Investigation of the population structure of *Legionella pneumophila* by analysis of tandem repeat copy number and internal sequence variation

Paolo Visca,1 Silvia D’Arezzo,2 Françoise Ramisse,3 Yevgeniy Gelfand,4,5 Gary Benson,4 Gilles Vergnaud,6,7 Norman K. Fry8 and Christine Pourcel6

1Dipartimento di Biologia, Università Roma Tre, Rome, Italy
2Istituto Nazionale per le Malattie Infettive ‘Lazzaro Spallanzani’ IRCCS, Rome, Italy
3Division of Analytical Microbiology, DGA CBRN Defence, Vert le Petit, France
4Laboratory for Biocomputing and Informatics, Boston University, Boston, MA, USA
5Department of Computer Science, Department of Biology, Program in Bioinformatics, Boston University, Boston, MA, USA
6Université Paris-Sud, Institut de Génétique et Microbiologie, Orsay, France
7DGA/MRIS, Mission pour la Recherche et l’Innovation Scientifique, Bagneux, France
8Respiratory and Systemic Infection Laboratory, Health Protection Agency Centre for Infections, London, UK

The population structure of the species *Legionella pneumophila* was investigated by multilocus variable number of tandem repeats (VNTR) analysis (MLVA) and sequencing of three VNTRs (Lpms01, Lpms04 and Lpms13) in selected strains. Of 150 isolates of diverse origins, 136 (86 %) were distributed into eight large MLVA clonal complexes (VACCs) and the rest were either unique or formed small clusters of up to two MLVA genotypes. In spite of the lower degree of genome-wide linkage disequilibrium of the MLVA loci compared with sequence-based typing, the clustering achieved by the two methods was highly congruent. The detailed analysis of VNTR Lpms04 alleles showed a very complex organization, with five different repeat unit lengths and a high level of internal variation. Within each MLVA-defined VACC, Lpms04 was endowed with a common recognizable pattern with some interesting exceptions. Evidence of recombination events was suggested by analysis of internal repeat variations at the two additional VNTR loci, Lpms01 and Lpms13. Sequence analysis of *L. pneumophila* VNTR locus Lpms04 alone provides a first-line assay for allocation of a new isolate within the *L. pneumophila* population structure and for epidemiological studies.

INTRODUCTION

The genus *Legionella* comprises more than 50 species widely distributed in water, where they multiply in protozoa (Gao & Kwaik, 2000). *Legionella pneumophila* is the most frequently isolated *Legionella* species from water-distribution networks and is also the one predominantly associated with disease (Fields et al., 2002). Within the species *L. pneumophila* a large genetic diversity is encountered and three subspecies have been defined, namely *L. pneumophila* subsp. pneumophila, *L. pneumophila* subsp. *fraseri* and *L. pneumophila* subsp. *pascullei* (Brenner et al., 1988). Comparing the six sequenced *L. pneumophila* genomes revealed that on average 25 % of the genome is dispensable (Cazalet et al., 2004; Chien et al., 2004; D’Auria et al., 2010; Schroeder et al., 2010; Steinert et al., 2007). At

Abbreviations: CC, clonal complex; Lpms, *Legionella pneumophila* mini-microsatellite; MLVA, multilocus variable number of tandem repeats analysis; MLVA-8, eight-loci MLVA; MULALBMA, multiple alignment by BLAST; SBA, sequence-based analysis; SBT, sequence-based typing; SB7, seven-loci SBT protocol; sg, serogroup; ST, sequence type; TRDB, Tandem Repeat Database; UPGMA, unweighted pair group method with arithmetic means; VACCs, MLVA clonal complexes; VNTR, variable number of tandem repeats.

The GenBank/EMBL/DBJ accession numbers for the sequences determined in this study are FR729481 to FR729558 and FR851977 to FR851991 (Lpms04), FR848159 to FR848205 (Lpms13) and FR848206 to FR848244 (Lpms01).

Eight supplementary figures and two supplementary tables are available with the online version of this paper.
present 15 serogroups (sgs) have been designated and there is evidence for the existence of additional ones (Helbig et al., 2007; Lück et al., 1995). A study performed in the UK showed that the diversity as estimated by different assays, including serogrouping and genetic analysis, was higher in environmental isolates than in clinical ones. In that study, three dominant genotypes were observed among clinical isolates (Harrison et al., 2007).

The genetic structure of the L. pneumophila species was analysed early on by Selander et al. (1985). Families were described on the basis of enzyme polymorphism (as assessed by multi-locus enzyme electrophoresis). These data were compatible with a clonal structure of L. pneumophila and showed that many clones were distributed worldwide. A sequence-based typing (SBT) scheme was later developed for L. pneumophila, utilizing genes that are subject to different evolutionary constraints (Gaia et al., 2005; Ratzow et al., 2007). Using six-loci SBT (Gaia et al., 2005), the structure of the population was investigated in clinical and environmental isolates (Edwards et al., 2008). Large-scale comparative genomic studies based upon DNA arrays confirmed the higher genetic diversity among environmental isolates with respect to clinical isolates and the presence of specific clones of L. pneumophila overrepresented in human disease or causing legionellosis worldwide (Gomez-Valero et al., 2009; Tijet et al., 2010). Pulsed field gel electrophoresis (PFGE) has also been extensively used to compare strains. A particular strain family, the ‘Paris’ lineage, originally characterized by a specific PFGE profile, has been detected since 1987 in the water distribution network in many towns in Europe, and is responsible for numerous cases of infection (Aurell et al., 2003; Cazalet et al., 2008; Lawrence et al., 1999). This lineage appears to show a strong association with sequence type (ST) 1 (1, 4, 3, 1, 1, 1, 1) (Reimer et al., 2010) and until now has been described in sg1 isolates only. Similarly the ‘Lorraine’ PFGE profile which appears to correspond to the monoclonal antibody subgroup Allentown ST47 has been shown to be frequently found in Europe (Ginevra et al., 2008; Harrison et al., 2009).

Despite its clonal population structure, genetic exchanges between strains occur in L. pneumophila, as shown by the analysis of the dotA, mip and rpoB genes (Bumbaugh et al., 2002; Ko et al., 2002). Molecular evolution of the virulence-related dotA gene in L. pneumophila has unravelled multiple recombinational exchanges within its coding region (Ko et al., 2003a). A mosaic structure has also been shown for pathogenicity islands (Ko et al., 2003b). Recent comparative genomic studies of L. pneumophila have confirmed the existence of a high rate of DNA exchange with both bacteria and eukaryotes (Gomez-Valero et al., 2009).

To better define particular lineages more frequently associated with human infection and to investigate the mechanism of their emergence, polyphasic genotyping studies are increasingly being employed, making use of genetic elements with different evolutionary rates (Keim et al., 2004). An interesting source of genetic polymorphism is provided by tandemly repeated sequences, called variable number of tandem repeats (VNTRs), whose number of repetitions varies at different rates depending on the different loci and even alleles (García-Yoldi et al., 2007). Molecular typing based on the analysis of repeat copy number at multiple VNTR loci, called multiple locus VNTR analysis (MLVA), is a genotyping method that is being used for strain comparison. It can also provide insights into population structure (reviewed by Vergnaud & Pourcel, 2009). MLVA using eight VNTRs (MLVA-8) has been shown to provide high-resolution L. pneumophila sg1 genotyping (Coil et al., 2008; Pourcel et al., 2003, 2007). In addition, sequence variations have been observed inside repeats for some loci (Pourcel et al., 2007), which might help trace the evolution of these elements and represent an additional and easy to run typing tool. Lpms04 is one such candidate since it is composed of differently sized tandem repeats, precluding its use in a first-line DNA-fragment-length-based genotyping protocol (Pourcel et al., 2003, 2007).

To further evaluate the efficiency of tandem repeats in providing a global picture of L. pneumophila, we first investigated the European Working Group on Legionella Infection (EWGLI) reference collection and a few additional reference strains. Clustering suggested by MLVA analysis was compared with SBT typing and we found that in spite of the inherent homoplasy associated with variations in the number of repetitions of VNTRs, the two independent approaches were highly congruent for the definition of clonal complexes. Comparison of MLVA clustering and analysis of internal sequence variation in three tandem repeats, including Lpms04, suggests the existence of recombination events.

METHODS

Bacterial strains. A total of 150 L. pneumophila isolates was investigated. Ninety-five of these isolates are from the European Union Legionella (EUL) collection, some of which had previously been genotyped by three different techniques, namely amplified fragment length polymorphism (AFLP), SBT and MLVA (Fry et al., 2002; Gaia et al., 2005; Pourcel et al., 2007). The 14 additional samples for which SBT results were known include reference strains for sg1, sg2 and sg4 to sg14 (of which two are members of the subspecies fraseri (Los Angeles and Dallas 1E) and a member of the subspecies pascaliei (strain U8W, purchased from the Institut Pasteur, Collection, 03-008 in Supplementary Table S1, available with the online version of this paper). Forty-one clinical and environmental isolates were collected over a period of 15 years by Cuong Tram at the Institut Pasteur as part of water-distribution system surveillance and in epidemiological investigations (Pourcel et al., 2003) (Supplementary Table S1). Where available, sg ST and SBT allelic profiles are shown for each isolate. For this study we previously described seven-loci SBT protocol (SBT7) was used (Ratzow et al., 2007). However, due to strain variation, some non-sg1 strains do not amplify with the neuA primers; these are denoted by ‘F’ for that locus (see http://www.hpa-bioinformatics.org.uk/legionella/legionella_sbt/php/sbt_homepage.php).
SBT data analysis. Sequences corresponding to the SBT loci were deduced in silico from sequenced genomes using published primers (Gaia et al., 2005; Ratzow et al., 2007) or recovered from the EWGLI SBT typing web page. SBT alleles were coded following the EWGLI conventions. SBT codes were clustered using Hamming’s distance (the categorical coefficient) and unweighted pair group method with arithmetic mean (UPGMA). Sequence comparisons were also run by concatenating six loci (excluding newa because of the absence of data for some strains). A composite dataset and minimum spanning tree was made using BioNumerics 6.5 (Applied-Maths) default parameters.

VNTRs typing. Eight VNTR loci, designated Lpms (for Legionella pneumophila mini-microsatellite) followed by a number were used as pneumophila was performed using Hamming’s distance and UPGMA.

Analysis of linkage disequilibrium, bootstrapping and congruence between different methods. Linkage disequilibrium was measured by using LIAN version 3.5 software (http://guanine.evolbio.mpg.de/) as described by Haubold & Hudson (2000). The Monte-Carlo simulation was run with 100 000 iterations. Bootstrap analyses were run using BioNumerics 6.5 with 500 simulations. Congruence between different experiments was measured by BioNumerics version 6.5 using Pearson correlation.

VNTR DNA sequence analysis. Selected Lpms01, Lpms04 and Lpms13 alleles were sequenced. Selection was based on their size and distribution by MLVA-8 clustering. Because of the degree of divergence in the DNA immediately flanking the repeat in Lpms04, oligonucleotide primer sequences were designed in the DNA region outside the gene encompassing Lpms04 (Table 1). The full-length sequences of PCR products were determined on both strands following DNA purification by the QiAquick PCR purification kit (Qiagen). Sequencing reactions were performed using the BigDye terminator technology (Applied Biosystems) according to the manufacturer’s recommendations, and products were analysed in an ABI 3100 capillary electrophoresis system equipped with the POP 4 matrix (Applied Biosystems). Data obtained with forward and reverse sequencing primers were combined. EMBL accession numbers are FR729481 to FR729558 and FR851977 to FR851991 (Lpms04), FR848159 to FR848205 (Lpms13) and FR848206 to FR848244 (Lpms01).

Alignment of nucleic acids and protein sequences. To compare VNTR sequences the Tandem Repeat Finder (TRF) (Benson, 1999) and the Tandem Repeat Database (TRDB) (Gelfand et al., 2007) were used (both tools available at http://tandem.bu.edu/) (Gelfand et al., 2007). Additional tools were developed in the present work allowing the production by TRDB of a matrix of pairwise distances between VNTR alleles (Benson, 2002). For this purpose, each VNTR allele is first transformed into a multiple alignment of its individual copies. Each column of the multiple alignment is then represented by a nucleotide composition which is a vector of counts of the individual nucleotide types in the column (A, C, G, T and gap). The sequence for a VNTR, in the same order as the columns of the multiple alignment, is termed a profile. Profiles were globally aligned using a Euclidean distance function for composition (Rao et al., 2005).

Multiple alignment by BLAST (MULALBLA), developed by the Biology of Extremophiles Laboratory http://www-archbac.u-psud.fr/MULALBLA/mulalbla.html, was used for amino acid sequence similarity searches. Prediction of hydrophobicity or hydropillicity scales was performed with the Protscal tool at http://www.expasy.org/. Prediction of secondary structure was made using GOR IV at http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor4.html (Garnier et al., 1996).

Southern blot analysis. Genomic DNA was purified using the DNeasy purification kit (Qiagen) as recommended by the manufacturer. The DNA concentration was estimated using an ND-1000

Table 1. Characteristics of VNTRs analysed in this study

<table>
<thead>
<tr>
<th>VNTR</th>
<th>Position Phil (kb)*</th>
<th>Motif (bp)</th>
<th>Primers</th>
<th>Associated gene†</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lpms01L</td>
<td>3231</td>
<td>45</td>
<td>ACGAGCATATGACAAAGGCTTGG</td>
<td>lpg2854</td>
<td>Signal transduction</td>
</tr>
<tr>
<td>Lpms01R</td>
<td>3231</td>
<td>45</td>
<td>CGGATCATCTGAGGCTATTG</td>
<td>lpg2793</td>
<td>Other regulatory functions</td>
</tr>
<tr>
<td>Lpms03L</td>
<td>3144</td>
<td>96</td>
<td>CAACCAATGAGAAGAAGCA</td>
<td>lpg2844</td>
<td>Hypothetical histidine-rich protein</td>
</tr>
<tr>
<td>Lpms03R</td>
<td>3144</td>
<td>96</td>
<td>AGGGTTGATGTCTCAG</td>
<td>lpg2844</td>
<td>Hypothetical histidine-rich protein</td>
</tr>
<tr>
<td>Lpms04ATGFor</td>
<td>3219</td>
<td>21; 48; 51; 54</td>
<td>CAGATAGGTTAATACG</td>
<td>lpg2844</td>
<td>Hypothetical histidine-rich protein</td>
</tr>
<tr>
<td>Lpms04SecRb</td>
<td>1645</td>
<td>24</td>
<td>AAGCTTGCTACAGTTGCTGG</td>
<td>lpg1488</td>
<td>Conserved protein, surface antigen</td>
</tr>
<tr>
<td>Lpms13L</td>
<td>930</td>
<td>39</td>
<td>CAATAGCATCGACTGAGCA</td>
<td>lpg0854</td>
<td>FleQ, transcriptional regulator</td>
</tr>
<tr>
<td>Lpms13R</td>
<td>930</td>
<td>39</td>
<td>TGGCCTGTATCTGGAAAGAC</td>
<td>lpg0854</td>
<td>FleQ, transcriptional regulator</td>
</tr>
<tr>
<td>Lpms17L</td>
<td>913</td>
<td>21</td>
<td>GAACTATCGAGAGGAGCAT</td>
<td>lpg1488</td>
<td>Conserved protein, surface antigen</td>
</tr>
<tr>
<td>Lpms19L</td>
<td>2578</td>
<td>125</td>
<td>GGGAGAAATGTTACGCTAGC</td>
<td>Intergenic</td>
<td></td>
</tr>
<tr>
<td>Lpms19R</td>
<td>2578</td>
<td>125</td>
<td>ACCACAGCAGTTGAAACATAT</td>
<td>Intergenic</td>
<td></td>
</tr>
<tr>
<td>Lpms33L</td>
<td>2661</td>
<td>125</td>
<td>AAAACCTGCTGACCCCGCTTGC</td>
<td>Intergenic</td>
<td></td>
</tr>
<tr>
<td>Lpms33R</td>
<td>2661</td>
<td>125</td>
<td>AAAACCTGCTGACCCCGCTTGC</td>
<td>Intergenic</td>
<td></td>
</tr>
<tr>
<td>Lpms34L</td>
<td>1428</td>
<td>18</td>
<td>CTGAAAAAGCTGTAGTTGTA</td>
<td>lpg1299</td>
<td>FimV, transmembrane Tfp pilus assembly</td>
</tr>
<tr>
<td>Lpms34R</td>
<td>1428</td>
<td>18</td>
<td>TATCAACCTGATCATCCCT</td>
<td>lpg1299</td>
<td>FimV, transmembrane Tfp pilus assembly</td>
</tr>
</tbody>
</table>

*Relative to the L. pneumophila Philadelphia-1 genome.
†Relative to the annotation of L. pneumophila Philadelphia-1 genome.
spectrophotometer (NanoDrop, Labtech). Samples (5 μg) of genomic DNA were digested with HindII, electrophoresed on a 1.5 % agarose gel and transferred onto a nylon membrane (Genescreen). Hybridization was performed using radiolabelled probes (Church & Gilbert, 1984). Probes encompassing the Lpms04 region from strains 81A3105 (ATCC 43736) (48 bp repeat size), 02-157 (51 bp repeat size) and 02-304 (21 bp repeat size), and the Lpms13 region from strain Philadelphia-1 (ATCC 33152), were generated by PCR using the primers shown in Table 1 and purified by PEG precipitation as described by Embley (1991).

RESULTS

SBT and MLVA clustering define congruent L. pneumophila clonal complexes

We first investigated the population structure of 107 strains that had been previously typed by a number of different methods, including SBT and MLVA-8 (Pourcel et al., 2007). This set included 95 strains from the EWGLI reference collection and 12 additional reference strains. SBT discriminated 58 genotypes (STs) among the 98 strains for which a full SBT7 dataset was available (neuA data are missing for nine isolates). SBT7 clustering (categorical coefficient, UPGMA) allowing a maximum of two allele differences out of the seven loci defined eight clusters (CC1 to CC8) including at least two different STs (the colour code reflects clusters), as shown in the minimum spanning tree in Fig. 1. Five clusters comprised two or three different STs and the three largest clusters CC1, CC2 and CC3 comprised respectively 10, 12 and five different STs. Eighteen STs were singletons. Two of these (ST47 and ST53) were represented by two isolates; one (ST62) was represented by four. The standardized index of linkage association for SBT7 genotypes was 0.23 (Table 2). The measured linkage disequilibrium was highly significant ($P<10^{-2}$). MLVA-8 based upon Lpms01, Lpms03, Lpms13, Lpms17, Lpms19, Lpms33, Lpms34 and Lpms35 defined 50 genotypes by clustering using the categorical coefficient and UPGMA. Fig. 2 shows the dendrogram generated by using the same colour code as in Fig. 1, which illustrates the high congruence of clustering defined by the two independent methods, SBT7 and MLVA-8 (calculated congruence 79 %). The partition mapping, also called the contingency table, of the SBT7 and MLVA-8 clustering further confirmed the congruence between SBT7 and MLVA-8 (data not shown). A bootstrap analysis was used to evaluate the significance of observed MLVA-8 clonal complexes (CCs), which were called VACC to differentiate them from SBT CCs. The largest SBT CCs, CC1 and CC2, corresponded to the two largest MLVA CCs, coded in bright green (VACC1) and red (VACC2) in Fig. 1. VACC1 included the ‘Paris’ reference strain (ST1; 1, 4, 3, 1, 1, 1, 1), whose genome has been sequenced, and the sg9 reference strain IN-23-G1-C2 or ATCC 35289 (ST390; 1, 4, 3, 28, 1, 1, 6). VACC2 included the Philadelphia-1 or ATCC 33152 strain (ST36; 3, 4, 1, 1, 14, 9, 1), the first L. pneumophila isolate associated with disease in humans.

For a comparison an analysis in which the full SBT sequence information was used for the six genes in 107 isolates (analysis called here sequence-based analysis – SBA— not including the contribution from neuA, for which nine isolates could not be typed) was performed (data not shown). A total of 237 nucleotide differences were present.

Fig. 1. SBT-based minimum spanning tree showing clonal complexes as colour code. Only connecting branches with a maximum length of 3 are shown. The arbitrary SBT clonal complex numbers are indicated by CCs, e.g. CC1. Numbers with no CC prefix are sequence types (STs) as assigned by the EWGLI SBT site.

http://mic.sgmjournals.org

2585
The congruence between SBA and SBT7 was 64%, only slightly higher than that between SBA and MLVA-8 (59%).

The standardized index of linkage association for MLVA-8 genotypes was 0.032 (Table 2). The detected linkage disequilibrium was significant \((P < 0.05\) for MLVA-8 and SBT7, respectively. Strain EUL031 or Corby (ST51; 6, 10, 15, 28, 9, 14, 6), respectively. Concord 3 or ATCC 35096 (sg8; 8, 1, 22, 30, 6, 10, F), and EUL137 or Corby (ST51; 6, 10, 15, 28, 9, 14, 6), respectively. Strain EUL031 (ST47; 5, 10, 22, 15, 6, 2, 6) with the Lorraine PFGE profile belonged to VACC8. VACC4 (or SBT CC4), with two genotypes, was not related to any reference strain or strain whose genome had been sequenced, and all the other isolates were distributed into smaller clusters or were unique.

**Internal sequence variations of the Lpms04 repeats**

Since previous studies have revealed a high degree of internal repeat variation of Lpms04 (Pourcel et al., 2003), we sequenced this locus in 105 isolates (denoted with asterisks in Fig. 2 and Supplementary Fig. S1) and evaluated its polymorphism with respect to the previously defined CCs. The tandem repeats were identified with the help of the Tandem Repeat Finder software (TRF) (Benson, 1999). Different repeat unit sizes were observed, but each allele was usually composed of a single type of repeat length, namely 54 bp, 51 bp, 48 bp or 21 bp (Supplementary Fig. S1 and Fig. 3a). Other repeats were occasionally observed, such as a 39 bp repeat in reference strain ATCC 33154 (Togus-1, sg2, ST39; 3,5,1,7,14,9,8) and a 42 bp repeat present at the beginning of the VNTR in CC2 members and in reference strains ATCC 33154 (Togus-1, sg2, ST39; 3,5,1,7,14,9,8) and ATCC 43290 (570-CO-H, sg12, ST187; 3, 10, 1, 28, 14, 9, 3). Alleles of EUL048 and EUL056 (both ST48; 5, 2, 22, 27, 6, 10, 12) showed the most complex organization, with a succession of 48 bp, 54 bp and 57 bp repeats. These three repeat units can be derived one from the other by simple

---

**Table 2. Analysis of the MLVA-8 and SBT7 data using LIAN 3.5 to test for linkage disequilibrium**

<table>
<thead>
<tr>
<th>Genotypes</th>
<th>MLVA-8</th>
<th>SBT7</th>
<th>MLVA-8 + SBT7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loci</td>
<td>50</td>
<td>65</td>
<td>72</td>
</tr>
<tr>
<td>Mean genetic diversity ((H))</td>
<td>0.6344 ± 0.1002</td>
<td>0.8738 ± 0.0081</td>
<td>0.7329 ± 0.0574</td>
</tr>
<tr>
<td>Observed mismatch variance ((V_d))</td>
<td>1.5875</td>
<td>1.8502</td>
<td>7.5298</td>
</tr>
<tr>
<td>Expected mismatch variance ((V_e))</td>
<td>1.2928</td>
<td>0.7690</td>
<td>2.2446</td>
</tr>
<tr>
<td>Standardized index of association ((I_s))</td>
<td>0.0326</td>
<td>0.2343</td>
<td>0.1682</td>
</tr>
<tr>
<td>Simulated 5% critical value ((L_{calc}))</td>
<td>1.4241</td>
<td>0.8141</td>
<td>2.4061</td>
</tr>
<tr>
<td>Calculated 5% critical value ((L_{para}))</td>
<td>1.4146</td>
<td>0.8122</td>
<td>2.3994</td>
</tr>
<tr>
<td>Simulated significance ((P_{calc}))</td>
<td>5.8 × 10^{-4}</td>
<td>&lt;1.00 × 10^{-5}</td>
<td>&lt;1.00 × 10^{-5}</td>
</tr>
<tr>
<td>Calculated significance ((P_{para}))</td>
<td>1.82 × 10^{-8}</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>Heterozygosity locus 1*</td>
<td>0.7380</td>
<td>0.8755</td>
<td></td>
</tr>
<tr>
<td>Heterozygosity locus 2</td>
<td>0.5200</td>
<td>0.8418</td>
<td></td>
</tr>
<tr>
<td>Heterozygosity locus 3</td>
<td>0.9037</td>
<td>0.8832</td>
<td></td>
</tr>
<tr>
<td>Heterozygosity locus 4</td>
<td>0.1151</td>
<td>0.8923</td>
<td></td>
</tr>
<tr>
<td>Heterozygosity locus 5</td>
<td>0.3698</td>
<td>0.9029</td>
<td></td>
</tr>
<tr>
<td>Heterozygosity locus 6</td>
<td>0.7992</td>
<td>0.8678</td>
<td></td>
</tr>
<tr>
<td>Heterozygosity locus 7</td>
<td>0.6833</td>
<td>0.8534</td>
<td></td>
</tr>
<tr>
<td>Heterozygosity locus 8</td>
<td>0.9461</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*The loci are listed in the following order: Lpms01, Lpms03, Lpms13, Lpms17, Lpms19, Lpms33, Lpms34 and Lpms35, and flA, pilE, asd, mip, mompsS, proA and neuA for MLVA-8 and SBT7, respectively.**
**Fig. 2.** MLVA-8-based dendrogram showing clonal complexes (VACC). The same colour code as in Fig. 1 is used. Bootstrap values >59% are shown.
internal duplication and deletion events (see EUL048 in Fig. 3a). For instance the 57 bp repeat derives from the 48 bp motif by the 9 bp (GCA)₃ addition. They are also related to the 42 bp motif present once or rarely twice at the beginning of the VNTR in VACC2 members. The 21 bp and 51 bp repeats are quite different from the latter and from each other, which raises the question of their origin. One possibility could be the acquisition of the different repeated sequences by recombination with other region(s) of the chromosome. To test this hypothesis, probes were prepared by PCR amplification of three Lpms04 variants showing different internal sequences (48 bp unit, 21 bp unit, 51 bp unit; Fig. 3a) and hybridized to Hinfl-digested genomic DNA from strains possessing the different kinds of repeats (Fig. 3b). Following probe stripping, the same membrane was repeatedly used for hybridization. According to sequencing data, Hinfl does not cut inside the repeats and therefore a single band is observed which contains the entire VNTR. The results obtained with a selection of 14 isolates showed that the 48 bp probe hybridized with both the 54 bp and 48 bp repeats that show internal sequence similarity (Fig. 3a) and are also related to the 42 bp and 57 bp repeats (Fig. 3a). On the other hand, the 21 bp and 51 bp repeats only hybridized with themselves (Fig. 3b). In addition to the band encompassing the VNTR, minor bands were observed which were probably due to hybridization with short sequences flanking the VNTR and present in the probe. Restriction enzyme (Hinfl) polymorphism differentiates strains with a 48 bp repeat, which show hybridization fragments <1 kb, from strains with 54 bp, 51 bp or 21 bp repeats, which show hybridization fragments >1 kb. The absence in some strains of a Hinfl site immediately flanking

---

Fig. 3. Lpms04 repeats variability. (a) Sequence of the tandem repeats in strain ATCC 43736 (48 bp unit), 02-157 (51 bp unit), 02-304 (21 bp unit) and EUL048 (48 bp, 54 bp and 57 bp units). (b) Southern blot analysis of Hinfl-digested genomic DNA in strains 02-300 (54 bp), 02-332 (48 bp), 02-304 (21 bp), 02-185 (54 bp), 02-321 (54 bp), 02-322 (48 bp), 02-323 (54 bp), 02-127 (54 bp), 02-138 (48 bp), 02-157 (51 bp), 02-158 (54 bp) and 02-139 (54 bp). The hybridization probes correspond to the Lpms04 tandem repeat in strain ATCC 43736 (48 bp), 02-157 (51 bp), 02-304 (21 bp) and the Lpms13 tandem repeat from strain ATCC 33152 as a control. Following probe stripping, the same membrane was repeatedly used for hybridization. The positions of the DNA size markers are shown on the right.
the VNTR on one side was confirmed by examination of genome sequences (data not shown). As a control for hybridization, the Lpms13 VNTR from strain ATCC 33152 (Philadelphia-1) was used as a probe and a band was observed in all the samples.

Generally, inside MLVA VACCs the Lpms04 repeat unit length was identical (Supplementary Fig. S1), except for a few strains. Strain 02-157 possesses a 51 bp repeat but clusters in a heterogeneous small group, separated from the typical and very similar VACC4 strains which all possess a 51 bp Lpms04 repeat (Supplementary Fig. S1). Strains 02-323 and 02-333 which have 54 bp repeats cluster with VACC3 strains possessing 48 bp repeats (Supplementary Fig. S1). Both EUL048 and EUL056 belong to VACC8 and their ST is also close to that of members of this group, although their Lpms04 shows three different repeat lengths of 48 bp, 54 bp and 57 bp. The other members of VACC8 have 48 bp repeats. These two strains were isolated in 1999 from the same patient in Spain at a 15 day interval (Fry et al., 1999).

In addition, repeats with similar motif lengths have important internal variations. To compare and classify internal variations we used a multiple sequence alignment software suited for tandem repeats available in TRDB. For the 48 bp tandem repeats, the sequences were aligned against the reference strain Philadelphia-1 (ATCC 33152) as shown in Supplementary Fig. S2, allowing identification of a typical pattern for the 12 reference strains with a 48 bp repeat. This pattern is produced by comparison of each repeat against a consensus sequence, where each difference is shown and coloured. Reference strains Los Angeles-1 ATCC 33156 sg4 (11, 14, 16, 25, 7, 13, F) and Dallas-1E ATCC 33216 sg5 (11, 14, 16, 18, 15, 13, F) both belong to the subsp. fraseri and harbour a similar Lpms04 locus, although they do not cluster by MLVA (Supplementary Fig. S1). The 54 bp Lpms04 repeats of reference strains ATCC 43283 (sg10) and ATCC 43703 (sg14) also showed a common pattern upon alignment (data not shown). We then examined whether a specific pattern could be recognized in each VACC. In Supplementary Fig. S3(a) the sequence of VACC2 strain Philadelphia-1 (box1) is compared with the sequence of sg13 reference strain ATCC 43736 (6, 10, 5, 10, 9, 1, 0) (box 2), EUL004 ST23 (box 3), EUL026 ST22 (box 4) and EUL029 ST19 (box 5), all of which belong to VACC3 and show extensive similarities in the repeat patterns. Similarly, in Supplementary Fig. S3(b), the Lpms04 sequences of the VACC1 strains EUL157 ST18 (box 2), ‘Paris’ (box 3) and ATCC 35289 sg9 ST390 (box 4) show extensive similarity. Similar alignments were performed with strains of the different VACCs, and in each VACC Lpms04 appeared to have a common recognizable pattern.

A new tool was developed on the TRDB website to allow the export of a similarity matrix between tandem repeat alleles. The resulting dendrogram for the Lpms04 alleles shows that clustering of isolates is in general agreement with MLVA-8 clustering although some exceptions are observed (see Supplementary Fig. S4).

**Variations and size homoplasy at Lpms01 and Lpms13 VNTR loci**

Sequencing of alleles of two additional VNTR loci, Lpms01 and Lpms13 (Tables 1 and 2) was performed to assess their internal variability and size homoplasy. Unlike Lpms04, the repeat size of Lpms01 and Lpms13 was always 45 bp and 24 bp, respectively. For Lpms01, comparison of motifs by multiple alignment revealed patterns which were highly similar inside clusters (see Supplementary Fig. S5). In addition, intercluster similarity could be observed, as for example between strains of VACC1 and VACC3. Lpms01 was almost identical in strains of the VACC9 (21 bp Lpms04) and VACC5 (54 bp Lpms04) clusters, in agreement with the clustering analysis (Supplementary Fig. S1). For Lpms13, assessment of intercluster resemblance was not straightforward, probably because of the higher instability of this marker as compared with Lpms01 (see Supplementary Fig. S6).

By comparing Lpms01, Lpms04 and Lpms13 VNTRs in different strains it was possible to observe the genetic proximity of strains isolated in different countries and the stability of the patterns (see Supplementary Fig. S7). This confirmed the observations made with the sole Lpms04 VNTR as shown for the VACC1 and VACC2 isolates (Supplementary Fig. S4). Interestingly, in other strains genetic exchanges were suggested by the VNTR sequence. In Supplementary Fig. S8, the patterns of three VNTRs in three VACC5 isolates [03-004 NP1 from a patient and 03-011 W1 from a water sample in a Parisian hospital (Pourec et al., 2003), reference strain 1586-SCT-H (ATCC 43703 sg14)] and the VACC10 strain Chicago-8 (ATCC 33823) sg7 are compared. Lpms01 and Lpms04 of the water and patient strains corresponded to those of the reference strains ATCC 33823 sg7 and ATCC 43703 sg14 respectively, whereas their Lpms13 alleles were similar and different from those of the two reference strains.

**DISCUSSION**

Phylogenetic analysis in *L. pneumophila* is a difficult task considering the enormous genetic differences observed in the accessory genome of this species, with more than 30 % unique DNA in some strains and the high level of recombination (Cazalet et al., 2008). However, the use of VNTR polymorphism at the level of both internal sequence variation and repeat copy number made it possible to group isolates within complexes which are highly congruent with SBT clustering results, and provided further insight into the relationships between these groups.

**Importance of the VNTRs for clustering analyses**

Typability of the VNTRs was excellent in this collection of strains with the exception of Lpms19, an intergenic sequence, which was not amplified in a few samples. Considering the degree of variability inside the *L. pneumophila* species, the
presence of selected VNTR loci in nearly all strains (Supplementary Table S2) argues for their importance for the biology of \textit{L. pneumophila}. Lpms33 and Lpms34 are intergenic elements with similar repeated sequence of unknown function. Lpms04 and five other tandem repeats (Lpms01, Lpms03, Lpms13, Lpms17 and Lpms35) are embedded in coding sequences, three of which probably encode a surface protein (Table 1). Lpms35 belongs to the gene encoding the Tfp pilus assembly protein FimV and its variability has been further studied (Coil et al., 2008).

Lpms04 is a very intriguing VNTR: (i) it shows internal sequence variability of the repeat, a feature common with other VNTRs in a variety of species, although at a lower level (here shown also for Lpms01 and Lpms13); (ii) some alleles are made up from a quite different repeat unit. To our knowledge, this is the first time that such a polymorphism in the length and sequence of the repeat has been reported within a single tandem repeat locus in any genome. The 57 bp, 54 bp and 48 bp repeats, within the EU1048 and EU1056 allele, are clearly related, a relatively rare but not exceptional situation observed in tandem repeats from a wide variety of genomes including the human genome (Buard & Vergnaud, 1994). Significant sequence similarity is observed between the 21 bp and the 51 bp repeats (Fig. 3), suggesting the derivation of repeat units one from the other and the stochastic amplification of the new variant. The 51 bp repeats often contain two copies of the 10 bp sequence ACGACGGATA separated by 20 bp. An internal recombination would produce a 21 bp repeat very highly similar to the motif present in strain 02-304 (Fig. 3a). However, there is no obvious link between the 48/54/57 bp repeats and the 21/51 bp repeats; one possibility to explain the existence of such different sequences at a single locus might be gene conversion with DNA from another species. Such genetic exchanges have been proposed at other loci on the basis of genomic sequence (Czalet et al., 2008). In addition several independent observations (MLVA-8 clustering, Lpms01 sequence and restriction site polymorphism, SBT sequence analysis) suggest that the VACC5 cluster with a 54 bp repeat is closest to the VACC9 cluster with a 21 bp repeat.

The ORF in the gene containing Lpms04 is preserved whatever the repeat length (always a multiple of three). The complete Lpms04 gene was sequenced from an isolate with a 51 bp Lpms04 sequence (02-157) and a 21 bp Lpms04 sequence (02-304) and the sequence was compared with that of sequenced genomes, Paris (NC_006368) (48 bp) and Philadelphia-1 (NC_002942 sg1) (48 bp) (Fig. 4a). The deduced proteins are well conserved in the region flanking the repeat. Interestingly, although the repeat regions are very different, the putative proteins clearly show the same hydrophobicity/hydrophilicity profile, suggesting that the function could be maintained despite the different primary structure (Fig. 4b). Prediction of secondary structure using GOR IV confirms that the different amino acids repeats do not modify the structure of the putative proteins (data not shown). A search for homologies with proteins in the NCBI database provided a single significant similarity score with the \textit{Mycoplasma pulmonis} VSAA lipoprotein, a surface protein whose sequence variation provides a mechanism for avoidance of the immune system (Shen et al., 2000).

**Insight into the population structure**

A comparison with the eBURST analysis of Edwards et al. (2008) using SBT data shows that clustering with MLVA correctly aggregates isolates with similar sequence types. However, the SBT CCs defined using our criteria differ slightly from those defined by Edwards et al. (2008) and for this reason we have not used the same nomenclature. Other eBURST analyses using SBT data also identify CCs but there is still no common designation for these (Kozak et al., 2009; Tijet et al., 2010). Inside CCs polymorphism is observed with the most informative VNTRs, although their variability is different in the different CCs: Lpms35 is stable in the VACC2 (Philadelphia-1) lineage but four alleles are found in the VACC1 (Paris) lineage. The mechanism by which the number of repeats increases or decreases is still largely unknown in tandem repeats with large repeat units and therefore the clustering analysis does not take into account the amplitude of the difference between two alleles. Detailed sequence analysis of the repeats suggests that contiguous repeats are duplicated, although mosaicism is sometimes observed, indicating recombination. When looking inside VACCs it appears that changes are often limited to the addition or deletion of one unit, in particular when looking at the most polymorphic VNTRs, Lpms35 (an 18 bp repeat) or Lpms13 (a 24 bp repeat). In VACC3 all the Lpms35 alleles have large repeat numbers (up to 32), suggesting that the repeats are also added or lost gradually and not by multiple repeats duplication or deletion. In \textit{Mycobacterium tuberculosis} it was suggested that some clades show a tendency towards the loss of repeats in modern strains as compared with ancestral clades (Arnold et al., 2006). \textit{L. pneumophila} cluster VACC2, which presents several VNTRs with small numbers of repeats, is reminiscent of this behaviour.

Sequencing of Lpms04 confirms the proximity of isolates inside clusters and provides additional information to the MLVA assay. A single MLVA-8 genotype is observed in VACC4 whereas sequencing of Lpms04 shows the presence of different numbers of a 51 bp repeat. In the two subsp. \textit{fraseri} strains ATCC 33156 (sg4) and ATCC 33216 (sg5) separated by both SBT and MLVA, the 48 bp Lpms4 VNTRs show clear resemblance. Sequencing of the tandem repeats and multiple alignments using TRDB is highly informative for visual assessment of resemblance. In addition the new developed matrix output option provides a more quantitative measure of the similarity of tandem repeat alleles. Therefore sequencing of Lpms04 might represent a very simple first-line genotyping assay for \textit{L. pneumophila} isolates. Such an assay has been developed in \textit{Staphylococcus aureus}, taking advantage of internal...
Fig. 4. Comparison of the putative protein encoded by the ORF bearing Lpms04. (a) Alignment of the protein sequence from four strains ATCC 33152 (Philadelphia-1 sg1), ‘Paris’, 02-157 and 02-304 using MULALBLA. (b) Prediction of hydrophobicity or hydrophilicity scales for the putative proteins using Protscale.
sequence variations within the gene encoding protein A (spa) (Frénay et al., 1996; Grundmann et al., 2010). Only a few tandem repeat loci are sufficiently variable in terms of repeat copy number and internal sequence variation to be of interest in that respect.

Recombination

Our results support the hypothesis that in L. pneumophila genetic exchanges are taking place in a population that is largely clonal. In most of the strains linkage disequilibrium is observed when looking at the tandem repeats alleles, and their variability can be explained by internal changes, either by variation of the number of repeats or by point mutations. Lpms01 and Lpms04 are separated by only 10 kb and likely to be transmitted together, although, interestingly, fewer internal differences were observed in Lpms01 than in Lpms04. However, in several strains we found evidence of genetic exchanges reflected by the presence of VNTRs characteristic of different CCs. The sg12 reference strain 570-CO-H, which belongs to VACC2, possesses an Lpms13 allele similar to that of VACC1 strains. The sg10 reference strain ATCC 43283 and the sg14 reference strain ATCC 43703 possess a 54 bp repeat in Lpms04 (Fig. 1) and their Lpms01 is identical (VACC5 Supplementary Fig. S1) but the sg10 ATCC 43283 Lpms13 sequence is similar to that of the sg7/sg8/sg9 reference strains ATCC 33823, 35096 and 35289 (data not shown). Likewise, in strains 03-011 and 03-004, which were respectively isolated from a water sample and a patient in a hospital during an outbreak (Pourcel et al., 2003), the same Lpms13 is observed, whereas Lpms01 and Lpms04 are similar to those of ATCC 33823 sg7 and ATCC 43703 sg14 reference strains (Supplementary Fig. S8). This suggests that the new arrangement is the result of recombination between two strains.

A recent SBT study analysing six genes in 335 samples described the clonal structure of L. pneumophila and proposed that mutation, rather than recombination, is the predominant mechanism behind genotypic differentiation in this species (Edwards et al., 2008). This was also the conclusion of work by Selander et al. (1985). However, the use of extended sequence-based genotyping with a total of 19 loci including 13 intergenic regions to investigate the variability of 31 environmental isolates confirmed that recombination is a common phenomenon (Coscollà & González-Candelas, 2007, 2009). Thus, recombination may play a role in the diversification of the species and formation of ecologically distinct types (Edwards et al., 2008).

Conclusion

Here, we have demonstrated that the typing of eight VNTRs (MLVA-8) and the sequencing of an additional tandem repeat provides valuable information for epidemiological studies and for identification of clonal complexes in L. pneumophila. In particular we believe that the polymorphism in Lpms04 can be very useful for classifying strains and for gaining insights into the population structure of L. pneumophila, although recombination at this locus might be elevated. A first-line genotyping assay similar to spa typing in S. aureus (Harmsen et al., 2003) could be proposed. MLVA clustering in conjunction with Lpms04 sequencing defines clusters for which no representative genome has been sequenced yet and which might present interesting characteristics. This is the case of VACC5 with a 54 bp Lpms04 VNTR, which contains many environmental strains. Finally, the particularly high degree of polymorphism inside Lpms04 suggests that the protein region it encodes might benefit from the plasticity provided by diversity at this locus. This gene might be an evolutionary hot-spot and the elucidation of its function merits additional studies.

ACKNOWLEDGEMENTS

This work was funded by the CNRS and by the Italian Ministry of Health (grant RFFS-2006-4-336506-P3 to P. V.). G. B. and Y. G. were supported in part by the US National Science Foundation (grant NSF IIS-1017621).

REFERENCES


Edited by: D. W. Ussery