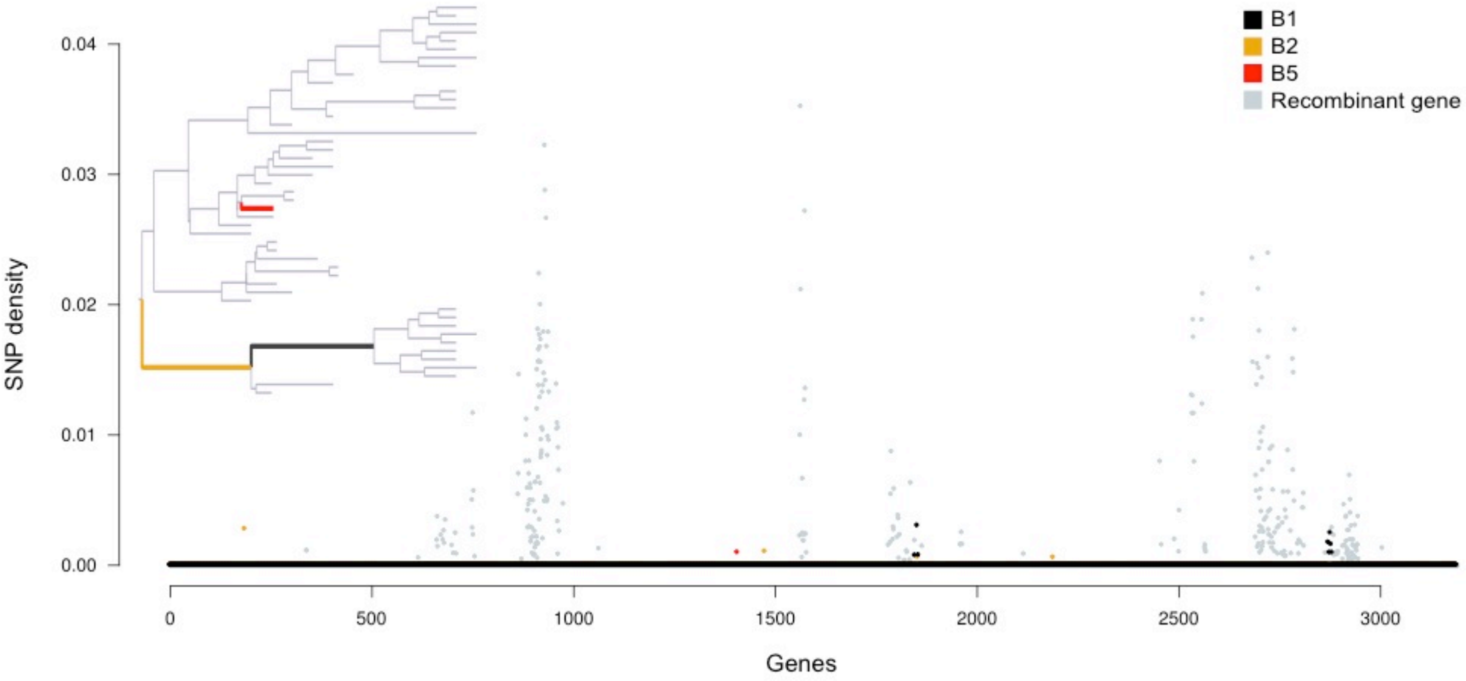


Figure S9 | Non-recombining core per-gene SNP density. The SNP density per non-recombinant gene is shown for branches B1, B2 and B5 taking into account the ancestral SNPs traced using Mesquite and the corresponding gene lengths. Branches B3 and B4 are not present in the topology of the 2,688-gene non-recombining MCC BEAST tree and thus are not represented in the plot. The 432 genes detected as recombinant are represented in the back as gray points for comparison purposes.

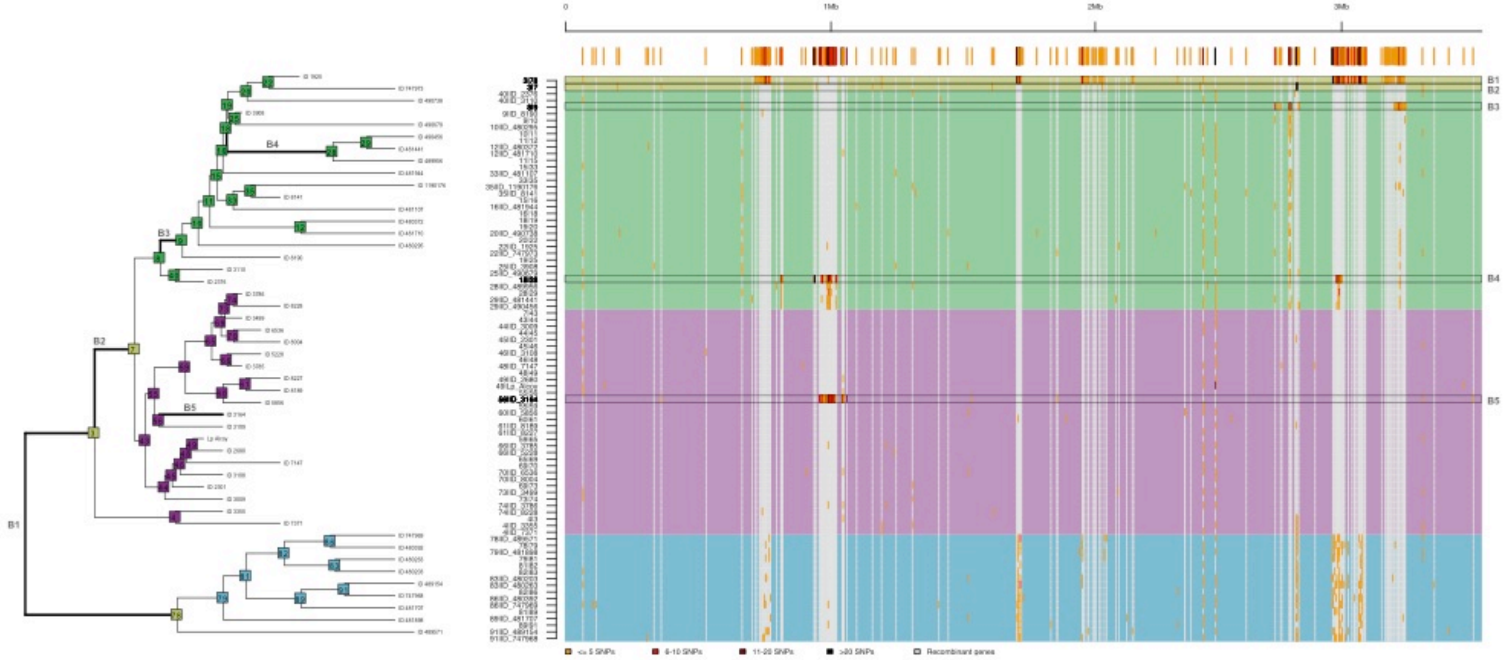


Figure S10 | Representation of the number of SNPs per gene per branch in the 3,120-gene ST578 core tree. The tree on the left contains the node numbers that identify the different branches (the notation for the five branches, B1–B5, that accumulate more SNPs is shown in gray). The graph on the right shows the total number of SNPs per gene as the header. The following rows represent the number of SNPs per gene in each of the branches. Background colors represent the clades in the tree where the corresponding branches are present. The five branches on which this study focuses are marked in bold with black rectangles. Gray vertical lines represent the 247 genes detected with a non-vertical phylogenetic signal.

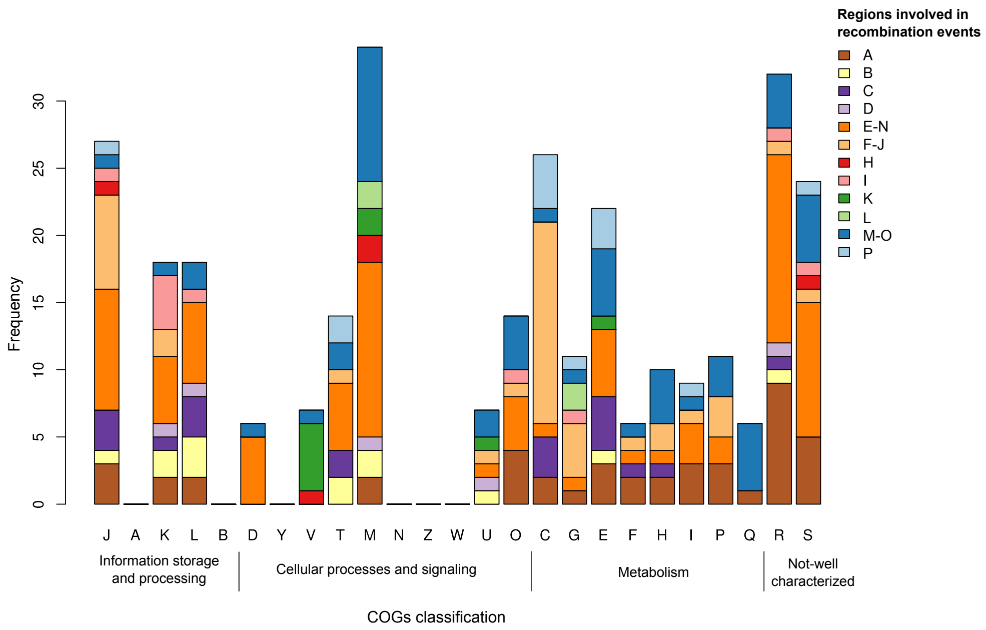


Figure S11 | Functional classification of the genes involved in recombination events. The distribution of the function of the genes involved in the detected recombination events over Cluster of Orthologous Genes (COGs) classification. These genes include virulence factors involved in the type IV secretion system (*sidA*, *sidF*, *lepA* and *lepB*) and genes involved in invasion and motility (*enhA*, *enhB*, *enhC* and *mviN*), regulation (*letA*) and iron acquisition (*feoA* and *feoB*). Moreover, with recombination events B and C, ST578B could have incorporated SNPs in genes involved in DNA damage repair and mutagenesis, such as *mutH*, *umuC* and *umuD*. Letters on the x axis correspond to the functional categories described in ftp://ftp.ncbi.nlm.nih.gov/pub/COG/COG/fun.txt.

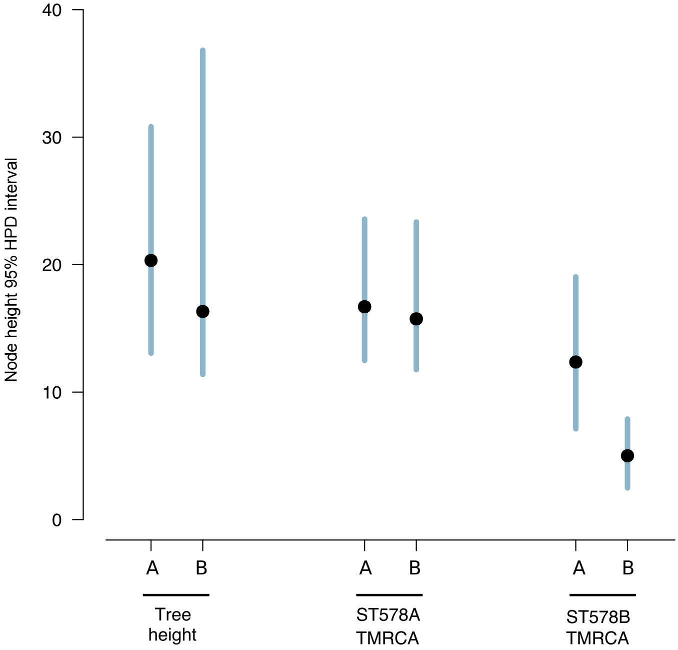


Figure S12 | Node height highest posterior density (HPD) intervals at 95%. Comparison between the mean time to MRCA estimates (black dots) and the 95% HPD intervals (blue dots) before (A) and after (B) removing recombination events for the whole ST578 tree and the two sublineages ST578A and ST578B independently.

Chapter 6

Metagenomic analysis of environmental biofilms from natural springs

Leonor Sánchez-Busó[1,2], Juan Miguel Calafat[3], Francisco Adrián[3], Iñaki Comas[1,2], Fernando González-Candelas[1,2].

1. Unidad Mixta Infección y Salud Pública FISABIO/Universitat de València. Avenida Cataluña, 21; 46020, Valencia, Spain.
2. CIBER en Epidemiología y Salud Pública, Valencia, Spain.
3. Centro de Salud Pública de Alcoy. DGSP. Alcoy, Spain.

Work in preparation for submission.

Abstract

Biofilms from natural water sources can act as reservoirs for a wide variety of microbial and eukaryotic communities. Some of the microorganisms that benefit from living in these ecosystems may be harmful for humans and result in risks for public health. High throughput sequencing approaches have the potential for fully characterizing and measuring the composition of those complex communities.

In this work, we describe a novel approach to characterize the microbial, viral and eukaryotic communities of 14 environmental biofilms sampled in natural springs around the locality of Alcoy (Alicante, Spain). The communities were dominated by *Bacteria*, with *Proteobacteria*, *Cyanobacteria, Actinobacteria* and *Bacteroidetes* phyla*,* representing 97% of the bacterial abundance. Frequent inhabitants of soil and freshwater ecosystems such as *Porphyrobacter, Xanthomonas, Rhodobacter* or *Sphingomonas* were the most common genera. *Halobacteria*, fungi, parasitic protists, algae and microviruses were found in the minority fraction. Despite the known endemism of the opportunistic pathogen *Legionella* in this area, it was only found in a relative abundance ranging 0.01-0.07%.

Metagenomic approaches rely on databases dependent on sequencing efforts. As a consequence, different methodologies for taxonomy assignment can provide different results. A comparison of our results with other existing methods (original Last Common Ancestor (LCA) algorithm, 16S rDNA, and MetaPhlAn) reveals the inherent bias associated to the databases they rely on.

Introduction

Many microorganisms live in communities. The relationships among them can be mutualistic or parasitic but, in any case, they usually involve physical interactions. These can be achieved by the formation of complex structures supported by a polysaccharide matrix called biofilms[37]. These are usually attached on many different water-related surfaces[31] and can result in a risk for public health as they can serve as reservoirs for pathogens[30,32].

Sequencing efforts based on high throughput strategies have allowed the study of different microbial communities, mainly related to the human body[303]. These methodologies are being increasingly applied to environmental samples, especially since the first in-depth analyses performed over the Sargasso Sea[316]. Projects such as the Global Ocean Sampling (GOS)[315] or TARA Oceans[301] introduced many of the current knowledge on the microbial community of oceanic and freshwater microbiota. Environmental biofilms from natural sources have been underrepresented in these studies, with most efforts focused on drinking-water biofilms under different conditions or pipe materials[326,334,470,471].

High throughput approaches are essential to get a good representation of the overall community[472,473]. But, even with high sequencing depths, the databases on which these metagenomic studies rely as well as the algorithms used are essential for extracting taxonomic information on the corresponding community. Genome databases are being used increasingly because they allow the correction of the counts assigned to specific organisms[312,474]. However, current sequencing projects do not fairly represent all the existing environmental diversity despite the large-scale genomic projects that aim at having a better representation of the tree of life[475]. Even so, it is very usual to find a high percentage of unassigned reads, which have been hypothesized to come from a yet undiscovered life domain[310].

In this study, we aimed at characterizing the microbial, viral and eukaryotic communities of biofilms sampled in canalization tubes of natural springs. These springs are located in the vicinity of Alcoy (Alicante, Spain), where *Legionella pneumophila* is known to have colonized the water distribution system[339]. In fact, legionellosis outbreaks and sporadic cases have been relatively frequent in this area. As previous efforts have found its detection in the area difficult[352], our aim includes the detection and quantification of the genus *Legionella*, and if possible, the ubiquitous *L. pneumophila,* in the sampled community.

Additionally, we propose a modified version of the Last Common Ancestor (LCA) algorithm, used by many metagenomic studies, for taxonomic classification. This proposal overcomes the assignment of reads to higher levels because of lateral gene transfer events or high gene conservation. This methodology is compared to the original LCA algorithm implemented in BLAST2LCA (see URLs), 16S rDNA extracted from the metagenomic reads and MetaPhlaN[476], showing the role of database composition in producing biased results in metagenomic analyses.

Materials and methods

Sample collection and processing

Fourteen biofilm samples were taken from canalization tubes of natural springs around the city of Alcoy (Alicante, Spain) in October 2013 (Fig. S1). Biofilm surfaces were gently scraped with a plastic spatula and samples were transported to our laboratory in 15 mL-Falcon tubes at 4ºC.

Part of the available supernatant (approximately 500 μL) was used for culturing in *Legionella*-specific GVPC media the same day of the sampling. Cultures were incubated at 37ºC during 10 days. *L. pneumophila*-like colonies were sub-cultured in BCYE with and without L-cysteine.

Approximately 0.20 g of biofilm material was taken from each sample and used for total DNA extraction using PowerBiofilm® DNA Isolation Kit (MoBio Laboratories, Inc.) following the manufacturer's instructions. RNA was removed incubating the extracted genetic material with 20 mg/mL RNAse A for 30 minutes at 37ºC. Purification of DNA was performed by ethanol precipitation using 3M sodium acetate (pH 5.2). After an evaporation step using Savant DNA SpeedVac (Thermo Scientific), DNA was suspended in nuclease-free water. DNA quantity and quality were tested using NanoDrop™ 1000 (Thermo Scientific) and Quant-iT™ PicoGreen® (Invitrogen). From 1 to 6 μg of total DNA were retrieved per sample and used for multiplex shotgun paired-sequencing with Illumina® MiSeq 2x300 bp.

Reads pre-processing

Paired-end reads were matched using FLASh v1.2[477]. Unpaired reads were trimmed dynamically from 3' ends to remove bases below Q20. Assembled pairs and those that remained unpaired were posteriorly filtered using prinseq_lite[478] to remove reads shorter than 100 bp and with an average quality below Q25. Pre- and post- processing of reads qualities were evaluated using FastQC v0.10.1 (see URLs).

Creation of a genome database

A BLAST database was created using 17,314 complete genomes and whole genome sequencing projects retrieved from NCBI (see Supplementary Note 1). In total, the database included 8,539 different species (3,159 bacteria, 382 archaeal, 4,544 viral and 454 from low-eukaryotes) (Fig. S2; Tables S1, S2 and S3). The complete taxonomic levels and the genome/draft-genome size were retrieved for each of the strains in order to create a database index. The average genome size for each species (n=8,539), genus (n=1,701) family (n=534), order (n=226), class (n=105) and phylum (n=65) in each taxonomic level was calculated for correction purposes.

Taxonomic classification

Determination of the taxonomic classification was performed from filtered reads using BLASTn v2.2.28+[451] against the genome database using an E-value threshold of < 1E-05 and an alignment length of at least 100 bp. BLAST output was filtered in order to leave only hits with a bitscore in the top 15% for each query. The complete taxonomic information for each of the hits from a read was retrieved using the Taxonomy database from NCBI (see URLs).

Assignments for all the hits from the same read were subjected to a modified version of the LCA algorithm (named Taxonomic Vote, TV) starting at the species level using in-house scripts in Perl and R[344]. The complete workflow is detailed in the Supplementary Note 2 (Fig. S3).

Reads with no hits against the genome database were blasted to the non-redundant nucleotide database (nr-nt) from NCBI. The same restrictions detailed above were applied to the output file. The transfer from GI identifiers to taxonomic assignation was also performed using the Taxonomy database from NCBI (see URLs) and the final assignment made using the TV method.

Diversity and richness estimates

Diversity estimates were calculated from the abundance data retrieved from the TV method using the “vegan” package in R[479]. Specifically, rarefaction curves[480] were calculated to account for the influence of sample size in the characterization of the community. The coverage of the true community composition was calculated using Good's estimate with the formula [1-(n/N)]x100, considering *n* the number of genera represented with just one read and *N* the total number of sequences in that sample[481]. Inverse Simpson[482,483], Shannon[484] and Chao1[485] indices were calculated for each sample in order to measure diversity and richness.

Jaccard and Bray-Curtis distances were calculated to cluster the biofilm samples with the same R package. In order to study a potential correlation between microbial relative abundances and geographical location, a Mantel test was performed using Pearson's correlation method and 1,000 permutations to assess statistical significance.

Comparison to other methods

Results from the taxonomic assignment of each read with the TV method were compared to those from other available methodologies. Firstly, bitscore-filtered tabular BLAST output files were processed by the command line-based software BLAST2LCA (see URLs). This program implements a strict LCA algorithm, in which 100% of the hits from a specific read must be concordant to each other at a specific level to perform a taxonomic assignment.

Secondly, reads from the small subunit of ribosomal DNA were extracted from the shotgun sequences using Metaxa v1.1.2[486]. Taxonomic assignment of 16S and 18S reads was performed with the *classify.seqs* instruction in Mothur v1.33.3[487] using the SILVA 16S/18S database[488] containing bacterial, archaeal and eukaryotic sequences. Information in the ribosomal RNA Operon Copy Number Database (rrnDB)[489] was used to correct the number of reads assigned to each taxonomic level by the number of rDNA copies in the corresponding genomes. If information of a specific genome was not available, the average number of copies from the immediate upper level contained in the database was used.

Finally, raw filtered reads were tested against clade-specific markers using MetaPhlAn v1.7.8[476]. The program was run using Bowtie2 v0.12.7[490] and the corresponding markers database set by the authors.

Results

Taxonomic assignment of shotgun reads

Shotgun sequencing from total DNA extracted from the biofilms output from 1.5 to 2.5 million of 300-bp paired-end reads per sample (Table S4). After pairing and filtering, samples were blasted against the custom reference genome database. The total number of reads assigned to any taxonomic level ranged from 10.6% to 33.6%. In order to search for matches against organisms not present in our database because no complete genome or genome project was available, unassigned reads were blasted against the NCBI non-redundant nucleotide database. The percentage of reads with a taxonomic assignment after the second search increased only an average of 0.5% in all samples except N2, which had an increase of 13% of assigned reads. Eventually, an average of 78% reads (range 63.5-89.1%) could not be assigned to any specific organism or taxonomic level (Table S4).

Description of the biofilm metagenomic community

The community of all samples was found dominated by *Bacteria* (range 97.9-99.5%). *Archaea* and low *Eukaryotes* were found in very low abundance (range 0.02-0.52% and range 0.03-0.67%, respectively), as well as *Viruses* (range 0.36-1.24%) (Fig. S4).

Thirty-five different bacterial phyla were found in all the samples, although only four of them explained 97% of the bacterial cumulative abundance (Fig. S5). This was found dominated by *Proteobacteria* (range 68.2-90.4% of the reads assigned to *Bacteria*), *Cyanobacteria* (range 0.7-22.1%), *Actinobacteria* (range 1.5-19.0%) and *Bacteroidetes* (range 0.8-17.1%). However, fluctuations of approximately 20% were found among samples. Principal component analysis on the bacterial relative abundance data showed most of the variance explained by the high proportion of *Proteobacteria* in all samples (Fig. 1A). However, PC2 showed a deviation of samples N2 and N9 towards *Cyanobacteria* (22.1% and 18.96%, respectively), samples N8 and N25 towards *Actinobacteria* (16.1% and 19.0%, respectively) and sample N4 towards *Bacteroidetes* (17.1%) (Fig. 1B).

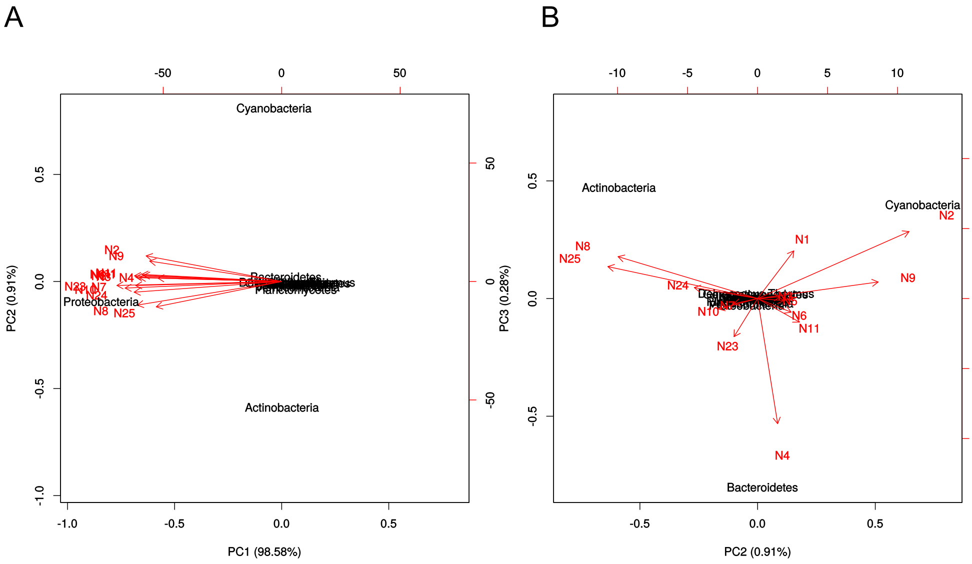


Figure 1 | Principal component analysis of the bacterial phyla relative abundances found in the biofilm metagenomes. (A) shows principal components 1 and 2, (B) shows principal components 2 and 3. The percentage of variance explained by each component is shown in the corresponding axes.

A similar composition of the bacterial community was also found at the class level, with 52 different clades. *Alpha-Proteobacteria* was the most abundant class, although some fluctuations were also seen (range 20.9-75.9%) (Fig. 2A). The main differences among the different samples were found at the genus level. Results showed 927 different genera predicted from the reads with taxonomy assigned to *Bacteria*, 176 of which had a relative abundance higher than 1%. N2 and N9 were found enriched in *Xanthomonas* (21.4% and 10.7%, respectively), whereas N10 and N23 were found enriched in *Porphyrobacter* (18.2% and 26.4%, respectively) (Fig. 2B and 3).

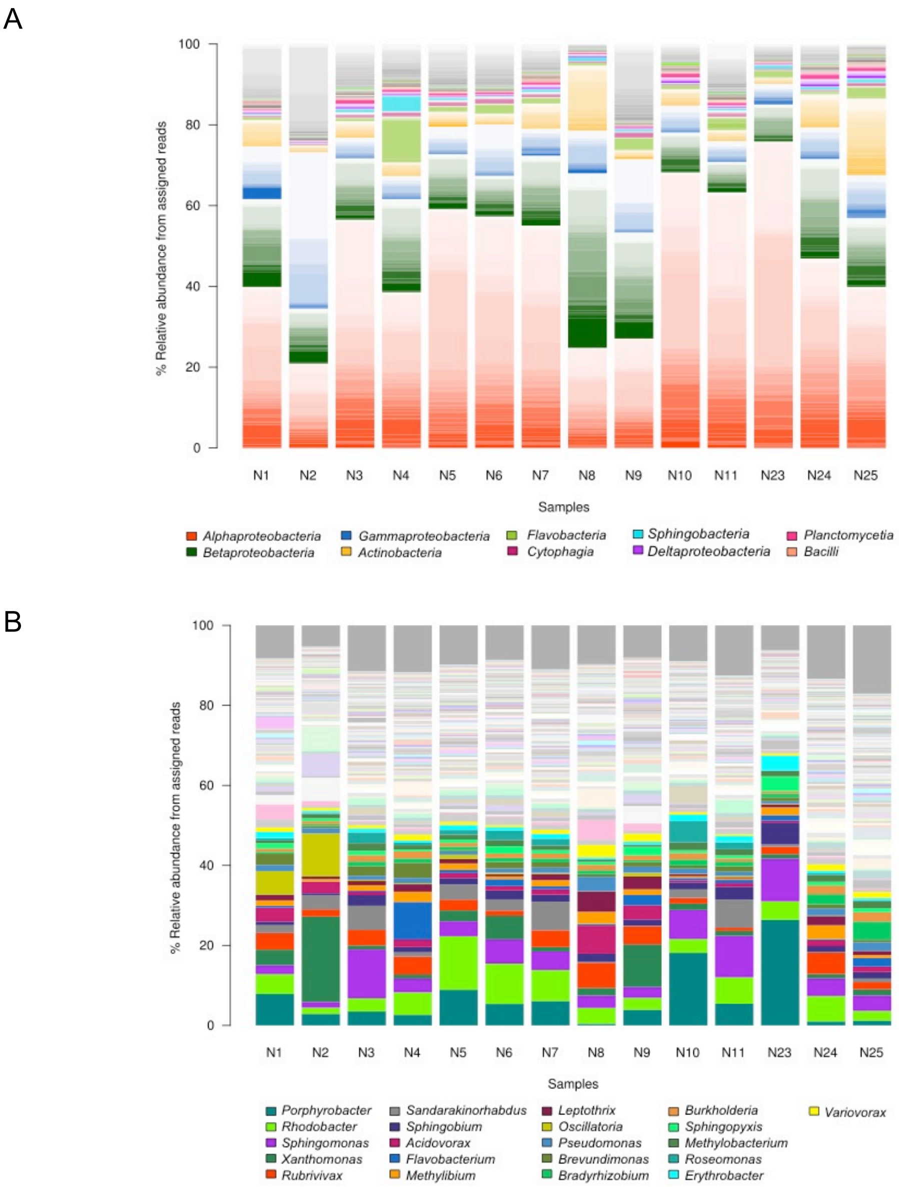


Figure 2 | Relative abundances of all the bacterial classes (A) and genera (B) detected in the metagenomes. (A) Class level: colours represent different levels above 1% of bacterial abundance estimated from assigned reads. Degradation shows the amount of different genera detected in the corresponding class. Legend shows the ten most abundant classes. (B) Genus level: dark colours represent 21 genera that explain 50% of the bacterial cumulative abundance. Transparent colours show other genera above 1% of relative abundance. The grey portion represents all the levels under the 1% fraction of relative abundance.

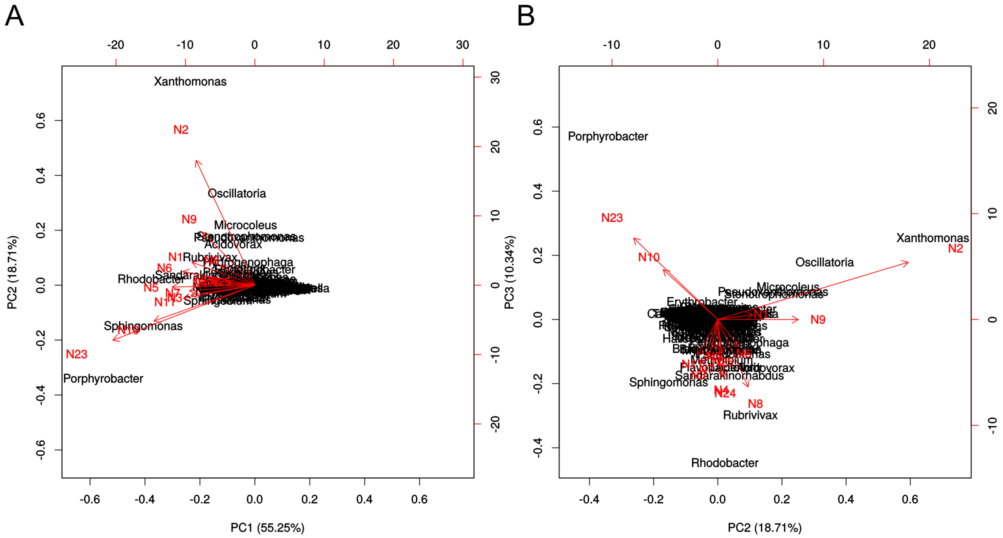


Figure 3 | Principal component analysis of the bacterial genera relative abundances above 1% found in all biofilm metagenomes. (A) shows principal components 1 and 2, (B) shows principal components 2 and 3. The percentage of variance explained by each component is shown in the corresponding axes.

The archaeal community detected included members of 3 different phyla, *Euryarchaeota* (range 31.5-98.2%), *Crenarchaeota* (range 0.0-5.7%) and *Thaumarchaeota* (range 1.6-68.4%), 8 classes and 80 genera. *Halobacteria* (range 26.5-79.1%) and *Methanomicrobia* (range 4.5-28.6%) were found as the most abundant archaeal classes in all samples (Fig. 4A). Genera from the *Halobacteria* class were found to represent more than 50% of the archaeal abundance (Fig. 4B). A similar composition was found with some particularities, such as the high abundance of the ammonia-oxidizing soil archaeon *Candidatus Nitrosoarchaeum* in N24 (64.5%) or *Halorubrum* in N10 (23.1%) (Fig. 5).

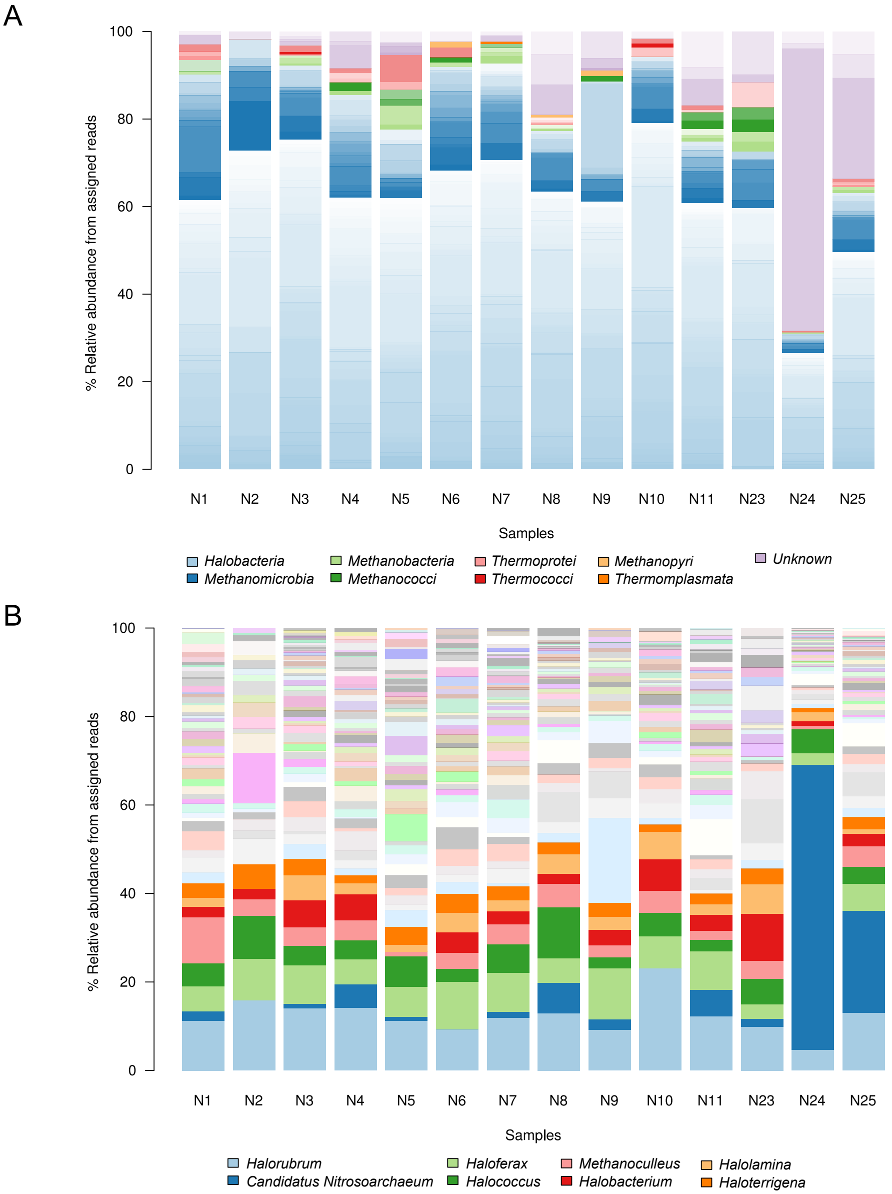


Figure 4 | Relative abundances of all archaeal classes (A) and genera (B) detected in the biofilm metagenomes. (A) Class level: colours represent the different levels estimated from assigned reads. Degradation shows the amount of different genera detected within the corresponding class. (B) Genus level: dark colours represent 8 genera that explain 50% of the bacterial cumulative abundance.

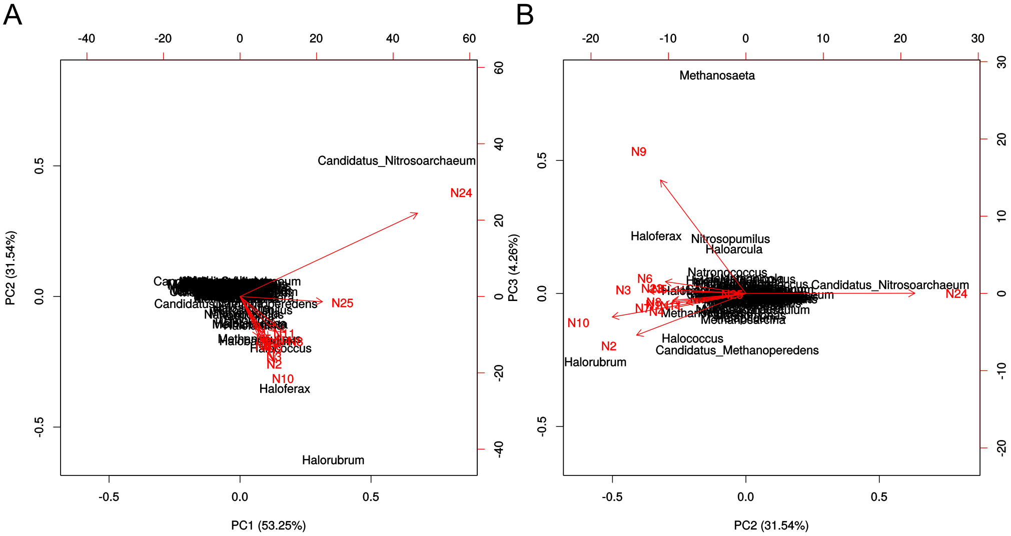


Figure 5 | Principal component analysis of the archaeal genera relative abundances above 1% found in all biofilm metagenomes. Panel (A) shows principal components 1 and 2, and panel (B) shows principal components 2 and 3. The percentage of variance explained by each component is shown in the corresponding axes.

The eukaryotic composition of the biofilms was found to be highly diverse, with 16 different phyla found, 5 of which represented 95% of the cumulative eukaryotic abundance (Fig. S5). The most abundant eukaryotic phyla were *Apicomplexa* (range 16.7-45.2%), *Ascomycota* (14.7-44.4%), *Basidiomycota* (range 0.1-10.7) and *Chlorophyta* (range 3.2-10.58) (Fig. 6). More specifically, 43 different classes including 253 genera were detected, 136 of them had a relative abundance above 1% of the reads assigned to low *Eukarya* (Fig. 7).

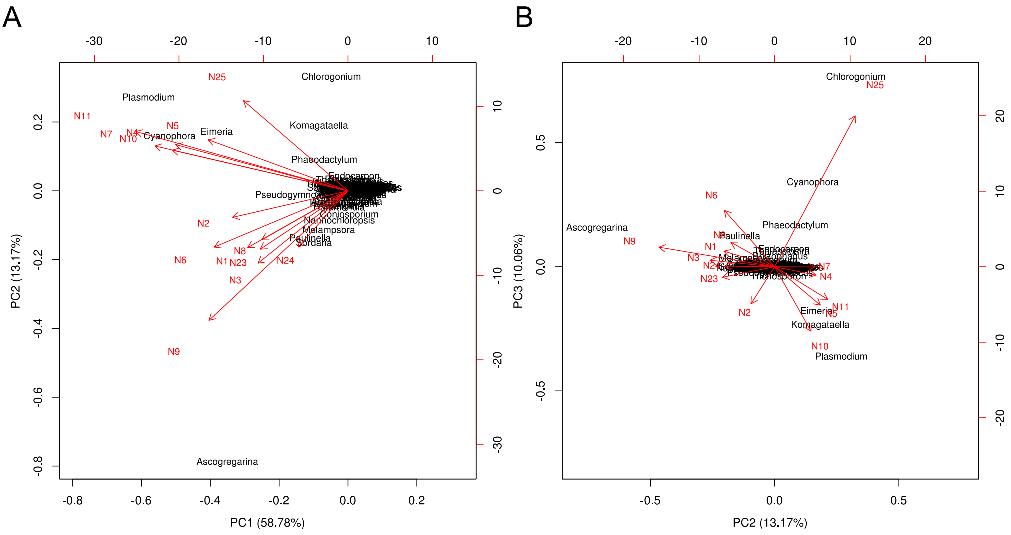


Figure 6 | Principal component analysis of the eukaryotic phyla relative abundances found in the biofilm metagenomes. Panel (A) shows principal components 1 and 2, and panel (B) shows principal components 2 and 3. The percentage of variance explained by each component is shown in the corresponding axes.

Fungi and Apicomplexan genera were found to dominate the eukaryotic community (Fig. 7B). More specifically, genus *Chlorogonium* was only found in N25 (29.9%) and genus *Ascogregarina* was specially enriched in sample N9 (26.3%) (Fig. 8).

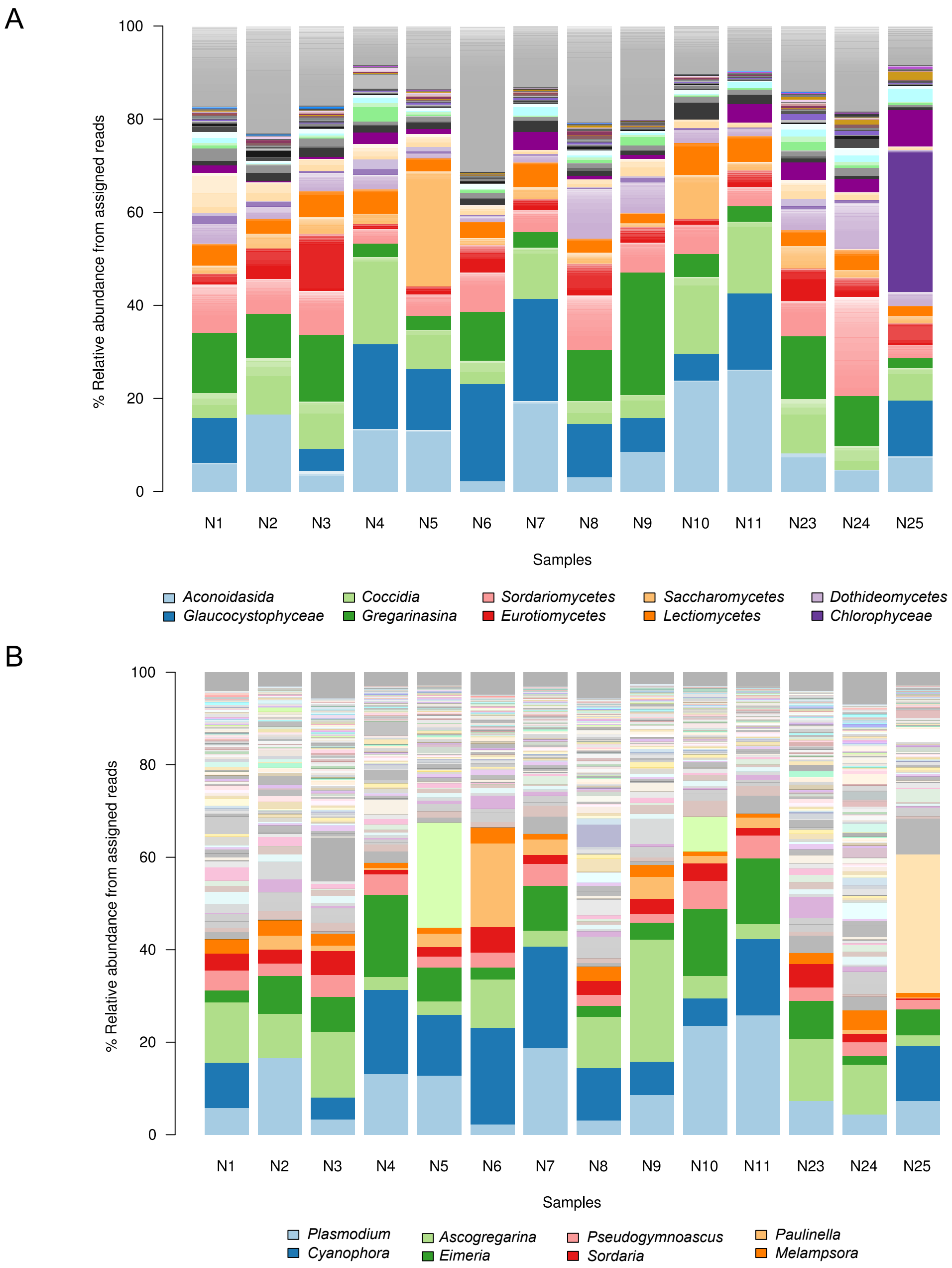


Figure 7 | Relative abundances of all eukaryotic classes (A) and genera (B) detected in the metagenomes. (A) Class level: colours represent different levels above 1% of eukaryotic abundance estimated from assigned reads. Degradation shows the amount of different genera detected within the corresponding class. Legend shows the ten most abundant classes. (B) Genus level: dark colours represent 8 genera that explain 50% of the eukaryotic cumulative abundance. Transparent colours show other genera above 1% of relative abundance. The grey portion represents all the levels under the 1% fraction of relative abundance.

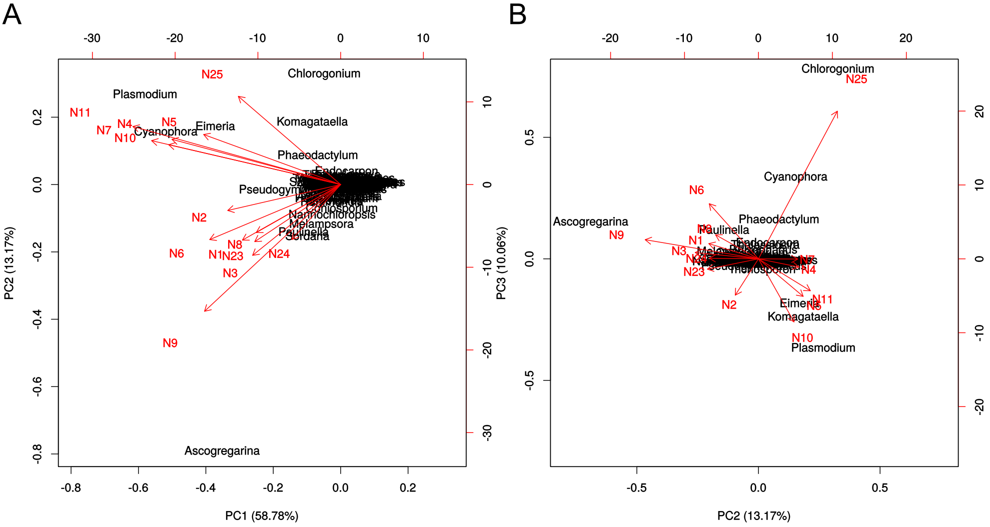


Figure 8 | Principal component analysis of the eukaryotic genera relative abundances above 1% found in all biofilm metagenomes. Panel (A) shows principal components 1 and 2, and panel (B) shows principal components 2 and 3. The percentage of variance explained by each component is shown in the corresponding axes.

Most of the viruses detected in all the samples were assigned as *Microvirus* (range 78.6-100% of the reads assigned to Viruses) (Fig. 9).

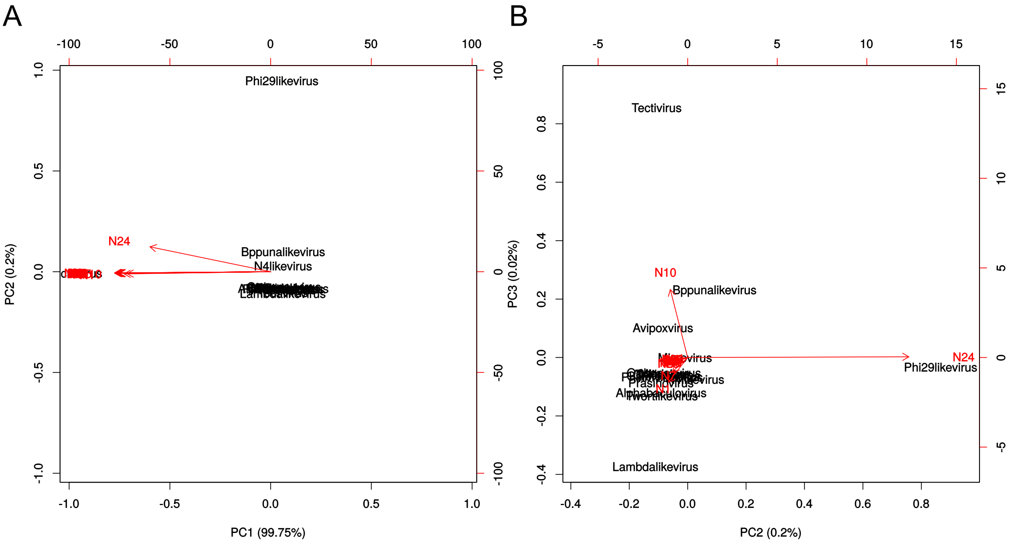


Figure 9 | Principal component analysis of the viral genera relative abundances in all biofilm metagenomes. Panel (A) shows principal components 1 and 2, and panel (B) shows principal components 2 and 3. The percentage of variance explained by each component is shown in the corresponding axes.

Reads unassigned to the reference genome database

An average of 78.7% reads remained unassigned after blasting against the reference genome database. These were re-tested against the more general NCBI non-redundant nucleotide database but, on average, less than 1% of the unassigned reads could be identified using this second strategy, except for sample N2, as stated above. Depending on the sample, more reads were assigned to either *Bacteria* (range 0.2-16.7%) or *Eukarya* (0.06-0.63%)*. Archaea* and *Viruses* remained the least represented (range 0.0001-0.0026% and 0.0001-0.0808%, respectively).

Thirty-five bacteria genera were found, 29 of them previously identified in the TV. However, other non-represented genera were detected, such as *Hymenobacter, Neorhizobium, Roseibacterium, Marichromatium, Kutzneria* or *Sulfuritalea.* Interestingly, samples N1 and N2 showed a significant deviation towards *Roseibacterium* (17.4% and 33.7% of the new assignments, respectively). Samples N8 and N25 showed certain bias towards *Marichromatium* and *Hymenobacter* (Fig. 10).

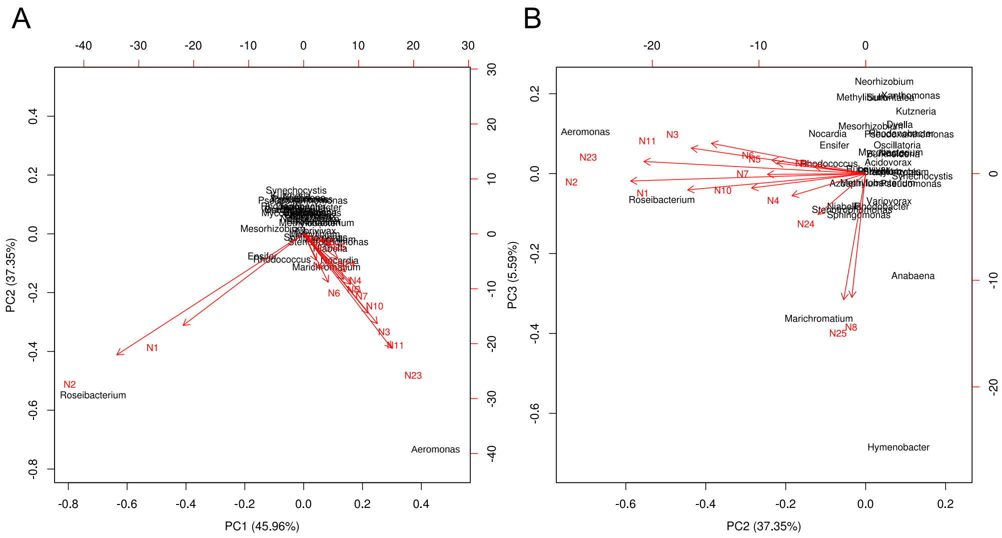


Figure 10 | Principal component analysis of the reads assigned to *Bacteria* after blasting the unassigned reads from the TV against the NCBI non-redundant nucleotide database. Panel (A) shows principal components 1 and 2, and panel (B) shows principal components 2 and 3. The percentage of variance explained by each component is shown in the corresponding axes.

The most represented eukaryotic genera were high eukaryotes such as the freshwater planarian *Schmidtea* (20.2% of the reads assigned in all samples), the plant genera *Vitis* (8.9%), *Ricinus* (6.8%) and *Gossypium* (6.2%) or the arthropod *Culex* (5.6%) (Fig. 11).

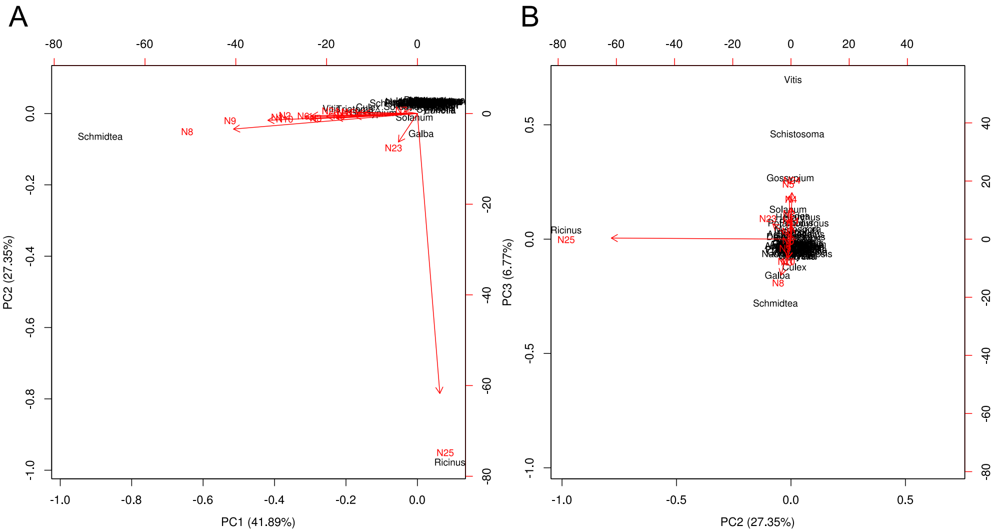


Figure 11 | Principal component analysis of the reads assigned to *Eukarya* after blasting the unassigned reads from the TV against the NCBI non-redundant nucleotide database. Panel (A) shows principal components 1 and 2, and panel (B) shows principal components 2 and 3. The percentage of variance explained by each component is shown in the corresponding axes.

Taxonomy diversity and richness

Rarefaction curves performed at the genus level revealed that sequencing depth was sufficient to get a good characterization of the communities under study (Fig. S6). Good's estimator of coverage[481] was estimated as 99.9% in all samples, which means that we are seeing a picture of almost all of the genera present in the analysed specimens. Inverse Simpson[482,483] and Shannon-Weaver[484] estimates were used to evaluate the taxonomic diversity of the biofilm samples. The inverse Simpson index revealed differences between samples, some of them with less number of genera with a comparable relative abundance (13.9 and 10.9 genera, respectively; Table S5) and others with a higher number (62.0 and 81.8 genera in N24 and N25, respectively).

However, as this index is biased towards the presence of a few abundant species, Shannon's index of diversity was also calculated. This estimator reduces the weight assigned to the dominant species, giving more importance to the rare ones. In this case, Shannon index was estimated to be more homogeneous between samples, with values ranging from 3.75 to 5.16 (Table S5). However, transformed to true diversity, these values refer to the presence of a range from 42 to 174 genera with the same relative abundance in the ecosystem.

The number of genera detected between samples ranged from 848 in N2 to 1,115 in N3. Taking into account the number of rare genera, the Chao1 index[485] estimated this number in all samples up to an average of 1,146 (Table S5). Evenness values of 0.54-0.74 showed a considerable proportion of equally abundant species. Finally, as an estimate of heterogeneity between samples or beta diversity, the total number of genera in all sites and the average richness per site was estimated and revealed low heterogeneity between the different biofilm samples (β = 0.23).

Legionella detection and similarity among samples

Direct microbiological cultures from the biofilm samples did not result in isolation of *Legionella*. Either no growth or overgrowth was observed in the different plates, thus making its identification impossible. Additionally, the previously reported results from the metagenomes detected genus *Legionella* in a very low proportion in the analysed communities, in a range between 0.01-0.07% of bacterial and total relative abundance.

When hits above 97% percentage of identity were considered, the species *L. pneumophila* and *L. drancourtii* were observed in six and five samples, respectively. However, the coverage of the corresponding reference genomes was estimated as only 0.0004X or below of the full genome length. No significant correlation was found between the relative abundance of *Legionella* as a genus and *L. pneumophila* as species with any other organism found in the community after multiple testing correction using the False Discovery Rate (FDR) method[386] (q-value > 0.05).

Relative abundances of genera were used to calculate Jaccard similarity and Bray-Curtis dissimilarity distances. Samples were clustered using these distances, revealing three main groups at a Jaccard index of about 0.6 (Fig. S7A) and a Bray-Curtis index of approximately 0.5 (Fig. S7B). As samples were taken from different natural springs surrounding the locality of Alcoy, we decided to test the correlation between the pairwise genetic and geographic distances that could explain the observed clustering pattern. For this, we performed a Mantel test. The results did not show a significant correlation between both matrices (Mantel statistic r = -0.0099; p-value = 0.508).

Modified vs original LCA algorithm

The results obtained with the TV procedure on the microbial and eukaryotic composition of biofilm communities were compared to those of other methods. Firstly, bitscore-filtered BLAST output files were subjected to the original LCA algorithm using BLAST2LCA (see URLs). Differences between both strategies reside mainly in the process of taxonomic assignment of the reads. While BLAST2LCA needs that all of the hits of a specific read call the same organism group at a given taxonomic level, the TV method is more flexible by considering that more than 50% of those hits should agree in the classification at a given level. As explained more extensively in Supplementary Note 2, the best hit is always taken into consideration. This program detected 1,204 different genera for all domains in the whole dataset, compared to 1,276 genera retrieved with the TV method. 1,201 genera were detected by both methods and represented from

98.6% to 99.6% of the relative abundance retrieved from the TV. The relative abundances of all genera detected by the TV and BLAST2LCA showed a high and significant correlation (Pearson's r = 0.98; p-value < 2.2E-16), which validated the results given by our approach. The 75 genera detected only with the TV method were mainly from the fungal phyla *Ascomycota* and *Basidiomycota*.

Comparison with extracted rDNA reads

Reads matching 12S, 16S and 18S genes were extracted from the metagenomes (Table S6) and analysed with Mothur[487]. Eukaryotic reads were mostly unclassified (54.6%), although some of them were assigned to three phyla: *Zea* (33.3%), *Tribolium* (10.9%), and *Mus* (1.1%). No low *Eukaryotes* were found, which could be a result of the small amount of rRNA reads we are working with or an underrepresentation of many eukaryotic phyla in the SILVA database. Twenty-one different bacterial and archaeal phyla, 39 classes and 267 genera were detected using the 16S rDNA reads.

Many common taxa were detected by the two methods: 16 phyla, 29 classes and 133 genera. However, taxa specific to each of them were also detected (Fig. S8). This result can be associated to the differences in database composition. The relative abundances at high taxonomic levels (phylum and class) showed a statistically significant correlation between the two approaches (Pearson's r = 0.99 and 0.98, respectively; p-value < 2.2E-16) (Fig. 12).

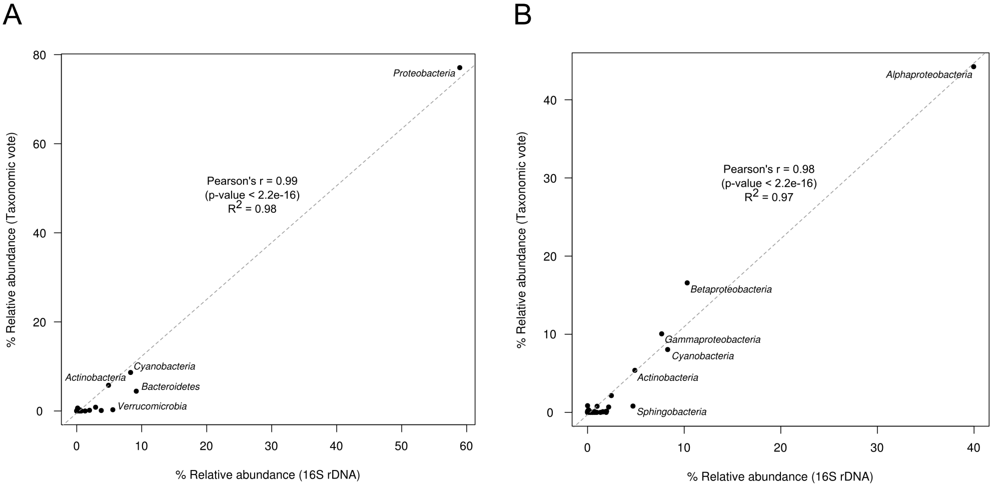


Figure 12 | Correlation between the bacterial relative abundance detected from the TV and 16S rDNA extracted reads at phylum (A) and class (B) levels.

From the information derived from 16S rDNA reads, the community was also found dominated by *Proteobacteria* (range 39.0-80.9%), *Cyanobacteria* (range 0.2-15.4%), *Actinobacteria* (range 0.7-21.4) and *Bacteroidetes* (range 2.1-26.1%). However, differences were observed in the relative abundance of genera, although the top most abundant organisms were *Porphyrobacter* (range 0.4-21.9%), *Sphingomonas* (range 0.9-19.1%), *Hydrogenophaga* (range 0.0-24.0%) and *Rhodobacter* (range 1.0-12.8%), which were also detected among the most abundant genera in the metagenomes by the TV method (Fig. 13). In this case, the correlation between genera obtained from metagenomes and 16S was modest but significant (Pearson's r = 0.66; p-value < 2.2E-16). Also, the genera defined by 16S rDNA represented only a 65.5% of the diversity found in the metagenomes.

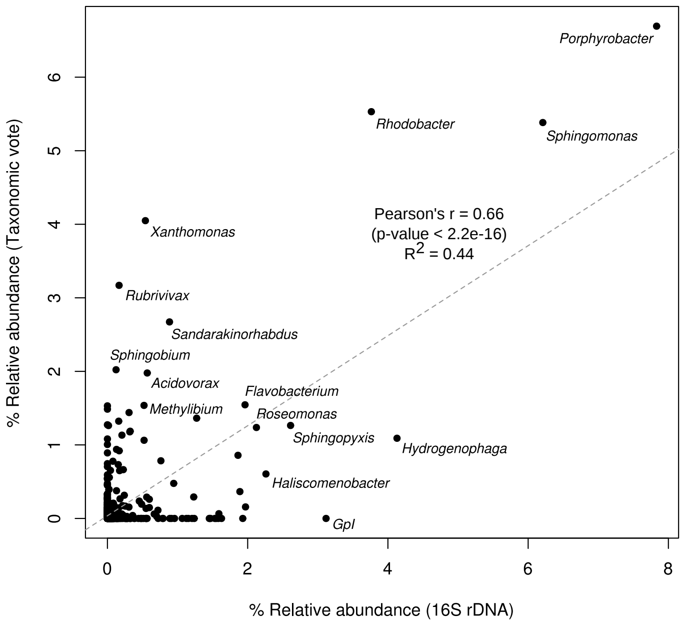


Figure 13 | Correlation between the relative abundance of bacterial genera detected by the TV method and 16S rDNA extracted reads.

Taxonomic vote versus MetaPhlAn

Filtered fastq reads were tested against the microbial clade-specific marker database in MetaPhlAn[476]. This software contains single-copy markers that are specific for different taxonomic levels. Fourteen different phyla encompassing 26 classes and 173 genera were detected (Fig. S8). These results also correlated significantly with those from the metagenomes (Pearson’s r = 0.98 and 0.92 for phylum and class levels, respectively; p-value < 2.2E-16) (Fig. 14).

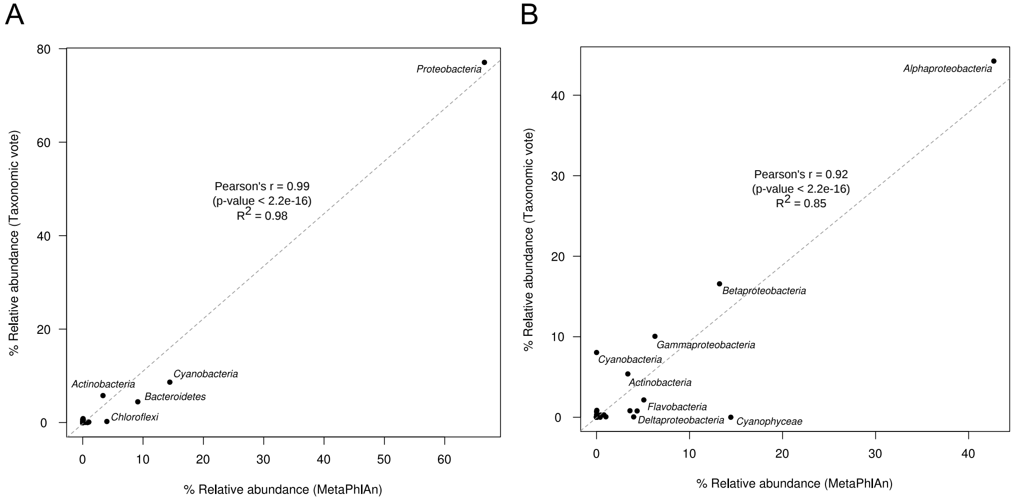


Figure 14 | Correlation between the bacterial relative abundance detected from the TV and MetaPhlAn at phylum (A) and class (B) levels.

The main differences between the MetaPhlAn results and those from the TV method were observed at the genus level (Pearson's r = 0.27; p-value < 2.2E-16) (Fig. 15). The most abundant organisms were *Burkholderia* (4.2-17.4%), *Sphingopyxis* (0.2-85.0%), *Citromicrobium* (0.0-65.0%) and *Flavobacterium* (0.0-23.7%), different from those detected from the metagenomes or 16S rDNA (Fig. S9). The genera detected using MetaPhlAn were estimated to represent only 63.5% of the genera detected using the whole metagenomes. However, the most abundant genus detected by the TV method, *Porphyrobacter*, was not found in MetaPhlAn. This is because there are no markers for this particular genus in the clade-specific markers database on which MetaPhlAn is based.

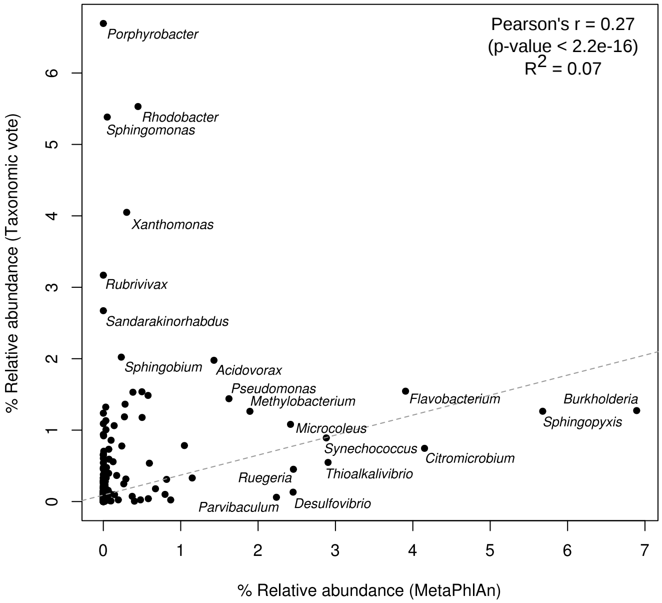


Figure 15 | Correlation between the relative abundance of bacterial genera detected from the TV method and MetaPhlAn.

In summary, a total of 62 genera were detected by the TV method and the three methods used for comparison, BLAST2LCA, 16S rDNA and MetaPhlAn (Fig. S8C). The relative abundances calculated using only the shared genera showed differences among methods, mainly 16S rDNA and MetaPhlAn. These can be related to either the algorithms, the number of reads, as it is the case of 16S rDNA, which were extracted from the metagenomes, or the corresponding database compositions (Fig. 16).

Species level

Taxonomic assignments with an identity equal or higher than 97% were used to explore the presence of particular eukaryotic and microbial species. All viral sequences that passed this threshold were assigned as phages.

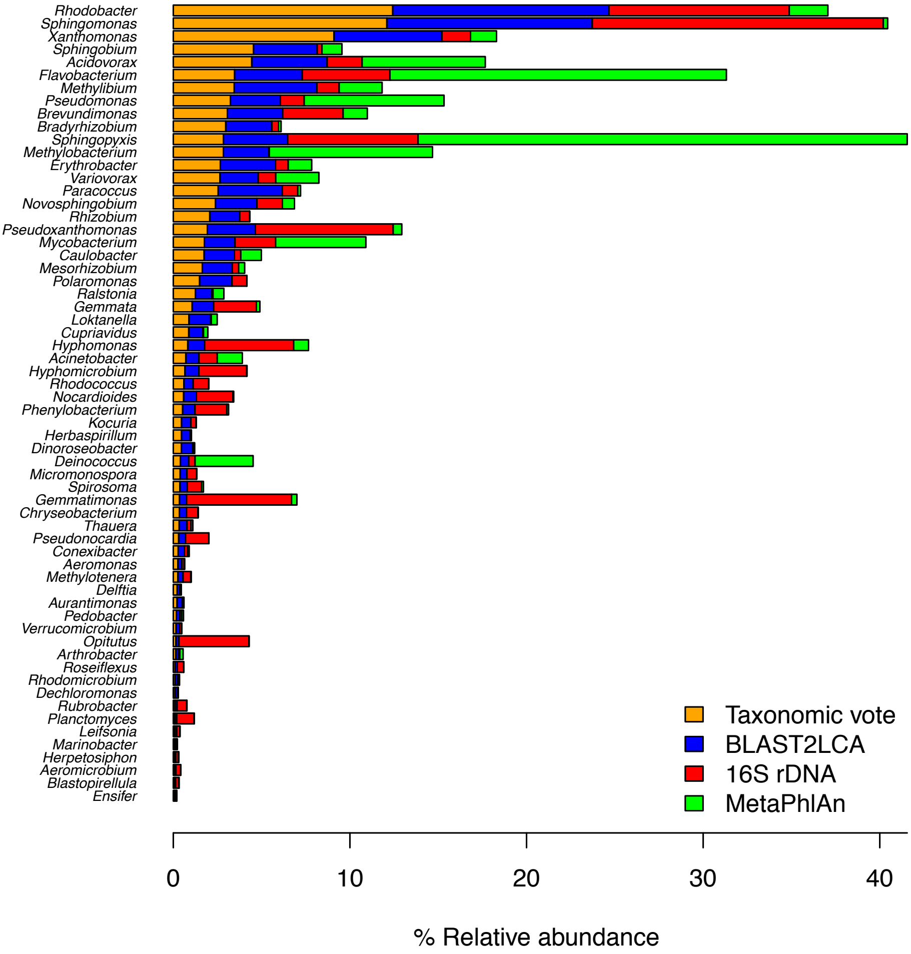


Figure 16 | Relative abundances of common bacterial genera calculated using 62 shared levels identified with the TV, BLAST2LCA, 16S rDNA reads and MetaPhlAn.

Bacterial sequences revealed 629 species, with a particular enrichment of samples N1 and N2 in *Oscillatoria nigro-viridis* (37.4% and 84.0%, respectively), N3 in *Gloeocapsa sp.* PCC 73106 (52.9%), N8 in *Pseudomonas putida* (27.3%) and N11, N23 and N24 in *Sphingomonas echinoides* (24.6%, 55.6% and 44.8%, respectively) (Fig. 17). Only a few archaeal species were assigned, such as *Halococcus morrhuae* (38.4%), *Halococcus saccharolyticus* (33.4%) and *Halococcus thailandensis* (28.3%) in N8 or Candidatus *Nitrosoarchaeum koreensis* in N24 (99.7%).

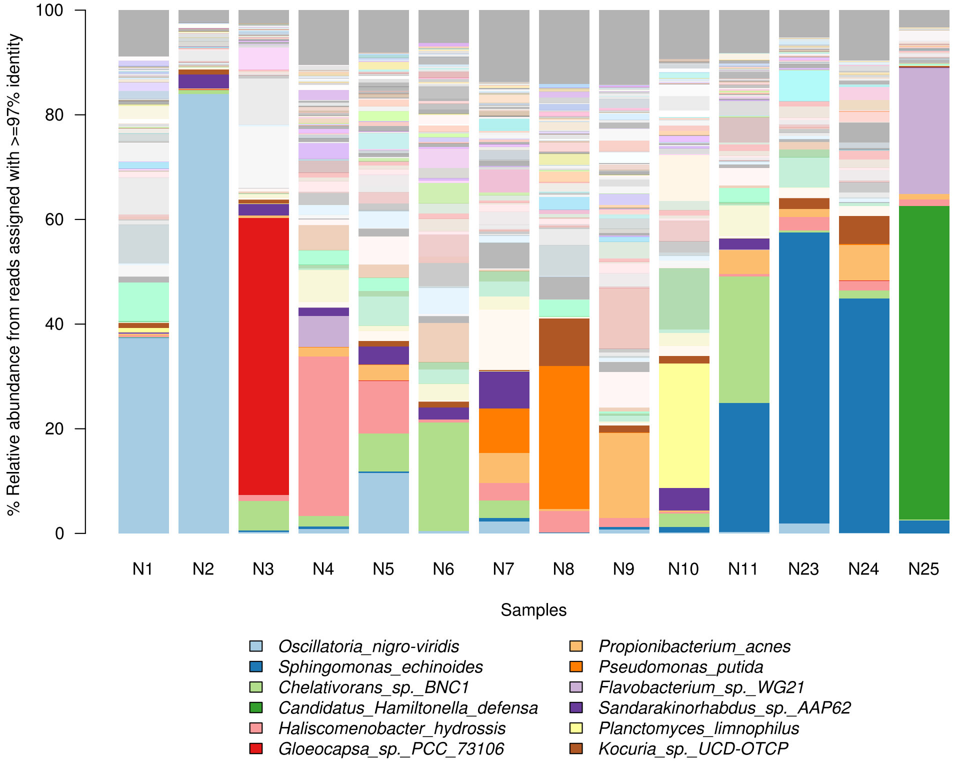


Figure 17 | Relative abundances of bacterial species estimated from reads assigned taxonomically with an identity equal or higher than 97%. Dark colours represent 12 species that explain 50% of the bacterial cumulative abundance estimated under the mentioned restrictions. Transparent colours show other species above 1% of relative abundance. The grey portion represents all species under the 1% fraction.

71 eukaryotic species were observed in the dataset. *Melampsora pinitorqua* was identified in all the samples (range 3.7-73.2%). Ten biofilms showed the presence of *Aureobasidium pullulans* (range 3.0-48.3%) and

seven were identified as containing *Plasmodium falciparum* (range 1.7-37.6%).

Clade-specific markers used by MetaPhlAn consist of single-copy genes that identify a specific taxonomic level while excluding ribosomal operons. A total of 233 different species were identified using this approach. Specifically, N4 was found enriched in *Sphingopyxis alaskensis* (46.0%)*,* N5 in *Flavobacterium psychrophilum* (54.2%), N6 in *Citromicrobium bathyomarinum* (39.4%), N7 in *Synechococcus elongatus* (31.6%) or N10 in *Parvibaculum lavamentivorans* (26.6%). Seventy-eight species were detected by both strategies, including the ones mentioned above except for *S. elongatus.* No significant correlation was found between the relative abundances found for those 77 species using the two approaches (Pearson's r = 0.09; p-value = 0.4031).

Discussion

Metagenomics is helping to increase our knowledge about the microbial communities of multiple natural environments, such as sediments, soil[491–493] or different water-related sources[300,301,319,320]. Huge efforts, such as the Earth Microbiome Project[302], are aimed at characterizing many different ecosystems and the interaction between them. However, many of these works have been performed using 16S rDNA PCR, which can result in biased results due to amplification[306,308]. High-throughput shotgun metagenomic approaches are encouraged in order to get a better, overall representation of the communities[472,473]. Nevertheless, despite recent and increasing sequencing efforts[475], genomic databases are still limited and an important part of the reads cannot be assigned yet[310].

Depending on the sequencing strategy, the taxonomic assignment and the analysis of reads can be performed following very different methods such as QIIME[494], Mothur[487], BLAST2LCA (see URLs), MEGAN[495], MetaPhlAn[476], etc. Many works use the BLAST algorithm[451] to assign a specific taxonomic level to each read or Operational Taxonomic Unit (OTU), which often have multiple hits. The Last Common Ancestor (LCA) algorithm has been widely used to get the most specific consensus taxon among all the hits from each query, and it has been implemented in softwares such as BLAST2LCA (see URLs) or MEGAN[495]. However, this procedure has several limitations because conserved genes or regions that have undergone lateral gene transfer can hit against very distant organisms from the same or even different domains of life. In these cases, taxonomic assignment can be narrowed only to phylum, domain or even remain unknown.

Here, we have proposed a modification of the traditional LCA algorithm, denoted as Taxonomic Vote (TV). Our main goal was to achieve a high quality evaluation of the composition of a metagenome at the lowest taxonomic level possible. First, we included a pre-filtering step of BLAST output files in order to remove spurious and distant hits, retaining only those that were not too distant from the top hit. With this, only the best hits are considered for taxonomic assignment. Next, we used a BLAST database with complete genome sequences from all the domains of cellular life and viruses (see Supplementary Note 1), so that matches are obtained initially only with high quality targets. An average genome size was estimated for each taxonomic level and used for database correction. The relative difference between the average identity score in the TV method and the highest one (best hit) is retained and denoted as ‘delta’. We have used a delta value of 10, but this can be easily modified for different sensitivity and specificity goals. The identity to the best hit is always considered in order to prevent its discarding when the match has similarity of 95% or higher. Finally, the assignment for each read is made to the lowest taxonomic level at which more than 50% of the retained hits agreed.

In this study, we aimed at characterizing the microbial, viral and eukaryotic community of biofilms created in the canalization tubes of natural springs in different spots around the same locality. These water sources are not treated for human consumption. The study of the microorganisms in these biofilms can help to know better this community and their potential threat for public health. Specifically, a predominance of *Bacteria* was found in all samples (97% of the assigned reads). Thirty-five different bacterial phyla, 52 classes and 927 genera were detected. However, only four of the phyla represented above 97% of the bacterial abundance: *Proteobacteria, Cyanobacteria, Actinobacteria* and *Bacteroidetes,* which presented fluctuations in their relative abundances in the different samples. The four phyla are known to be common in natural environments (water, soil, sediments, etc.), participating in essential processes such as nitrogen fixation or the degradation of organic material[496]. *Cyanobacteria* are frequent symbionts of plants and algae[497,498] and, thus, very commonly found in water-related sources. Although these four phyla include many different pathogens, the biofilms described in this work were enriched in genera frequently found in soil and freshwater ecosystems, such as *Porphyrobacter*, *Xanthomonas, Rhodobacter* and *Sphingomonas*[499,500]. Interestingly, *Xanthomonas* is a well-known plant pathogen[501]. Other genera not represented in the genome database were found in the NCBI non-redundant nucleotide database, such as the alphaproteobacteria *Roseibacterium, the* gammaproteobacteria *Marichromatium* or *Hymenobacter* from the *Bacteroidetes* group, which has been previously isolated from natural sources such as sand, soil or freshwater[502–504]. At the species level, the cyanobacteria *Oscillatoria nigro-viridis* and *Gloeocapsa sp. PCC 73106* were identified, as well as the metabolically versatile *Sphingomonas echinoides* and *Pseudomonas putida*[505,506]*.*

A few phyla dominated the identified *Eukaryotes*, such as the parasitic protists in *Apicomplexa*, the green algae *Chlorophyta* and representatives of the *Dikarya* subkingdom of *Fungi*, including *Ascomycota* and *Basidiomycota* phyla. However, a relatively high diversity of eukaryotic taxa was identified, with a total of 16 phyla, 43 classes and 253 genera. Parasitic protozoan genera such as *Plasmodium* and *Ascogregarina* were detected in the community. Although *Plasmodium* is commonly known as the causal agent of malaria, only 1-4 reads were assigned at a 97% or higher identity to the species *Plasmodium falciparum*. This cannot be conclusive of its presence in the community under study.

Archaea and *Viruses* were found as the minority fraction, mostly below 1% of relative abundance. Eighty different archaeal genera encompassing 3 phyla were identified, with the class *Halobacteria* representing half of the archaeal diversity. This contrasts with the fact that sampling was performed in freshwater, as the metabolism of *Halobacteria* requires high salt concentrations[507]. Most of the sequences assigned to *Viruses* were identified as *Microvirus*, known as proviruses prone to integrate in the genomes of the phylum *Bacteroidetes*, and thus frequently found in the human gut[508]. However, they have been also detected in marine environments, such as in ecosystems studied in the GOS project[315].

The proportion of eukaryotes and viruses could also have been affected by the laboratory procedures preceeding the sequencing. Although the extraction kit was adapted to biofilm samples, the process could have not been enough for the lysis of all eukaryotic cells. The low proportion of viruses could have been a result of that not all viruses can be detected using these procedures, as it is the case of RNA-based viruses, for instance.

From 848 to 1,115 bacterial genera were detected in the fourteen samples, and the Chao1 richness parameter[485] estimated the total number of genera in 1,146. These results and the rarefaction curves showed that most of the diversity likely present in the environment under study has been included and identified. This is because the number of estimated bacterial genera for some of the samples is very close to the maximum predicted by the Chao1 estimator (Good's estimate of coverage = 99.9%). However, some heterogeneity, especially in the least abundant taxa, was found among samples, showing that biofilms from natural springs from very close locations may have similar compositions, although the relative abundances of many organisms may differ. No significant association was found between the compositional distance among the different samples and the geographical coordinates of the sampling points.

The results obtained with the TV method were compared to those from other available methods. First, the original LCA algorithm as implemented in BLAST2LCA (see URLs) is more conservative than the modified version of the TV method. In cases of horizontal gene transfer or highly conserved genes, the LCA assignment may select upper taxonomic levels, such as class, phylum or even kingdom. Specifically, we detected 1,201 genera in common between both methods, and 75 were detected only by the TV method. Interestingly, these were found mainly in fungal phyla. However, the correlation between the genera relative abundances obtained through the two methods was found very high (Pearson's r = 0.98; p-value < 2.2E-16), showing that our method is perfectly valid for taxonomic assignment.

Additionally, we compared the TV with other popular methods for describing microbial communities, such as 16S rDNA taxonomic assignment using Mothur[487] and MetaPhlAn[476]. Bacterial genera identified by these two methods only represented 65.5% and 63.5% of the bacterial abundance detected by the TV method, respectively. High correlation values were found between the TV and 16S rDNA and MetaPhlAn relative abundances at higher levels (phylum and class), while lower correlations were obtained at the genus level. The differences between all the methods used derive from the composition of the databases on which they are based. 16S rDNA provides the identification of organisms for which no complete genome or sequencing projects are available and, as such, may not be present in our reference genome database. This is the case of the genus *Armatimonadetes* or different *Cyanobacteria* genera. MetaPhlAn is based on a clade-specific marker database developed from whole genome sequences[476]. Although the database has been updated to take environmental bacteria into account, it does not contain markers for all the organisms for which sequencing projects are currently available yet. For instance, *Porphyrobacter*, a bacterial genus found as one of the most abundant in all biofilm samples in this study, was not identified following this approach.

No significant evidence was found for the presence of organisms that may pose a risk for public health, although very low coverages (under 1% of genome coverage) were found for pathogens such as *Escherichia*, *Legionella*, *Mycobacterium* or protozoa that may act as reservoirs for pathogens such as *Acanthamoeba*. Only *Xanthomonas,* a plant pathogen, was detected at a maximum of 4% coverage. Interestingly, despite the known endemism of *Legionella pneumophila* in the Alcoy area[339,352], the genus *Legionella* was only observed at a very low relative abundance (0.01-0.07%).

In summary, a very diverse microbial and eukaryotic composition was found in the environmental biofilm samples under study, with a high proportion of low abundance genera. Although subtle differences in the relative abundances of some organisms were found among samples, their structure and composition of most abundant taxa, usual in the freshwater microbiota, was rather homogeneous. A recent work[509] discussed whether similar environments would be expected to have a comparable composition whereas abundances would be dependent on environmental selective pressures. Further work is necessary in order to confirm the presence of microorganisms in freshwater biofilms that could lead to risks for public health.

URLs

NCBI Taxonomy database. ftp://ftp.ncbi.nih.gov/pub/taxonomy/
FastQC. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
BLAST2LCA. https://github.com/emepyc/Blast2lca

Supplementary material

Supplementary Note 1

Genome database

Reference database creation

A reference genome database was created containing Bacterial, Archaeal, Lower Eukaryotic and Viral data for blasting the metagenomes. Bacterial (and some Archaeal) complete and draft genomes were downloaded directly from the 'Bacteria' and 'Bacteria_DRAFT' folders in the NCBI FTP site (http://ftp.ncbi.nlm.nih.gov/genomes/). EFetch from NCBI E-utilities[510] was used to search and automatically download genome data from Low Eukaryotes[474], Viruses and to complete the Archaeal genome data. The exact search patterns were the following:

Archaea	`Archaea[ORGN]+NOT+mRNA[FILTER]+AND+genome[TITLE]`
Viruses	`Viruses[ORGN]+AND+genome[TITLE]+NOT+Bacteria[ORGN]+NOT+mRNA[FILTER]`
Low Eukaryotes	`Eukaryota[ORGN]+NOT+Bilateria[ORGN]+NOT+Streptophyta[ORGN]+NOT+mitochondrial[TITLE]+NOT+mitochondrion[TITLE]+NOT+chloroplast[TITLE]+NOT+plastid[TITLE]+NOT+mRNA[FILTER]+AND+genome[TITLE]`

The raw data was filtered to remove plasmidic, mitochondrial and chloroplastic information. The resulting genome data was used to create a BLAST+ v2.2.28 database with makeblastdb[451].

Database indexing

A database index was created containing all the different organisms and their taxonomic classification. The length of each complete genome was annotated. For draft genomes contained in more than one contig, the sum of all contig lengths was annotated as representative of the genomic content of that organism. An average genome length was calculated per species using the information from the corresponding strains. The information about the number of genomes for each species available in the database was also annotated (Table S1).

Average genome lengths were extrapolated for each taxonomic level using an arithmetic mean. For instance, being *Legionella* a genus with 5 different species with a pre-calculated average genome size for each of them, the average genome length for the corresponding genus was calculated as the arithmetic mean among those, as shown below. This process was applied recursively to each taxonomic level in the custom database.

Species	**Average size (bp)**
Legionella drancourtii	4,152,613
Legionella longbeachae	4,081,771
Legionella pneumophila	3,838,671
Legionella shakespearei	3,512,682
Legionella tunisiensis	3,374,751

Average size for the genus *Legionella*:

3,792,098 bp

Hence, apart from the main file shown in Table S1, the index database contains six additional subtables containing this information for all the taxonomic levels, from domain to genus (Table S2).

The average genome size for each level was used in the study in order to correct the number of BLAST hits assigned to a specific level by the amount of genetic material of that level contained in the database. Specifically, coverage values were calculated by taking into consideration the average read length (379 bp).

Database contents

The reference genome database contains 17,314 genomes and whole genome sequencing projects and includes 8,539 species from *Bacteria* (n=3,159), *Archaea* (n=382), *Viruses* (n=4,544) and Low *Eukaryotes* (n=454). The database contained 1,701 genera, 534 families, 226 orders, 105 classes and 65 phyla. Depending on the sequencing effort over the different organisms, some levels are over or underrepresented in the database. The correction applied on the counts assigned to each taxonomic level is intended to overcome this problem (Table S3 and Fig. S1).

Supplementary Note 2

Taxonomic vote

Taxonomic assignment of reads was performed after blasting (BLASTn v2.2.28+)[451] against the custom genome reference database. The whole taxonomic classification was obtained from all filtered hits among the 15% of the maximum bit-score for each read using the Taxonomy database from NCBI (ftp://ftp.ncbi.nih.gov/pub/taxonomy/). A consensus assignment for all the hits from the same read was retrieved using a modified version of the Last Common Ancestor (LCA) algorithm (named Taxonomic Vote, TV; Fig. S2).

Specifically, taxonomy was assigned up to the level for which more than 50% of the hits agreed. However, information about the maximum and average percentage of identity, E-value, and bitscores for the reads involved in the TV and the best hit were retained. Cases in which the best hit was discarded during the vote were marked as discrepancies. The difference between the average percentage of identity of the hits involved in the vote and that from the best hit was calculated and designated as ‘delta’. If the best hit had a percentage of identity higher or equal to 95%, the taxonomy from the best hit was assigned. If the best hit had a percentage of identity below 95% and the delta value was below 10%, the assignment resulting from the TV was selected. Reads in which the best hit was below 95% identity and the delta value was equal or higher than 10% were left unassigned (Fig. S3).

Table S1 | Example of the format of the genome database index file. The average genome size and the number of genomes available from the corresponding species are annotated in the last two columns.

Kingdom	Phylum	Class	Order	Family	Genus	Species	Avge size (bp)	Freq.
Archaea	*Crenarchaeota*	*Thermoprotei*	*Sulfolobales*	*Sulfolobaceae*	*Acidianus*	*Acidianus hospitalis*	4,303,957	1
Archaea	*Crenarchaeota*	*Thermoprotei*	*Acidilobales*	*Acidilobaceae*	*Acidilobus*	*Acidilobus saccharovorans*	2,992,906	1
Bacteria	*Proteobacteria*	*Gamma-proteobacteria*	*Aeromonadales*	*Aeromonadaceae*	*Aeromonas*	*Aeromonas hydrophila*	4,824,132	9
Bacteria	*Proteobacteria*	*Gamma-proteobacteria*	*Aeromonadales*	*Aeromonadaceae*	*Aeromonas*	*Aeromonas salmonicida*	4,815,164	2
Eukaryota	*Ascomycota*	*Eurotiomycetes*	*Eurotiales*	*Thermoascaceae*	*Byssochlamys*	*Byssochlamys spectabilis*	28,937,452	1
Eukaryota	*Ascomycota*	*Saccharomycetes*	*Saccharomycetales*	*Debaryomycetaceae*	*Candida*	*Candida albicans*	16,225,487	13
Viruses	*unknown*	*unknown*	*unknown*	*Mimiviridae*	*Mimivirus*	*Acanthamoeba polyphaga mimivirus*	2,869,525	2
Viruses	*unknown*	*unknown*	*unknown*	*Phycodnaviridae*	*Chlorovirus*	*Acanthocystis turfacea Chlorella virus 1*	576,094	1

Table S2 | Example of the ‘Genus’ subtable showing the amount of genomic material in the database and the average genome size for the genera represented in the database.

Genus	Total bp	Number of species	Average genome size (bp)
Hydra	2,780,169,611	1	2,780,169,611
Phytophthora	1,346,471,473	11	122,406,497
Fusarium	858,217,747	9	95,357,527
Nematostella	713,227,170	1	713,227,170
Streptomyces	683,279,371	78	8,759,991
Symbiodinium	609,476,485	1	609,476,485
Leishmania	559,743,881	16	34,983,992
Aspergillus	477,562,791	10	47,756,279
Puccinia	452,530,572	3	150,843,524
Tetrahymena	440,965,617	4	110,241,404
Pseudomonas	424,316,078	69	6,149,508
Acropora	416,342,532	1	416,342,532
Colletotrichum	411,648,997	6	68,608,166

Table S3 | Top five most abundant organisms in the reference genome database at four taxonomic levels.

Rank	Phylum	Class	Genus	Species
Bacteria				
1	*Proteobacteria*	*Gammaproteobacteria*	*Streptococcus*	*Escherichia coli*
2	*Firmicutes*	*Bacilli*	*Escherichia*	*Salmonella enterica*
3	*Actinobacteria*	*Actinobacteria*	*Staphylococcus*	*Streptococcus agalactie*
4	*Spirochaetes*	*Alphaproteobacteria*	*Salmonella*	*Streptococcus pneumoniae*
5	*Bacteroidetes*	*Epsilonproteobacteria*	*Vibrio*	*Helicobacter pylori*
Archaea				
1	*Euryarchaeota*	*Halobacteria*	*Methanobrevibacter*	*Methanobrevibacter smithii*
2	*Crenarchaeota*	*Thermoprotei*	*Sulfolobus*	*Sulfolobus islandicus*
3	*Thaumarchaeota*	*Methanobacteria*	*Halorubrum*	*Methanococcus maripaludis*
4	*Nanoarchaeota*	*Methanomicrobia*	*Haloferax*	*Haloquadratum walsbyi*
5	*Parvarchaeota*	*Thermococci*	*Thermococcus*	*Sulfolobus acidocaldarius*
Low Eukaryota				
1	*Ascomycota*	*Saccharomycetes*	*Saccharomyces*	*Saccharomyces cerevisiae*
2	*Basidiomycota*	*Eurotiomycetes*	*Plasmodium*	*Plasmodium falciparum*
3	*Apicomplexa*	*Sordariomycetes*	*Fusarium*	*Saccharomyces kudriavzevii*
4	*Microsporidia*	*Aconoidasida*	*Phytophthora*	*Fusarium oxysporum*
5	*Chlorophyta*	*Agaricomycetes*	*Leishmania*	*Pseudogymnoascus pannorum*
Viruses				
1	-	-	*Norovirus*	*Norwalk virus*
2	-	-	*Begomovirus*	*Severe acute respiratory syndrome-related coronavirus*
3	-	-	*Betacoronavirus*	*Hepatitis C virus*
4	-	-	*Enterovirus*	*Rotavirus A*
5	-	-	*Mastadenovirus*	*Rhinovirus A*

Table S4 | Statistics on the number of reads assigned to any taxonomic level after blasting against the reference genome database and after blasting the unassigned reads to the NCBI non-redundant nucleotide database (nr-nt).

Sample	Number of paired-end reads	Assigned after taxonomic vote	% Assigned reads	Total assigned after nr-nt blastn	% Total assigned reads	% Total unassigned
N1	1,714,525	576,414	33.62	606,924	35.40	64.60
N2	1,773,670	413,991	23.34	646,805	36.47	63.53
N3	2,309,668	376,505	16.30	387,161	16.76	83.24
N4	1,999,774	263,704	13.19	270,706	13.54	86.46
N5	1,958,330	335,666	17.14	343,425	17.54	82.46
N6	1,477,726	306,657	20.75	314,362	21.27	78.73
N7	2,173,611	298,131	13.72	308,055	14.17	85.83
N8	2,028,900	639,896	31.54	658,651	32.46	67.54
N9	1,843,292	501,266	27.19	513,975	27.88	72.12
N10	2,195,139	382,384	17.42	390,812	17.80	82.20
N11	2,516,046	267,603	10.64	274,250	10.90	89.10
N23	1,514,040	210,644	13.91	217,954	14.40	85.60
N24	1,517,947	286,720	18.89	296,019	19.50	80.50
N25	1,742,231	206,609	11.86	222,732	12.78	87.22

Table S5 | Taxonomy diversity and richness estimates.

	Diversity		Richness			Evenness
	Shannon index	Inverse Simpson index	Number of genera	Chao1	Standard error (Chao1)	
N1	4,53	41,17	1015	1113,88	21,95	0,65
N2	3,79	13,89	848	1017,36	35,35	0,56
N3	4,74	36,43	1115	1197,07	19,62	0,68
N4	4,82	44,24	1079	1200,16	26,03	0,69
N5	4,62	29,89	1026	1127,93	23,52	0,67
N6	4,64	37,82	990	1131,75	29,27	0,67
N7	4,81	43,89	1071	1165,46	20,62	0,69
N8	4,51	39,39	1068	1140,60	17,33	0,65
N9	4,51	37,77	994	1077,39	19,24	0,65
N10	4,39	21,21	1076	1156,36	18,77	0,63
N11	4,74	35,74	1097	1164,19	16,34	0,68
N23	3,75	10,98	993	1174,01	35,74	0,54
N24	5,02	62,02	1070	1191,41	27,07	0,72
N25	5,16	81,80	1078	1203,35	27,56	0,74

Table S6 | Number of ribosomal reads extracted from the metagenomes with Metaxa.

Sample	Bacteria	Archaea	Eukaryota	Chloroplast	Mitochondria	Uncertain	Total
N1	1,031	0	104	38	25	4	1,202
N2	969	1	112	14	23	3	1,122
N3	1,130	0	276	12	26	8	1,452
N4	843	0	858	134	70	23	1,928
N5	1,000	1	333	8	16	2	1,360
N6	861	0	94	9	16	3	983
N7	807	1	871	33	71	7	1,790
N8	1,249	2	161	34	36	8	1,490
N9	1,112	0	49	47	31	6	1,245
N10	1,155	0	147	12	25	6	1,345
N11	997	2	462	16	33	3	1,513
N23	451	0	1,022	20	37	4	1,534
N24	676	2	145	34	34	3	894
N25	521	2	869	219	80	21	1,712

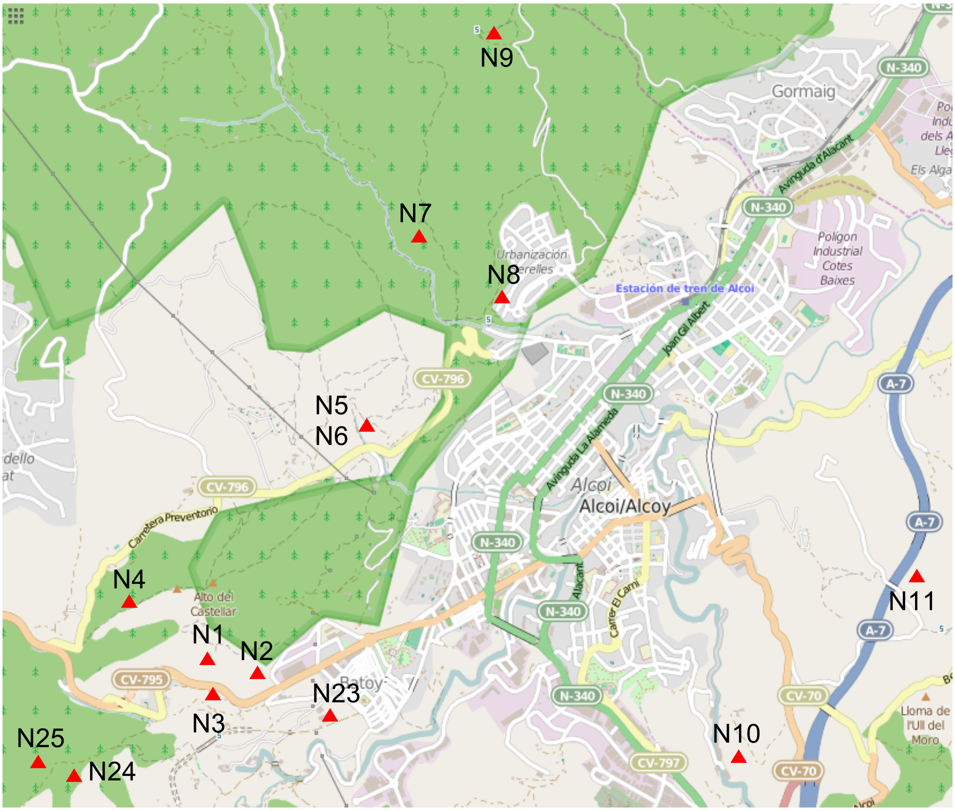


Figure S1 | Map of the locality of Alcoy and its surroundings. Red triangles mark points where biofilm samples were taken from natural springs (labels represent the sample names).

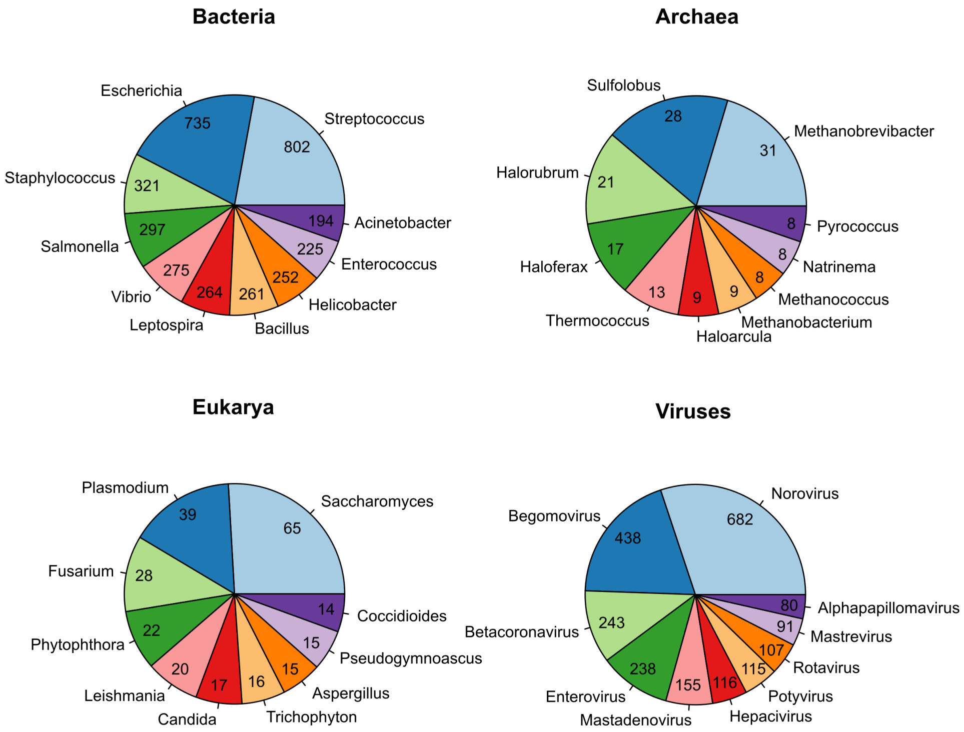


Figure S2 | Top ten genera with the highest number of genomes represented in the reference database for each domain. The specific number of genera per level is shown inside the pie chart.

1. Filtered blast tabular output:

Hits >100 bp

Bitscore on the top 15%

Read	Hits	% ID	Alignment length
Read1	Hit1	99	379
Read1	Hit2	99	379
Read1	Hit3	98	376
Read1	Hit4	96	320
Read1	Hit5	90	318

...

E-value	Bitscore
1E-153	499
1E-153	499
1E-140	448
1E-70	435
1E-22	424

2. Taxonomic assignment for each hit

Start at the species level

Hit	Kingdom	Phylum	Class	Order	Family	Genus	Species
1	*Bacteria*	*Proteobacteria*	*Alpha-proteobacteria*	*Rodospirillales*	*Acetobacteraceae*	*Acetobacter*	*Acetobacter pasteurianus*
2	*Bacteria*	*Proteobacteria*	*Alpha-proteobacteria*	*Rodospirillales*	*Acetobacteraceae*	*Acetobacter*	*Acetobacter pasteurianus*
3	*Bacteria*	*Proteobacteria*	*Gamma-proteobacteria*	*Aeromonadales*	*Aeromonadaceae*	*Aeromonas*	*Aeromonas fluvialis*
4	*Bacteria*	*Proteobacteria*	*Alpha-proteobacteria*	*Rodospirillales*	*Acetobacteraceae*	*Granulibacter*	*Granulibacter bethesdensis*
5	*Bacteria*	*Proteobacteria*	*Alpha-proteobacteria*	*Rodospirillales*	*Rhodobacteraceae*	*Gemmobacter*	*Gemmobacter aquaticus*

3. TV: > 50% of the hits should agree with the assignment

60% *Acetobacteraceae*

40% *Acetobacter*

40% *Acetobacter pasteurianus*

Consensus taxonomic classification

From TV:

Bacteria	*Proteobacteria*	*Alpha-proteobacteria*	*Rodospirillales*	*Acetobacteraceae*

From LCA:

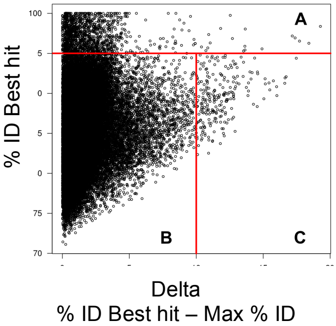
Bacteria	*Proteobacteria*

4. Refinement step: Has the best hit been discarded during the vote?

NO

YES

The result from the TV is accepted



A. The assignment from the best hit is accepted when % ID Best hit >= 95%

B. The result from the TV is accepted when % ID Best hit < 95% and Delta <10%

C. Discarded

Figure S3 | Schematic representation of the taxonomic vote (TV) method.

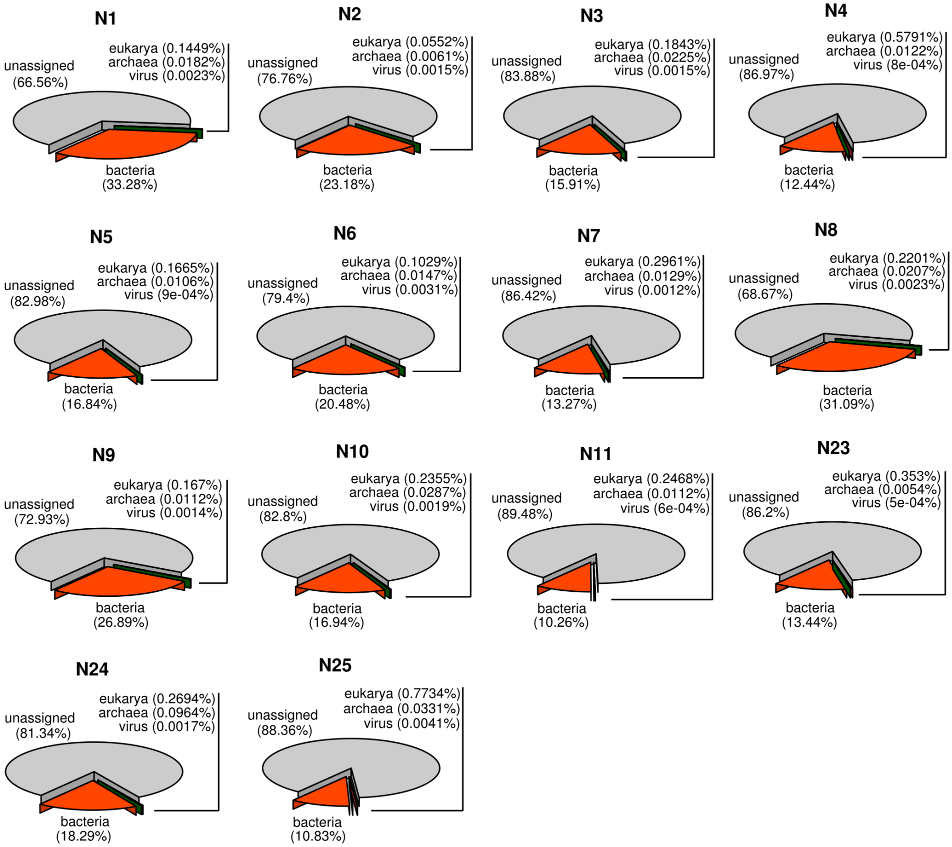


Figure S4 | Proportion of reads assigned to *Bacteria*, *Archaea*, low *Eukarya* and *Viruses* as well as unassigned ones for each biofilm sample after applying the TV method.

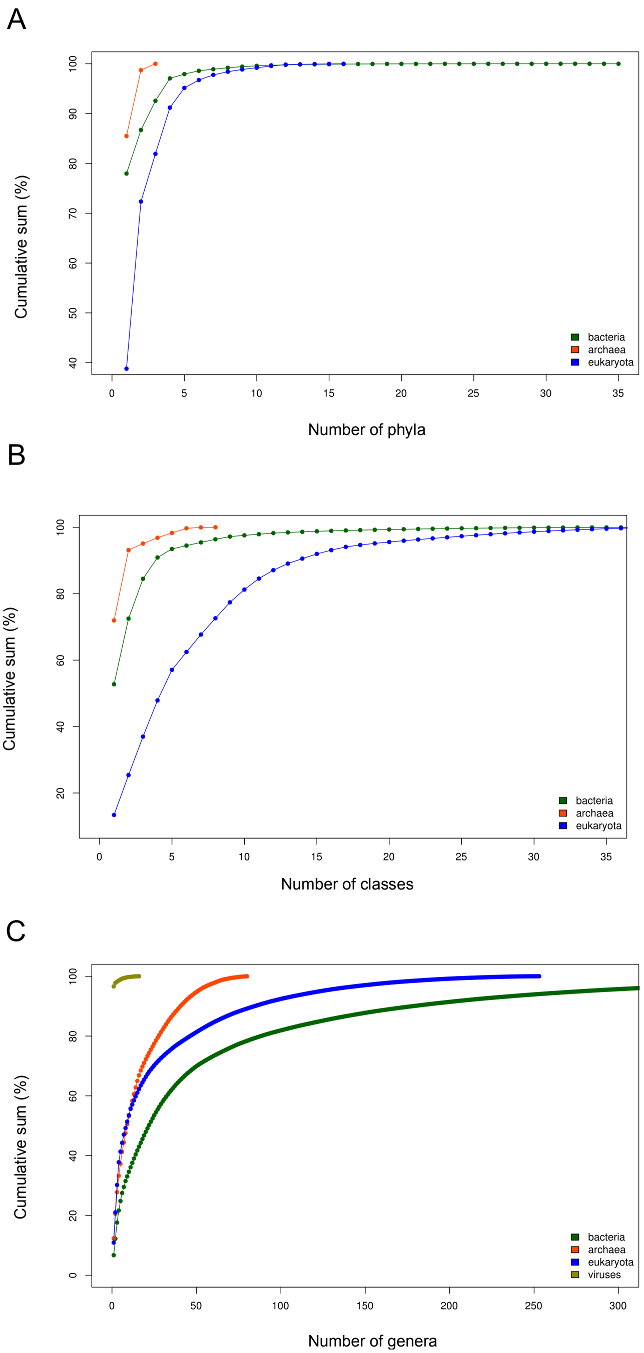


Figure S5 | Cumulative sum of the all the phyla (A), classes (B) and genera (C) detected in the dataset.

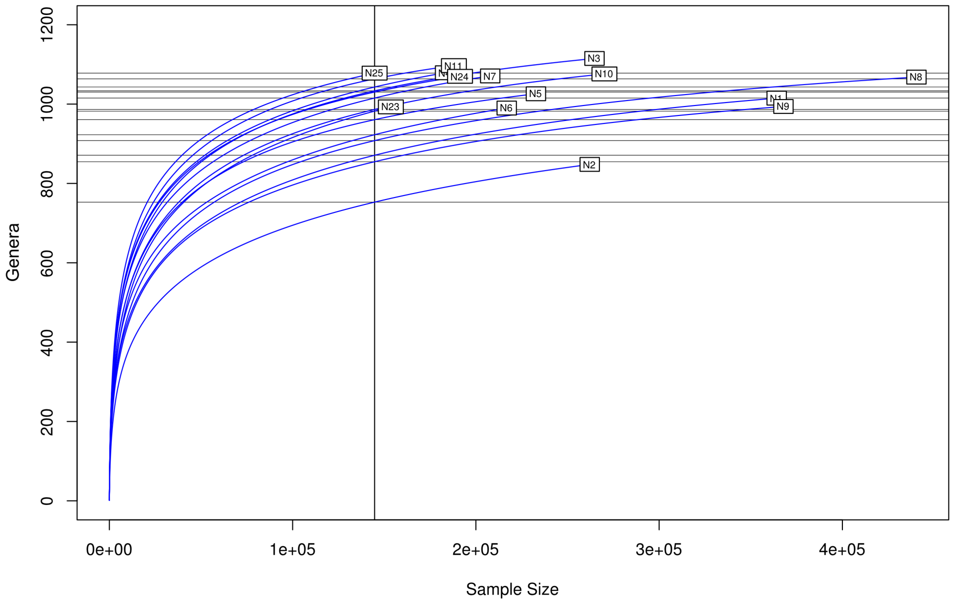


Figure S6 | Rarefaction curves calculated for the fourteen biofilm metagenomes.

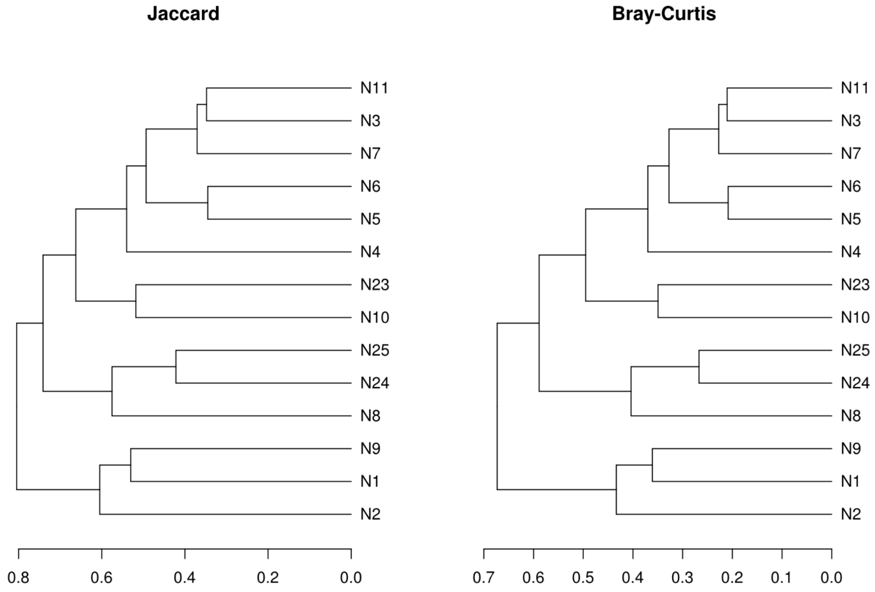


Figure S7 | Clustering of the fourteen biofilm samples using Jaccard and Bray-Curtis distances.

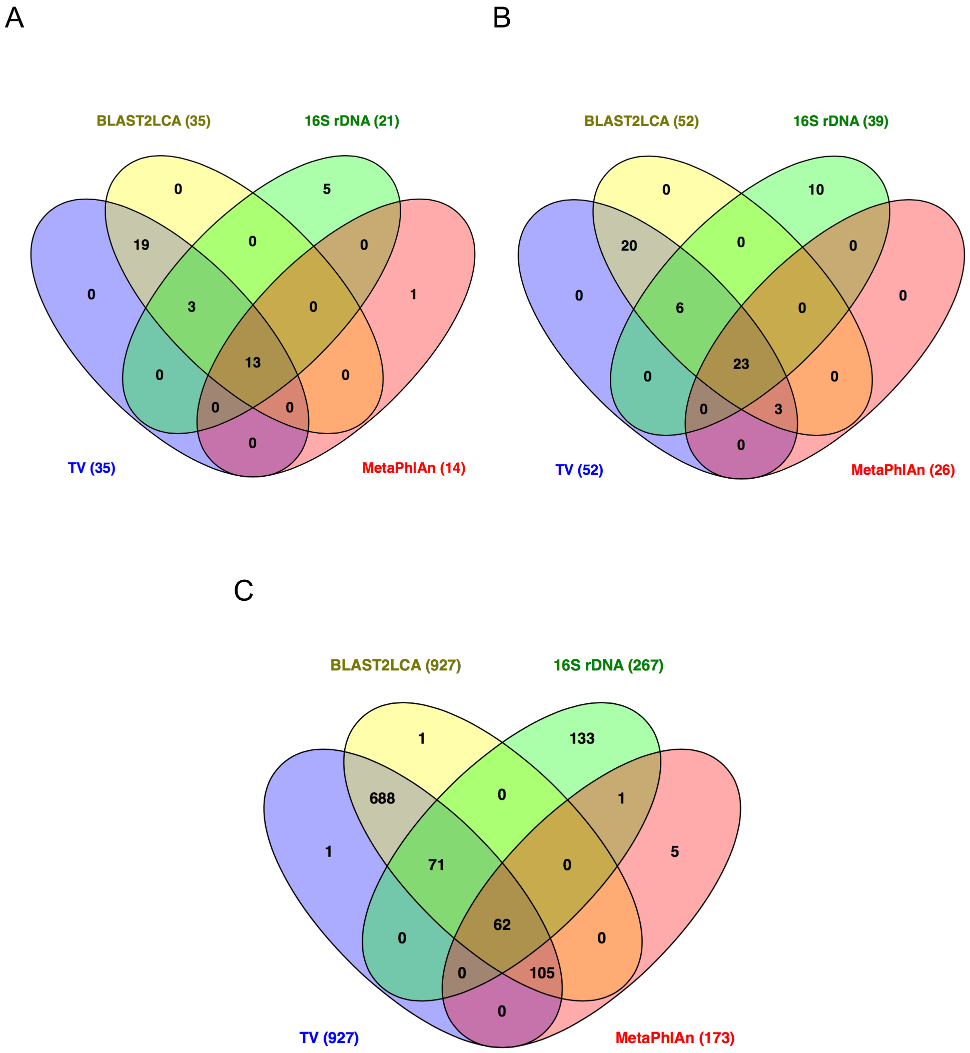


Figure S8 | Venn diagrams representing the number of shared and unique bacterial phyla (A), classes (B) and genera (C) detected by the taxonomic vote (TV), BLAST2LCA, the 16S rDNA extracted reads and MetaPhlAn. Plots were obtained with VENNY[511].

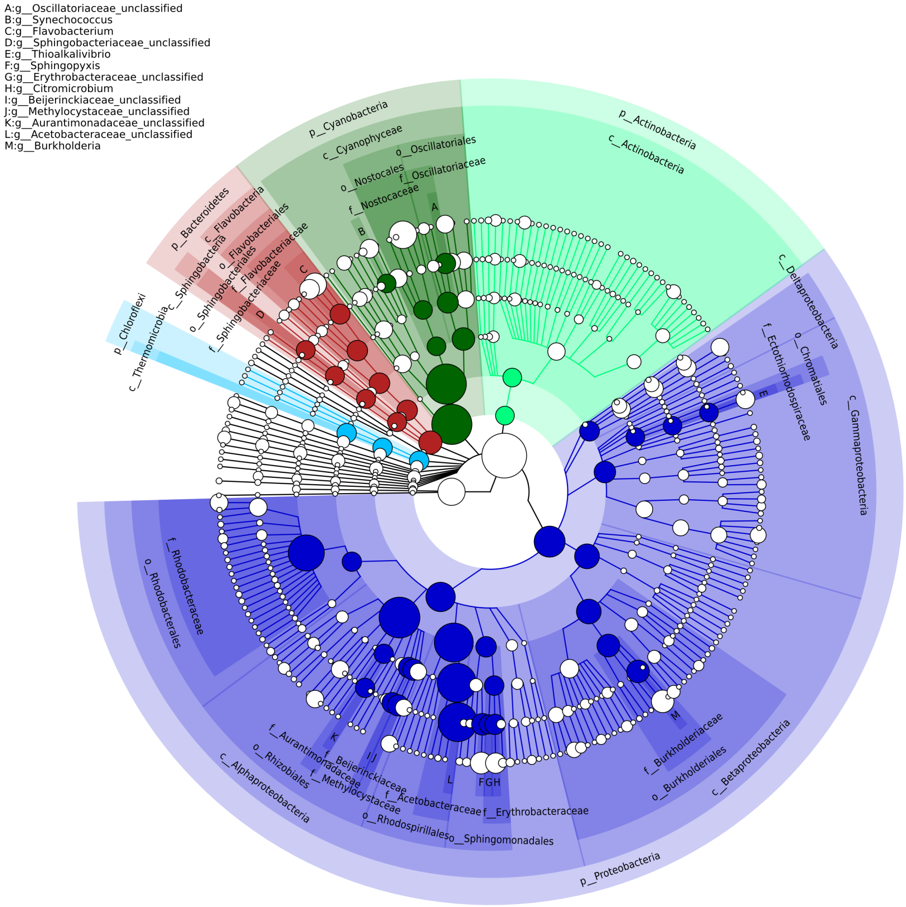


Figure S9 | Composition of the microbial community from MetaPhlAn. Taxonomic levels from phylum to genus are shown in different colours. Circle diameters represent the scaled relative abundance of each level.

Discussion

Infectious diseases still represent one of the major causes of mortality for human populations. Many human pathogens are considered a risk for public health because of their high transmissibility and/or morbidity. These include organisms from different domains of life, mainly bacteria, fungi, parasites and viruses. For most of them, the source of the infection igniting an epidemic problem is difficult to ascertain, and molecular (gene or genome based) approaches represent a very useful help in establishing the link between source and infected people. The analyses of subsequent transmissions among persons, either directly or through intermediate vectors, can also benefit from these approaches. The same information obtained for and from molecular epidemiology analyses can be used to gain insight into the evolutionary dynamics and processes of the organisms involved, thus establishing a mutual reinforcement cycle between basic and applied science. This thesis represents our contribution to this molecular epidemiology and evolutionary biology joint area for a recently emerged human pathogen, *Legionella pneumophila*.

Human-to-human transmission can be produced through different routes, from physical contact in sexually-transmitted diseases such as gonorrhoea[512], caused by *Neisseria gonorrhoeae*, to aerosol inhalation from the expectoration of an infected patient in the case of tuberculosis disease[513], caused by *Mycobacterium tuberculosis*. Nevertheless, there are organisms for which infection among individuals has not been described yet. In those cases, human infection often results in a dead-end for the pathogen. Examples are cholera[514], caused by *Vibrio cholerae*, which is normally transmitted by contaminated water or food, and direct person-to-person transmission is very rare[515] or Legionnaires' disease[89], caused by *Legionella pneumophila*, in which human infection is strictly caused by inhalation of contaminated aerosols.

Although the biological mechanisms of pathogenicity for many organisms are fully characterized, their epidemiological structure and dynamics remain poorly understood. This is especially true in the case of microorganisms that do not involve person-to-person transmission, such as *L. pneumophila*, in which human infection is even considered as accidental[10]. Although the potential reservoirs have been described[14,96,97,118,406] in detail, the environmental control of this pathogen still remains a challenge for public health officials[89]. Fortunately, epidemiological studies have benefited from the introduction of microbiological and, especially, genetic typing tools, which have improved the capability of discrimination between strains[196]. These typing tools, with the help of the more recent genomic sequencing[516], are essential in the study of genetic variability of microorganisms. This genetic information can be used to analyse the existing populations and results can be translated to public health, often with a short-term and even immediate application[228,352].

In this thesis, different aspects of *L. pneumophila* epidemiology have been covered. First, Chapters 1 and 2 shed light into the genetic variability and population of this species in a specific area compared to the overall reported diversity. Then, a sensible detection method for *L. pneumophila* from biofilm samples was developed and described in Chapter 3. This procedure was later applied to a real-time analysis of a legionellosis outbreak, as described in Chapter 4, showing its rapid applicability to public health investigations. Finally, high throughput sequencing technologies have been applied at two levels. First, a genome population study of outbreak-related *L. pneumophila* ST578s in the locality of Alcoy showed intra sequence-type variability and clear patterns of recombination, as described in Chapter 5. Second, the microbial, eukaryotic and viral community of biofilms from natural springs in Alcoy is described in a metagenomic study in

Chapter 6, showing the low proportion of *Legionella* in the natural environment in a supposedly endemic area.

Human infection by *L. pneumophila* is considered as accidental[10]. Its life is strictly associated to environmental water-related sources[5], and it is known to replicate actively in biofilms[21]. In order to control infections by this bacterium, it is essential to control its population in the environment. To do so, many physical and chemical strategies are applied to the potential reservoirs, such as deep cleaning or treatment with heat and biocides[130]. However, *Legionella* can survive after these by persisting inside cystic amoebae[27]. Thus, due to the evident difficulty in controlling the associated populations in the environment, it is very important to know how they are distributed in order to revise the standard measures. Previous studies confirmed the prevalence of *L. pneumophila* over other *Legionella* species in water and biofilm samples[390,393–395], although little is known about the distribution of *L. pneumophila* STs.

Our work focuses in the area of Comunidad Valenciana (CV, Eastern part of Spain), the second region that accrues most LD cases in the country. In Chapter 1, we have shown the presence of more than 100 different *L. pneumophila* STs in the overall community during a 16-year period spanning 23 different areas. But, in order to estimate the level of diversity, we decided to compare the observed variability in CV with that from a global dataset[221], containing information from 28 countries, mainly European. A comparable level of diversity was found in CV and the global dataset, which pointed to low geographical genetic differentiation, in agreement with a previous, albeit at a much smaller scale, study by our group[338]. Interestingly, 13 localities irrigated with water from a common watershed already shown 30 different STs[341], as noted in Chapter 2, where the within-group diversity also resulted comparable to that among groups when studying the different SBT loci separately.

The broad distribution of the *L. pneumophila* diversity found in Chapters 1 and 2, especially in the ST1 strains, contrasts with the more frequent observation of specific STs in particular areas, such as ST578s in Alcoy[117], ST47 in the United Kingdom[197,353] or ST23 in Italy[226]. In fact, 33% of the observed variance was found to be an effect of geographical structure, as noted in Chapter 1. However, approximately 25% of that variability was attributed to a temporal effect, confirming the well-known seasonality as a relevant factor affecting *L. pneumophila* populations[6,12,122,393,517]. Higher infection rates are often found in the warmer months of the year (June – October)[6,122], when temperatures are optimal for *Legionella* growth and aerosols are more likely to be produced. These conditions, jointly with other uncontrolled factors, as weather conditions[96] or the recently hypothesized transport vehicles[97], can help to its dispersal to very distant areas. This would explain the broad distribution of STs found in Chapters 1 and 2 of this thesis.

It is interesting to emphasize that, despite the high variability of *L. pneumophila*, a small subgroup of environmental STs was detected in clinical samples. Specifically, only 22 of the 102 profiles detected in CV in Chapter 1 were observed in patients. This trend has also been reported in other studies[197,225], pointing to the observation that environmental strains that cause infections may have particular genomic features or that selection may be acting on genes related to pathogenesis and virulence in this subgroup, increasing their fitness in the human host. However, the genome analysis presented in Chapter 5 showed no differentiation between clinical and environmental strains[339], although further genome analyses are needed to better investigate this assumption.

In CV, Alcoy (Alicante) is the locality in which more cases are reported and *Legionella* is considered as endemic in the area. This fact has made this city one of the focuses in the present thesis. Surveillance programs of *L. pneumophila* in high-risk installations are performed periodically as part of the preventive measures undertaken by public health authorities. However, efforts are more intense during the investigation of an outbreak or sporadic case, especially at the suspected reservoirs. This assumption led us to organize an extraordinary sampling programming with the first aim of studying the diversity of *L. pneumophila* in an endemic area in the absence of any epidemiological alert, as described in Chapter 3.

Water and biofilm samples were taken from hydrants, pipes dead-ends, nozzles and deposits in different points of the water distribution system in Alcoy. Samples were subjected to *Legionella*-specific microbiological culture, currently still the standard method for detecting the bacteria both from clinical and environmental sources[131,132,134,136]. However, *Legionella* is known to be able to enter in a viable but non-cultivable state (VBNC)[11,151] under changing conditions, complicating its identification in the environment. This has encouraged the introduction of molecular amplification by PCR and sequencing for the detection and typing of this bacterium[191,192,228,518]. Specifically, quantitative PCR is being applied increasingly because of its higher specificity than the traditional amplification[191,192]. Nevertheless, the low amounts of DNA retrieved from both types of samples complicated the set-up of a real-time PCR. This led us to develop a sensitive method based on a touchdown cycling step previous to the traditional amplification (TD-PCR) as well as the use of a high fidelity polymerase coupled with DMSO as adjuvant[352].

The results, detailed in Chapter 3, revealed the presence of *L. pneumophila* in only 15/60 samples considering three methods: culture, TD-

PCR from water and TD-PCR from biofilm samples. Only four strains were isolated by culture, revealing ST578 in only one case despite its well-known endemism in the area[117,339]. Direct sequencing from water and biofilm samples revealed difficulties in assigning a specific allele to the SBT scheme due to the presence of multiple strains, which was confirmed by cloning PCR products. Another interesting result from this study was the finding that *L. pneumophila* detection from biofilm samples clearly outperformed in ten-fold that from water samples. This organism is known to multiply within protozoa, mainly attached to biofilms in natural and built environments[20,21,335]. However, detection methods are often focused only in water. Despite the low number of positive samples, our results indicate that, under non-peak conditions, *L. pneumophila* is probably confined within biofilms, which would explain the low detection rate in water samples. This is an important and rapidly applicable finding for public health authorities, to which we are encouraging to revise the standard sampling measures in order to optimize the detection procedures.

Local public health authorities have been incorporating the method described in Chapter 3 in routine samplings and outbreak investigations since the project was developed. As an example, we were involved in the detection and genetic analysis of *L. pneumophila* from clinical and environmental samples taken at an outbreak occurred in a hotel in the locality of Calpe (Alicante, Spain) (Chapter 4). *L. pneumophila* detection from biofilms sampled at rooms and public facilities of the hotel was essential to link all the clinical cases to an environmental reservoir. Culturing revealed ST23 as the main causative strain of the outbreak, which was also found in water related to the spa pool by the same method. However, in the SBT analysis of clinical samples five patients showed mixed patterns, revealing the presence of more than one strain causing infection. ST578 was found by direct TD-PCR in biofilm samples from the hotel rooms, and most of the non-ST23 alleles in clinical samples matched the ST578 profile, as shown in Chapter 4. Other works have also revealed the possibility of mixed infections in LD cases[229–231].

But, despite the higher rate of *L. pneumophila* detection from biofilms, this kind of sample also presents some drawbacks. Apart from the low amount of DNA recovered in the extraction from swabs, already mentioned, Sanger sequencing PCR products from biofilms usually gives mixed assignations. This occurs due to the underlying diversity in the environment that results in the amplification of multiple strains at the same time. Cloning is needed in those cases in order to elucidate the specific alleles but, in the case of SBT, it is not possible to resolve the combinations of the true profiles. Cultures do not have this problem as SBT is performed from a single isolate.

Moreover, quantification remains an issue for public health officers, as well as knowing whether the *Legionella* cells are alive and able to produce infection. Specific methods for quantitative PCR approaches from water samples with the incorporation of ethidium monoazide (EMA) or propidium monoazide (PMA)[193,194] are being developed. These agents destruct DNA from dead cells by permeating through the affected membrane, allowing only the amplification of living cells, which will have intact membranes. Another issue that needs to be taken into consideration is that *Legionella* is a common inhabitant of water-related habitats. This fact makes very probable to find it during monitoring programs, although it has not caused a clinical case. In those cases, a positive result is indicative of the need of a deep cleaning. In the case of existing infections, its combination with epidemiological data is essential to clarify whether a specific facility has been causing the problem. Errors in the assignment of a risk facility as the reservoir of an ongoing outbreak could lead to serious economical and socio-political consequences.

Public health has started to benefit from the current high-throughput sequencing strategies that allow retrieving information about the complete genome of specific pathogens[275,292]. The huge amount of information obtained has already been proved invaluable to some outbreak investigations[281,283,284,444], to assess the effect of vaccine introductions[447] and even to reconstruct the global and local spread and transmission of pathogens[257,448]. Even, these studies can change our view of the history of major diseases, as it has been the case of *Mycobacterium tuberculosis*, which was thought to have a zoonotic origin for many years. However, genomic analyses confirmed that animal-adapted strains indeed derive from the human-adapted form[519], although a punctual zoonotic transfer from seals has been recently found from Peruvian human skeletons[520].

Most of the published works regarding pathogen genome sequencing have been focused on infectious agents with person-to-person transmission[255,257,283,443,448], but strictly environmental pathogens as in the case of *Legionella* are clearly underrepresented[286,442]. In Chapter 5 of this thesis, we present the first genome analysis of strains from multiple legionellosis outbreaks[339]. We focused the work in the locality of Alcoy, where 18 different outbreaks occurred in an 11-year period. This is the first study to analyse the structure of legionellosis outbreaks and has a profound implication for future public health interventions.

First, strains clustered in the phylogenetic tree by sequence type, despite their year of isolation or source (clinical or environmental). In the case of ST578 isolates, two sublineages were found, and those strains from the same outbreak showed high intra-ST variability and did not cluster together, as would be expected from a point source outbreak in which there is an outgrowth of a specific clone. This led us to conclude that legionellosis outbreaks in the endemic area of Alcoy may be caused by different strains that infect susceptible people under the appropriate conditions. Thus, the concept of a point-source outbreak does not apply in this case, where the accumulation of clinical cases in the city could be attributed to multiple sources. Besides, a Bayesian phylodynamic analysis from the ST578 core genome data revealed that sublineage A could have colonized the city in the early 90s, and sublineage B could have been introduced from uncontrolled external sources a few years later[117]. These results led us to conclude that SBT analysis of outbreaks, despite showing identity between putative source and clinical samples, could be misleading and we encourage the introduction of genome sequencing for the investigation of outbreaks in these cases because of its higher discrimination power.

This capability of differentiation between strains would help the epidemiological investigation in revealing the potential reservoir(s) and even the introduction of external variability that could re-colonize the area. In fact, we hypothesized that ST578B was introduced in the city in 2009, possibly by a paving machine[117] as detailed in Chapter 5, because it had not been found in the city before. Interestingly, this sublineage was also isolated in 2010 from a different source, indicating that it had already colonized the area. Genome monitoring of these sublineages could help to understand how they are evolving and to develop an intra-ST typing tool that could be applied to differentiate between both sublineages. However, decreasing costs in whole-genome sequencing make its routine application increasingly affordable.

Genome sequencing was also applied to five ST23 isolates from the Calpe outbreak described in Chapter 4. In this case, WGS revealed higher variability among those strains than that expected from a clonal outbreak. The results from this study, and the finding of ST23 in adjacent installations, support the conclusion that a population of ST23 strains was already present in the area and colonized the hotel during its construction and opening to the public. So, in this case, the intra-ST variability retrieved from genome sequencing was essential in order to reconstruct the colonization history of the hotel that ultimately led to the outbreak.

WGS projects reveal huge amounts of information about variability. However, the observed genetic variation can have other sources apart from point mutation. Many bacteria have developed strategies to exchange and incorporate genomic and plasmid segments from other organisms[435,436]. This mechanism allows the acquisition of genes or variants that provide them resistance to different stresses[438–441]. When donor and recipient are organisms from the same species, this process is known as recombination, otherwise, horizontal gene transfer (HGT) is more commonly used. Patterns of recombination and HGT are very visible in genomic analyses, as they are often represented as the accumulation of variation in particular regions.

Previous studies working with SBT data have already unveiled the exchange of genetic material among *L. pneumophila* strains[225,267], despite the initial hypothesis supporting clonality[261,262]. Chapters 1 and 2 of this thesis also provided evidence of recombination that involved SBT loci. Interestingly, Chapter 2 showed that five out of seven SBT segments were involved in intergenic recombination (all except *fliC* and *mompS*). The involvement of typing loci in transfer events is also supported by the genomic analysis in Chapter 5, in which a ST578 strain was typed as ST51 because of recombination of the *mip* locus. Chapter 5 revealed that 16 recombination events had occurred in a short-time scale of about 20 years in only five branches of the ST578 core genome phylogenetic tree, as estimated through Bayesian analysis. These events, spanning an average of 35.7 kb, accrued approximately 98% of the observed variability. Markers such as host-cell apoptosis inhibitors were found among the transferred segments. These results support the idea that strains involved in clinical cases undergo recombination events that could help them increase their fitness for and/or during human infection.

However, the picture that we get from the genomic analysis is only part from the actual existing diversity. This is because the study in Chapter 5 relies on strains that could be isolated through microbiological culture and it is well known that *Legionella* can live in a viable but non-cultivable state[11]. Previous works already support the presence of multiple strains of *L. pneumophila* in infected patients[229–231] and results from previous chapters in this thesis have shown mixed patterns also in water and biofilm samples. This observation supports the idea of the existence of much more variability in environmental sources than the one that we obtain from traditional microbiological cultures. Currently, high throughput sequencing approaches can get a more accurate picture through metagenomic studies.

In Chapter 6, we have used WGS to characterize the microbial, eukaryotic and viral community of biofilms from natural springs taken in the surrounding areas of the locality of Alcoy. As *L. pneumophila* has caused many outbreaks and sporadic cases, *Legionella* was expected in a considerable proportion in the environment. However, the genus *Legionella* was actually found in the minority fraction, in a relative abundance between 0.01-0.07%. This low proportion could explain the difficulties in finding *L. pneumophila* during routine surveillance programs. Although using molecular amplification by PCR, this low abundance is probably under the detection limit for the technique.

Taxonomic assignment was performed using a modified version of the Last Common Ancestor (LCA) algorithm that we denoted Taxonomic Vote, as detailed in Chapter 6. It is well known that metagenomic analyses can have multiple biases, which can come from the laboratory procedures, the sequencing process or the posterior analysis of sequences. Different methods for taxonomic classification use different databases and we have shown that they can determine the final result. In the Taxonomic Vote method, the top hits for each read are evaluated in order to get the most complete assignation possible by discarding those that could come from HGT events and which could compromise the final assignment.

Regarding the global composition and abundance, the general community was found dominated by *Proteobacteria, Cyanobacteria, Actinobacteria* and *Bacteroidetes*, with organisms frequently found in freshwater ecosystems as the most abundant genera, as detailed in Chapter 6. *Archaea*, *Eukaryotes* and *Viruses* were found around 1% of relative abundance, meaning that *Bacteria* are the leading components of the community. A similar pattern was observed for all the samples, although differences in relative abundances were found. No correlation was observed between genetic and geographic distances among the biofilms from different sampling points. This homogeneity could be explained because the sampling was performed in a very localized region or because we were comparing similar environments[509]. In order to confirm the later hypothesis, further work comparing the composition of these ecosystems in other areas will be needed.

In summary, the results obtained in the different chapters of this thesis provide new insights into the molecular epidemiology of *L. pneumophila*, how it is distributed in the environment and at which levels. The introduction of high throughput sequencing approaches allowed an in-depth analysis of genomic variability and recombination that involved clinical and environmental strains equally. These results have immediate applications to public health, such as the availability of a sensitive detection method from biofilm samples in real-time investigations. But also other applications, such as the population structure and dynamics of the pathogen as well as the information about its genome evolution, can be used to better understand and model its spread and to help in predicting and readily controlling the occurrence of outbreaks.

Conclusions

- The population structure of *L. pneumophila* reflects its high genetic variability. A broad distribution of STs was found in a reduced area, the Comunidad Valenciana, but also in a global dataset representing the overall (mainly European) diversity. In spite of these results, geographical and temporal differentiation was found, with some STs probably better adapted to specific locations, as it is the case of ST578 in Alcoy. Clinical strains were observed as a subset of the existing cultivable environmental diversity.

- *L. pneumophila* detection from biofilm samples is a very promising method as it outperforms that from water specimens. However, the method needs to be optimized to accommodate the needs of public health investigators: to quantify the number of cells and establish if they are active in the microbial population and able to produce disease.

- Biofilm testing during the investigation of an outbreak has been proved as a rapid and sensitive method for *L. pneumophila* detection. Our results allowed a prompt intervention of public health authorities during the outbreak occurred in a hotel in Calpe (2012), minimizing the number of cases.

- Genome analysis of ST578 outbreak-related strains revealed intra-ST variability and even the re-colonization by a sublineage that has been hypothesized as introduced from external sources. Public health authorities can benefit from such studies in order to have enough discrimination power for detecting the potential reservoir(s) and identify new incoming sources, as well as reconstructing the colonization history of a reservoir.

- The results from genome sequencing also indicated that legionellosis outbreaks in the endemic area of Alcoy could not have a single source, but they may be caused by different strains infecting susceptible people simultaneously when environmental conditions are appropriate.

- Recombination is the main driver in the evolution of *L. pneumophila* ST578 in Alcoy. Intergenic recombination events involving SBT loci have been found. Furthermore, above 98% of the genomic variability observed among outbreak-related ST578s in Alcoy has been accumulated by this process. Recombinant segments carry markers of resistance to external stresses and adaptation to intracellular life.

- Biofilms formed at natural springs in the area of Alcoy were found dominated by *Proteobacteria, Cyanobacteria, Actinobacteria* and *Bacteroidetes*, with organisms frequently found in the freshwater microbiota as the most abundant genera. The genus *Legionella* was observed in a very low relative abundance (0.01-0.07%). This result could explain that, despite ST578 causing recurrent outbreaks in the city, under normal conditions routine surveillance programs have difficulties in detecting its presence.

References

1. Brenner DJ, Steigerwalt AG, McDade JE. Classification of the Legionnaires' Disease Bacterium: *Legionella pneumophila*, genus *novum*, species *nova*, of the Family *Legionellaceae*, familia *nova*. Ann Intern Med. 1979;90(4):656.

2. Brenner DJ, Steigerwalt AG, Epple P, Bibb WF, Mckinney RM, Starnes RW, *et al. Legionella pneumophila* Serogroup Lansing 3 Isolated from a Patient with Fatal Pneumonia, and Descriptions of *L. pneumophila* subsp. *pneumophila* subsp. nov., *L. pneumophila* subsp. *fraseri* subsp. nov., and *L. pneumophila* subsp. *pascullei* subsp. nov. J Clin Microbiol. 1988;26(9):1695–703.

3. Reingold AL, Thomason BM, Brake BJ, Thacker L, Wilkinson HW, Kuritsky JN. *Legionella* Pneumonia in the United States: The Distribution of Serogroups and Species Causing Human Illness. J Infect Dis. 1984;149(5):819–819.

4. Yu VL, Plouffe JF, Pastoris MC, Stout JE, Schousboe M, Widmer A, *et al.* Distribution of *Legionella* species and serogroups isolated by culture in patients with sporadic community-acquired legionellosis: an international collaborative survey. J Infect Dis. 2002;186:127–8.

5. Fields BS, Benson RF, Besser RE. *Legionella* and Legionnaires' disease: 25 years of investigation. Clin Microbiol Rev. Am Soc Microbiol; 2002;15(3):506.

6. Beauté J, Zucs P, de Jong B, Network on behalf of the ELDS, Network ELDS. Legionnaires' disease in Europe, 2009-2010. Euro Surveill. 2013;18(10):20417.

7. Joly JR, Chen YY, Ramsay D. Serogrouping and subtyping of *Legionella pneumophila* with monoclonal antibodies. J Clin Microbiol. 1983 Nov;18(5):1040–6.

8. Joly JR, McKinney RM, Tobin JO, Bibb WF, Watkins ID, Ramsay D. Development of a standardized subgrouping scheme for *Legionella pneumophila* serogroup 1 using monoclonal antibodies. J Clin Microbiol. 1986;23(4):768–71.

9. Brenner DJ. Classification of *Legionellaceae*. Current status and remaining questions. Isr J Med Sci. 1986;22:620–32.

10. Mekkour M, Driss EKB, Tai J, Cohen N. *Legionella pneumophila*: An Environmental Organism and Accidental Pathogen. Int J Sci Technol. 2013;2(2):187–96.

11. Al-Bana BH, Haddad MT, Garduño RA. Stationary phase and mature infectious forms of *Legionella pneumophila* produce distinct viable but non-culturable cells. Environ Microbiol. 2014;16(2):382–95.

12. Fliermans CB, Cherry WB, Orrison LH, Smith J, Tison DL, Pope DH. Ecological Distribution of *Legionella pneumophila.* Appl Environ Microbiol. 1981;41(1):9–16.

13. Hughes MS, Steele TW. Occurrence and Distribution of *Legionella* Species in Composted Plant Materials. Appl Environ Microbiol. 1994;60(6):2003–5.

14. Casati S, Conza L, Bruin J, Gaia V. Compost facilities as a reservoir of *Legionella pneumophila* and other *Legionella* species. Clin Microbiol Infect. 2010;16(7):945–7.

15. Velonakis EN, Kiousi IM, Koutis C, Papadogiannakis E, Babatsikou F, Vatopoulos a. First isolation of *Legionella* species, including *L. pneumophila* serogroup 1, in Greek potting soils: possible importance for public health. Clin Microbiol Infect. 2010 Jun;16(6):763–6.

16. Koide M, Arakaki N, Saito A. Distribution of *Legionella longbeachae* and other *legionellae* in Japanese potting soils. J Infect Chemother. 2001;7(4):224–7.

17. Currie SL, Beattie TK, Knapp CW, Lindsay DSJ. *Legionella* spp. in UK composts--a potential public health issue? Clin Microbiol Infect. 2014;20(4):O224–9.

18. Steele TW, Lanser J, Sangster N. Isolation of *Legionella longbeachae* Serogroup 1 from potting mixes. Appl Environ Microbiol. 1990;56(1):49–53.

19. Steele T, Moore C, Sangster N. Distribution of *Legionella longbeachae* serogroup 1 and other *legionellae* in potting soils in Australia. Appl Environ Microbiol. 1990;56(10):2984–8.

20. Declerck P, Behets J, van Hoef V, Hoef V, Ollevier F. Replication of *Legionella pneumophila* in Floating Biofilms. Curr Microbiol. 2007;55(5):435–40.

21. Declerck P. Biofilms: the environmental playground of *Legionella pneumophila*. Environ Microbiol. 2010;12(3):557–66.

22. Stone BJ, Kwaik YA. Natural Competence for DNA Transformation by *Legionella pneumophila* and Its Association with Expression of Type IV Pili. J Bacteriol. 1999;181(5):1395–402.

23. Katz SM, Hammel JM. The effect of drying, heat, and pH on the survival of *Legionella pneumophila*. Ann Clin Lab Sci. 1987;17(3):150–6.

24. Fields BS. The molecular ecology of *legionellae*. Trends Microbiol. 1996 Jul;4(7):286–90.

25. Rowbotham TJ. Preliminary report on the pathogenicity of *Legionella pneumophila* for freshwater and soil amoebae. J Clin Pathol. Assoc Clin Pathol; 1980;33(12):1179–83.

26. Barbaree JM, Fields BS, Feeley JC, Gorman GW, Martin WT. Isolation of Protozoa from Water Associated with a Legionellosis Outbreak and Demonstration of Intracellular Multiplication of *Legionella pneumophila.* Appl Environ Microbiol. 1986;51(2):422–4.

27. Richards AM, Dwingelo JE Von, Price CT, Kwaik YA. Cellular microbiology and molecular ecology of *Legionella*-amoeba interaction. Virulence. 2013;4(4):1–8.

28. Ohno A, Kato N, Sakamoto R, Kimura S, Yamaguchi K. Temperature-dependent parasitic relationship between *Legionella pneumophila* and a free-living amoeba (*Acanthamoeba castellanii)*. Appl Environ Microbiol. 2008;74(14):4585–8.

29. Kilvington S, Price J. Survival of *Legionella pneumophila* within cysts of *Acanthamoeba polyphaga* following chlorine exposure. J Appl Bacteriol. 1990;68(5):519–25.

30. Donlan RM, Costerton JW. Biofilms: Survival Mechanisms of Clinically Relevant Microorganisms. Clin Microbiol Rev. 2002;15(2):167–93.

31. Donlan R. Biofilms: microbial life on surfaces. Emerg Infect Dis. 2002;8(9):881–90.

32. Hall-Stoodley L, Costerton JW, Stoodley P. Bacterial biofilms: from the natural environment to infectious diseases. Nat Rev Microbiol. 2004;2(2):95–108.

33. Watnick P, Kolter R. Biofilm, City of Microbes. J Bacteriol. 2000;182(10):2675–9.

34. Lehtola MJ, Laxander M, Miettinen IT, Hirvonen A, Vartiainen T, Martikainen PJ. The effects of changing water flow velocity on the formation of biofilms and water quality in pilot distribution system consisting of copper or polyethylene pipes. Water Res. 2006;40(11):2151–60.

35. Liu Z, Lin Y, Stout J. Effect of flow regimes on the presence of *Legionella* within the biofilm of a model plumbing system. J Appl Microbiol. 2006;101(2):437–42.

36. Ciesielski CA, Blaser MJ, Wang WL. Role of stagnation and obstruction of water Role of Stagnation and Obstruction of Water Flow in Isolation of *Legionella pneumophila* from Hospital Plumbing. Appl Environ Microbiol. 1984;48(5):984–7.

37. Costerton JW, Lewandowski Z, Caldwell DE, Korber DR, Lappin-Scott HM, Biofilms M. Microbial Biofilms. Lappin-Scott HM, Costerton JW, editors. Annu Rev Microbiol. Cambridge: Cambridge University Press; 1995;49:711–45.

38. Sidhu JPS, Toze SG. Human pathogens and their indicators in biosolids: a literature review. Environ Int. Elsevier B.V.; 2009;35(1):187–201.

39. Sharma S, Sachdeva P, Virdi JS. Emerging water-borne pathogens. Appl Microbiol Biotechnol. 2003;61(5-6):424–8.

40. Girones R, Ferrús M. Molecular detection of pathogens in water–the pros and cons of molecular techniques. Water Res. 2010;44:4325–39.

41. Piao Z, Sze CC, Barysheva O, Iida K, Yoshida S. Temperature-regulated formation of mycelial mat-like biofilms by *Legionella pneumophila*. Appl Environ Microbiol. 2006;72(2):1613–22.

42. Mampel J, Spirig T, Weber SS, Haagensen JAJ, Molin S, Hilbi H, *et al.* Planktonic Replication Is Essential for Biofilm Formation by *Legionella pneumophila* in a Complex Medium under Static and Dynamic Flow Conditions. Appl Environ Microbiol. 2006;72(4):2885–95.

43. Stoodley P, Wilson S, Hall-Stoodley L, Boyle JD, Lappin-Scott HM, Costerton JW. Growth and detachment of cell clusters from mature mixed-species biofilms. Appl Environ Microbiol. 2001;67(12):5608–13.

44. Stoodley P, Cargo R, Rupp CJ, Wilson S, Klapper I. Biofilm material properties as related to shear-induced deformation and detachment phenomena. J Ind Microbiol Biotechnol. 2002;29(6):361–7.

45. Storey M V, Ashbolt NJ, Stenström TA. Biofilms, thermophilic amoebae and *Legionella pneumophila*--a quantitative risk assessment for distributed water. Water Sci Technol. 2004;50(1):77–82.

46. Berk SG, Ting RS, Turner GW, Ashburn RJ. Production of Respirable Vesicles Containing Live *Legionella pneumophila* Cells by Two Acanthamoeba spp . Appl Environ Microbiol. 1998;64(1):279–86.

47. Edelstein P. Comparative study of selective media for isolation of *Legionella pneumophila* from potable water. J Clin Microbiol. 1982;15(3):506–26.

48. Steinert M, Hentschel U, Hacker J. *Legionella pneumophila*: an aquatic microbe goes astray. FEMS Microbiol Rev. 2002;26:149–62.

49. Horwitz MA. Formation of a novel phagosome by the Legionnaires' Disease bacterium (*Legionella pneumophila*) in human monocytes. J Exp Med. 1983;158:1319–31.

50. Horwitz MA, Silverstein SC. Legionnaires' disease bacterium (*Legionella pneumophila*) multiples intracellularly in human monocytes. J Clin Invest. 1980;66(3):441–50.

51. Nash TW, Libby DM, Horwitz MA. Interaction between the legionnaires' disease bacterium (*Legionella pneumophila*) and human alveolar macrophages. Influence of antibody, lymphokines, and hydrocortisone. J Clin Invest. 1984;74(3):771–82.

52. Chien M, Morozova I, Shi S, Sheng H, Chen J, Gomez SMSM, et al. The Genomic Sequence of the Accidental Pathogen *Legionella pneumophila*. Science. 2004;305(5692):1966–8.

53. Newton HJ, Ang DKY, van Driel IR, Hartland EL. Molecular pathogenesis of infections caused by *Legionella pneumophila.* Clin Microbiol Rev. 2010;23(2):274–98.

54. Segal GIL, Shuman HA. *Legionella pneumophila* Utilizes the Same Genes To Multiply within *Acanthamoeba castellanii* and Human Macrophages. Infect Immun. 1999;67(5):2117–24.

55. Horwitz MA. Phagocytosis of the Legionnaires' disease bacterium (*Legionella pneumophila*) occurs by a novel mechanism: engulfment within a pseudopod coil. Cell. Elsevier; 1984;36(1):27–33.

56. Cirillo JD, Falkow S, Tompkins LS. Growth of *Legionella pneumophila* in *Acanthamoeba castellanii* enhances invasion. Infect Immun. 1994 Aug;62(8):3254–61.

57. Isberg RR, O'Connor TJ, Heidtman M. The *Legionella pneumophila* replication vacuole: making a cosy niche inside host cells. Nat Rev Microbiol. 2009;7(1):13–24.

58. Price CT, Al-Khodor S, Al-Quadan T, Santic M, Habyarimana F, Kalia A, *et al*. Molecular mimicry by an F-box effector of *Legionella pneumophila* hijacks a conserved polyubiquitination machinery within macrophages and protozoa. PLOS Pathog. 2009;5(12):e1000704.

59. Dorer MS, Kirton D, Bader JS, Isberg RR. RNA interference analysis of *Legionella* in *Drosophila* cells: exploitation of early secretory apparatus dynamics. PLOS Pathog. 2006;2(4):e34.

60. Robinson CG, Roy CR. Attachment and fusion of endoplasmic reticulum with vacuoles containing *Legionella pneumophila*. Cell Microbiol. 2006;8(5):793–805.

61. Sadosky AB, Wiater LA, Shuman HA. Identification of *Legionella pneumophila* Genes Required for Growth within and Killing of Human Macrophages. Infect Immun. 1993;61(12):5361–73.

62. Marra A, Steven J, Horwitzt MA, Shuman HA. Identification of a *Legionella pneumophila* locus required for intracellular multiplication in human macrophages. Proc Natl Acad Sci. 1992;89:9607–11.

63. Segal G, Shuman HA. Characterization of a New Region Required for Macrophage Killing by *Legionella pneumophila*. Infect Immun. 1997;65(12):5057–66.

64. Berger KH, Merriam JJ, Isberg RR. Altered intracellular targeting properties associated with mutations in the *Legionella pneumophila dotA* gene. Mol Microbiol. 1994;14(4):809–22.

65. Segal G, Purcell M, Shuman HA. Host cell killing and bacterial conjugation require overlapping sets of genes within a 22-kb region of the *Legionella pneumophila* genome. Proc Natl Acad Sci. 1998;95:1669–74.

66. Vogel JP. Conjugative Transfer by the Virulence System of *Legionella pneumophila*. Science. 1998;279(5352):873–6.

67. Ensminger AW, Isberg RR*. Legionella pneumophila* Dot/Icm translocated substrates: a sum of parts. Curr Opin Microbiol. 2009;12(1):67–73.

68. Gomez-Valero L, Rusniok C, Rolando M, Neou M, Dervins-Ravault D, Demirtas J, *et al.* Comparative analyses of *Legionella* species identifies genetic features of strains causing Legionnaires' disease. Genome Biol. 2014;15(11):505.

69. Zhu W, Banga S, Tan Y, Zheng C, Stephenson R, Gately J, et al. Comprehensive identification of protein substrates of the Dot/Icm type IV transporter of *Legionella pneumophila*. PLoS One. 2011;6(3):e17638.

70. O'Connor TJ, Adepoju Y, Boyd D, Isberg RR, Connor TJO. Minimization of the *Legionella pneumophila* genome reveals chromosomal regions involved in host range expansion. Proc Natl Acad Sci. 2011;108(36):14733.

71. O'Connor T, Boyd D, Dorer MSM, Isberg RRR, Connor TJO. Aggravating genetic interactions allow a solution to redundancy in a bacterial pathogen. Science. 2012;338(December):1440–4.

72. Cazalet C, Rusniok C, Brüggemann H, Zidane N, Magnier A, Ma L, *et al.* Evidence in the *Legionella pneumophila* genome for exploitation of host cell functions and high genome plasticity. Nat Genet. 2004;36(11):1165–73.

73. Cianciotto NP, Eisenstein BI, Mody CH, Toews GB, Englebergl NC. A *Legionella pneumophila* Gene Encoding a Species-Specific Surface Protein Potentiates Initiation of Intracellular Infection. Infect Immun. 1989;57(4):1255–62.

74. Cianciotto NP, Eisenstein BI, Mody CH, Engleberg NC. A Mutation in the *mip* Gene Results in an Attenuation of *Legionella pneumophila* Virulence. J Infect Dis. 1990;162(1):121–6.

75. Cianciotto NP, Fields BS. *Legionella pneumophila mip* gene potentiates intracellular infection of protozoa and human macrophages. Proc Natl Acad Sci. 1992;89(June):5188–91.

76. Laguna RK, Creasey EA, Li Z, Valtz N, Isberg RR. A *Legionella pneumophila*-translocated substrate that is required for growth within macrophages and protection from host cell death.. 2006 Dec 5 p. 18745–50.

77. Case CL, Kohler LJ, Lima JB, Strowig T, Zoete MRD, Flavell RA, *et al.* Caspase-11 stimulates rapid flagellin-independent pyroptosis in response to *Legionella pneumophila*. Proc Natl Acad Sci. 2013;110(5):1851–6.

78. Zamboni DS, Kobayashi KS, Kohlsdorf T, Ogura Y, Long EM, Vance RE, *et al*. The Birc1e cytosolic pattern-recognition receptor contributes to the detection and control of *Legionella pneumophila* infection. Nat Immunol. 2006;7(3):318–25.

79. Molmeret M, Bitar DM, Han L, Kwaik YA. Disruption of the phagosomal membrane and egress of *Legionella pneumophila* into the cytoplasm during the last stages of intracellular infection of macrophages and Acanthamoeba polyphaga. Infect Immun. 2004;72(7):4040–51.

80. Alli OAT, Gao L, Pedersen LL, Zink S, Radulic M, Doric M*, et al.* Temporal Pore Formation-Mediated Egress from Macrophages and Alveolar Epithelial Cells by *Legionella pneumophila.* Infect Immun. 2000;68(11):6431–40.

81. Gao L, Kwaik YA. The mechanism of killing and exiting the protozoan host *Acanthamoeba polyphaga* by *Legionella pneumophila*. Environ Microbiol. 2000;2(1):79–90.

82. Fraser DW, Tsai TR, Orenstein W, Parkin WE, Beecham HJ, Sharrar RG, et al. Legionnaires' disease: description of an epidemic of pneumonia. N Engl J Med. 1977;297(22):1189–97.

83. McDade JE, Shepard C, Fraser DW, Tsai TR, Redus MA, Dowdle WR, *et al.* Isolation of a Bacterium and Demonstration of Its Role in Other Respiratory Disease. N Engl J Med. 1977;297(22):1197–203.

84. McDade JE, Brenner DJ, Bozeman FM. Legionnaires' disease bacterium isolated in 1947. Ann Intern Med. 1979;90:659–61.

85. Osterholm MT, Chin TDY, Osborne DO, Dull HB, Dean AG, Fraser DW, *et al.* A 1957 outbreak of Legionnaires' Disease associated with a meat packing plant. Am J Epidemiol. 1983;117(1):60–7.

86. Thacker S, Bennett J, Tsai T, Fraser D, McDade J, Shepard C, *et al.* An Outbreak in 1965 of Severe Respiratory Illness Caused by the Legionnaires' Disease Bacterium. J Infect Dis. 1978;138(4):512–9.

87. Terranova W, Cohen ML, Fraser DW. 1974 Outbreak of Legionnaires' Disease Diagnosed in 1977. Clinical and Epidemiological Features. Lancet. 1978;312(8081):122–4.

88. Glick TH, Gregg MB, Berman B, Mallison G, Rhodes WW, Kassanoff I. Pontiac fever. An epidemic of unknown etiology in a health department: I. Clinical and epidemiologic aspects. Am J Epidemiol. 1978;107(2):149–60.

89. Phin N, Parry-Ford F, Harrison T, Stagg HR, Zhang N, Kumar K, *et al.* Epidemiology and clinical management of Legionnaires' disease. Lancet Infect Dis. Elsevier Ltd; 2014;14(10):1011–21.

90. Boer J Den, Peeters M, Ketel R van. A large outbreak of Legionnaires' disease at a flower show, the Netherlands, 1999. Emerg Infect Dis. 2002;8(1):37–43.

91. Den Boer JW, Nijhof J, Friesema I. Risk factors for sporadic community-acquired Legionnaires' disease. A 3-year national case-control study. Public Health. 2006;120(6):566–71.

92. Tsakris A, Souliou E, Antoniadis A. In-vitro activity of antibiotics against *Legionella pneumophila* isolates from water systems. J Antimicrob Chemother. 1999;44:693–5.

93. Nielsen K, Bangsborg JM, Høiby N. Susceptibility of *Legionella* species to five antibiotics and development of resistance by exposure to erythromycin, ciprofloxacin, and rifampicin. Diagn Microbiol Infect Dis. 2000 Jan;36(1):43–8.

94. Alexandropoulou IG, Parasidis TA, Konstantinidis TG, Constantinidis TC, Panopoulou M. Antibiotic Susceptibility Surveillance of Environmental *Legionella* Strains: Application of the E-Test to Bacteria Isolated From Hospitals in Greece. J Infect Dis Ther. 2014;02(01):1–2.

95. Onody C, Matsiota-Bernard P, Nauciel C. Lack of resistance to erythromycin, rifampicin and ciprofloxacin in 98 clinical isolates of *Legionella pneumophila.* J Antimicrob Chemother. 1997;39:815–6.

96. Garcia-Vidal C, Labori M, Viasus D, Simonetti A, Garcia-Somoza D, Dorca J, *et al.* Rainfall is a risk factor for sporadic cases of *Legionella pneumophila* pneumonia. PLoS One. 2013;8(4):e61036.

97. Sakamoto R, Ohno a, Nakahara T, Satomura K, Iwanaga S, Kouyama Y, *et al*. Is driving a car a risk for Legionnaires' disease? Epidemiol Infect. 2009;137(11):1615–22.

98. Dennis P, Lee J. Differences in aerosol survival between pathogenic and non-pathogenic strains of *Legionella pneumophila* serogroup 1. J Appl Bacteriol. 1988;65(2):135–41.

99. Straus WL. Risk Factors for Domestic Acquisition of Legionnaires Disease. Arch Intern Med. 1996;156(15):1685. 011

100. Stout J, Yu V, Yee Y, Vaccarello S, Diven W, Lee T. *Legionella pneumophila* in residential water supplies: environmental surveillance with clinical assessment for Legionnaires' disease. Epidemiol Infect. 1992;109:49–57.

101. Stout JE, Yu VL, Muraca P, Joly J, Troup N, Tompkins LS. Potable water as a cause of sporadic cases of community-acquired Legionnaires' Disease. N Engl J Med. 1992;326(3):151–5.

102. Jong B de, Hallström L, Robesyn E, Ursut D, Zucs P, ELDSNet. Travel-associated Legionnaires' disease in Europe, 2010. Euro Surveillance. 2013. p. pii: 20498.

103. Rota M, Scaturro M, Fontana S, Foroni M, Boschetto G, Trentin L, *et al.* Cluster of travel-associated Legionnaires disease in Lazise, Italy, July to August 2011. Euro Surveill. 2011 Jan;16(40):1–3.

104. Burnsed LJ, Hicks LA, Smithee LMK, Fields BS, Bradley KK, Pascoe N, *et al.* A large, travel-associated outbreak of legionellosis among hotel guests: utility of the urine antigen assay in confirming Pontiac fever. Clin Infect Dis. 2007 Jan 15;44(2):222–8.

105. Gotz HM, Tegnell A, De Jong B, Broholm KA, Kuusi M, Kallings I, *et al.* A whirlpool associated outbreak of Pontiac fever at a hotel in Northern Sweden. Epidemiol Infect. 2001;126:241–7.

106. Erdogan H, Arslan H. Colonization of *Legionella* Species in Hotel Water Systems in Turkey. J Travel Med. 2007;14(6):369–73.

107. Vanaclocha H, Guiral S, Morera V, Calatayud MA, Castellanos M, Moya V, *et al.* Preliminary report: outbreak of Legionnaires disease in a hotel in Calp, Spain, update on 22 February 2012. Euro Surveill. 2012 Jan;17(8):31–3.

108. Mouchtouri V, Velonakis E, Tsakalof A, Kapoula C, Goutziana G, Vatopoulos A, *et al.* Risk factors for contamination of hotel water distribution systems by *Legionella* species. Appl Environ Microbiol. 2007;73(5):1489–92.

109. Leoni E, Luca G De, Legnani PP, Sacchetti R, Stampi S, Zanetti F. *Legionella* waterline colonization: detection of *Legionella* species in domestic, hotel and hospital hot water systems. J Appl Microbiol. 2005;98:373–9.

110. Codony F, Alvarez J, Oliva JM, Ciurana B, Company M, Camps N, *et al*. Factors Promoting Colonization by *Legionellae* in Residential Water Distribution Systems: an Environmental Case-Control Survey. Eur J Clin Microbiol Infect Dis. 2002;21(10):717–21.

111. Krøjgaard LH, Krogfelt KA, Albrechtsen H-J, Uldum SA. Cluster of Legionnaires disease in a newly built block of flats, Denmark, December 2008 - January 2009. Euro Surveill. 2011 Jan;16(1).

112. Ditommaso S, Giacomuzzi M, Rivera SRA, Raso R, Ferrero P, Zotti CM. Virulence of *Legionella pneumophila* strains isolated from hospital water system and healthcare-associated Legionnaires' disease in Northern Italy between 2004 and 2009. BMC Infect Dis. 2014;14(1):483.

113. Viasus D, Di Yacovo S, Garcia-Vidal C, Verdaguer R, Manresa F, Dorca J, *et al.* Community-acquired *Legionella pneumophila* pneumonia: a single-center experience with 214 hospitalized sporadic cases over 15 years. Medicine (Baltimore). 2013;92(1):51–60.

114. Haupt TTE, Heffernan RRT, Kazmierczak JJ, Nehls-Lowe H, Rheineck B, Powell C, *et al.* An outbreak of Legionnaires disease associated with a decorative water wall fountain in a hospital. Infect Control Hosp Epidemiol. 2012;33(2):185–91.

115. Kool J, Buchholz U, Peterson C, Brown E, Benson RF, Pruckler J, *et al.* Strengths and limitations of molecular subtyping in a community outbreak of Legionnaires' disease. Epidemiol Infect. 2000;125:599–608.

116. Pankhurst CL, Coulter WA. Do contaminated dental unit waterlines pose a risk of infection? J Dent. 2007;35(9):712–20.

117. Coscollá M, Fenollar J, Escribano I, González-Candelas F. Legionellosis outbreak associated with asphalt paving machine, Spain, 2009. Emerg Infect Dis. 2010;16(9):1381–7.

118. Phin N, Cresswell T, Parry-Ford F, Team TIC. Case of Legionnaires' disease in a neonate following a home birth in a heated birthing pool, England, June. Euro Surveill. 2014;(June).

119. Dominguez A, Alvarez J. Factors influencing the case-fatality rate of Legionnaires' disease. Int J Tuberc Lung Dis. 2009;13(3):407–12.

120. Joseph CA, Ricketts KD. Legionnaires disease in Europe 2007-2008. Euro Surveill. 2010 Feb;15(8):19493.

121. Benin AL, Benson RF, Besser RE. Trends in legionnaires disease, 1980–1998: declining mortality and new patterns of diagnosis. Clinical Infectious Diseases. 2002;35:1039-46.

122. De Jong B, Coulomb, Hallström LP, Takkinen J, Ursut D, Zucs P. Legionnaires' disease in Europe 2012. 2012 p. 1–32. ¡

123. Hicks LA, Garrison LE, Nelson GE, Hampton LM. Legionellosis - United States, 2000-2009. 2011 p. 1083–6.

124. Ng V, Tang P, Jamieson F. Laboratory-based evaluation of legionellosis epidemiology in Ontario, Canada, 1978 to 2006. BMC Infect Dis. 2009;9:68.

125. Graham FF, White PS, Harte DJG, Kingham SP. Changing epidemiological trends of legionellosis in New Zealand, 1979-2009. Epidemiol Infect. 2012;140(8):1481–96.

126. Guest C, O'Brien E. A review of national legionellosis surveillance in Australia, 1991 to 2000. Commun Dis Intell. 2002;26(3):461–8.

127. NNDSS ARWG. Australia's notifiable disease status, 2010: Annual report of the National Notifiable Diseases Surveillance System. Commun Dis Intell. 2012;36(1):1–69.

128. Ozeki Y, Yamada F, Saito A, Kishimoto T, Tanno S, Nakamura Y. Seasonal Patterns of Legionellosis in Saitama, 2005-2009. Jpn J Infect Dis. 2012;65(4):330–3.

129. Lam M. Epidemiology and Control of Legionellosis, Singapore. Emerg Infect Dis. 2011;17(7):1209–15.

130. (WHO) WHO. *Legionella* and the prevention of legionellosis. 2007.

131. Kozak NA, Lucas CE, Winchell JM. Identification of *Legionella* in the environment. Methods Mol Biol. 2013;954:3–25.

132. Blyth CC, Adams DN, Chen SCA. Diagnostic and typing methods for investigating *Legionella* infection. N S W Public Health Bull. 2009;20(10):157.

133. Feeley JC, Gorman JGW, Weaver RE, Mackel DONC, Smith HW. Primary Isolation Media for Legionnaires Disease Bacterium. J Clin Microbiol. 1978;8(3):320–5.

134. Feeley JC, Gibson RJ, Gorman GW, Langford NC, Rasheed JK, Mackel DC, *et al.* Charcoal-yeast extract agar: primary isolation medium for *Legionella pneumophila.* J Clin Microbiol. 1979;10(4):437–41.

135. Edelstein PH. Improved Semiselective Medium for Isolation of *Legionella pneumophila* from Contaminated Clinical and Environmental Specimens. J Clin Microbiol. 1981;14(3):298–303.

136. Lück PC, Igel L, Helbig JH, Kuhlisch E, Jatzwauk L. Comparison of commercially available media for the recovery of *Legionella* species. Int J Hyg Environ Health. 2004;207(6):589–93.

137. Boulanger C, Edelstein P. Precision and accuracy of recovery of *Legionella pneumophila* from seeded tap water by filtration and centrifugation. Appl Environ Microbiol. 1995;61(5):1805–9.

138. Ta AC, Stout JE, Yu VL, Wagener MM. Comparison of Culture Methods for Monitoring *Legionella* Species in Hospital Potable Water Systems and Recommendations for Standardization of Such Methods. J Clin Microbiol. 1995;33(8):2118–23.

139. Mathieu L, Robine E, Deloge-Abarkan M, Ritoux S, Pauly D, Hartemann P, *et al. Legionella* Bacteria in Aerosols: Sampling and Analytical Approaches Used during the Legionnaires Disease Outbreak in Pas-de-Calais. The Journal of Infectious Diseases. 2006. p. 1333–5.

140. Ishimatsu S, Miyamoto H, Hori H, Tanaka I, Yoshida S. Sampling and detection of *Legionella pneumophila* aerosols generated from an industrial cooling tower. Ann Occup Hyg. 2001 Aug;45(6):421–7.

141. Wilkinson HW, Fikes BJ. Slide agglutination test for serogrouping *Legionella pneumophila* and atypical *Legionella*-like organisms. J Clin Microbiol. 1980 Jan;11(1):99–101.

142. Cherry W, Pittman B, Harris PP, Hebert GA, Thomason BM, Thacker L, *et al.* Detection of Legionnaires disease bacteria by direct immunofluorescent staining. J Clin Microbiol. 1978;8(3):329–38.

143. Thacker W, Wilkinson HW, Benson RF. Comparison of slide agglutination test and direct immunofluorescence assay for identification of *Legionella* isolates. J Clin Microbiol. 1983;18(5):1113–8.

144. Helbig JH, Bernander S, Pastoris MC, Etienne J, Gaia V, Lauwers S, *et al.* Pan-European Study on Culture-Proven Legionnaires' Disease: Distribution of *Legionella pneumophila* Serogroups and Monoclonal Subgroups. Eur J Clin Microbiol Infect Dis. 2002;21(10):710–6.

145. Gaia V, Casati S, Tonolla M. Rapid identification of *Legionella* spp. by MALDI-TOF MS based protein mass fingerprinting. Syst Appl Microbiol. 2011;34(1):40–4.

146. Van Belkum A, Tassios PT, Dijkshoorn L, Haeggman S, Cookson B, Fry NK, *et al.* Guidelines for the validation and application of typing methods for use in bacterial epidemiology. Clin Microbiol Infect. 2007 Oct;13(3):1–46.

147. Svarrer CW, Uldum SA. The occurrence of *Legionella* species other than *Legionella pneumophila* in clinical and environmental samples in Denmark identified by *mip* gene sequencing and matrix-assisted laser desorption ionization time-of-flight mass spectrometry. Clin Microbiol Infect. 2012;18(10):1004–9.

148. Ng DLK, Koh BB, Tay L, Heng BH. Comparison of polymerase chain reaction and conventional culture for the detection of *legionellae* in cooling tower waters in Singapore. Lett Appl Microbiol. 1997 Mar;24(3):214–6.

149. Villari P, Motti E, Farullo C, Torre I. Comparison of conventional culture and PCR methods for the detection of *Legionella pneumophila* in water. Lett Appl Microbiol. 1998 Aug;27(2):106–10.

150. Paszko-Kolva C, Shahamat M, Colwell RR. Long-term survival of *Legionella pneumophila* serogroup 1 under low-nutrient conditions and associated morphological changes. FEMS Microbiol Lett. 1992;102(1):45–55.

151. Hussong D, Colwell RR, O'Brien M, Weiss E, Pearson AD, Weiner RM, *et al.* Viable *Legionella pneumophila* Not Detectable by Culture on Agar Media. Nat Biotechnol. 1987;5(9):947–50.

152. Oliver JD. Recent findings on the viable but nonculturable state in pathogenic bacteria. FEMS Microbiol Rev. 2010;34(4):415–25.

153. Oliver J. The viable but nonculturable state in bacteria. J Microbiol. 2005 Feb;43 Spec No:93–100.

154. Steinert M, Emödy L, Amann R, Hacker J. Resuscitation of viable but nonculturable *Legionella pneumophila* Philadelphia JR32 by *Acanthamoeba castellanii.* Appl Environ Microbiol. 1997 May;63(5):2047–53.

155. Helbig JH, Uldum SA, Bernander S, Lück PC, Wewalka G, Abraham B, *et al.* Clinical Utility of Urinary Antigen Detection for Diagnosis of Community-Acquired, Travel-Associated, and Nosocomial Legionnaires' Disease. J Clin Microbiol. 2003;41(2):838–40.

156. Jarraud S, Descours G, Ginevra C, Lina G, Etienne J. Identification of *Legionella* in clinical samples. Methods Mol Biol. 2013;954:27–56.

157. Alleron L, Khemiri A, Koubar M, Lacombe C, Coquet L, Cosette P, *et al*. VBNC *Legionella pneumophila* cells are still able to produce virulence proteins. Water Res. 2013;47(17):6606–17.

158. Rowbotham TJ. Isolation of *Legionella pneumophila* from clinical specimens via amoebae, and the interaction of those and other isolates with amoebae. J Clin Pathol. 1983;36:978–86.

159. La Scola B, Mezi L, Weiller PJ, Raoult D. Isolation of *Legionella anisa u*sing an amoebic coculture procedure. J Clin Microbiol. 2001;39(1):365–7.

160. Brown S, Bibb W, McKinney R. Retrospective examination of lung tissue specimens for the presence of *Legionella* organisms: comparison of an indirect fluorescent-antibody system with direct. J Clin Microbiol. 1984;19(4):468–72.

161. Blackmon JA, Chandler FW, Cherry WB, England III AC, Feeley JC, Hicklin MD, *et al.* Legionellosis. Am J Pathol. 1981;103(3):429–65.

162. Wilkinson HW, Farshy CE, Fikes BJ, Cruce DD, Yealy LP. Measure of Immunoglobulin G-, M-, and A-specific titers against *Legionella pneumophila* and inhibition of titers against nonspecific, gram-negative bacterial antigens in the indirect immunofluorescence test. J Clin Microbiol. 1979;10(5):685–9.

163. Dufresne SF, Locas MC, Duchesne A, Restieri C, Ismaïil J, Lefebvre B, *et al*. Sporadic Legionnaires' disease: the role of domestic electric hot-water tanks. Epidemiol Infect. 2012;140(01):172–81.

164. Ratcliff R, Donnellan S, Lanser JA, Manning PA, Heuzenroeder MW. Interspecies sequence differences in the Mip protein from the genus *Legionella*: implications for function and evolutionary relatedness. Mol Microbiol. 1997;25(6):1149–58.

165. Ratcliff RM, Lanser JA, Manning PA Heuzenroeder MW. Sequence-based classification scheme for the genus *Legionella* targeting the *mip* gene. J Clin Microbiol. 1998 Jun;36(6):1560–7.

166. Riffard S, Vandenesch F, Reyrolle M, Etienne J. Distribution of *mip*-related sequences in 39 species (48 serogroups) of *Legionellaceae*. Epidemiol Infect. 1996;117:501–6.

167. Fry NK, Afshar B, Bellamy W, Underwood AP, Ratcliff RM, Harrison TG. Identification of *Legionella* spp. by 19 European reference laboratories: results of the European Working Group for *Legionella* Infections External Quality Assessment Scheme using DNA sequencing of the macrophage infectivity potentiator gene and dedicated online tools. Clin Microbiol Infect. 2007;13(11):1119–24.

168. Haroon A, Koide M, Higa F, Tateyama M, Fujita J. Identification of *Legionella pneumophila* serogroups and other *Legionella* species by mip gene sequencing. J Infect Chemother. 2012;18(2):276–81.

169. Ludwig W, Stackebrandt E. A phylogenetic analysis of *Legionella*. Arch Microbiol. 1983;135(1):45–50.

170. Hookey J, Saunders N, Fry NK, Birtles RJ, Harrison TG. Phylogeny of *Legionellaceae* based on small-subunit ribosomal DNA sequences and proposal of *Legionella lytica* comb. nov. for *Legionella*-like amoebal pathogens. Int J Syst Bacteriol. 1996;(28):526–31.

171. Fry NK, Warwick S, Saunders N, Embley T. The use of 16S ribosomal RNA analyses to investigate the phylogeny of the family *Legionellaceae*. J Gen Microbiol. 1991 Jul;137:1215–22.

172. Ko KS, Lee HK, Park M, Yun Y, Woo S, Miyamoto H, *et al.* Application of RNA Polymerase β-Subunit Gene (*rpoB*) Sequences for the Molecular Differentiation of *Legionella* Species. J Clin Microbiol. 2002;40(7):2653–8.

173. Rubin C-J, Thollesson M, Kirsebom LA, Herrmann B. Phylogenetic relationships and species differentiation of 39 *Legionella* species by sequence determination of the RNase P RNA gene *rnpB*. Int J Syst Evol Microbiol. 2005;55(Pt 5):2039–49.

174. Ratcliff RM. Sequence-based identification of *Legionella*. Methods Mol Biol. 2013;954:57–72.

175. Grattard F, Ginevra C, Riffard S, Ros A, Jarraud S, Etienne J, *et al.* Analysis of the genetic diversity of *Legionella* by sequencing the 23S-5S ribosomal intergenic spacer region: from phylogeny to direct identification of isolates at the species level from clinical specimens. Microbes Infect. 2006;8(1):73–83.

176. Feddersen A, Meyer H-GW, Matthes P, Bhakdi S, Husmann M. *GyrA* sequence-based typing of *Legionella.* Med Microbiol Immunol. 2000;189(1):7–11.

177. Cloud JL, Carroll KC, Pixton P, Erali M, Hillyard DR. Detection of *Legionella* species in respiratory specimens using PCR with sequencing confirmation. J Clin Microbiol. 2000;38(5):1709–12.

178. Rantakokko-jalava K, Jalava J. Development of conventional and real-time PCR assays for detection of *Legionella* DNA in respiratory specimens. J Clin Microbiol. Am Soc Microbiol; 2001;39(8):2904–10.

179. Nazarian EJ, Bopp DJ, Saylors A, Limberger RJ, Musser KA. Design and implementation of a protocol for the detection of *Legionella* in clinical and environmental samples. Diagn Microbiol Infect Dis. 2008;62(2):125–32.

180. Maurin M, Hammer L, Gestin B, Timsit JF, Rogeaux O, Delavena F, *et al.* Quantitative real-time PCR tests for diagnostic and prognostic purposes in cases of legionellosis. Clin Microbiol Infect. 2010;16(4):379–84.

181. Haroon A, Koide M, Higa F, Hibiya K, Tateyama M, Fujita J. Repetitive element-polymerase chain reaction for genotyping of clinical and environmental isolates of *Legionella* spp. Diagn Microbiol Infect Dis. Elsevier Inc.; 2010;68(1):7–12.

182. Mérault N, Rusniok C, Jarraud S, Gomez-Valero L, Cazalet C, Marin M, *et al.* Specific real-time PCR for simultaneous detection and identification of *Legionella pneumophila* serogroup 1 in water and clinical samples. Appl Environ Microbiol. 2011;77(5):1708–17.

183. Mentasti M, Fry NK, Afshar B, Palepou-Foxley C, Naik FC, Harrison TG. Application of *Legionella pneumophila*-specific quantitative real-time PCR combined with direct amplification and sequence-based typing in the diagnosis and epidemiological investigation of Legionnaires' disease. Eur J Clin Microbiol Infect Dis. Springer Berlin / Heidelberg; 2012;1–12.

184. Benitez AJ, Winchell JM. Clinical application of a multiplex real-time PCR assay for simultaneous detection of *Legionella* species, *Legionella pneumophila*, and *Legionella pneumophila* serogroup 1. J Clin Microbiol. 2013;51(1):348–51.

185. Yáñez MA, Carrasco-Serrano C, Barbera VM, Catalán V, Ya MA, Catala V. Quantitative Detection of *Legionella pneumophila* in Water Samples by Immunomagnetic Purification and Real-Time PCR Amplification of the *dotA* Gene. Appl Environ Microbiol. 2005;71(7):3433–41.

186. Stojek NS, Wójcik-Fatla A, Dutkiewicz J. Efficacy of the detection of *Legionella* in hot and cold water samples by culture and PCR. II. Examination of native samples from various sources. Ann Agric Environ Med. 2012;19(2):295–8.

187. Wójcik-Fatla A, Stojek NM, Dutkiewicz J. Efficacy of the detection of *Legionella* in hot and cold water samples by culture and PCR. I. Standardization of methods. Ann Agric Environ Med. 2012;19(2):289–93.

188. Guillemet TA, Lévesque B, Gauvin D, Brousseau N, Giroux J-P, Cantin P. Assessment of real-time PCR for quantification of *Legionella* spp. in spa water. Lett Appl Microbiol. 2010;51(6):639–44.

189. Yaradou DF, Hallier-Soulier S, Moreau S, Poty F, Hillion Y, Reyrolle M, *et al.* Integrated real-time PCR for detection and monitoring of *Legionella pneumophila* in water systems. Appl Environ Microbiol. 2007;73(5):1452–6.

190. Lee J V, Lai S, Exner M, Lenz J, Gaia V, Casati S, *et al.* An international trial of quantitative PCR for monitoring *Legionella* in artificial water systems. J Appl Microbiol. 2011;110(4):1032–44.

191. Joly P, Falconnet PA, André J, Weill N, Reyrolle M, Vandenesch F, *et al.* Quantitative real-time *Legionella* PCR for environmental water samples: data interpretation. Appl Environ Microbiol. Am Soc Microbiol; 2006;72(4):2801.

192. Wellinghausen N, Frost C, Marre R. Detection of *legionellae* in hospital water samples by quantitative real-time LightCycler PCR. Appl Environ Microbiol. Am Soc Microbiol; 2001;67(9):3985.

193. Chen N-T, Chang C-W. Rapid quantification of viable *legionellae* in water and biofilm using ethidium monoazide coupled with real-time quantitative PCR. J Appl Microbiol. 2010;109(2):623–34.

194. Nocker A, Cheung C-Y, Camper AK. Comparison of propidium monoazide with ethidium monoazide for differentiation of live vs. dead bacteria by selective removal of DNA from dead cells. J Microbiol Methods. 2006;67(2):310–20.

195. Joseph C. Investigation of outbreaks: epidemiology. Methods Mol Biol. 2013;954:73–86.

196. Lück C, Fry NK, Helbig JH, Jarraud S, Harrison TG. Typing methods for *legionella*. Methods Mol Biol. 2013;954:119–48.

197. Harrison TG, Afshar B, Doshi N, Fry NK, Lee J V. Distribution of *Legionella pneumophila* serogroups, monoclonal antibody subgroups and DNA sequence types in recent clinical and environmental isolates from England and Wales (2000-2008). Eur J Clin Microbiol Infect Dis. 2009;28(7):781–91.

198. Amemura-Maekawa J, Kikukawa K, Helbig JH, Kaneko S, Suzuki-Hashimoto A, Furuhata K, *et al.* Distribution of Monoclonal Antibody Subgroups and Sequence-Based Types among *Legionella pneumophila* Serogroup 1 Isolates Derived from Cooling Tower Water, Bathwater, and Soil in Japan. Appl Environ Microbiol. 2012;78(12):4263–70.

199. Selander RK, Caugant DA, Ochman H, Musser JM, Gilmour MN, Whittam TS. Methods of Multilocus Enzyme Electrophoresis for Bacterial Population Genetics and Systematics. Appl Environ Microbiol. 1986;51(5):873–84.

200. Lück PC, Köhler J, Maiwald M, Helbig JH. DNA Polymorphisms in Strains of *Legionella pneumophila* Serogroups 3 and 4 Detected by Macrorestriction Analysis and Their Use for Epidemiological Investigation of Nosocomial Legionellosis. Appl Environ Microbiol. 1995;61(5):2000–3.

201. Lück PC, Helbig JH, Hagedorn H-J, Ehret W. DNA Fingerprinting by Pulsed-Field Gel Electrophoresis To Investigate a Nosocomial Pneumonia Caused by *Legionella bozemanii* Serogroup 1. Appl Environ Microbiol. 1995;61(7):2759–61.

202. Ott M, Bender L, Marre R, Hacker J. Pulsed field electrophoresis of genomic restriction fragments for the detection of nosocomial *Legionella pneumophila* in hospital water supplies. J Clin Microbiol. 1991;29(4):813–5.

203. Schoonmaker D, Heimberger T, Birkhead G. Comparison of ribotyping and restriction enzyme analysis using pulsed-field gel electrophoresis for distinguishing *Legionella pneumophila* isolates obtained during a nosocomial outbreak. J Clin Microbiol. 1992;30(6):1491–8.

204. De Zoysa AS, Harrison TG. Molecular typing of *Legionella pneumophila* by pulsed-field gel electrophoresis with SfiI and comparison of this method with restriction fragment-length polymorphism analysis. J Med Microbiol. 1999;48:269–78.

205. Van Belkum A, Maas H, Verbrugh H, Van Leeuwen N. Serotyping, ribotyping, PCR-mediated ribosomal 16S–23S spacer analysis and arbitrarily primed PCR for epidemiological studies on *Legionella pneumophila.* Res Microbiol. 1996;147(5):405–13.

206. Gomez-Lus P, Fields B, Benson RF, Martin WT, O'Connor SP, Black CM. Comparison of arbitrarily primed polymerase chain reaction, ribotyping, and monoclonal antibody analysis for subtyping *Legionella pneumophila* serogroup 1. J Clin Microbiol. 1993;31(7):1940–2.

207. Belkum A Van, Struelens M, Quint W. Typing of *Legionella pneumophila* strains by polymerase chain reaction-mediated DNA fingerprinting. J Clin Microbiol. 1993;31(8):2198–200.

208. Saunders NA, Harrison TG, Haththotuwa A, Taylor AG. A comparison of probes for restriction fragment length polymorphism (RFLP) typing of *Legionella pneumophila* serogroup 1 strains. J Med Microbiol. 1991;35(3):152–8.

209. Saunders NA, Harrison TG, Haththotuwa A, Kachwalla N, Taylor AG. A method for typing strains of *Legionella pneumophila* serogroup 1 by analysis of restriction fragment length polymorphisms. J Med Microbiol. 1990;31:45–55.

210. Harrison TG, Doshi N, Fry NK, Joseph CA. Comparison of clinical and environmental isolates of *Legionella pneumophila* obtained in the UK over 19 years. Clin Microbiol Infect. 2007 Jan;13(1):78–85.

211. Valsangiacomo C, Baggi F, Gaia V, Balmelli T, Peduzzi R, Piffaretti J-C. Use of amplified fragment length polymorphism in molecular typing of *Legionella pneumophila* and application to epidemiological studies. J Clin Microbiol. 1995;33(7):1716–9.

212. Benin AL, Benson RF, Arnold KE, Fiore AE, Cook PG, Williams LK, *et al.* An outbreak of travel-associated Legionnaires disease and Pontiac fever: the need for enhanced surveillance of travel-associated legionellosis in the United States. J Infect Dis. 2002;185(2):237–43.

213. Georghiou P, Doggett A, Kielhofner MA, Stout JE, Watson DA, Lupski JR, *et al.* Molecular fingerprinting of *Legionella* species by repetitive element PCR. J Clin Microbiol. 1994;32(12):2989–94.

214. Visca P, D'Arezzo S, Ramisse F, Gelfand Y, Benson G, Vergnaud G, *et al.* Investigation of the population structure of *Legionella pneumophila* by analysis of tandem repeat copy number and internal sequence variation. Microbiology. 2011;157(Pt 9):2582–94.

215. Gaia V, Fry NK, Harrison TG, Peduzzi R. Sequence-Based Typing of *Legionella pneumophila* Serogroup 1 Offers the Potential for True Portability in Legionellosis Outbreak Investigation. J Clin Microbiol. 2003;41(7):2932–9.

216. Gaia V, Fry NK, Afshar B, Lück PC, Meugnier H, Etienne J, *et al.* Consensus sequence-based scheme for epidemiological typing of clinical and environmental isolates of *Legionella pneumophila*. J Clin Microbiol. Am Soc Microbiol; 2005;43(5):2047.

217. Ratzow S, Gaia V, Helbig JH, Fry NK, Lück PC, Helbig HJ. Addition of *neuA*, the gene encoding N-acylneuraminate cytidylyl transferase, increases the discriminatory ability of the consensus sequence-based scheme for typing *Legionella pneumophila* serogroup 1 strains. J Clin Microbiol. 2007;45(6):1965–8.

218. Mentasti M, Underwood A, Lück C, Kozak-Muiznieks NA, Harrison TG, Fry NK. Extension of the *Legionella pneumophila* sequence-based typing scheme to include strains carrying a variant of the N-acylneuraminate cytidylyltransferase gene. Clin Microbiol Infect. 2014;20(7):O435–41.

219. Fry NK, Alexiou-Daniel S, Bangsborg JM, Bernander S, Pastoris MC, Etienne J, Forsblom B, *et al.* A multicenter evaluation of genotypic methods for the epidemiologic typing of *Legionella pneumophila* serogroup 1: results of a pan-European study. Clin Microbiol Infect. 1999;5:462–77.

220. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, *et al.* Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci. 1998;95(6):3140–5.

221. ESGLI, HPA, ECDC. *Legonella pneumophila* Sequence-Based Typing. Available from: http://www.hpa-bioinformatics.org.uk/legionella/legionella_sbt/php/sbt_homepage.php

222. Garcia-Nuñez M, Quero S, Catini S, Pedro-Botet ML, Mateu L, Sopena N, *et al.* Comparative molecular and antibody typing during the investigation of an outbreak of Legionnaires' disease. J Infect Chemother. 2013;19(5):896–901.

223. Chasqueira MJ, Rodrigues L, Nascimento M, Marques T. Sequence-based and monoclonal antibody typing of *Legionella pneumophila* isolated from patients in Portugal during 1987-2008. Euro Surveill. 2009;14(28):1–4.

224. Kanatani J-I, Isobe J, Kimata K, Shima T, Shimizu M, Kura F, *et al.* Close Genetic Relationship between *Legionella pneumophila* Serogroup 1 Isolates from Sputum Specimens and Puddles on Roads, as Determined by Sequence-Based Typing. Appl Environ Microbiol. 2013;79(13):3959–66.

225. Coscollá M, González-Candelas F. Comparison of clinical and environmental samples of *Legionella pneumophila* at the nucleotide sequence level. Infect Genet Evol. 2009;9(5):882–8.

226. Fontana S, Scaturro M, Rota MC, Caporali MG, Ricci ML. Molecular typing of *Legionella pneumophila* serogroup 1 clinical strains isolated in Italy. Int J Med Microbiol. 2014;304(5-6):597–602.

227. Ginevra C, Lopez M, Forey F, Reyrolle M, Meugnier H, Vandenesch F, *et al.* Evaluation of a nested-PCR-derived sequence-based typing method applied directly to respiratory samples from patients with Legionnaires' disease. J Clin Microbiol. 2009;47(4):981–7.

228. Coscollá M, González-Candelas F. Direct sequencing of *Legionella pneumophila* from respiratory samples for sequence-based typing analysis. J Clin Microbiol. 2009;47(9):2901–5.

229. Coscollá M, Fernández C, Colomina J, Sánchez-Busó L, González-Candelas F. Mixed infection by *Legionella pneumophila* in outbreak patients. Int J Med Microbiol. 2014;304:307–13.

230. Yiallouros PK, Papadouri T, Karaoli C, Papamichael E, Zeniou M, Pieridou-bagatzouni D, *et al.* First outbreak of nosocomial *Legionella* infection in term neonates caused by a cold mist ultrasonic humidifier. Clin Infect Dis. 2013;57(1):48–56.

231. Wewalka G, Schmid D, Harrison TG, Uldum SA, Lück C. Dual infections with different *Legionella* strains. Clin Microbiol Infect. 2014;20(1):O13–9.

232. Ginevra C, Jacotin N, Diancourt L, Guigon G, Arquilliere R, Meugnier H, *et al. Legionella pneumophila* sequence type 1/Paris pulsotype subtyping by spoligotyping. J Clin Microbiol. 2012;50(3):696–701.

233. Ko KSK, Lee HKH, Park MMMM, Lee K, Woo S, Yun Y, *et al.* Population Genetic Structure of *Legionella pneumophila* Inferred from RNA Polymerase Gene (*rpoB*) and DotA Gene (*dotA*) Sequences. J Bacteriol. 2002;184(8):2123–30.

234. Sciences ML, Albert-Weissenberger C, Cazalet C, Buchrieser C. *Legionella pneumophila* - a human pathogen that co-evolved with fresh water protozoa. Cell Mol Life Sci. 2007;64(4):432–48.

235. Steinert M, Heuner K, Buchrieser C, Albert-Weissenberger C, Glöckner G. *Legionella* pathogenicity: genome structure, regulatory networks and the host cell response. Int J Med Microbiol. 2007;297(7-8):577–87.

236. D'Auria G, Jiménez-Hernández N, Peris-Bondia F, Moya A, Latorre A. *Legionella pneumophila* pangenome reveals strain-specific virulence factors. BMC Genomics. 2010;11(181):1–13.

237. Amaro F, Gilbert JA, Owens S, Trimble W, Shuman HA. Whole-Genome Sequence of the Human Pathogen *Legionella pneumophila* Serogroup 12 Strain 570-CO-H. J Bacteriol. 2012;194(6):1613–4.

238. Khan MA, Knox N, Prashar A, Alexander D, Abdel-Nour M, Duncan C, *et al.* Comparative Genomics Reveal That Host-Innate Immune Responses Influence the Clinical Prevalence of *Legionella pneumophila* Serogroups. PLoS One. 2013;8(6):e67298.

239. Ma J, He Y, Hu B, Luo Z. Genome sequence of an environmental isolate of the bacterial pathogen *Legionella pneumophila*. Genome Announc. 2013;1(3):2010–1.

240. Schroeder GN, Petty NK, Mousnier A, Harding CR, Vogrin AJ, Wee B, *et al. Legionella pneumophila* strain 130b possesses a unique combination of type IV secretion systems and novel Dot/Icm secretion system effector proteins. J Bacteriol. 2010;192(22):6001–16.

241. Reuter S, Harrison TG, Köser CU, Ellington MJ, Smith GP, Parkhill J, *et al.* A pilot study of rapid whole-genome sequencing for the investigation of a *Legionella* outbreak. BMJ Open. 2013;3(1).

242. Underwood AP, Jones G, Mentasti M, Fry NK, Harrison TG. Comparison of the *Legionella pneumophila* population structure as determined by sequence-based typing and whole genome sequencing. BMC Microbiol. 2013;13:302.

243. Gil R, Silva FJ, Peretó J, Moya A. Determination of the core of a minimal bacterial gene set. Microbiol Mol Biol Rev. 2004 Sep;68(3):518–37, table of contents.

244. Gomez-Valero L, Buchrieser C. Genome dynamics in *Legionella*: the basis of versatility and adaptation to intracellular replication. Cold Spring Harb Perspect Med. 2013;3(6).

245. Tetz V. The pangenome concept: a unifying view of genetic information. Med Sci Monit. 2005;11(7):24–30.

246. Cazalet C, Jarraud S, Ghavi-Helm Y, Kunst F, Glaser P, Etienne J, *et al.* Multigenome analysis identifies a worldwide distributed epidemic *Legionella pneumophila* clone that emerged within a highly diverse species. Genome Res. 2008;18(3):431–41.

247. Méric G, Yahara K, Mageiros L, Pascoe B, Maiden MCJ, Jolley KA, *et al.* A reference pan-genome approach to comparative bacterial genomics: identification of novel epidemiological markers in pathogenic *Campylobacter*. PLoS One. 2014;9(3):e92798.

248. Lassalle F, Campillo T, Vial L, Baude J, Costechareyre D, Chapulliot D, *et al*. Genomic species are ecological species as revealed by comparative genomics in Agrobacterium tumefaciens. Genome Biol Evol. 2011;3(0):762–81.

249. Park J, Zhang Y, Buboltz AAM, Zhang X, Schuster SC, Ahuja U, *et al.* Comparative genomics of the classical *Bordetella* subspecies: the evolution and exchange of virulence-associated diversity amongst closely related pathogens. BMC Genomics. 2012;13:545.

250. Takuno S, Kado T, Sugino RP, Nakhleh L, Innan H. Population genomics in bacteria: a case study of *Staphylococcus aureus*. Mol Biol Evol. 2012;29(2):797–809.

251. Wilson DJ. Insights from genomics into bacterial pathogen populations. PLOS Pathog. 2012;8(9):e1002874.

252. Holden MTG, Hsu L-Y, Kurt K, Weinert L a, Mather AE, Harris SR, *et al.* A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic. Genome Res. 2013;23(4):653–64.

253. Price LB, Stegger M, Hasman H, Aziz M, Larsen J, Andersen PS, *et al. Staphylococcus aureus* CC398: host adaptation and emergence of methicillin resistance in livestock. MBio. 2012;3(1):e00305–11.

254. Yamamoto T, Takano T, Higuchi W, Iwao Y, Singur O, Reva I, *et al.* Comparative genomics and drug resistance of a geographic variant of ST239 methicillin-resistant *Staphylococcus aureus* emerged in Russia. PLoS One. 2012;7(1):e29187.

255. Harris SRS, Clarke INI, Seth-Smith H, Solomon AW, Cutcliffe LT, Marsh P, *et al.* Whole-genome analysis of diverse *Chlamydia trachomatis* strains identifies phylogenetic relationships masked by current clinical typing. Nat Genet. 2013;44(4):413–9, S1.

256. Mather AE, Reid SWJ, Maskell DJ, Parkhill J, Fookes MC, Harris SR, *et al.* Distinguishable Epidemics of Multidrug-Resistant *Salmonella Typhimurium* DT104 in Different Hosts. Science. 2013;341:1514–7.

257. He M, Miyajima F, Roberts P, Ellison L, Pickard DJ, Martin MJ, *et al*. Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. Nat Genet. 2013;45(1):109–13.

258. Selander RK, Beltran P. Evolutionary genetic relationships of clones of *Salmonella* serovars that cause human typhoid and other enteric fevers. Infect Immun. 1990;58(7):2262–75.

259. Bygraves JA, Maiden MCJ. Analysis of the clonal relationships between strains of *Neisseria meningitidis* by pulsed field gel electrophoresis. J Gen Microbiol. 1992;138(3):523–31.

260. Caugant D, Mocca L. Genetic structure of *Neisseria meningitidis* populations in relation to serogroup, serotype, and outer membrane protein pattern. J Bacteriol. 1987;169(6):2781–92.

261. Selander RK, McKINNEY RM, Whittam TS, Bibb WF, Brenner DJ, Nolte FS, *et al*. Genetic structure of populations of *Legionella pneumophila.* J Bacteriol. 1985;163(3):1021–37.

262. Edwards MT, Fry NK, Harrison TG. Clonal population structure of *Legionella pneumophila* inferred from allelic profiling. Microbiology. 2008;154(Pt 3):852–64.

263. Smith JM, Smith NH, Rourke MO, Spratt BG. How clonal are bacteria? Proc Natl Acad Sci. 1993;90:4384–8.

264. Feil EJ, Spratt BG. Recombination and the population structures of bacterial pathogens. Annu Rev Microbiol. 2001;55:561–90.

265. Glöckner G, Albert-Weissenberger C, Weinmann E, Jacobi S, Schunder E, Steinert M, *et al.* Identification and characterization of a new conjugation/type IVA secretion system (trb/tra) of *Legionella pneumophila* Corby localized on two mobile genomic islands. Int J Med Microbiol. 2008;298(5-6):411–28.

266. Bellanger X, Payot S, Leblond-Bourget N, Guédon G. Conjugative and mobilizable genomic islands in bacteria: evolution and diversity. FEMS Microbiol Rev. 2014;38(4):720–60.

267. Coscollá M, González-Candelas F. Population structure and recombination in environmental isolates of *Legionella pneumophila*. Environ Microbiol. 2007;9(3):643–56.

268. Coscollá M, Comas I, González-Candelas F. Quantifying nonvertical inheritance in the evolution of *Legionella pneumophila.* Mol Biol Evol. 2011;28(2):985–1001.

269. Costa J, Teixeira PG, d'Avó AF, Júnior CS, Veríssimo A. Intragenic Recombination Has a Critical Role on the Evolution of *Legionella pneumophila* Virulence-Related Effector *sidJ*. PLoS One. 2014;9(10):e109840.

270. Vos M, Didelot X. A comparison of homologous recombination rates in bacteria and archaea. ISME J. 2009;3(2):199–208.

271. Gomez-Valero L, Rusniok C, Jarraud S, Vacherie B, Rouy ZZ, Barbe V, *et al.* Extensive recombination events and horizontal gene transfer shaped the *Legionella pneumophila* genomes. BMC Genomics. 2011;12(1):536.

272. Siefert JL. Defining the mobilome. Methods Mol Biol. 2009;532:13–27.

273. Flynn K, Swanson M. Integrative Conjugative Element ICE-βox Confers Oxidative Stress Resistance to *Legionella pneumophila* In Vitro and in Macrophages. MBio. 2014;5(3):e01091–14.

274. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004;14(7):1394–403.

275. Robinson ER, Walker TM, Pallen MJ. Genomics and outbreak investigation: from sequence to consequence. Genome Med. 2013 Apr 29;5(4):36.

276. Sabat AJ, BudimirA, Nashev D, Sá-Leão R, van Dijl JM, Laurent F, *et al.* Overview of molecular typing methods for outbreak detection and epidemiological surveillance. Euro Surveill. 2013 Jan;18(4):20380.

277. Sanger F, Nicklen S, Coulson AAR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci. National Acad Sciences; 1977;74(12):5463–7.

278. Arens M. Methods for subtyping and molecular comparison of human viral genomes. Clin Microbiol Rev. 1999;12(4):612–26.

279. Loman NJ, Misra R V, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, *et al.* Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol. 2012;30(5):434–9.

280. Schuetz P, Diep BA. Use of whole-genome sequencing for outbreak investigations. Lancet Infect Dis. 2013;13(2):99–101.

281. Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, *et al.* Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. PLoS One. 2011;6(7):e22751.

282. Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, *et al.* Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. N Engl J Med. 2011;365(8):709–17.

283. Harris SR, Cartwright EJP, Török ME, Holden MTG, Brown NM, Ogilvy-Stuart AL, *et al*. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study. Lancet Infect Dis. 2012;13(2):130–6.

284. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, *et al.* Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med. 2011 Mar;364(8):730–9.

285. Bryant JM, Schürch AC, van Deutekom H, Harris SR, de Beer JL, de Jager V, *et al.* Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. BMC Infect Dis. 2013;13(1):110.

286. Bryant J, Grogono D, Greaves D. Whole-genome sequencing to identify transmission of *Mycobacterium abscessus* between patients with cystic fibrosis: a retrospective cohort study. Lancet Infect Dis. 2013;381(9877):1551–60.

287. Snitkin EES, Zelazny AMA, Thomas PJ, Stock F, Henderson DK, Palmore TN, *et al. T*racking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. Sci Transl Med. 2012;4(148):148ra116.

288. Morelli G, Song Y, Mazzoni CJ, Eppinger M, Roumagnac P, Wagner DM, *et al. Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. Nat Genet. 2010;42(12):1140–3.

289. Monot M, Honoré N, Garnier T, Zidane N, Sherafi D, Paniz-Mondolfi A, *et al*. Comparative genomic and phylogeographic analysis of *Mycobacterium leprae*. Nat Genet. 2009;41(12):1282–9.

290. Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, *et al.* Evidence for several waves of global transmission in the seventh *Cholera* pandemic. Nature. 2011;477(7365):462–5. act

291. Fricke WF, Rasko DA. Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. Nat Rev Genet. 2014;15(1):49–55.

292. Gilmour MW, Graham MWGM, Reimer A, Van Domselaar G, Domselaar ARG Van. Public health genomics and the new molecular epidemiology of bacterial pathogens. Public Health Genomics. 2013;16(1-2):25–30.

293. McAdam PR, Vander Broek CW, Lindsay DSJ, Ward MJ, Hanson MF, Gillies M, *et al.* Gene flow in environmental *Legionella pneumophila* leads to genetic and pathogenic heterogeneity within a Legionnaires' disease outbreak. Genome Biology. 2014. p. 504.

294. Harb OS, Gao L, Kwaik YA. From protozoa to mammalian cells: a new paradigm in the life cycle of intracellular bacterial pathogens. Environ Microbiol. 2000;2(3):251–65.

295. Amann R, Ludwig W, Schleifer K. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev. 1995;59(1):143–69.

296. Rodriguez-Valera F. Environmental genomics, the big picture? FEMS Microbiol Lett. 2004;231(2):153–8.

297. Tanaka T, Kawasaki K, Daimon S, Kitagawa W, Yamamoto K, Tamaki H, *et al.* A hidden pitfall in agar media preparation undermines cultivability of microorganisms. Appl Environ Microbiol. 2014;80(24):7659–66.

298. Streit WR, Schmitz R a. Metagenomics--the key to the uncultured microbes. Curr Opin Microbiol. 2004;7(5):492–8.

299. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature. 2004;428(March):37–43.

300. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, *et al*. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol. 2007;5(3):e16.

301. Karsenti E, Acinas SG, Bork P, Bowler C, De Vargas C, Raes J, *et al.* A holistic approach to marine eco-systems biology. PLoS Biol. 2011;9(10):e1001177.

302. Gilbert JA, Jansson JK, Knight R. The Earth Microbiome project: successes and aspirations. BMC Biol. 2014;12(1):69.

303. Lewis CM, Obregón-Tito A, Tito RY, Foster MW, Spicer PG. The Human Microbiome Project: lessons from human genomics. Trends Microbiol. 2012;20(1):1–4.

304. Belda-Ferre P, Alcaraz LD, Cabrera-Rubio R, Romero H, Simón-Soro A, Pignatelli M, *et al*. The oral metagenome in health and disease. ISME J. 2012;6(1):46–56.

305. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, *et al.* Enterotypes of the human gut microbiome. Nature. 2011;473(7346):174–80.

306. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, *et al.* Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol. 2014;12(9):635–45.

307. Weisburg W, Barns S. 16S ribosomal DNA amplification for phylogenetic study. J Bacteriol. 1991;173(2):697–703.

308. Wintzingerode F v, Göbel UB, Stackebrandt E. Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. FEMS Microbiol Rev. 1997;21:213–29.

309. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014;12(1):87.

310. Woyke T, Rubin EM. Searching for new branches on the tree of life. Science. 2014;346(6210):698–9.

311. Ladoukakis E, Kolisis FN, Chatziioannou AA. Integrative workflows for metagenomic analysis. Front Cell Dev Biol. 2014;2(November):1–11.

312. Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P. A bioinformatician's guide to metagenomics. Microbiol Mol Biol Rev. 2008;72(4):557–78.

313. Chen K, Pachter L. Bioinformatics for whole-genome shotgun sequencing of microbial communities. PLoS Comput Biol. 2005;1(2):106–12.

314. Yooseph S, Nealson KH, Rusch DB, McCrow JP, Dupont CL, Kim M, *et al.* Genomic and functional adaptation in surface ocean planktonic prokaryotes. Nature. 2010;468(7320):60–6.

315. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, *et al.* The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol. 2007;5(3):e77.

316. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. Science. 2004;304(5667):66–74.

317. Schmidt T, DeLong E, Pace N. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. J Bacteriol. 1991;173(14):4371–8.

318. Dupont CL, Larsson J, Yooseph S, Ininbergs K, Goll J, Asplund-Samuelsson J, *et al.* Functional tradeoffs underpin salinity-driven divergence in microbial community composition. PLoS One. 2014;9(2):e89549.

319. Xie W, Wang F, Guo L, Chen Z, Sievert SM, Meng J, *et al.* Comparative metagenomics of microbial communities inhabiting deep-sea hydrothermal vent chimneys with contrasting chemistries. ISME J. 2011;5(3):414–26.

320. Ghai R, Hernandez CM, Picazo A, Mizuno CM, Ininbergs K, Díez B, *et al.* Metagenomes of Mediterranean coastal lagoons. Sci Rep. 2012;2:490.

321. Jiménez DJ, Andreote FD, Chaves D, Montaña JS, Osorio-Forero C, Junca H, *et al.* Structural and functional insights from the metagenome of an acidic hot spring microbial planktonic community in the Colombian Andes. PLoS One. 2012;7(12):e52069.

322. Fernández AB, Vera-Gargallo B, Sánchez-Porro C, Ghai R, Papke RT, Rodriguez-Valera F, *et al.* Comparison of prokaryotic community structure from Mediterranean and Atlantic saltern concentrator ponds by a metagenomic approach. Front Microbiol. 2014;5(May):196.

323. Fernández AB, Ghai R, Martin-Cuadrado A-B, Sánchez-Porro C, Rodriguez-Valera F, Ventosa A. Prokaryotic taxonomic and metabolic diversity of an intermediate salinity hypersaline habitat assessed by metagenomics. FEMS Microbiol Ecol. 2014;88(3):623–35.

324. Chao Y, Ma L, Yang Y, Ju F, Zhang X-X, Wu W-M, *et al.* Metagenomic analysis reveals significant changes of microbial compositions and protective functions during drinking water treatment. Sci Rep. 2013;3:3550.

325. Hwang C, Ling F, Andersen GL, LeChevallier MW, Liu W-T. Microbial community dynamics of an urban drinking water distribution system subjected to phases of chloramination and chlorination treatments. Appl Environ Microbiol. 2012;78(22):7856–65.

326. Gomez-Alvarez V, Revetta RP, Santo Domingo JW. Metagenomic analyses of drinking water receiving different disinfection treatments. Appl Environ Microbiol. 2012;78(17):6095–102.

327. Yooseph S, Andrews-Pfannkoch C, Tenney A, McQuaid J, Williamson S, Thiagarajan M, *et al*. A metagenomic framework for the study of airborne microbial communities. PLoS One. 2013;8(12):e81862.

328. Schmeisser C, Stockigt C, Raasch C, Wingender J, Timmis KN, Wenderoth DF, *et al*. Metagenome Survey of Biofilms in Drinking-Water Networks. Appl Environ Microbiol. 2003;69(12):7298–309.

329. Neria-González I, Wang ET, Ramírez F, Romero JM, Hernández-Rodríguez C. Characterization of bacterial community associated to biofilms of corroded oil pipelines from the southeast of Mexico. Anaerobe. 2006;12(3):122–33.

330. Hong P-Y, Hwang C, Ling F, Andersen GL, LeChevallier MW, Liu W-T. Pyrosequencing analysis of bacterial biofilm communities in water meters of a drinking water distribution system. Appl Environ Microbiol. 2010;76(16):5631–5.

331. Kelly JJ, Minalt N, Culotti A, Pryor M, Packman A. Temporal Variations in the Abundance and Composition of Biofilm Communities Colonizing Drinking Water Distribution Pipes. PLoS One. 2014;9(5):e98542.

332. Buse HY, Lu J, Lu X, Mou X, Ashbolt NJ. Microbial diversities (16S and 18S rRNA gene pyrosequencing) and environmental pathogens within drinking water biofilms grown on the common premise plumbing materials unplasticized polyvinylchloride and copper. FEMS Microbiol Ecol. 2014;88:280-295.

333. Shaw JL a, Monis P, Fabris R, Ho L, Braun K, Drikas M, *et al*. Assessing the impact of water treatment on bacterial biofilms in drinking water distribution systems using high-throughput DNA sequencing. Chemosphere. 2014;117:185–92.

334. Lu J, Buse H, Gomez-Alvarez V, Struewing I, Santo Domingo J, Ashbolt NJ. Impact of drinking water conditions and copper materials on downstream biofilm microbial communities and *Legionella pneumophila* colonization. J Appl Microbiol. 2014;117(3):905–18.

335. Declerck P, Behets J, Margineanu A, van Hoef V, Keersmaecker B De, Ollevier F. Replication of *Legionella pneumophila* in biofilms of water distribution pipes. Microbiol Res. Elsevier; 2009;164(6):593–603.

336. White PS, Graham FF, Harte DJG, Baker MG, Ambrose CD, Humphrey ARG. Epidemiological investigation of a Legionnaires' disease outbreak in Christchurch, New Zealand: the value of spatial methods for practical public health. Epidemiol Infect. 2013;141(4):789–99.

337. Keramarou M, Evans MR. A community outbreak of Legionnaires' disease in South Wales, August-September 2010. Euro Surveill. 2010;15(42):1–4.

338. Coscollá M, Gosalbes MJ, Catalán V, González-candelas F. Genetic variability in environmental isolates of *Legionella pneumophila* from Comunidad Valenciana (Spain). Environ Microbiol. 2006;8(6):1056–63.

339. Sánchez-Busó L, Comas I, Jorques G, González-Candelas F. Recombination drives genome evolution in outbreak-related *Legionella pneumophila* isolates. Nat Genet. 201;46(11):1205–11.

340. Instituto de Salud Carlos III. Brotes de Legionelosis notificados a la Red Nacional de Vigilancia Epidemiológica en el periodo 1999-2011. Available from: http://www.isciii.es/ISCIII/es/contenidos/fd-servicios-cientifico-tecnicos/fd-vigilancias-alertas/fd-enfermedades/legionelosis.shtml

341. Sánchez-Busó L, Coscollá M, Pinto-Carbó M, Catalán V, González-Candelas F. Genetic Characterization of *Legionella pneumophila* Isolated from a Common Watershed in Comunidad Valenciana, Spain. PLoS One. 2013;8(4):e61564.

342. Staden R. The Staden Sequence Analysis Package. Mol Biotechnol. 1996;5:233–41.

343. Francisco AP, Vaz C, Monteiro PT, Melo-Cristino J, Ramirez M, Carriço JA. PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods. BMC Bioinformatics. 2012 Jan;13:87.

344. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2014. Available from: http://www.r-project.org

345. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006;22(21):2688–90.

346. Letunic I, Bork P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. 2011;39(Web Server issue):W475–8.

347. Bandelt H-JJ, Forster P, Röhl A. Median-Joining Networks for Inferring Intraspecific Phylogenies. Mol Biol Evol. 1999;16(1):37–48.

348. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009;25(11):1451–2.

349. Excoffier L, Lischer HEL. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour. 2010;10(3):564–7.

350. Corander J, Marttinen P, Sirén J, Tang J. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics. 2008;9:539.

351. Excoffier L, Smouse PE, Quattro JM. Analysis of Molecular Variance Inferred From Metric Distances Among DNA Haplotypes: Application to Human Mitochondrial DNA Restriction Data. Genetics. Genetics Soc America; 1992;131(2):479–91.

352. Sánchez-Busó L, Olmos MP, Camaró ML, Adrián F, Calafat JM, González-Candelas F. Phylogenetic analysis of environmental *Legionella pneumophila* isolates from an endemic area (Alcoy, Spain). Infect Genet Evol. 2015;30:45–54.

353. Vekens E, Soetens O, Mendonça R De, Echahidi F, Roisin S, Deplano A, *et al.* Sequence-based typing of *Legionella pneumophila* serogroup 1 clinical isolates from Belgium between 2000 and 2010. Euro Surveill. 2012;17(43):1–6.

354. Amemura-Maekawa J, Kura F, Chang B, Watanabe H. *Legionella pneumophila* serogroup 1 isolates from cooling towers in Japan form a distinct genetic cluster. Microbiol Immunol. Wiley-Blackwell; 2005;49(12):1027–33.

355. Gomez-Valero L, Rusniok C, Buchrieser C. *Legionella pneumophila*: population genetics, phylogeny and genomics. Infect Genet Evol. 2009;9(5):727–39.

356. Bozue JA, Johnson W. Uptake by coiling phagocytosis and inhibition of phagosome-lysosome fusion. Infect Immun. 1996;64(2):668–73.

357. Grist NR. Legionnaires' Disease and the Traveller. Ann Intern Med. 1979;90(4):563.

358. Bartlett C, Swann R, Casal J, Canada Royo G, Taylor A. Recurrent Legionnaires' disease from a hotel water system. *Legionella* Proceedings of the 2nd International Symposium, Washington, DC Washington, American Society for Microbiology. 1984. p. 237–9.

359. Farhat C, Mentasti M, Jacobs E, Fry NK, Lück C. The N-acylneuraminate cytidyl transferase gene, *neuA* is heterogenous in *Legionella pneumophila* strains but can be used as marker for epidemiological typing in the consensus sequence-based tying scheme. J Clin Microbiol. 2011;49(12):4052-8.

360. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004;5:113.

361. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.

362. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 2011;28(10):2731–9.

363. Posada D. jModelTest: Phylogenetic Model Averaging. Mol Biol Evol. 2008;25:1253–6.

364. Akaike H. A new look at the estatistical model identification. IEEE Trans Automat Contr. 1974;19:716–23.

365. Nei M. Molecular Evolutionary Genetics. Molecular Evolutionary Genetics. 1987.

366. Tajima F. Evolutionary Relationship of DNA Sequences in Finite Populations. Genetics. Genetics Soc America; 1983;105(2):437–60.

367. Tajima F. Simple Methods for Testing the Molecular Evolutionary Clock Hypothesis. Genetics. Genetics Soc America; 1993;135(2):599–607.

368. Shimodaira H, Hasegawa M. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. Mol Biol Evol. 1999 Jan;16(8):1114–6.

369. Strimmer K, Rambaut A. Inferring confidence sets of possibly misspecified gene trees. Proc R Soc B Biol Sci. The Royal Society; 2002;269(1487):137–42.

370. Shimodaira H. An approximately unbiased test of phylogenetic tree selection. Syst Biol. 2002;51(3):492–508.

371. Schmidt HA, Strimmer K, Vingron M, von Haeseler A. TREE-PUZZLE : Maximum Likelihood Phylogenetic Analysis Using Quartets and Parallel Computing Parallelization. Bioinformatics. 2002 Mar;18(3):3–6.

372. Shimodaira H, Hasegawa M. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics. 2001;17(12):1246–7.

373. Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics. 2010;26(19):2462–3.

374. Martin DP, Rybicky E. RDP: detection of recombination amongst aligned sequences. Bioinformatics. 2000;16:562–3.

375. Padidam M, Sawyer S, Fauquet CM. Possible emergence of new geminiviruses by frequent recombination. Virology. 1999;265:218–25.

376. Martin DP, Williamson C, Posada D. RDP2: recombination detection and analysis from sequence alignments. Bioinformatics. Oxford University Press; 2005;21(2):260–2.

377. Maynard Smith J. Analyzing the mosaic structure of gene. J Mol Evol. 1992;34:126–9.

378. Posada D, Crandall KA. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc Natl Acad Sci. 2001;98:13757–62.

379. Gibbs M, Armstrong J, Gibbs A. Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. Bioinformatics. 2000;16:573–82.

380. Boni MF, Posada D, Feldman MW. An exact nonparametric method for inferring mosaic structure in sequence triplets. Genetics. 2007;176(2):1035–47.

381. Pritchard JK, Stephens M, Donelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155:945–59.

382. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 2003 Aug;164(4):1567–87.

383. Jakobsson M, Rosenberg N. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics. 2007;23(14):1801–6.

384. Rosenberg NAN. Distruct: a Program for the Graphical Display of Population Structure. Mol Ecol Notes. 2004;4(1):137–8.

385. Publishing B, Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol. 2005;14(8):2611–20.

386. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. J R Stat Soc. 1995;47(1):289–300.

387. Benjamini Y, Yekutieli D. The control of the False Discovery Rate in multiple testing under dependency. Ann Stat. 2001;29(4):1165–88.

388. Fliermans CB, Cherry WB, Orrison LH, Thacker L. Isolation of *Legionella pneumophila* from Nonepidemic-Related Aquatic Habitats. Appl Environ Microbiol. 1979;37(6):1239–42.

389. Costa J, Tiago I, da Costa MS, Veríssimo A. Presence and persistence of *Legionella* spp. in groundwater. Appl Environ Microbiol. 2005 Feb;71(2):663–71.

390. Wullings BA, Kooij D Van Der. Occurrence and Genetic Diversity of Uncultured *Legionella* spp. in Drinking Water Treated at Temperatures below 15^{o}C. Appl Environ Microbiol. 2006;72(1):157–66.

391. Sheehan K, Henson J, Ferris M. *Legionella* species diversity in an acidic biofilm community in Yellowstone National Park. Appl Environ Microbiol. 2005;71(1):507–11.

392. Declerck P, Behets J, Hoef V Van, Ollevier F, van Hoef V. Detection of *Legionella* spp. and some of their amoeba hosts in floating biofilms from anthropogenic and natural aquatic environments. Water Res. 2007;41(14):3159–67.

393. Parthuisot N, West NJ, Lebaron P, Baudart J. High diversity and abundance of *Legionella* spp. in a pristine river and impact of seasonal and anthropogenic effects. Appl Environ Microbiol. 2010;76(24):8201–10.

394. Carvalho FRS, Foronda AS, Pellizari VH. Detection of *Legionella pneumophila* in water and biofilm samples by culture and molecular methods from man-made systems in São Paulo-Brazil. Brazilian J Microbiol. 2007;38:743–51.

395. Wullings BA, Bakker G, van der Kooij D. Concentration and diversity of uncultured *Legionella* spp. in two unchlorinated drinking water supplies with different concentrations of natural organic matter. Appl Environ Microbiol. 2011;77(2):634–41.

396. Lyautey E, Lu Z, Lapen DR, Wilkes G, Scott A, Berkers T, *et al.* Distribution and diversity of *Escherichia coli* populations in the South Nation River drainage basin, eastern Ontario, Canada. Appl Environ Microbiol. 2010;76(5):1486–96.

397. Chen B, Zheng W, Yu Y, Huang W, Zheng S, Zhang Y, *et al.* Class 1 integrons, selected virulence genes, and antibiotic resistance in *Escherichia coli* isolates from the Minjiang River, Fujian Province, China. Appl Environ Microbiol. 2011;77(1):148–55.

398. Goto DK, Yan T. Genotypic diversity of *Escherichia coli* in the water and soil of tropical watersheds in Hawaii. Appl Environ Microbiol. 2011;77(12):3988–97.

399. Lyautey E, Lapen DR, Wilkes G, McCleary K, Pagotto F, Tyler K, *et al.* Distribution and characteristics of *Listeria monocytogenes* isolates from surface waters of the South Nation River watershed, Ontario, Canada. Appl Environ Microbiol. 2007;73(17):5401–10.

400. Didelot X, Bowden R, Street T, Golubchik T, Spencer C, McVean G, *et al.* Recombination and Population Structure in *Salmonella enterica.* PLoS Genet. 2011;7(7):e1002191.

401. Batté M, Appenzeller B, Grandjean D, Fass S, Gauthier V, Jorand F, *et al*. Biofilms in Drinking Water Distribution Systems. Rev Environ Sci Bio/Technology. 2003;2(2-4):147–68.

402. Lück PC, Ecker C, Reischl U, Linde H-J, Stempka R. Culture-independent identification of the source of an infection by direct amplification and sequencing of *Legionella pneumophila* DNA from a clinical specimen. J Clin Microbiol. 2007;45(9):3143–4.

403. Lehtola MJ, Torvinen E, Kusnetsov J, Pitkänen T, Maunula L, von Bonsdorff C-H, *et al*. Survival of *Mycobacterium avium*, *Legionella pneumophila*, *Escherichia coli*, and *caliciviruses* in drinking water-associated biofilms grown under high-shear turbulent flow. Appl Environ Microbiol. 2007;73(9):2854–9.

404. Moritz MM, Flemming H, Wingender J. Integration of *Pseudomonas aeruginosa* and *Legionella pneumophila* in drinking water biofilms grown on domestic plumbing materials. Int J Hyg Environ Health. 2010;213(3):190–7.

405. Taylor M, Ross K, Bentham R. *Legionella*, protozoa, and biofilms: interactions within complex microbial systems. Microb Ecol. 2009 Oct;58(3):538–47.

406. Wingender J, Flemming H-C. Biofilms in drinking water and their role as reservoir for pathogens. Int J Hyg Environ Health. 2011;214(6):417–23.

407. Delgado-Viscogliosi P, Simonart T, Parent V, Marchand G, Dobbelaere M, Pierlot E, *et al.* Rapid method for enumeration of viable *Legionella pneumophila* and other *Legionella* spp. in water. Appl Environ Microbiol. 2005;71(7):4086.

408. Valster RM, Wullings BA, van der Kooij D. Detection of protozoan hosts for *Legionella pneumophila* in engineered water systems by using a biofilm batch test. Appl Environ Microbiol. 2010;76(21):7144–53.

409. Guerrieri E, Bondi M, Sabia C, de Niederhäusern S, Borella P, Messi P. Effect of bacterial interference on biofilm development by *Legionella pneumophila*. Curr Microbiol. 2008;57(6):532–6.

410. Lau HY, Ashbolt NJ. The role of biofilms and protozoa in *Legionella* pathogenesis: implications for drinking water. J Appl Microbiol. 2009;107(2):368–78.

411. Murga R, Forster TS, Brown E, Pruckler JM, Fields BS, Donlan RM. Role of biofilms in the survival of *Legionella pneumophila* in a model potable-water system. Microbiology. 2001 Nov;147(Pt 11):3121–6.

412. Alleron L, Merlet N, Lacombe C, Frère J. Long-term survival of *Legionella pneumophila* in the viable but nonculturable state after monochloramine treatment. Curr Microbiol. 2008;57(5):497–502.

413. Cooper IR, Hanlon GW. Resistance of *Legionella pneumophila* serotype 1 biofilms to chlorine-based disinfection. J Hosp Infect. Elsevier Ltd; 2010;74(2):152–9.

414. Moore MR, Pryor M, Fields B, Lucas C, Phelan M, Besser RE. Introduction of Monochloramine into a Municipal Water System: Impact on Colonization of Buildings by *Legionella* spp. Appl Environ Microbiol. 2006;72(1):378–83.

415. Fernández JA, López P, Orozco D, Merino J. Clinical Study of an Outbreak of Legionnaire’ s Disease in Alcoy, Southeastern Spain. Eur J Clin Microbiol Infect Dis. 2002;21:729–35.

416. López P, Chinchilla A, Andreu M, Pelaz C, Sastre J. El laboratorio de microbiología clínica en el brote de *Legionella* spp. en la comarca de Alcoy: rentabilidad de las diferentes técnicas diagnósticas. Enferm Infecc Microbiol Clin. 2001;19:435–8.

417. Suenaga E, Nakamura H. Evaluation of three methods for effective extraction of DNA from human hair. J Chromatogr. 2005;820(1):137–41.

418. Don RH, Cox PT, Wainwright BJ, Baker K, Mattick JS. "Touchdown" PCR to circumvent spurious priming during gene amplification. Nucleic Acids Res. 1991;19(14):4008.

419. Hecker KH, Roux KH. High and Low Annealing Temperatures Increase Both Specificity and Yield in Touchdown and Stepdown PCR. Biotechniques. 1996;20(3):478–85.

420. Korbie DJ, Mattick JS. Touchdown PCR for increased specificity and sensitivity in PCR amplification. Nat Protoc. 2008;3(9):1452–6.

421. Feazel LM, Baumgartner LK, Peterson KL, Frank DN, Harris JK, Pace NR. Opportunistic pathogens enriched in showerhead biofilms. Proc Natl Acad Sci. 2009 Sep 22;106(38):16393–9.

422. Ferré MRS, Arias C, Oliva JM, Pedrol A, García M, Pellicer T, *et al.* A community outbreak of Legionnaires' disease associated with a cooling tower in Vic and Gurb, Catalonia (Spain) in 2005. Eur J Clin Microbiol Infect Dis. 2009;28(2):153–9.

423. Goutziana G, Mouchtouri VA, Karanika M, Kavagias A, Stathakis NE, Gourgoulianis K, *et al.* *Legionella* species colonization of water distribution systems, pools and air conditioning systems in cruise ships and ferries. BMC Public Health. 2008;8:390.

424. Borella P, Guerrieri E, Marchesi I, Bondi M, Messi P. Water ecology of *Legionella* and protozoan: environmental and public health perspectives. Biotechnol Annu Rev. Elsevier; 2005;11(05):355–80.

425. Meyer RD. Legionnaires' Disease: Unusual Clinical and Laboratory Features. Ann Intern Med. 1980;93(2):240.

426. Muder R, Yu V, Vickers R, Rihs J, Shonnard J. Simultaneous infection with *Legionella pneumophila* and Pittsburgh pneumonia agent. Clinical features and epidemiologic implications. Am J Med. 1983;74(4):609–14.

427. Buchbinder S, Leitritz L, Trebesius K, Banas B, Heesemann J. Mixed lung infection by *Legionella pneumophila* and *Legionella gormanii* detected by fluorescent in situ hybridization. Infection. 2004;32(4):242–5.

428. Matsui M, Fujii S, Shiroiwa R, Amemura-Maekawa J, Chang B, Kura F, *et al.* Isolation of *Legionella rubrilucens* from a pneumonia patient co-infected with *Legionella pneumophila.* J Med Microbiol. 2010 59(Pt 10):1242–6.

429. Zerbino DR. Using the Velvet de novo assembler for short-read sequencing technologies. Curr Protoc Bioinforma. 2010;Chapter 11:Unit 11.5.

430. Rzhetsky A, Nei M. Theoretical foundation of the minimum-evolution method of phylogenetic inference. Mol Biol Evol. 1993;10(5):1073–95.

431. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012;29(8):1969–73.

432. Harris SR, Feil EJ, Holden MTG, Quail M a, Nickerson EK, Chantratita N, *et al*. Evolution of MRSA during hospital transmission and intercontinental spread. Science. 2010;327(5964):469.

433. Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, *et al.* Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet. 2014;46(3):305–9.

434. Lévesque S, Plante P-L, Mendis N, Cantin P, Marchand G, Charest H, *et al.* Genomic characterization of a large outbreak of *Legionella pneumophila* serogroup 1 strains in Quebec City, 2012. PLoS One. 2014;9(8):e103852.

435. Lawrence JG, Retchless AC. The interplay of homologous recombination and horizontal gene transfer in bacterial speciation. Methods Mol Biol. 2009;532:29–53.

436. Levin BR, Cornejo OE. The Population and Evolutionary Dynamics of Homologous Gene Recombination in Bacteria. PLoS Genet. Public Library of Science; 2009;5(8):16.

437. Townsend JP, Bøhn T, Nielsen KM. Assessing the probability of detection of horizontal gene transfer events in bacterial populations. Front Microbiol. 2012;3(February):27.

438. Levin B, Bergstrom C. Bacteria are different: Observations, interpretations, speculations, and opinions about the mechanisms of adaptive evolution in prokaryotes. Proc Natl Acad Sci. 2000 Jun 20;97(13):6981–5.

439. Roumagnac P, Weill F-X, Dolecek C, Baker S, Brisse S, Chinh NT, *et al.* Evolutionary history of *Salmonella typhi*. Science. 2006;314(5803):1301–4.

440. Baquero F, Tedim AP, Coque TM. Antibiotic resistance shaping multi-level population biology of bacteria. Front Microbiol. 2013;4(March):15.

441. Levy SB, Marshall B. Antibacterial resistance worldwide: causes, challenges and responses. Nat Med. 2004;10(12 Suppl):S122–9.

442. Chin C-S, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR, *et al.* The origin of the Haitian *Cholera* outbreak strain. N Engl J Med. Massachusetts Medical Society; 2011;364(1):33–42.

443. Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, *et al.* Rapid pneumococcal evolution in response to clinical interventions. Science. 2011 Jan;331(6016):430–4.

444. Köser CUC, Holden MTGM, Ellington MJ, Cartwright EJP, Brown NM, Ogilvy-Stuart AL, *et al.* Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. N Engl J Med. 2012;366(24):2267–75.

445. Coetzee N, Duggal H, Hawker J, Ibbotson S, Harrison TG, Phin N, *et al.* An outbreak of Legionnaires' disease associated with a display spa pool in retail premises, Stoke-on-Trent, United Kingdom, July 2012. Euro Surveill. 2012 Jan;17(37):1–4.

446. McCormick D, Thorn S, Milne D, Evans C, Stevenson J, Llano M, *et al.* Public health response to an outbreak of Legionnaires' disease in Edinburgh, United Kingdom, June 2012. Euro Surveill. 2012 Jan;17(28):1–4.

447. Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J, *et al.* Population genomics of post-vaccine changes in pneumococcal epidemiology. Nat Genet. 2013;45(6):656–63.

448. Holt KE, Baker S, Weill F-X, Holmes EC, Kitchen A, Yu J, *et al. Shigella sonnei* genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. Nat Genet. 2012;44(9):1056–9.

449. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.

450. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18(5):821–9.

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

452. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al.* The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.

453. Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, Wong CH, *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 2012;40(22):11189–201.

454. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One. 2010;5(6):e11147.

455. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972–3.

456. Vilella AJ, Blanco-Garcia A, Hutter S, Rozas J. VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. Bioinformatics. 2005;21(11):2791–3.

457. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004;20(2):289–90.

458. Bucci C, Lavitola A, Salvatore P, Del Giudice L, Massardo DR, Bruni CB, *et al*. Hypermutation in pathogenic bacteria: frequent phase variation in meningococci is a phenotypic trait of a specialized mutator biotype. Mol Cell. 1999 May;3(4):435–45.

459. Weigand MR, Sundin GW. General and inducible hypermutation facilitate parallel adaptation in *Pseudomonas aeruginosa* despite divergent mutation spectra. Proc Natl Acad Sci. 2012;109(34):13680–5.

460. Hanage WP, Fraser C, Tang J, Connor TR, Corander J. Hyper-Recombination, Diversity, and Antibiotic Resistance in *Pneumococcus*. Science. 2009;324:1454–7.

461. Thaipadungpanit J. A dominant clone of *Leptospira interrogans* associated with an outbreak of human leptospirosis in Thailand. PLoS Negl Trop Dis. 2007;1(1):e56.

462. Nicolas P, Mondot S, Achaz G, Bouchenot C, Bernardet J-F, Duchaud E. Population structure of the fish-pathogenic bacterium *Flavobacterium psychrophilum*. Appl Environ Microbiol. 2008;74(12):3702–9.

463. Hanage WP, Fraser C, Spratt BG. Fuzzy species among recombinogenic bacteria. BMC Biol. 2005;3:6.

464. Brochet M, Rusniok C, Couve E, Dramsi S, Poyart C, Trieu-Cuot P, *et al.* Shaping a bacterial genome by large chromosomal replacements, the evolutionary history of *Streptococcus agalactiae.* Proc Natl Acad Sci. 2008;105(41):15961–6.

465. He M, Sebaihia M, Lawley TD, Stabler RA, Dawson LF, Martin MJ, *et al. E*volutionary dynamics of *Clostridium difficile* over short and long time scales. Proc Natl Acad Sci. 2010;107(16):7527–32.

466. Velmurugan K, Chen B, Miller JL, Azogue S, Gurses S, Hsu T, *et al. Mycobacterium tuberculosis nuoG* is a virulence gene that inhibits apoptosis of infected host cells. PLOS Pathog. 2007;3(7):e110.

467. Miller JL, Velmurugan K, Cowan MJ, Briken V. The type I NADH dehydrogenase of *Mycobacterium tuberculosis* counters phagosomal NOX2 activity to inhibit TNF-alpha-mediated host cell apoptosis. PLOS Pathog. 2010;6(4):e1000864.

468. Blomgran R, Desvignes L, Briken V, Ernst JD. *Mycobacterium tuberculosis* inhibits neutrophil apoptosis, leading to delayed activation of naive CD4 T cells. Cell Host Microbe. 2012;11(1):81–90.

469. Doolittle WF, Zhaxybayeva O. On the origin of prokaryotic species. Genome Res. 2009;19(5):744–56.

470. Delafont V, Brouke A, Bouchon D, Moulin L, Héchard Y, He Y. Microbiome of free-living amoebae isolated from drinking water. Water Res. 2013;47(19):6958–65.

471. Yu J, Kim D, Lee T. Microbial diversity in biofilms on water distribution pipes of different materials. Water Sci Technol. 2010;61(1):163–71.

472. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, *et al.* Comparative metagenomics of microbial communities. Science. 2005;308(5721):554–7.

473. Tringe SG, Rubin EM. Metagenomics: DNA sequencing of environmental samples. Nat Rev Genet. 2005;6(11):805–14.

474. Martin J, Sykes S, Young S, Kota K, Sanka R, Sheth N, *et al.* Optimizing read mapping to reference genomes to determine composition and species prevalence in microbial communities. PLoS One. 2012;7(6):e36427.

475. Kyrpides NC, Hugenholtz P, Eisen J a, Woyke T, Göker M, Parker CT, *et al.* Genomic encyclopedia of bacteria and archaea: sequencing a myriad of type strains. PLoS Biol. 2014;12(8):e1001920.

476. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods. 2012;9(8):811–4.

477. Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 2011;27(21):2957–63.

478. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;27(6):863–4.

479. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, *et al.* vegan: Community Ecology Package. R package version 2.2-1. 2015. Available from: http://cran.r-project.org/package=vegan

480. Hurlbert SH. The nonconcept of species diversity: a critique and alternative parameters. Ecology. 1971;52:577–86.

481. Good I. The Population Frequencies of Species and the Estimation of Population Parameters. Biometrika. 1953;40(3):237–64.

482. Hill MO. Diversity and evenness: a unifying notation and its consequences. Ecology. 1973. p. 427–32.

483. Simpson EH. Measurement of Diversity. Nature. 1949;163(4148):688–688.

484. Shannon CE, Weaver W. A mathematical theory of communication. Bell Syst Tech J. 1948;27:379–656.

485. Chao A, Shen T-J. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. Environ Ecol Stat. 2003;10:429–43.

486. Bengtsson J, Eriksson KM, Hartmann M, Wang Z, Shenoy BD, Grelet G-A, *et al.* Metaxa: a software tool for automated detection and discrimination among ribosomal small subunit (12S/16S/18S) sequences of archaea, bacteria, eukaryotes, mitochondria, and chloroplasts in metagenomes and environmental sequencing datasets. Antonie Van Leeuwenhoek. 2011;100(3):471–5.

487. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, *et al.* Introducing *mothur*: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75(23):7537–41.

488. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41(Database issue):D590–6.

489. Stoddard SF, Smith BJ, Hein R, Roller BRK, Schmidt TM. rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. Nucleic Acids Res. 2014;43(November 2014):593–8.

490. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.

491. Monier J, Demanèche S, Delmont TO, Mathieu A, Vogel TM, Simonet P. Metagenomic exploration of antibiotic resistance in soil. Current Opinion in Microbiology. 2011. p. 229–35.

492. Gibbons SM, Jones E, Bearquiver A, Blackwolf F, Roundstone W, Scott N, *et al.* Human and environmental impacts on river sediment microbial communities. PLoS One. 2014;9(5):e97435.

493. Gittel A, Bárta J, Kohoutová I, Mikutta R, Owens S, Gilbert J, *et al.* Distinct microbial communities associated with buried soils in the Siberian tundra. ISME J. 2014;8(4):841–53.

494. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, *et al.* QIIME allows analysis of high-throughput community sequencing data. Nature Methods. Nature Publishing Group; 2010. p. 335–6.

495. Huson DH, Mitra S, Ruscheweyh H-J, Weber N, Schuster SC. Integrative analysis of environmental sequences using MEGAN4. Genome Res. 2011;21(9):1552–60.

496. Madigan MT, Martinko JM, Bender KS, Buckley DH, Stahl DA, Brock T. Brock Biology of Microorganisms, 14th Edition. Benjamin Cummings; 2014.

497. Rai AN, Bergman B, Rasmussen U, editors. Cyanobacteria in Symbiosis. Kluwer Academic Publishers; 2002.

498. Hilton JA, Foster RA, Tripp HJ, Carter BJ, Zehr JP, Villareal TA. Genomic deletions disrupt nitrogen metabolism pathways of a cyanobacterial diatom symbiont. Nat Commun. 2013;4:1767.

499. Yutin N, Suzuki MT, Teeling H, Weber M, Venter JC, Rusch DB, *et al*. Assessing diversity and biogeography of aerobic anoxygenic phototrophic bacteria in surface waters of the Atlantic and Pacific Oceans using the Global Ocean Sampling expedition metagenomes. Environ Microbiol. 2007;9(6):1464–75.

500. Hong C, Si Y, Xing Y, Li Y. Illumina MiSeq sequencing investigation on the contrasting soil bacterial community structures in different iron mining areas. Environ Sci Pollut Res Int. 2015 (In press).

501. Santos CR, Hoffmam ZB, de Matos Martins VP, Zanphorlin LM, de Paula Assis LH, Honorato RV, *et al.* Molecular mechanisms associated with xylan degradation by *Xanthomonas* plant pathogens. J Biol Chem. 2014;289(46):32186–200.

502. Hirsch P, Ludwig W, Hethke C, Sittig M, Hoffmann B, Gallikowski CA. *Hymenobacter roseosalivarius* gen. nov., sp. nov. from continental Antartica soils and sandstone: bacteria of the *Cytophaga*/*Flavobacterium*/*Bacteroides* line of phylogenetic descent. Syst Appl Microbiol. 1998;21(3):374–83.

503. Shivali K, Ramana VV, Ramaprasad EVV, Sasikala C, Ramana CV. *Marichromatium litoris* sp. nov. and *Marichromatium chrysaorae* sp. nov. isolated from beach sand and from a jelly fish (*Chrysaora colorata*). Syst Appl Microbiol. 2011;34(8):600–5.

504. Suzuki T, Mori Y, Nishimura Y. *Roseibacterium elongatum* gen. nov., sp. nov., an aerobic, bacteriochlorophyll-containing bacterium isolated from the west coast of Australia. Int J Syst Evol Microbiol. 2006;56(Pt 2):417–21.

505. Marques S, Ramos JL. Transcriptional control of the *Pseudomonas putida* TOL plasmid catabolic pathways. Mol Microbiol. 1993;9(5):923–9.

506. Shin SC, Kim SJ, Ahn DH, Lee JK, Park H. Draft genome sequence of *Sphingomonas echinoides* ATCC 14820. J Bacteriol. 2012;194(7):1843.

507. Gupta RS, Naushad S, Baker S. Phylogenomic Analyses and Molecular Signatures for the Class *Halobacteria* and its Two Major Clades: A Proposal for Division of the Class *Halobacteria* into an emended order *Halobacteriales* and Two New Orders, *Haloferacales* ord. nov. and *Natrialbales* ord. n. Int J Syst Evol Microbiol. 2014;65(Pt 3):1050-69.

508. Krupovic M, Forterre P. *Microviridae* goes temperate: microvirus-related proviruses reside in the genomes of *Bacteroidetes*. PLoS One. 2011;6(5):e19893.

509. Lynch MDJ, Neufeld JD. Ecology and exploration of the rare biosphere. Nat Rev Microbiol. 2015;13(4):217–29.

510. Sayers E. The E-utilities In-Depth: Parameters, Syntax and More.. Entrez Programming Utilities Help. Bethesda (MD): National Center for Biotechnology Information (US). 2009.

511. Oliveros JC. Venny. An interactive tool for comparing lists with Venn's diagrams. 2007. Available from: http://bioinfogp.cnb.csic.es/tools/venny/index.html

512. Creighton S. Gonorrhoea. BMJ Clin Evid. 2014;1604.

513. Lawn SD, Zumla AI. Tuberculosis. Lancet. 2011;378(9785):57–72.

514. Harris JB, LaRocque RC, Qadri F, Ryan ET, Calderwood SB. *Cholera*. Lancet. 2012;379(9835):2466–76.

515. Nelson EJ, Harris JB, Morris JG, Calderwood SB, Camilli A. *Cholera* transmission: the host, pathogen and bacteriophage dynamic. Nat Rev Microbiol. 2009;7(10):693–702.

516. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010;11(1):31–46.

517. Fisman DN, Lim S, Wellenius GA, Johnson C, Britz P, Gaskins M, *et al.* It's Not the Heat, It's the Humidity: Wet Weather Increases Legionellosis Risk in the Greater Philadelphia Metropolitan Area. J Infect Dis. 2005;192:2066–73.

518. Portillo ME, Reina G, Fernández-Alonso M, Leiva J, Portillo ME, Reina G, *et al.* Real-time PCR for early detection and monitoring of *Legionella pneumophila* in clinical specimens. Enferm Infecc Microbiol Clin. 2010;28(8):562–3.

519. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, *et al.* Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat Genet. 2013;45(10):1176-82.

520. Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I*, et al.* Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. Nature. 2014;514:494-7.