Clonal population structure of *Legionella pneumophila* inferred from allelic profiling

Martin T. Edwards,^{1,2} Norman K. Fry¹ and Timothy G. Harrison¹

Correspondence

Norman K. Fry

Norman.Fry@HPA.org.uk

¹Respiratory and Systemic Infection Laboratory, Health Protection Agency Centre for Infections, London, UK

²Statistics, Modelling and Bioinformatics Department, Health Protection Agency Centre for Infections, London, UK

The population structure of *Legionella pneumophila* was investigated by analysing nucleotide sequences from six loci (*flaA*, *pilE*, *asd*, *mip*, *mompS* and *proA*) of 335 globally distributed isolates from clinical and environmental sources over a 29-year period (1977–2006). Data were obtained from unrelated isolates from Europe ($n=270$), Japan ($n=31$), Canada ($n=7$), the USA ($n=24$) and Australia ($n=1$). The country of origin of two strains was unknown. Analysis of these isolates indicated significant linkage disequilibrium between the six loci. Application of six sequence-based recombination detection tests did not reveal evidence of recombination, but estimates of rates of recombination and mutation made by a seventh test suggested that recombination could have occurred at a rate similar to, but probably lower than, that of mutation. Genealogies inferred under models with and without recombination were congruent with each other, providing no definitive evidence regarding recombination, and were in agreement with sequence clusters identified by graph methods. Further evidence supporting the distinct nature of two of the three subspecies of *L. pneumophila*, subsp. *fraseri* and subsp. *pascuella*, was also found. The ratios of non-synonymous to synonymous nucleotide polymorphisms for each of the allele sets were examined and revealed that the putative virulence loci *mompS* and *pilE* are under diversifying pressure, while the allelic regions of three other loci linked to virulence (*flaA*, *proA* and *mip*) do not appear to be.

Received 8 August 2007

Revised 31 October 2007

Accepted 9 December 2007

INTRODUCTION

Following the recognition of the aetiological agent of Legionnaires' disease in 1977 (McDade *et al.*, 1977), phenotypic and genotypic analyses of *Legionella pneumophila* have mainly focused on the short-term epidemiology of the bacterium. In this scenario, the aim has been to demonstrate potential environmental sources of infection, thereby allowing timely intervention, in order to prevent further infection. Many such methods have been developed and applied with the purpose of discriminating between

epidemiologically unrelated strains of *L. pneumophila*, particularly those belonging to serogroup (sg) 1 (Edelstein *et al.*, 1986; Fry *et al.*, 1999; Saunders *et al.*, 1990; Schoonmaker *et al.*, 1992; van Ketel *et al.*, 1984).

The utility of the multi-locus sequence typing (MLST) approach, in which sequences from five to 10 house-keeping genes are determined, in the identification of major bacterial lineages associated with invasive disease was first described in 1998 (Enright & Spratt, 1998; Maiden *et al.*, 1998). Since then, this robust and portable technique (or variations of it) has been applied to the study of genetic diversity and clonal expansion, and to the long-term epidemiological analysis of microbial populations (see <http://pubmlst.org/> and <http://www.mlst.net/>) (Giske *et al.*, 2006; Paraskevopoulos *et al.*, 2006; Vassileva *et al.*, 2006).

Recently, a similar approach, termed sequence-based typing (SBT), developed by members of the European Working Group for Legionella Infections (EWGLI), was applied to the epidemiological typing of clinical and environmental isolates of *L. pneumophila* (Gaia *et al.*, 2003, 2005). Sequences from defined regions of six genes, *flaA*, *pilE*, *asd*, *mip*, *mompS* and *proA*, which encode the flagellum subunit (Heuner *et al.*, 1995), type IV pilin

Abbreviations: BK, Bron-Kerbosch; EWGLI, European Working Group for Legionella Infections; MLST, multi-locus sequence typing; MST, minimum spanning tree; SBT, sequence-based typing; SLV, single-locus variant.

Three supplementary figures showing likelihood mappings, the polymorphic region of *proA* corresponding to the putative substrate binding site of the enzyme, and the genealogy inferred using ClonalFrame with estimates made of recombination parameters, as well as three supplementary tables showing additional strain information for *L. pneumophila* isolates, the geographical distribution of a profile clustered using the clique identification approach, and a partial distance matrix used to create G_{av} and G_{nv} , are available with the online version of this paper.

(Stone & Abu Kwaik, 1998), aspartate- β -semialdehyde dehydrogenase (Harb & Abu Kwaik, 1998), macrophage infectivity potentiator (Engleberg *et al.*, 1989), a major outer-membrane protein of ~29 kDa (Ehret & Ruckdeschel, 1985; accession no. AF078136, E. Christoph & W. Ehret, unpublished data) and zinc metalloprotease protein (Black *et al.*, 1990), respectively, are used to create allelic profiles in a pre-determined order, e.g. 1,4,3,1,1,1. This technique has been used in national and international outbreak investigation, and has been shown to be highly robust (Amemura-Maekawa *et al.*, 2005; Harrison *et al.*, 2006; Scaturro *et al.*, 2005; Young *et al.*, 2005). As of 24 July 2007, the total number of distinct allelic profiles in the EWGLI SBT database for *L. pneumophila* was 181, and the number of designated allelic variants was as follows: *flaA*, 18; *pilE*, 24; *asd*, 25; *mip*, 30; *mompS*, 35; and *proA*, 26 (authors' unpublished data).

The rationale of the MLST scheme, and the multilocus enzyme electrophoresis (MLEE) approach on which it was based, is to exploit the very slow accumulation of mutations in the population likely to be selectively neutral. Although the number of individual alleles within the population may be relatively small, acceptable levels of discrimination are achieved by combining many loci. The classic MLST schemes use house-keeping genes, which are not considered to be under diversifying pressure. In developing the SBT scheme for *L. pneumophila*, the aim was to maximize the discrimination between strains. A wide range of genes was therefore examined, including some frequently used in the classic MLST schemes, but also those coding for surface-expressed proteins and/or putative virulence genes. In addition to those described above, four other genes likely to be under stabilizing pressure have been considered in previous studies: *acn*, *groES*, *groEL* and *recA* (Gaia *et al.*, 2003). The *acn* gene encodes a major iron-containing protein of *L. pneumophila* which demonstrates aconitase activity (Mengaud & Horwitz, 1993), and *recA* encodes the RecA protein of *L. pneumophila* (Zhao & Dreyfus, 1990). The *groESL* genes (also referred to as *hsp10/60* and *htpA/B*) encode the 10 kDa (GroES) and 58 kDa (GroEL) common antigens in *L. pneumophila* (Hoffman *et al.*, 1990; Sampson *et al.*, 1990). However, all four of these were rejected from inclusion in the SBT typing scheme, as they did not offer sufficient discrimination.

The *L. pneumophila* SBT scheme formally described by Gaia *et al.* (2005) contains five genes which were considered as likely to be under diversifying pressure (*flaA*, *pilE*, *mip*, *mompS*, *proA*) and one considered as under stabilizing pressure (*asd*). The *mip* gene encodes an immunophilin of the FK506 binding protein (FKBP) class, and although its precise role is not fully defined, Mip is an outer-membrane protein important in the intracellular cycle of *Legionella* (Cianciotto & Fields, 1992).

The aims of this study were to determine the population structure of *L. pneumophila* through evidence of linkage disequilibrium (LD), sequence clustering, and any

recombination; and to clarify whether the assumptions on positive selection for some of the loci are justified given the data.

METHODS

Strains. Data were obtained from *L. pneumophila* isolates from the following sources: (i) the EU *Legionella* culture collection (Fry *et al.*, 1999, 2000, 2002; Gaia *et al.*, 2003, 2005); (ii) the authors' strain collection following investigation of legionellosis (community-acquired, nosocomial or travel-associated) in the UK from 1977 to 2006; (iii) the National Collection of Type Cultures, Health Protection Agency Centre for Infections, London, UK; or from other published data (Amemura-Maekawa *et al.*, 2005; Benson *et al.*, 2006; Bernander *et al.*, 2006; Coscollá & González-Candelas, 2007; Fendukly *et al.*, 2007; Gilmour *et al.*, 2007; Ratzow *et al.*, 2007; Wong *et al.*, 2006). Only data from isolates considered epidemiologically unrelated to any other, or those subsequently shown to be unrelated to the index case isolate in each cluster, were included. This resulted in a dataset consisting of 335 *L. pneumophila* isolates from clinical and environmental sources, including 280 serogroup 1, 45 from other serogroups, and 10 whose serogroup was not known (Table 1), from Europe ($n=270$), Japan ($n=31$), Canada ($n=7$), the USA ($n=24$) and Australia ($n=1$). The country of origin of two strains was unknown.

Allelic profiling. The six gene fragments *flaA*, *pilE*, *asd*, *mip*, *mompS* and *proA* were amplified and sequenced as described previously (Gaia *et al.*, 2005). Designation of alleles was according to the EWGLI SBT database (<http://www.hpa.org.uk/cfi/bioinformatics/dbases.htm#EWGLI>), and the combination of the alleles observed is represented as an ordered numerical vector. Strain data are summarized in Table 1 and further information is provided in Supplementary Table S1.

DNA sequence analysis. Linkage disequilibrium was evaluated using the standardized index of association (I_A^S) method (Haubold & Hudson, 2000) implemented in the START2 package (Jolley *et al.*, 2001). The I_A^S is calculated from the ratio of the variance of the observed mismatches in the test set (V_D) to the variance expected for a state of linkage equilibrium (V_e), scaled by the number of loci used in the analysis. The significance of the difference between V_D and V_e is estimated by resampling the input dataset without replacement (100 000 times in this study) and calculating the values of V_D for each replicate of resampled data. The frequency with which V_D exceeds V_e can be used to provide an estimate as to the significance of the difference. Further evidence of recombination was sought using six additional recombination detection tests, implemented under the RDP package version 2.08: RDP (Martin & Rybicki, 2000) detects recombined fragments by statistical analysis of the composition of aligned sequence triplets; GENECONV (Padidam *et al.*, 1999) seeks aligned segments for which a pair of sequences are sufficiently similar to be suggestive of recombination; Bootscan (Salminen *et al.*, 1995) detects recombination breakpoint using bootstrapping of phylogenetic trees generated from sequence fragments; MaxChi (Smith, 1992) compares the composition of windows within pairs of aligned sequences in order to identify recombination breakpoints; Chimaera (Posada & Crandall, 2001) is a variant of the MaxChi method; and SiScan (Gibbs *et al.*, 2000) is a method for measuring variations in the relatedness of aligned nucleotide sequences; the significance of detected variations is tested in the program using Monte Carlo randomization procedures.

The non-synonymous to synonymous nucleotide substitution ratio (d_N/d_S) of each SBT locus fragment was estimated using the modified Nei and Gojobori method (Nei & Gojobori, 1986) using START2; estimates of positive/purifying selection for individual amino acid residues at each locus were carried out using Selecton (Stern *et al.*,

2007) with default settings; the statistical significance of instances of positive selection was determined using a likelihood ratio test against results from a null model without positive selection. Rates of recombination and mutation were estimated using the LDhat package (McVean *et al.*, 2002).

Phylogenetic analysis. Genealogies were created using ClonalFrame (Didelot & Falush, 2007). Two trees were created using all 127 unique allelic profiles, one estimating parameters of recombination and the other with recombination parameters fixed at zero. Simulations were run for 200 000 iterations, with 25 % discarded as burn-in. Three runs were performed for each model, with convergence tested using the ClonalFrame package, and a 50 % majority rule tree created for each.

To examine the phylogenetic resolution within the data from each of the loci, sequences were assessed using likelihood mapping (Strimmer & von Haeseler, 1997) under TREE-PUZZLE (Schmidt *et al.*, 2002). An explanation of this and presentation of data are shown in Supplementary Fig. S1.

Graph analyses. Minimum spanning trees (MSTs) were created using Kruskal's algorithm (Kruskal, 1956) applied to graphs $G=(V,E)$, where $V=\{\text{allelic profiles}\}$ and $E=\{\text{relationships between vertices having an edit distance less than some threshold}\}$; for the purposes of this study an edit is defined as a non-identity in an alignment of the two sequences or allelic profile vector. For analyses using the number of loci differences as the edit distance, an edge $e(u,v) \in E$ is created between vertices u and v if the number of locus variations is less than four; that is, only single, double and triple locus variants are connected. The graph using this criterion was designated G_{av} (allele variants). Where each nucleotide polymorphism between two alleles is counted as an edit event, the threshold for edge creation was established at seven polymorphisms across all six loci. This threshold was determined by establishing the left-hand 2.5 percentile of the distribution of polymorphisms between all pairs of profiles; this graph is designated G_{nv} (nucleotide variants). Each $e(u,v) \in E$ was weighted with the appropriate edit distance (allele variants or summed nucleotide polymorphisms). Clusters of related profiles were identified using the Bron-Kerbosch (BK) algorithm (Bron & Kerbosch, 1973) for finding the set of complete subgraphs (cliques) within the profile graph. In brief, a cluster of profiles is reported if each profile in the cluster is connected by an edge to every other profile. For the purposes of this analysis, in cases where two cliques share one or more edges, these are merged into a single cluster. Graph analyses were implemented in Java using the JUNG 1.7.6 graph library [<http://jung.sourceforge.net/>].

RESULTS AND DISCUSSION

Properties of the *L. pneumophila* unrelated dataset

Table 2 summarizes the properties of the allele collections used in this study. The variability over each of the loci was broadly in line with the values determined by Gaia *et al.* (2005), with inconsistencies between the values being explained by differences in the number of alleles and the resulting changes in the number of polymorphic sites. Relationships, established with graph algorithms and the BURST algorithm, between some of the allelic profiles of this dataset are shown diagrammatically in Fig. 1; profiles that did not belong to any minimum spanning tree (did not have any single-, double- or triple-locus variants) have been excluded from the figure. It is noteworthy that two

strains known to belong to *L. pneumophila* subspecies *fraseri* (Los Angeles-1 and Dallas 1E) (Brenner *et al.*, 1988) have very distinct allelic profiles (index nos 15 and 16; 11,14,16,18,15,13 and 11,14,16,25,7,13, respectively), and together with two other variants form their own cluster. Of the three strains tested that belong to *L. pneumophila* subspecies *pascallei*, all share an identical profile (index no. 26; 14,18,8,18,28,19) and have no single-, double- or triple-locus variants within the dataset, and hence are not represented in Fig. 1.

Allele frequency and nucleotide sequence-based techniques support a clonal population for *L. pneumophila*

In this study, two approaches have broadly been used to assess the extent of recombination and the clonal nature of *L. pneumophila*. The linkage disequilibrium analysis compared the variance of the pair-wise differences between the allelic profiles in the dataset to the variance of a null distribution assuming linkage equilibrium, and sequence data were exploited using the recombination detection package RDP.

The standardized index of association I_A^S (Haubold & Hudson, 2000; Ruiz-Garbajosa *et al.*, 2006) is a commonly used measure of linkage disequilibrium in multi-locus datasets. The I_A^S is zero when complete linkage equilibrium is observed for a collection of loci. In these cases, the observed frequency of allele combinations approaches the expected frequency given the individual allele frequencies over the dataset; that is, the occurrences of the alleles are independent of each other. Significant ($\alpha=1 \times 10^{-5}$) linkage disequilibrium was detected in the dataset of 335 isolates, with an I_A^S of 0.4949. It has been observed that in most populations, recombination must occur at least 20 times more frequently than mutation in order for a hypothesis of linkage equilibrium to be accepted (Hudson, 1994; Smith *et al.*, 1993). Since this is rather a broad range, the ratio of recombination rate to mutation rate (ρ/θ) was estimated using the LDhat package in an attempt to quantify more precisely the extent of recombination, if any, over the SBT loci (Table 3). The estimates of ρ implied that recombination had occurred at a low rate for some of the loci, although the point estimates of ρ were usually lower than the estimates of θ . To clarify this position, evidence for distinct recombination events was sought using the set of six methods implemented in RDP v2.08 (see Methods). Only one significant ($\alpha=0.01$) instance of recombination was predicted: the SiScan method predicted an event between *pilE*(3) and *pilE*(18), which have a total of 22 nucleotide differences between the alleles. However, since no other method identified this or any other recombination event, this seems most likely to be an artefact of the method.

These results provide support for the hypothesis that the loci used in the SBT scheme are in a state of linkage disequilibrium, and that mutation rather than recombination is

Table 1. Summary of allelic profiles of the 335 unrelated *L. pneumophila* strains showing serogroup and frequency

NK, Not known.

Index	Profile	Frequency	Serogroup
1	1,10,1,10,14,9	1	1
2	1,10,19,1,9,4	1	1
3	1,10,3,1,1,1	1	1
4	1,4,1,1,14,9	1	1
5	1,4,3,1,1,1	67	1,8,9,NK
6	1,4,3,1,14,1	1	1
7	1,4,3,19,1,1	2	4/5*
8	1,4,3,28,1,1	1	9
9	1,4,3,4,1,1	3	1
10	1,6,3,1,1,1	2	1
11	1,6,3,1,14,1	1	1
12	10,12,7,3,16,18	2	1
13	11,14,16,1,15,13	1	1
14	11,14,16,16,15,13	2	1
15	11,14,16,18,15,13	1	5
16	11,14,16,25,7,13	2	2/10/13*,4
17	11,4,1,1,14,9	1	1
18	12,15,3,10,18,1	1	1
19	12,17,11,10,5,12	1	11
20	12,8,11,20,5,12	1	1
21	12,8,11,5,20,12	1	1
22	12,9,2,5,27,20	1	NK
23	12,9,2,5,3,17	1	1
24	12,9,26,5,26,17	2	1
25	13,11,4,9,23,16	1	8
26	14,18,8,18,28,19	2	NK
27	2,10,15,5,9,4	1	1
28	2,10,18,10,2,1	1	1
29	2,10,3,10,9,4	2	1
30	2,10,3,12,17,14	1	1
31	2,10,3,28,21,4	1	3
32	2,10,3,28,9,14	1	10
33	2,10,3,28,9,4	3	3,10
34	2,10,9,13,2,1	1	1
35	2,10,9,13,2,5	1	1
36	2,12,3,6,8,14	1	1
37	2,19,5,10,18,1	1	NK
38	2,2,18,15,2,1	1	1
39	2,3,18,10,2,1	1	1
40	2,3,18,11,2,10	1	1
41	2,3,18,13,25,5	1	1
42	2,3,18,15,2,1	5	1
43	2,3,18,15,4,1	1	1
44	2,3,3,15,2,6	1	1
45	2,3,6,10,2,1	2	1
46	2,3,6,25,2,1	1	1
47	2,3,9,10,1,1	1	1
48	2,3,9,10,2,1	10	1
49	2,4,3,1,21,4	1	1
50	2,4,3,10,9,4	1	1
51	2,6,17,14,12,8	1	7
52	2,6,17,15,12,8	1	1
53	2,6,17,6,13,3	1	1
54	2,6,17,6,3,13	1	1

Index	Profile	Frequency	Serogroup
55	2,6,21,12,12,8	1	1
56	3,10,1,10,14,9	2	1
57	3,10,1,12,14,9	2	1
58	3,10,1,14,14,9	1	1
59	3,10,1,28,14,9	3	6,8,12
60	3,10,1,3,14,9	7	1,6
61	3,10,1,5,14,9	1	1
62	3,10,15,3,21,7	1	1
63	3,10,3,1,14,9	1	1
64	3,13,1,10,14,9	1	1
65	3,13,1,17,14,9	1	16
66	3,13,1,25,14,9	1	1
67	3,13,1,28,14,9	7	6,12
68	3,4,1,1,1,9	3	1
69	3,4,1,1,14,9	14	1
70	3,4,1,14,14,9	1	1
71	3,5,1,7,14,9	1	2
72	3,6,1,14,14,9	2	1
73	4,6,11,3,11,12	1	1
74	4,7,11,3,11,12	14	1
75	4,7,9,3,11,12	1	1
76	4,8,11,10,10,12	2	1
77	4,8,11,25,29,12	1	NK
78	5,1,22,10,6,10	1	1
79	5,1,22,15,6,10	1	1
80	5,1,22,26,6,10	6	1
81	5,1,22,30,6,10	11	4/5/8*,2-14†,8,14
82	5,1,22,5,6,10	3	1
83	5,10,22,15,6,2	20	1
84	5,2,22,10,6,25	1	9
85	5,2,22,15,6,10	1	1
86	5,2,22,27,6,10	6	1
87	6,10,14,10,2,1	1	1
88	6,10,14,10,2,3	2	1
89	6,10,15,10,9,14	1	1
90	6,10,15,12,13,14	1	1
91	6,10,15,13,21,14	3	1
92	6,10,15,13,4,14	1	1
93	6,10,15,24,17,14	1	1
94	6,10,15,28,21,3	1	NK
95	6,10,15,28,4,4	1	1
96	6,10,15,28,9,14	3	1
97	6,10,15,3,21,14	1	1
98	6,10,15,3,21,4	1	2-14†
99	6,10,17,3,4,14	3	1,6
100	6,10,17,6,9,4	2	1
101	6,10,19,28,19,4	2	1,6
102	6,10,19,3,19,4	2	1
103	6,10,2,28,9,4	1	1
104	6,10,21,12,9,4	2	1
105	6,10,21,6,9,4	1	1
106	6,10,22,3,21,3	1	6
107	6,10,23,12,18,4	1	1
108	6,10,23,3,33,4	1	1
109	6,10,3,3,21,7	1	1
110	6,10,5,10,9,1	1	13
111	6,16,14,3,21,14	1	1
112	6,4,14,12,2,3	1	1

Table 1. cont.

Index	Profile	Frequency	Serogroup
113	6,4,3,1,1,1	1	1
114	6,6,15,28,4,14	2	1
115	7,10,1,3,14,9	1	1
116	7,10,17,10,5,4	1	1
117	7,6,17,20,13,11	1	3
118	7,6,17,3,11,11	1	NK
119	7,6,17,3,13,11	9	1,6
120	7,6,17,3,14,11	1	1
121	7,6,17,3,24,11	1	1
122	7,6,3,20,13,11	1	10
123	7,6,3,8,13,11	1	5
124	8,1,22,30,6,10	1	8
125	8,10,3,15,18,1	10	1
126	8,10,3,15,6,1	1	1
127	9,6,3,10,22,15	1	1

*Denotes cross-reaction with these serogroups.

†Denotes that isolate gave a positive reaction with a test reagent for these serogroups.

the predominant mechanism behind genotypic differentiation in *L. pneumophila*. This view supports that of earlier work by Selander *et al.* (1985) on electrophoretic types, in which it was concluded that *Legionella* has a globally distributed clonal population, and by the more recent work of Coscollá *et al.* (2006), in which no evidence of recombination was found in a collection of 25 environmental isolates from Spain using three SBT loci (*flaA*, *proA* and *mompS*), and extended regions around these loci. A later study of 31 isolates (the same set as before but with six additional ones) (Coscollá & González-Candelas, 2007) found contradictory evidence of recombination at 19 loci of *L. pneumophila* (the six SBT loci and 13 intergenic regions found to be conserved across the then three available *L. pneumophila* genomes), since one method found no evidence of recombination, while another method did find such evidence.

Allelic profile partition by graph clustering techniques

Graph algorithmic techniques applied to the graphs G_{av} and G_{nv} (see Methods; Supplementary Table S3) can efficiently delineate sequence clusters, which could equate to clonal complexes when a reasonable sequence identity cut-off is applied. Two graph algorithms were evaluated: minimum spanning trees (MSTs); and clique identification using the BK algorithm. The resulting partitions are compared to results obtained using the eBURST application (Feil *et al.*, 2004), since this is considered the standard approach. The BURST algorithm expands sequence clusters by adding the nearest profile (where distance is the number of variant loci between one profile within the cluster and a candidate profile outside the cluster), and this is equivalent to nearest-neighbour (single linkage) clustering. Only single-locus variants (SLVs) are eligible to be merged, and consequently the subtrees of the minimum spanning trees found in G_{av} that are connected by SLV edges are identical to the groups found in the data by the eBURST application. These MSTs/BURST groups can be seen in Fig. 1; since the MSTs and BURST results are identical, MSTs are not discussed further. In contrast to BURST, the clique-finding approach is equivalent to complete linkage clustering, where a profile is added to a cluster only if it is below some maximum distance to every other member of the sequence cluster. The sequence clusters in this work are not strictly cliques, since they are composed of true cliques merged together on the basis of one or more shared edges. To acknowledge this distinction, sequence clusters identified by this method will be referred to as BK clusters; see Methods and Fig. 2 for a detailed example. Using a complete linkage methodology avoids the generation of clusters that are composed of chains of outliers where the mean pair-wise distance between all members of the cluster might actually be quite large.

In Fig. 1, BURST groups are identified as connected components with only solid black edges connecting profiles; they are minimum spanning trees on the graph of SLVs. One large BURST group {1,4,17,56,57,58,

Table 2. Summary of locus variability

Locus	Polymorphisms*	Alleles*	Length (bp)	Variability*†	$k‡$	$d_N/d_S§$
<i>flaA</i>	32 (23)	14 (11)	182	17.6 % (12.6 %)	0.045	0.047
<i>pilE</i>	55 (46)	19 (10)	333	16.5 % (13.8 %)	0.050	0.165
<i>asd</i>	51 (35)	20 (13)	473	10.8 % (7.4 %)	0.022	0.048
<i>mip</i>	34 (12)	25 (13)	402	8.5 % (3.0 %)	0.015	0.061
<i>mompS</i>	56 (43)	30 (16)	352	16.0 % (12.2 %)	0.036	0.202
<i>proA</i>	38 (32)	21 (15)	405	9.4 % (7.9 %)	0.025	0.024

*Numbers in parentheses are from Gaia *et al.* (2005).

†The proportion of the fragment length represented by polymorphic positions.

‡The average number of pair-wise differences between alleles of that locus, normalized by fragment length.

§Estimates of nucleotide substitution rates; ratio of the rate of non-synonymous (d_N) to synonymous (d_S) mutations.

Fig. 1. Minimum spanning tree of G_{av} ; relationships between allelic profiles of *L. pneumophila*. Solid lines indicate that the connected vertices are SLVs; dashed lines indicate double-locus variants; dotted lines, triple-locus variants. All vertices connected by solid lines were reported by eBURST as clonal complexes. Vertices enclosed by dashed lines are clusters predicted using the BK algorithm applied to G_{nv} . Disconnected profiles (not single-, double- or triple-locus variants of any profile in the dataset) are not presented. Cluster 7 is split because of constraints on representing the graph without crossing edges. See Table 1 for profile index.

59,60,61,64,65,66,68,69,70,72, and 115} is split into three BK clusters {6, 9, and 10}, and examination of Fig. 3 supports this partition. The figure shows a distance matrix for the majority of the members of the BURST group (profiles 1,4,17 and 115 were removed for clarity) represented as a contour map. The contour map is symmetrical along its diagonal (bottom-left to top-right) axis; the distances in nucleotide differences summed over all six loci between pairs of profiles are shown in shades of grey. When one considers that this matrix covers a single BURST group, it is clear that there is finer structure that is lost when nucleotide edit distance is ignored. The clearly

delineated darker islands in Fig. 3 correspond with the BK clusters 6,9 and 10; these areas represent collections of profiles in which all pair-wise distances between members of the cluster are small. Around these islands is a landscape of low sequence identity which undermines the definition of these allelic profiles as a single sequence cluster.

Mapping sequence clusters to phenotypic characteristics of *L. pneumophila* is a difficult process. The organism is an opportunistic pathogen of humans, and disease is always associated with an environmental source. Since infection of a human host is a terminal process, in the sense that both successful and unsuccessful infection results in the death of the pathogen, there is no mechanism by which the outcome of infection exists as a selective pressure. Selection resulting in discrete sequence clusters is therefore likely to be the result of complex factors in the natural environment, in either a biofilm community or a host amoeba. However, it is clear that sequence clusters do exist in these data, and that the choice of method used to delineate these clusters critically impacts their ability to represent the population. Defining sequence clusters as cliques on the graph of allelic profiles may not be applicable to all datasets, but in the case of these data, this approach better represents relationships within the data.

The BK clusters were used to investigate whether extant *L. pneumophila* exist in geographically distinct families. The 335 allelic profiles used in this study were gathered from globally distributed sources, although there is an obvious

Table 3. Estimates of the ratio of the rate of recombination (ρ) to the rate of mutation (θ)

ρ_{upper} and ρ_{lower} refer to the upper and lower bounds of the 95 % confidence interval, respectively. Upper and lower bounds of θ are not available.

Locus	ρ	θ	ρ_{lower}	ρ_{upper}	ρ/θ	ρ_{lower}/θ	ρ_{upper}/θ
<i>flaA</i>	6.95	9.43	5.18	9.03	0.74	0.55	0.96
<i>pilE</i>	10.45	13.45	6.39	15.53	0.78	0.48	1.15
<i>asd</i>	19.1	13.81	15.53	23.23	1.38	1.12	1.68
<i>mip</i>	11.86	8.83	8.11	16.37	1.34	0.92	1.85
<i>mompS</i>	10.7	12.62	6.8	14.83	0.85	0.54	1.18
<i>proA</i>	5.2	10.28	4.61	5.84	0.51	0.45	0.57

Fig. 2. Cluster 4 (see Fig. 1) decomposed into the individual cliques identified by the BK algorithm. Vertices (circles containing a profile index) are connected by edges (bi-directional arrows) of varying thickness (indicating number of variant loci) labelled with the total number of nucleotide differences between the connect profiles (digits in boxes). The complete cluster is formed through merger of the cliques (duplicate vertices or parallel edges are not permitted).

excess of European samples. Of the 12 BK clusters predicted, none was composed of profiles from a single country (see Supplementary Table S2). Only four BURST

groups were not sub- or supersets of BK clusters: {13,14,15}; {73,74,75}; {117,118,119,120,121,122,123}; and {125,126}. None of these BURST groups contained profiles

Fig. 3. Contour map of a (symmetrical) pair-wise distance matrix. Distances are number of mismatched positions in a pair-wise alignment of concatenated sequences of the two profiles. The string of profiles presented here forms the greater part of a single BURST group (profile 56–72); the dark islands represent pair combinations with a low number of mismatches within this group. The BK algorithm finds three clusters within the group; in Fig. 1 these clusters are 9 (squares), 10 (stars) and 6 (triangles). Each point (square, star or triangle) represents high similarity between two profiles of a cluster.

which originated from a single country. The BURST group containing the *L. pneumophila* subsp. *fraseri* samples {13,14,15} was distributed over Italy, England, US Virgin Islands, Spain and the USA. Therefore, it appears that this subspecies is also widely distributed. The allelic profiles examined in this work include the data used in a recent study on *L. pneumophila* strains of Japanese origin (Amemura-Maekawa *et al.*, 2005). In that study, only four of 15 unique profiles {5, 12, 20, 36, 40, 48, 64, 74, 92, 95, 100, 102, 111, 114, 120} were also found in Europe; Fig. 1 and Supplementary Table S2 show that an additional five of these profiles belong to BK clusters that contain strains from a wide variety of locations outside Japan, and therefore these strains are also likely to occur outside Japan.

The BK clusters support the conclusion of a clonal population for *L. pneumophila* by showing that within this sizeable sample, allelic profiles exist as populations of strains differing between each other by a few polymorphisms that represent genetic drift. Fig. 2 is a detailed exploration of one BK cluster, revealing that profile 48 (2,3,9,10,2,1) is a likely progenitor of this cluster, since it is only a single nucleotide (bold type) different from 39 (2,3,18,10,2,1), 47 (2,3,9,10,1,1) and 45 (2,3,6,10,2,1); it is worthy of note that BURST predicted the founder of this cluster as profile 39. This cluster is also an excellent example of clonal expansion, with clear drift by accumulation of single polymorphisms at different loci.

Genealogies inferred from the data corroborate sequence clusters

A comparison was made of genealogies created by ClonalFrame both with and without explicit modelling of recombination. The clades assigned by these processes are very similar, and the profiles that form the BK clusters are predominantly monophyletic in both genealogies. A feature common to both trees is the early separation of the *fraseri* and *pascuelli* subspecies clusters from the remainder of the strains. Fig. 4 shows the genealogy of the profiles inferred with recombination parameters fixed at zero, and the only significant difference between this and the with-recombination genealogy is that branch lengths are longer (except at the leaves), reflecting the amount of time required to acquire the individual mutations compared to atomic recombination events (see Supplementary Fig. S3 for a tree modelled with recombination). Virtually every branch point of the with-recombination tree indicated a recombination event, but since the genealogies were congruent, it cannot be said that either model (with or without recombination) is the correct one.

The data suggest that clonal complexes (sequence clusters) exist in *L. pneumophila*, and that some contribution by recombination in the formation of these complexes cannot be ruled out. Clonal complexes have then expanded through the accumulation of mutations. Presumably, if

Fig. 4. Coalescent inferred using ClonalFrame with recombination parameters fixed at zero. Numbers at leaves are the profiles index (Table 1) and interior nodes are labelled with the BK cluster number (Fig. 2), where appropriate.

recombination had played a part in the formation of the individual populations, then this would equally be observed within those populations. Since linkage disequilibrium is observed, this suggests either that recombination does not occur or that periodic selection has at some point minimized diversity within populations following subsequent recombination. This position is reinforced by the low number of polymorphisms seen between strains of a sequence cluster. In addition to low intra-complex recombination, a low rate of inter-complex recombination might be explained by the clones being separated soon after their inception by some physical barrier, either on a geographical scale or due to the clones occupying distinct

ecological niches. Since the arguments above suggest that profiles are currently globally distributed, it could be concluded that if geographical separation was the barrier to recombination after formation of the clonal complexes, then this barrier has recently been overcome: the 'geotype + Boeing' model (Cohan & Perry, 2007). The geotype + Boeing model seems unlikely in this case, since the effective global dissemination of *L. pneumophila* strains would have happened too recently for the observed genetic drift to become evident. In addition to this empirical evidence, it is difficult to hypothesize how the advent of extensive international travel could efficiently seed an environmental organism, which does not transmit between humans, to the extent that geographical separation is no longer apparent. It seems more appropriate then to view the sequence clusters as representing clonal complexes adapted to specific ecologies. The presence of ecologically distinct complexes complements the conclusions of Cohan *et al.* (2006), in which analysis of a large collection of *mip* alleles from several *Legionella* species suggested 11 ecologically diverse clusters (stable ecotypes) of *L. pneumophila* within that dataset.

The likelihood then is that stable populations have been established, possibly through recombination, and following adaptation have been maintained in genetic isolation due to constraints on passage between ecological niches. Strain (1,4,3,1,1,1) was suggested by Amemura-Maekawa *et al.* (2005) to be specific to cooling towers (in Japan, at least), and this could be one such stable ecology, following the opportunistic colonization of an artificial habitat that mimics a natural environment either in physical properties or, more likely, in the composition of the greater microbial community in these towers. That is, the characteristics of the biofilm formed in cooling towers may be critical to the colonization of this environment by this strain, due either to environmental characteristics such as temperature and humidity, or to the ease of seeding of these environments with early colonizers or amoebal hosts.

This study has found no irrefutable data to confirm recombination in *L. pneumophila*. However, *L. pneumophila* is naturally competent (Stone & Abu Kwaik, 1999) and evidence of recombination at the *dotA* locus has been presented in earlier studies (Bumbaugh *et al.*, 2002; Ko *et al.*, 2003). The *dotA* gene was investigated as a potential locus during the development of the EWGLI SBT scheme, but was rejected on the basis of variability at the primer binding sites and the instability of the locus in terms of indels in the coding sequence (CDS) (authors' unpublished data; Bumbaugh *et al.*, 2002). Coscollá & González-Candelas (2007) did not find evidence of intragenic recombination in the loci (including the SBT loci) of their small environmental study, but confidently asserted that recombination was found between intergenic regions. Depending on the presence of regulatory sequence in these intergenic regions, this might represent a mechanism of adapting transcriptional activity and establishing functional differences between strains.

Evaluation of the effects of selective processes on the evolution of the SBT loci

In contrast to the conventional MLST schemes, which are composed exclusively of house-keeping genes considered to be under stabilizing pressure (see Introduction), the loci chosen for the *L. pneumophila* SBT scheme contain a majority of loci (*flaA*, *pilE*, *mompS* and *proA*) assumed to be under diversifying pressure. These loci were expected to display a greater degree of nucleotide polymorphism and so provide for greater discrimination. For alleles of loci under stabilizing selection, substitutions resulting in a non-synonymous codon are less likely to confer a selective advantage and therefore not to persist in the population; this would result in lower d_N/d_S ratios. The results of these analyses (Table 2) do not fully reflect the assumption that several of the SBT loci are under positive selection. As expected, however, the surface membrane components *mompS* (which is the most diverse locus) and *pilE* have the largest d_N/d_S ratios, and the Selecton method indicates statistically significant evidence of positive selection in *mompS* ($\alpha=0.05$) and *pilE* ($\alpha=0.01$). The remaining loci were not indicated as having residues under significant positive selection by this method. The two other loci expected to exhibit evidence of positive selection, *proA* and *flaA*, displayed unexpected characteristics: the d_N/d_S ratio of *proA* was much smaller than for any other loci (0.024, compared to the next smallest 0.047 for *flaA*); and *flaA* appears as stable as the house-keeping gene *asd*. The d_N/d_S value calculated for the *proA* fragment contradicts the assumption that this locus, a secreted virulence factor, is under positive selective pressure, and suggested that non-synonymous substitutions predominate in the population. Given this apparent disparity from the rest of the d_N/d_S results, a more detailed examination was undertaken. The mature protein product of *proA* is a secreted metalloprotease which has been shown to have broad specificity and cytotoxic activity (Black *et al.*, 1990; Moffat *et al.*, 1994). In the SBT scheme, the fragment of *proA* that is used to define the allele (reference sequence M31884; positions 1134–1538) corresponds approximately to amino acids 139–276 of the mature protein. Examination of the allelic variation of the *proA* fragment revealed that, of the 38 polymorphic sites, 10 are conspicuously clustered in a 28 nt region corresponding to amino acids 167–176 (see Supplementary Fig. S2). Only one substitution within this region is non-synonymous, and this results in the highly conservative substitution of valine by isoleucine. According to GenBank Entrez annotation CAH1473, the SBT region lies within a predicted thermolysin metalloprotease α -helical domain (Conserved Domain Database designation Peptidase_M4_C; Marchler-Bauer *et al.*, 2005). Fortuitously, this domain is represented in the Protein Data Bank (Berman *et al.*, 2000) by the crystal structure of a *Bacillus cereus* protein (1NPC). Examination of this structure revealed that the nine amino acids of the highly polymorphic but stable region (amino acid residues 167–176) fold to an α -helix in the mature protein. This helix forms the floor of the substrate binding

site of the domain, and along its length two conserved histidine residues coordinate a zinc ion that is known to be essential for activity (Dreyfus & Iglewski, 1986). The apparent stability of the *proA* fragment might be understood in the light of the observation that non-synonymous mutations in this region could eliminate the ability of the helix to bind the zinc ion, or disrupt the α -helical structure and, therefore, the substrate binding specificity or activity of the enzyme. This might explain why the substitutions are silent, but does not explain the high frequency of the substitutions in this region, as compared to the rest of the sequence. Interestingly, this pattern of clustered polymorphisms has been observed in a homologous domain (from the autolysin family), and this was attributed to local, small-scale recombination with a highly similar bacteriophage protein (Whatmore & Dowson, 1999). This suggests the possibility of a phage role in the evolution of this virulence factor. These data show that structural constraints may have impacted the local d_N/d_S ratio of the *proA* fragment, and, since only fragments of the SBT loci are available for analysis, a value presented in Table 2 may not be representative of an entire locus. Furthermore, the assignment of a single d_N/d_S ratio to an entire structure, rather than to distinct functional domains which may be composed of sequences that are not contiguous in the primary sequence, may be inappropriate, since each domain may be under different selective pressure. For example, the specificity or activity of an enzyme may usefully be optimized to an environment by positive selection, while the requirement for a particular co-factor might remain constant, resulting in varying rates of accumulation of synonymous and non-synonymous substitutions.

The positive selective pressure observed at the *pilE* and *mompS* loci (indicated by the Selecton results and d_N/d_S ratios) is not seen at the locus of the other outer-membrane protein in the SBT set, *flaA*. This may be due to the location of the fragment analysed in the structure of the mature protein, or could be a more general reflection of necessary conservation of function. The *flaA* product is necessary for lysis of macrophages (Molofsky *et al.*, 2005) following the replicative phase and preceding the infective phase of *Legionella*. Although this does not benefit the microbe after infection of a human host, since human to human transmission has not been observed, the process is critical to the life cycle of *Legionella* in its amoebal host, and this may provide the reproductive benefit necessary for the observed sequence conservation. However, since both the length-normalized average nucleotide differences between alleles, and the variability values (see Table 2) for *flaA* are similar to those of *pilE* and *mompS*, it seems likely that the observations on selective pressure at this locus are due to some structural constraint, as seen in the *proA* fragment.

The *mip* gene product plays a critical role in the pathogenicity of *L. pneumophila*, but its exact role is unclear, as the substrate of the Mip protein peptidyl-prolyl *cis/trans* isomerase activity (Fischer *et al.*, 1992) has yet to be determined. Viewing Mip simply as an enzyme that

must be constrained in order to maintain correct function might explain the similar d_N/d_S ratio of *mip* (0.061) to that of the house-keeping *asd* locus (0.048). However, as noted by Bumbaugh *et al.* (2002), the *mip* locus displays considerable diversity over *Legionella* species and therefore the *mip* product retains an appropriate function over a diverse range of sequences. Although the Selecton method did detect possible instances of positive selection at five of the 132 amino acid residues of the *mip* fragment, these were shown not to be significant by the likelihood ratio test.

In conclusion, significant linkage disequilibrium between the six SBT loci was found, with contradictory results making definitive statements difficult with respect to the history of recombination in this organism. The allelic profiling analyses provide evidence of a clonal nature for *L. pneumophila*, though genealogies inferred from the sequence data suggest that clones may initially have formed through recombination events. The same partitions are also predicted when mutation alone is modelled, though at a slower rate. Estimates of ratios of recombination to mutation rates were low, though not zero, indicating a greater role for mutation overall. None of the methods of recombination detection implemented under the RDP package indicated recombination, and coalescent genealogies inferred using models both with and without recombination are congruent with regard to sequence clusters delineated using graphical techniques. Therefore, it can be said with some confidence that these data indicate that frequent recombination does not play a major role in diversification of this species, though it may have had a role in the formation of ecologically distinct types in the past. Systematic detailed sampling of defined environments would contribute greatly to answering questions on whether *L. pneumophila* exists in distinct ecologies, and to what extent and by what mechanism sequence variation occurs within any such communities. With a considered sampling scheme it might then also be possible to correlate genotypes with ecological phenotypes. Regarding selective pressures acting on the SBT loci, two of the putative virulence loci (outer-membrane proteins *mompS* and *pilE*) were found to be under diversifying pressure, while three other loci linked to virulence (*flaA*, *proA* and *mip*) did not reveal evidence of diversifying selection in the regions under examination.

Recently, an additional gene, *neuA*, encoding *N*-acylneuraminate cytidyl transferase, was proposed (Ratzow *et al.*, 2007) as an addition to the Gaia *et al.* (2005) scheme on the basis that it offers increased discrimination of serogroup 1 strains. Further studies are planned to assess the contribution of this new target in describing the population of this fascinating species.

ACKNOWLEDGEMENTS

We thank Anthony Underwood and Jon Green for critical reading of the manuscript, and Dick Hudson for helpful discussion of the

Standardized Index of Association metric. M. T. E. was funded by the European Centre for Disease Prevention and Control (contract no. ECD 368).

REFERENCES

- Amemura-Maekawa, J., Kura, F., Chang, B. & Watanabe, H. (2005). *Legionella pneumophila* serogroup 1 isolates from cooling towers in Japan form a distinct genetic cluster. *Microbiol Immunol* **49**, 1027–1033.
- Benson, R. F., Lucas, C. E., Brown, E. W., Cowgill, K. D. & Fields, B. S. (2006). Molecular comparison of isolates from a recurring outbreak of Legionnaires' disease spanning 22 years. Chapter 37 in *Legionella: State of the Art 30 Years after Its Recognition*, pp. 139–145. Edited by N. P. Cianciotto, Y. Abu Kwaik, P. H. Edelstein, B. S. Fields, D. F. Geary, T. G. Harrison, C. A. Joseph, R. M. Ratcliff, J. E. Stout & M. S. Swanson. Washington, DC: American Society for Microbiology.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242.
- Bernander, S., Claesson, B. E. B., Hjelm, E., Svensson, N. & Hjorth, M. (2006). Serologic study of an outbreak of Legionnaires' disease: variation of sensitivity associated with the subgroup of *Legionella pneumophila* sg 1 antigen used and evidence of concurrent reactivity to other atypical pneumonia agents. Chapter 17 in *Legionella: State of the Art 30 Years after Its Recognition*, pp. 63–67. Edited by N. P. Cianciotto, Y. Abu Kwaik, P. H. Edelstein, B. S. Fields, D. F. Geary, T. G. Harrison, C. A. Joseph, R. M. Ratcliff, J. E. Stout & M. S. Swanson. Washington, DC: American Society for Microbiology.
- Black, W. J., Quinn, F. D. & Tompkins, L. S. (1990). *Legionella pneumophila* zinc metalloprotease is structurally and functionally homologous to *Pseudomonas aeruginosa* elastase. *J Bacteriol* **172**, 2608–2613.
- Brenner, D. J., Steigerwalt, A. G., Eppe, P., Bibb, W. F., McKinney, R. M., Starnes, R. W., Colville, J. M., Selander, R. K., Edelstein, P. H. & Moss, C. W. (1988). *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. *pasculli* subsp. nov. *J Clin Microbiol* **26**, 1695–1703.
- Bron, C. & Kerbosch, J. (1973). Finding all cliques of an undirected graph. *Commun ACM* **16**, 575–577.
- Bumbaugh, A. C., McGraw, E. A., Page, K. L., Selander, R. K. & Whittam, T. S. (2002). Sequence polymorphism of *dotA* and *mip* alleles mediating invasion and intracellular replication of *Legionella pneumophila*. *Curr Microbiol* **44**, 314–322.
- Cianciotto, N. P. & Fields, B. S. (1992). *Legionella pneumophila* *mip* gene potentiates intracellular infection of protozoa and human macrophages. *Proc Natl Acad Sci U S A* **89**, 5188–5191.
- Cohan, F. M. & Perry, E. B. (2007). A systematics for discovering the fundamental units of bacterial diversity. *Curr Biol* **17**, R373–R386.
- Cohan, F. M., Koeppl, A. & Krizanc, D. (2006). Sequence-based discovery of ecological diversity within *Legionella*. Chapter 88 in *Legionella: State of the Art 30 Years after Its Recognition*, pp. 367–376. Edited by N. P. Cianciotto, Y. Abu Kwaik, P. H. Edelstein, B. S. Fields, D. F. Geary, T. G. Harrison, C. A. Joseph, R. M. Ratcliff, J. E. Stout & M. S. Swanson. Washington, DC: American Society for Microbiology.
- Coscollá, M. & González-Candelas, F. (2007). Population structure and recombination in environmental isolates of *Legionella pneumophila*. *Environ Microbiol* **9**, 643–656.
- Coscollá, M., Gosalbes, M. J., Catalan, V. & González-Candelas, F. (2006). Genetic variability in environmental isolates of *Legionella pneumophila* from Comunidad Valenciana (Spain). *Environ Microbiol* **8**, 1056–1063.
- Didelot, X. & Falush, D. (2007). Inference of bacterial microevolution using multilocus sequence data. *Genetics* **175**, 1251–1266.
- Dreyfus, L. A. & Iglewski, B. H. (1986). Purification and characterization of an extracellular protease of *Legionella pneumophila*. *Infect Immun* **51**, 736–743.
- Edelstein, P. H., Nakahama, C., Tobin, J. O., Calarco, K., Beer, K. B., Joly, J. R. & Selander, R. K. (1986). Paleoepidemiologic investigation of Legionnaires' disease at Wadsworth Veterans Administration Hospital by using three typing methods for comparison of legionellae from clinical and environmental sources. *J Clin Microbiol* **23**, 1121–1126.
- Ehret, W. & Ruckdeschel, G. (1985). Molecular weight of the major outer membrane protein of *Legionella pneumophila*. *Eur J Clin Microbiol* **4**, 592–593.
- Engleberg, N. C., Carter, C., Weber, D. R., Cianciotto, N. P. & Eisenstein, B. I. (1989). DNA sequence of *mip*, a *Legionella pneumophila* gene associated with macrophage infectivity. *Infect Immun* **57**, 1263–1270.
- Enright, M. C. & Spratt, B. G. (1998). A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. *Microbiology* **144**, 3049–3060.
- Feil, E. J., Li, B. C., Aanensen, D. M., Hanage, W. P. & Spratt, B. G. (2004). eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J Bacteriol* **186**, 1518–1530.
- Fendukly, F., Bernander, S. & Hanson, H.-S. (2007). Nosocomial Legionnaires' disease caused by *Legionella pneumophila* serogroup 6: implication of the sequence-based typing method (SBT). *Scand J Infect Dis* **39**, 213–216.
- Fischer, G., Bang, H., Ludwig, B., Mann, K. & Hacker, J. (1992). *Mip* protein of *Legionella pneumophila* exhibits peptidyl-prolyl-*cis/trans* isomerase (PPIase) activity. *Mol Microbiol* **6**, 1375–1383.
- Fry, N. K., Alexiou-Daniel, S., Bangsberg, J. M., Bernander, S., Castellani Pastoris, M., Etienne, J., Forsblom, B., Gaia, V., Helbig, J. H. & other authors (1999). A multicenter evaluation of genotypic methods for the epidemiological typing of *Legionella pneumophila* serogroup 1: results of a pan-European study. *Clin Microbiol Infect* **5**, 462–477.
- Fry, N. K., Bangsberg, J. M., Bernander, S., Etienne, J., Forsblom, B., Gaia, V., Hasenberger, P., Lindsay, D., Papoutsis, A. & other authors (2000). Assessment of intercentre reproducibility and epidemiological concordance of *Legionella pneumophila* serogroup 1 genotyping by amplified fragment length polymorphism analysis. *Eur J Clin Microbiol Infect Dis* **19**, 773–780.
- Fry, N. K., Bangsberg, J. M., Bergmans, A., Bernander, S., Etienne, J., Franzin, L., Gaia, V., Hasenberger, P., Baladrón Jiménez, B. & other authors (2002). Designation of the European Working Group on Legionella Infection (EWGLI) amplified fragment length polymorphism types of *Legionella pneumophila* serogroup 1 and results of intercentre proficiency testing using a standard protocol. *Eur J Clin Microbiol Infect Dis* **21**, 722–728.
- Gaia, V., Fry, N. K., Harrison, T. G. & Peduzzi, R. (2003). Sequence-based typing of *Legionella pneumophila* serogroup 1 offers the potential for true portability in legionellosis outbreak investigation. *J Clin Microbiol* **41**, 2932–2939.
- Gaia, V., Fry, N. K., Afshar, B., Lück, P. C., Meugnier, H., Etienne, J., Peduzzi, R. & Harrison, T. G. (2005). Consensus sequence-based scheme for epidemiological typing of clinical and environmental isolates of *Legionella pneumophila*. *J Clin Microbiol* **43**, 2047–2052.
- Gibbs, M. J., Armstrong, J. S. & Gibbs, A. J. (2000). Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* **16**, 573–582.

- Gilmour, M. W., Bernard, K., Tracz, D. M., Olson, A. B., Corbett, C. R., Burdz, T., Ng, B., Wiebe, D., Broukhanski, G. & other authors (2007). Molecular typing of a *Legionella pneumophila* outbreak in Ontario, Canada. *J Med Microbiol* **56**, 336–341.
- Giske, C. G., Libisch, B., Colinon, C., Scoulica, E., Pagani, L., Füzi, M., Kronvall, G. & Rossolini, G. M. (2006). Establishing clonal relationships between VIM-1-like metallo- β -lactamase-producing *Pseudomonas aeruginosa* strains from four European countries by multilocus sequence typing. *J Clin Microbiol* **44**, 4309–4315.
- Harb, O. S. & Abu Kwaik, Y. (1998). Identification of the aspartate- β -semialdehyde dehydrogenase gene of *Legionella pneumophila* and characterization of a null mutant. *Infect Immun* **66**, 1898–1903.
- Harrison, T. G., Fry, N. K., Afshar, B., Bellamy, W., Doshi, N. & Underwood, A. P. (2006). Typing of *Legionella pneumophila* and its role in elucidating the epidemiology of Legionnaires' Disease. Chapter 25 in *Legionella: State of the Art 30 Years after Its Recognition*, pp. 94–99. Edited by N. P. Cianciotto, Y. Abu Kwaik, P. H. Edelstein, B. S. Fields, D. F. Geary, T. G. Harrison, C. A. Joseph, R. M. Ratcliff, J. E. Stout & M. S. Swanson. Washington, DC: American Society for Microbiology.
- Haubold, B. & Hudson, R. R. (2000). LIAN 3.0: detecting linkage disequilibrium in multilocus data. *Bioinformatics* **16**, 847–848.
- Heuner, K., Bender-Beck, L., Brand, B. C., Lück, P. C., Mann, K.-H., Marre, R., Ott, M. & Hacker, J. (1995). Cloning and genetic characterization of the flagellum subunit gene (*flaA*) of *Legionella pneumophila* serogroup 1. *Infect Immun* **63**, 2499–2507.
- Hoffman, P. S., Houston, L. & Butler, C. A. (1990). *Legionella pneumophila* *htpAB* heat shock operon: nucleotide sequence and expression of the 60-kilodalton antigen in *L. pneumophila*-infected HeLa cells. *Infect Immun* **58**, 3380–3387.
- Hudson, R. R. (1994). Analytical results concerning linkage disequilibrium in models with genetic transformation and conjugation. *J Evol Biol* **7**, 535–548.
- Jolley, K. A., Feil, E. J., Chan, M.-S. & Maiden, M. C. J. (2001). Sequence type analysis and recombinational tests (START). *Bioinformatics* **17**, 1230–1231.
- Ko, K. S., Hong, S. K., Lee, H. K., Park, M.-Y. & Kook, Y.-H. (2003). Molecular evolution of the *dotA* gene in *Legionella pneumophila*. *J Bacteriol* **185**, 6269–6277.
- Kruskal, J. B. (1956). On the shortest spanning sub-tree of a graph and travelling salesman problem. *Proc Am Math Soc* **7**, 48–50.
- Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K. & other authors (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* **95**, 3140–3145.
- Marchler-Bauer, A., Anderson, J. B., Cherukuri, P. F., DeWeese-Scott, C., Geer, L. Y., Gwadz, M., He, S., Hurwitz, D. I., Jackson, J. D. & other authors (2005). CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res* **33**, D192–D196.
- Martin, D. & Rybicki, E. (2000). RDP: detection of recombination amongst aligned sequences. *Bioinformatics* **16**, 562–563.
- McDade, J. E., Shepard, C. C., Fraser, D. W., Tsai, T. R., Redus, M. A. & Dowdle, W. R. (1977). Legionnaires' disease: isolation of a bacterium and demonstration of its role in other respiratory disease. *N Engl J Med* **297**, 1197–1203.
- McVean, G., Awadalla, P. & Fearnhead, P. (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160**, 1231–1241.
- Mengaud, J. M. & Horwitz, M. A. (1993). The major iron-containing protein of *Legionella pneumophila* is an aconitase homologous with the human iron-responsive element-binding protein. *J Bacteriol* **175**, 5666–5676.
- Moffat, J. F., Edelstein, P. H., Regula, D. P., Cirillo, J. D. & Tompkins, L. S. (1994). Effects of an isogenic Zn-metalloprotease-deficient mutant of *Legionella pneumophila* in a guinea-pig pneumonia model. *Mol Microbiol* **12**, 693–705.
- Molofsky, A. B., Shetron-Rama, L. M. & Swanson, M. S. (2005). Components of the *Legionella pneumophila* flagellar regulon contribute to multiple virulence traits, including lysosome avoidance and macrophage death. *Infect Immun* **73**, 5720–5734.
- Nei, M. & Gojorori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**, 418–426.
- Padidam, M., Sawyer, S. & Fauquet, C. M. (1999). Possible emergence of new geminiviruses by frequent recombination. *Virology* **265**, 218–225.
- Paraskevopoulos, C., Bordenstein, S. R., Wernegreen, J. J., Werren, J. H. & Bourtzis, K. (2006). Toward a *Wolbachia* multilocus sequence typing system: discrimination of *Wolbachia* strains present in *Drosophila* species. *Curr Microbiol* **53**, 388–395.
- Posada, D. & Crandall, K. A. (2001). Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc Natl Acad Sci U S A* **98**, 13757–13762.
- Ratzow, S., Gaia, V., Helbig, J. H., Fry, N. K. & Lück, P. C. (2007). Addition of *neuA*, the gene encoding N-acetylneuraminate cytidyl transferase, increases the discriminatory ability of the consensus sequence-based scheme for typing *Legionella pneumophila* serogroup 1 strains. *J Clin Microbiol* **45**, 1965–1968.
- Ruiz-Garbajosa, P., Bonten, M. J. M., Robinson, D. A., Top, J., Nallapareddy, S. R., Torres, C., Coque, T. M., Cantón, R., Baquero, F. & other authors (2006). Multilocus sequence typing scheme for *Enterococcus faecalis* reveals hospital-adapted genetic complexes in a background of high rates of recombination. *J Clin Microbiol* **44**, 2220–2228.
- Salminen, M. O., Carr, J. K., Burke, D. S. & McCutchan, F. E. (1995). Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. *AIDS Res Hum Retroviruses* **11**, 1423–1425.
- Sampson, J. S., O'Connor, S. P., Holloway, B. P., Plikaytis, B. B., Carlone, G. M. & Mayer, L. W. (1990). Nucleotide sequence of *htpB*, the *Legionella pneumophila* gene encoding the 58-kilodalton (kDa) common antigen, formerly designated the 60-kDa common antigen. *Infect Immun* **58**, 3154–3157.
- Saunders, N. A., Harrison, T. G., Haththotuwa, A., Kachwalla, N. & Taylor, A. G. (1990). A method for typing strains of *Legionella pneumophila* serogroup 1 by analysis of restriction fragment length polymorphisms. *J Med Microbiol* **31**, 45–55.
- Scaturro, M., Losardo, M., De Ponte, G. & Ricci, M. L. (2005). Comparison of three molecular methods used for subtyping of *Legionella pneumophila* strains isolated during an epidemic of legionellosis in Rome. *J Clin Microbiol* **43**, 5348–5350.
- Schmidt, H. A., Strimmer, K., Vingron, M. & von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**, 502–504.
- Schoonmaker, D., Heimberger, T. & Birkhead, G. (1992). 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* **30**, 1491–1498.
- Selander, R. K., McKinney, R. M., Whittam, T. S., Bibb, W. F., Brenner, D. J., Nolte, F. S. & Pattison, P. E. (1985). Genetic structure of populations of *Legionella pneumophila*. *J Bacteriol* **163**, 1021–1037.

- Smith, J. M. (1992).** Analyzing the mosaic structure of genes. *J Mol Evol* **34**, 126–129.
- Smith, J. M., Smith, N. H., O'Rourke, M. & Spratt, B. G. (1993).** How clonal are bacteria? *Proc Natl Acad Sci U S A* **90**, 4384–4388.
- Stern, A., Doron-Faigenboim, A., Erez, E., Martz, E., Bacharach, E. & Pupko, T. (2007).** Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. *Nucleic Acids Res* **35**, W506–W511.
- Stone, B. J. & Abu Kwaik, Y. (1998).** Expression of multiple pili by *Legionella pneumophila*: identification and characterization of a type IV pilin gene and its role in adherence to mammalian and protozoan cells. *Infect Immun* **66**, 1768–1775.
- Stone, B. J. & Abu Kwaik, Y. (1999).** Natural competence for DNA transformation by *Legionella pneumophila* and its association with expression of type IV pili. *J Bacteriol* **181**, 1395–1402.
- Strimmer, K. & von Haeseler, A. (1997).** Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proc Natl Acad Sci U S A* **94**, 6815–6819.
- van Ketel, R. J., ter Schegget, J. & Zanen, H. C. (1984).** Molecular epidemiology of *Legionella pneumophila* serogroup 1. *J Clin Microbiol* **20**, 362–364.
- Vassileva, M., Torii, K., Oshimoto, M., Okamoto, A., Agata, N., Yamada, K., Hasegawa, T. & Ohta, M. (2006).** Phylogenetic analysis of *Bacillus cereus* isolates from severe systemic infections using multi-locus sequence typing scheme. *Microbiol Immunol* **50**, 743–749.
- Whatmore, A. M. & Dowson, C. G. (1999).** The autolysin-encoding gene (*lytA*) of *Streptococcus pneumoniae* displays restricted allelic variation despite localized recombination events with genes of pneumococcal bacteriophage encoding cell wall lytic enzymes. *Infect Immun* **67**, 4551–4556.
- Wong, S., Pabbaraju, K., Burk, V. F., Broukhanski, G. C., Fox, J., Louie, T., Mah, M. W., Bernard, K. & Tilley, P. A. (2006).** Use of sequence-based typing for investigation of a case of nosocomial legionellosis. *J Med Microbiol* **55**, 1707–1710.
- Young, M., Smith, H., Gray, B., Huang, B., Barten, J., Towner, C., Plowman, S., Afshar, B., Fry, N. & other authors (2005).** The public health implications of a sporadic case of culture-proven Legionnaires' disease. *Aust N Z J Public Health* **29**, 513–517.
- Zhao, X. & Dreyfus, L. A. (1990).** Expression and nucleotide sequence analysis of the *Legionella pneumophila* *recA* gene. *FEMS Microbiol Lett* **58**, 227–231.

Edited by: F. A. Rainey