

1 **A supervised statistical learning approach for accurate *Legionella***
2 ***pneumophila* source attribution during outbreaks**

3

4 **Andrew H. Buultjens^{1,2}, Kyra Y. L. Chua³, Sarah L. Baines¹, Jason Kwong^{1,2,3}, Wei Gao², Zoe Cutcher^{4,5},**
5 **Stuart Adcock⁴, Susan Ballard¹, Takehiro Tomita¹, Nela Subasinghe¹, Glen Carter^{2,3}, Sacha J. Pidot^{2,3},**
6 **Lucinda Franklin⁴, Torsten Seemann^{3,6}, Anders Gonçalves Da Silva^{1,3}, Benjamin P. Howden^{1,2,3,*}, Timothy P.**
7 **Stinear^{2,3,*}**

8

9 **Affiliations:**

10 ¹Department of Microbiology and Immunology at the Peter Doherty Institute for Infection and Immunity, The
11 University of Melbourne, Victoria, Australia

12 ²Doherty Applied Microbial Genomics, The Peter Doherty Institute for Infection and Immunity, Victoria, Australia

13 ³Microbiological Diagnostic Unit Public Health Laboratory at the Peter Doherty Institute for Infection and Immunity,
14 The University of Melbourne, Victoria, Australia.

15 ⁴Health Protection Branch, Department of Health and Human Services, Victoria, Australia

16 ⁵National Centre for Epidemiology and Population Health, Australian National University, Canberra, Australia

17 ⁶Victorian Life Sciences Computational Initiative, The University of Melbourne, Victoria, Australia

18

19

20 ***To whom correspondence may be addressed:** tstinear@unimelb.edu.au, bhowden@unimelb.edu.au

21

22 **Running title:** Genomics and machine learning for *L. pneumophila* source tracking

23 **Abstract**

24 Public health agencies are increasingly relying on genomics during Legionnaires' disease investigations.
25 However, the causative bacterium (*Legionella pneumophila*) has an unusual population structure with
26 extreme temporal and spatial genome sequence conservation. Furthermore, Legionnaires' disease
27 outbreaks can be caused by multiple *L. pneumophila* genotypes in a single source. These factors can
28 confound cluster identification using standard phylogenomic methods. Here, we show that a statistical
29 learning approach based on *L. pneumophila* core genome single nucleotide polymorphism (SNP)
30 comparisons eliminates ambiguity for defining outbreak clusters and accurately predicts exposure sources
31 for clinical cases. We illustrate the performance of our method by genome comparisons of 234 *L.*
32 *pneumophila* isolates obtained from patients and cooling towers in Melbourne, Australia between 1994
33 and 2014. This collection included one of the largest reported Legionnaires' disease outbreaks, involving
34 125 cases at an aquarium. Using only sequence data from *L. pneumophila* cooling tower isolates and
35 including all core genome variation, we built a multivariate model using discriminant analysis of principal
36 components (DAPC) to find cooling tower-specific genomic signatures, and then used it to predict the origin
37 of clinical isolates. Model assignments were 93% congruent with epidemiological data, including the
38 aquarium Legionnaires' outbreak and three other unrelated outbreak investigations. We applied the same
39 approach to a recently described investigation of Legionnaires' disease within a UK hospital and observed
40 model predictive ability of 86%. We have developed a robust means to breach *L. pneumophila* genetic
41 diversity extremes and provide objective source attribution data for outbreak investigations.

42

43 **Significance**

44 Microbial outbreak investigation is moving to a paradigm where phylogenomic trees and core genome
45 multilocus typing schemes are sufficient to identify infection sources with high certainty. We show by
46 studying 234 *Legionella pneumophila* genomes collected over 21 years, that it is critically important to have
47 a detailed understanding of local bacterial population diversity or risk misidentifying an outbreak source.
48 We propose statistical learning approaches that can accommodate all core genome variation and link
49 clinical *L. pneumophila* isolates back to environmental sources, at both the inter- and intra-institutional
50 levels, eliminating the ambiguity of inferring transmission from phylogenies. This information is critical in
51 outbreak investigations, particularly for *L. pneumophila*, which spreads via aerosols causing Legionnaires'
52 disease.

53

54 **Introduction**

55 Legionellae are Gram-negative bacteria that replicate within free-living aquatic amoebae and are present in
56 aquatic environments worldwide. These bacteria can proliferate in man-made water systems and cause
57 large outbreaks of pneumonia known as Legionnaires' disease when contaminated water is aerosolized and
58 inhaled¹. The majority of human infections are caused by *Legionella pneumophila* serogroup 1². Public
59 health investigations of Legionnaires' disease outbreaks are typically supported by molecular typing
60 methods to establish the likely source of the bacteria and the extent of the outbreak. Investigations usually
61 proceed with the assumption that a single *Legionella* genotype is responsible for an environmental point
62 source reservoir³. Traditional molecular typing methods described for fingerprinting Legionellae include
63 pulsed-field gel electrophoresis (PFGE) and sequence-based typing (SBT)⁴. Increasingly, whole genome
64 sequencing (WGS) is being employed to investigate individual *Legionella* outbreaks and the insights
65 obtained from these high-resolution comparisons are challenging our expectations regarding common-
66 source outbreaks, which usually are characterized by a single strain or genotype⁵⁻⁹. It is becoming evident
67 that outbreaks can be caused by multiple co-circulating *L. pneumophila* genotypes^{5,10} and that *L.*
68 *pneumophila* core genomes can be surprisingly conserved across space and time^{8,11,12}.

69

70 Melbourne is in the state of Victoria and it is the second largest city in Australia with a population
71 approaching five million inhabitants, and considered the ninth largest city in the Southern Hemisphere.
72 Legionellosis has been a notifiable disease in Victoria since 1979 and there are 50-100 cases reported each
73 year, most occurring in the greater metropolitan region of Melbourne¹³. The Microbiological Diagnostic
74 Unit Public Health Laboratory (MDU PHL) is Victoria's State Reference Laboratory for the characterization
75 and typing of *Legionella* spp. The laboratory's collection includes isolates from a particularly noteworthy
76 outbreak at the Melbourne Aquarium in April 2000. This was the largest single episode of Legionellosis
77 reported in Australia¹⁴, approximately three months after the aquarium was opened to visitors, with
78 construction of the site completed in December 1999. It resulted in 125 confirmed cases, with positive
79 cultures obtained from 11 patients. Our isolate collection also spanned 28 other potential legionellosis
80 outbreaks or infection clusters, for which at least one culture isolate had been obtained.

81

82 In this study, we used comparative genomics to explore the population structure of 234 *Legionella*
83 *pneumophila* isolates recovered from human and environmental sources submitted to the MDU PHL in
84 Melbourne over a 21-year period. This collection included 11 clinical and 14 environmental isolates from
85 the Aquarium outbreak and 42 clinical and 50 environmental isolates from 28 other likely point source case
86 clusters. We also assessed genomic data from a recently described investigation of Legionnaires' cases at a
87 UK hospital⁸. The aim of this project was to develop a robust genomic approach that would surmount the
88 unusual population structure of

89 *L. pneumophila* and assist identification of case clusters and source tracking efforts during Legionnaires'
90 disease outbreak investigations.

91

92 **Materials and Methods**

93 **Bacterial strains, growth conditions, case definitions.** *Legionella pneumophila* serogroup 1 isolates were
94 resuscitated from -80°C storage and assessed. Duplicate isolates from the same patient were excluded
95 from the study. Isolates were cultured for 48-72 h at 37°C on BCYE agar and re-confirmed serogroup 1 by
96 latex agglutination (Oxoid). Metadata collected on all isolates included year of isolation and country or city
97 of isolation. Cases resident in the state of Victoria, Australia, were assessed by the Victorian State
98 Government public health unit in accordance with national guidelines and an outbreak investigation was
99 initiated when common exposures were reported by different cases whose onset dates occurred within a
100 two-week window. ([http://www.health.gov.au/internet/main/publishing.nsf/content/cdna-song-](http://www.health.gov.au/internet/main/publishing.nsf/content/cdna-song-legionella.htm)
101 [legionella.htm](http://www.health.gov.au/internet/main/publishing.nsf/content/cdna-song-legionella.htm), accessed 31 August 2015). In this manner, we were able to determine the human cases
102 epidemiologically linked to each other. Many of the outbreaks/infection clusters contained a greater
103 number of cases than there were isolates as the diagnosis of Legionellosis was made by culture-
104 independent methods. Complete, closed genomes of *L. pneumophila* that were publicly available were
105 obtained from GenBank for inclusion in the analysis (Table S1).

106

107 **Sequence based typing.** This was performed as previously described according to the European
108 Legionnaires' Disease Surveillance Network (ELDSNet) method
109 (http://bioinformatics.phe.org.uk/legionella/legionella_sbt/php/sbt_homepage.php, accessed 31 August
110 2015)¹⁵.

111

112 **DNA sequencing.** DNA libraries were prepared using the NexteraXT DNA preparation kit (Illumina) and
113 whole genome sequencing was performed on the NextSeq platform (Illumina) with 2x150 bp chemistry. For
114 single molecule real-time (SMRT) sequencing (Pacific Biosciences), genomic DNA was extracted from
115 agarose plugs using the CDC Pulsenet Protocol to allow for recovery of high molecular weight, intact DNA
116 (<http://www.cdc.gov/pulsenet/pathogens>, accessed 31 August 2015). Size-selected 10kb DNA libraries
117 were prepared according to manufacturers' instructions and sequenced on the RS II platform (Pacific
118 Biosciences) using P6-C4 chemistry. All sequence reads and the completed genome are available (GenBank
119 BioProject ID: PRJEB13594)

120

121 **Legionella pneumophila serogroup 1 isolate Lpm7613 assembly and closure.** A high quality finished ST 30
122 reference genome was established for *L. pneumophila* serogroup 1 clinical isolate Lpm7613 using the
123 SMRT® Analysis System v2.3.0.140936 (Pacific Biosciences). Raw sequence data were *de novo* assembled
124 using the HGAP v3 protocol with a genome size of 4 Mb. Polished contigs were error corrected using Quiver
125 v1. The resulting assembly was then checked using BridgeMapper v1 in the SMRT® Analysis System, and the
126 consensus sequence corrected with short-read Illumina data, using the program Snippy
127 (<https://github.com/tseemann/snippy>). Whole genome annotation was performed using Prokka ¹⁶,
128 preferentially using the *L. pneumophila* Paris strain annotation ¹⁷. BRIG was used to visualize BLASTn
129 DNA:DNA comparisons of *L. pneumophila* Lpm7613 against other *L. pneumophila* genomes ¹⁸.
130 Nomenclature of the genomic islands demonstrated in *L. pneumophila* Lpm7613 was based on previously
131 described islands ¹⁹. CRISPR databases were used to search for CRISPR sequences
132 (<http://crispi.genouest.org> and <http://crispr.u-psud.fr/Server/>, accessed 14 February 2016).

133

134 **Variant detection and phylogenetic analysis.** The genomes of ten publicly available complete *L.*
135 *pneumophila* genomes (Table S1) were shredded to generate short *in silico* sequence reads of 250bp and all
136 244 *L. pneumophila* reads sets were mapped against the Lpm7613 reference genome using Snippy v3.2. An
137 alignment file from pairwise comparisons of core genome SNPs (with inferred recombining sites removed)
138 was used as input to FastTree v2.1.8 with double precision ²⁰ to infer a maximum likelihood phylogenetic
139 tree using the general time reversible model of nucleotide substitution. Branch support was estimated
140 using 1,000 bootstrap replicates. Resulting trees were visualized in FigTree v1.4.2
141 (<http://tree.bio.ed.ac.uk/software/figtree/>). Single nucleotide polymorphism (SNP) differences between
142 isolates were tabulated and visualized using a custom R-script (<https://github.com/MDU->
143 [PHL/pairwise_snp_differences](https://github.com/MDU-PHL/pairwise_snp_differences)). The core genome SNPs were also used as the input into a hierarchical
144 Bayesian analysis of population structure (hierBAPS) using iterative clustering to a depth of 10 levels and a
145 pre-specified maximum of 20 clusters ²¹.

146

147 **Recombination analysis.** Recombination detection was performed using ClonalFrameML ²², taking as input
148 a full genome alignment (included invariant sites) prepared using Snippy as above and the ML phylogeny as
149 a guide tree with polytomies removed from the FastTree tree using a custom python script
150 (https://github.com/kwongj/nw_multi2bifurcation). Results were visualized using a custom Python script to
151 render separate and superposable images of extant and ancestral inferred recombination regions
152 (<https://github.com/kwongj/cfml-maskrc>).

153

154 **Phylogeographic analysis.** Variant detection for the 64 environmental genomes was undertaken by running
155 snippy-core. Core SNPs were used to reconstruct a phylogenomic tree with FastTree that was overlaid upon
156 a base map in GenGIS²³. Victorian population mesh data was downloaded from the Australia Bureau of
157 Statistics webpage
158 (<http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.001July%202016?OpenDocument>)
159 and Local Government Area data was downloaded from the Victorian Government Data Directory webpage
160 (<https://www.data.vic.gov.au/data/dataset/lga-geographical-profiles-2014-beta/resource/f6c49074-0679-4c79-a0db-04dac8eda364>).
161

162

163 **DAPC model building using core SNPs.** DAPC was implemented in the R package *adegenet* v2.0.1²⁴. For
164 input, we used a matrix of single nucleotide polymorphisms (SNP) for all genomes originating from disease
165 clusters that possessed at least one environmental and clinical representative (Table S1). SNP detection was
166 undertaken by running Snippy and sites that were recombinogenic and or invariant among the
167 environmental genomes were discarded. The input matrix was used to generate a *genind* object in
168 *adegenet* that was subsequently separated into two subset *geninds*. The first subset consisted of
169 environmental isolates (hereon referred to as the training set) and the second *genind* consisted of the
170 clinical isolates from groups in common with validation set groups (hereon referred to as the validation
171 set). The training set was used to develop a DAPC model using the training set disease clusters (Table S1) to
172 define discriminant functions²⁵. The ability of the model to predict the source attribution of the validation
173 set genomes was simulated across the first to the 20th principal components, allowing an optimal number
174 of principal components to be identified. The optimized model was then used to predict the groupings of
175 validation set clinical isolate genomes.

176

177 **DAPC model building using cgMLST variation.** In order to detect variants within the recently described
178 cgMLST regions, reads were mapped to the Lp_Philadelphia chromosome (NC_002942.5) using snippy. SNP
179 profiles from within the cgMLST regions were reduced to allelic integers, with all genes containing zero
180 coverage or uncertain base-calls, excluded. Allelic integers were concatenated into a matrix and, using the
181 same DAPC model-building method as mentioned above, models were established using the training set
182 genomes and used to predict the groupings of the validation set clinical isolate genomes.

183

184 **Results**

185 **Isolates and epidemiology.** There were 234 *Legionella pneumophila* serogroup 1 (Lpn-SG 1) isolates
186 obtained across a 21-year period between 1994 and 2014. Initial MLST analysis indicated that 180 isolates
187 (77%) belonged to ST30. The collection comprised 180 clinical isolates of respiratory origin (sputum or
188 bronchoscopy specimens) and 64 environmental isolates recovered from cooling tower water samples. All
189 isolates were collected in the state of Victoria with the exception of six isolates from patients who were
190 exposed elsewhere. Further information for each isolate is available in Table S1, including NCBI SRA
191 accession numbers. One hundred and ten of the 234 isolates were epidemiologically associated with 29
192 formally investigated case clusters or outbreaks, designated as outbreaks A-AC (Table S1). The majority of
193 these cases occurred within a 42 km radius of Melbourne city center and over a 16-year period. Outbreak
194 A, the Melbourne Aquarium outbreak, was the largest¹⁴.

195

196 **Complete genome sequence of *Legionella pneumophila* serogroup 1 isolate Lpm7613.** Before this study,
197 there were no closed, fully assembled ST30 *L. pneumophila* genomes. Thus, to ensure identification of
198 maximum genetic variation among this dominant ST in our collection, we first established a ST30 reference
199 genome sequence, selecting a clinical isolate from the Melbourne Aquarium outbreak (Lpm7613). The
200 finished genome consisted of a single circular 3,261,562 bp chromosome (38.3% GC) and a 129,875 bp
201 circular plasmid (pLpm7613) (Fig. S1). Although the chromosome indicated this genome belonged to the
202 same lineage as *L. pneumophila* Philadelphia (see analysis below), the plasmid shared 100% nucleotide
203 identity with pLPP reported in *L. pneumophila* Paris, but 2kb shorter in length¹⁷. A total of 2,891
204 chromosomal protein-coding sequences (CDS), 43 tRNA genes and nine rRNA loci were predicted using
205 Prokka¹⁶. CRISPR-Cas regions were not detected²⁶.

206

207 **Assessment of *L. pneumophila* population structure.** Sequence reads from the other 233 genomes and the
208 ten selected publicly available completed genomes were mapped to the chromosome of reference strain
209 Lpm7613. Approximately 90% of the Lpm7613 genome was present in all genomes (*i.e.*, core), with 188,049
210 variable core nucleotide positions identified. Population structure analyses using an unsupervised Bayesian
211 hierarchical clustering approach revealed six distinct groups (BAPS groups) (Fig. 1A). Comparison of intra-
212 and inter-BAPS group pairwise SNP distances confirmed the validity of these clusters and highlighted the
213 extensive genetic variation among this Lpn-SG1 population (Fig. S2). The exceptions were BAPS groups 3
214 and 4, which classified isolates across two clades, and is likely explained by recombination. Most striking,
215 however, was the lack of diversity within the 186 genomes comprising BAPS group 5 (hereafter referred to
216 as BAPS-5), with a median core SNP distance of only 5 SNPs (IQR 3 – 7). Isolates dispersed across time and
217 space (including isolates from England, New South Wales, South Australia and Tasmania) were scattered
218 throughout the entire phylogeny. All 180 ST30 isolates were encompassed by BAPS-5, as was ST37 *L.*

219 *pneumophila* Philadelphia (Philadelphia, USA), ST211 *L. pneumophila* ATCC 43290 (Denver, USA) and ST733
220 *L. pneumophila* Thunder Bay (Ontario, Canada). The median inter-BAPS group distances ranged between
221 27,506 to 63,136 SNPs (Fig. S2), highlighting that there is also substantial genetic diversity among Lpn-SG1
222 isolates circulating in Melbourne.

223

224 A rooted maximum likelihood phylogeny of the population was then inferred using the 181,633 non-
225 recombining core SNP loci. The phylogenomic tree reflected the BAPS clusters with BAPS-5 forming a
226 distinct, well-supported lineage (Fig. 1A). The separation of the three North American reference isolates
227 from the Melbourne ST30 isolates is suggestive of contemporaneous global dispersal of this BAPS-5 lineage
228 (Fig. 1A). All BAPS groups displayed monophyletic origins with the exception of BAPS-3 and BAPS-4. BAPS-3
229 had a single isolate of paraphyletic origin that shared a most recent common ancestor (MRCA) with BAPS-2
230 while BAPS-4 contained two paraphyletic sub-clades, one of which shared a MRCA with the majority of
231 BAPS-3 isolates.

232

233 **Impact of recombination.** Recombination is a driving force in the evolution of the Legionellae^{5,7,27-30}.
234 Therefore, to further understand the structure and evolution of this Lpn-SG1 population we assessed the
235 impact of DNA exchange. There was evidence of extensive recombination among isolates across BAPS
236 groups 1-4, and 6 with approximately 3% of variable nucleotide sites impacted relative to the Lpm7613
237 reference chromosome. The detection of two paraphyletic groups (BAPS-3 and BAPS-4) is likely explained
238 by ancestral recombination among the component sub-clades. In comparison, there was little
239 recombination evident among BAPS-5 isolates (Fig. S3), in accord with the core SNP phylogeny described
240 above, and suggesting the relatively recent emergence of this *L. pneumophila* lineage.

241

242 **Genomic molecular epidemiology of local outbreaks.** We next compared only the 180 ST30 genomes to
243 our Lpm7613 reference genome, and again confirmed the very restricted genomic diversity within this
244 lineage (median core SNP distance was 6 SNPs (IQR 4 – 9), with five outlier genomes, impacted by
245 recombination. (Fig. 1B, Fig. S3). Within this reconstructed core-genome ST30-specific phylogeny, many but
246 not all epidemiologically-related isolates formed distinct, well-supported, monophyletic clades. In some
247 instances, epidemiologically-associated isolates spanned multiple clades (outbreaks A, B, C, D and K) (Fig.
248 1B). In addition, Outbreak A (the Melbourne aquarium outbreak), which was previously considered to
249 represent infections caused by a single clone (Table S1)¹⁴, actually contained five distinct genotypes (A1-
250 A5) (Fig 1B, Fig. 2).

251

252 The analysis of environmental surveillance isolates provided an ideal means to gain insights into the
253 diversity within potential reservoirs of Lpn-SG1 - diversity that might enable prospective source tracking. A
254 phylogeographic analysis was therefore undertaken to assess the relationship between 64 environmental
255 Melbourne metropolitan isolates against their 11 sampling locations. Based on variation in core SNPs,
256 substantial geographical structure was observed, with the majority of isolates from common outbreak
257 codes tightly clustering in the phylogeny (Fig. 2). As for the Aquarium isolates, we observed evidence for
258 relatively high genomic diversity (Fig. 2). The existence of phylogeographic structure among the
259 environmental isolates suggested it might be possible to use the genome data and build predictive
260 diagnostic models to assist source identification efforts.

261

262 **A multivariate statistical model for source attribution.** To enhance resolution and try to detect outbreak-
263 specific genomic signals, a supervised statistical learning approach called Discriminant Analysis of Principal
264 Components (DAPC)²⁵ was employed. DAPC is a linear discriminant analyses (LDA) that accommodates
265 discrete genetic-based predictors by first transforming the genetic data into continuous Principal
266 Components (PC) and builds predictive classification models. The PCs are used to build discriminant
267 functions (DF) under the constraint that they must minimize within group variance, and maximize variance
268 between groups. Infection clusters are specified *a priori* from the epidemiological findings, and training
269 isolates are used to define the discriminant functions. The model can then be used to estimate the
270 posterior probability of membership for unknown (*e.g.* clinical) isolates to each pre-specified infection
271 cluster given the training data. Here, we used 43 of the 64 environmental isolates as our training set
272 (isolates originating from epidemiologically defined infection clusters that possessed at least one
273 environmental and clinical representative), under the assumption that each outbreak was caused by
274 exposure to a point source of Lpn-SG1. We used core genome SNPs to build our classifier, in order to
275 maximize our discriminatory power³¹.

276

277 Outbreak-associated environmental Lpn-SG1 were grouped *a priori* into training set groups based on the
278 origin of the cooling towers from which they were isolated (see methods). The DFs were then used to
279 classify 15 clinical isolates that had been independently assigned based on epidemiological data to the
280 training set groups, hereon referred to as the validation genomes (Table S1). The input matrix for DAPC was
281 an alignment of 714 non-recombinogenic SNPs variable among the 43 environmental genomes. A model
282 was trained using the first four principal components (PC), as this was found to be optimal (see methods).
283 Our DAPC model captured 98% of the variation among the 43 environmental isolates (Fig. 3). When
284 classifying our clinical validation genomes, we found that in 93% of the isolates there was a match between
285 our model's assignment and that proposed by the epidemiological data (Fig. 3). These data show that

286 despite the high level of genome conservation, it is possible to utilize signature differences in core genome
287 SNPs to build predictive probabilistic classification models. The single discrepancy between model
288 predictions and epidemiological groupings was an infection cluster C genome that was predicted as
289 originating from the Aquarium. Interestingly, cluster C was located closest to the Melbourne Aquarium at a
290 distance of approximately 500 metres. Given the proximity of clusters A and C, these data may indicate
291 cooling towers seeded from a common *L. pneumophila* source. In order to appraise the utility of this
292 method beyond a large urban setting and the ST30 genotype, we built a sister model using 31 ST1
293 environmental *L. pneumophila* genomes from a previously published hospital investigation in Essex, UK,
294 and used it to predict the origins of seven nosocomial clinical isolates (Table S1)⁸. Here, the model was
295 trained using the first 15 PCs, as this was found to be optimal. The Essex DAPC model captured 98% of the
296 variation among the 31 environmental isolates and, as with the Melbourne disease clusters, the model
297 performed very well. For 86% of the isolates there was a match between the model's ward assignment and
298 the originating wards suggested by epidemiology. Again, a single discrepancy occurred with a ward G
299 genome predicted to originate from ward A. Wards A and G were co-located on the same corner and level of
300 a common building, again suggesting a common *L. pneumophila* source⁸. As before, isolates from a
301 common source would be miss-assigned by the model, owing to the lack of location-specific genomic
302 variants.

303

304 **Core genome multilocus sequence typing (cgMLST) has reduced discrimination.** In order to evaluate the
305 utility of the recently described cgMLST scheme for source tracking^{32,33}, we trained new DAPC models for
306 both the Melbourne and Essex hospital datasets using a matrix of allelic integers derived from SNP profiles
307 of the 1,529 cgMLST loci. When using the first one and seven PCs, we observed only 60% and 71%
308 concordance between our model's assignment and that predicted by the epidemiological data for the
309 Melbourne and Essex hospital datasets, respectively (Fig. 2,3).

310

311 Discussion

312 In this study, we have retrospectively examined a large collection of 234 clinical and environmental isolates
313 Lpn-SG1 isolates spanning 29 defined outbreaks. Isolates were collected over wide temporal and spatial
314 scales and detailed genomic comparisons revealed wide extremes of Lpn-SG1 genetic diversity within and
315 between outbreak clusters; a phenomenon not fully appreciated from previous genomic investigations that
316 have sampled less extensively and focused on single outbreaks^{5,6,34}. Most striking in our collection was the
317 high sequence conservation and dominance of a single genotype (BAPS-5, ST30), shared by 77% of isolates
318 with a median core SNP distance of only 5 SNPs across 21 years. In agreement with our findings, two recent

319 population genomic investigations of Lpn-SG1 also describe the unusual restriction in core genome diversity
320 ^{8,12}.

321

322 Based on our previous experience with other bacterial pathogens ³⁵ and reports in the literature of
323 Legionnaires' disease outbreak investigations using genomics ^{34,36} - we expected to be able to use Lpn-SG1
324 genomic comparisons and develop genetic rule-in or rule-out criteria to guide outbreak assessment and
325 source attribution. For example, we recently proposed a 'traffic-light' system for *Listeria monocytogenes*
326 based on SNP difference cutoffs of 'likely related', 'possibly-related' and 'not-related' ³⁵. This approach has
327 also been proposed for *L. pneumophila* ³². A comparison of genotyping approaches using 335 Lpn isolates,
328 including 106 from the European Society for Clinical Microbiology Study Group's *Legionella Typing Panel*,
329 proposed an escalating, hierarchical approach to genotyping, beginning with an extended 50-gene MLST
330 scheme up to a 1529-gene cgMLST ^{32,33}.

331

332 The analysis of the population structure of Lpn-SG1 presented here indicates that SNP-based typing with
333 threshold cut-offs, whether they are based on seven genes, 50 genes, 1500 genes or whole genomes will
334 not necessarily provide sufficient discriminative power. These genotyping approaches will be confounded
335 by the presence of (i) indistinguishable Lpn-SG1 genotypes present in unrelated cases and (ii) polyclonal
336 outbreaks. Our retrospective analysis of the Melbourne Aquarium outbreak illustrates clearly both these
337 issues, where five distinct subtypes were recovered from 25 clinical and environmental isolates (Fig. 1C and
338 1D). There is a growing awareness of single source, polyphyletic Lpn-SG1 outbreaks ^{8,10,37,38}. These data all
339 point to the need for a different approach in order to use molecular epidemiology and genomics in support
340 of *Legionella* outbreak investigations.

341

342 We address this issue by exploiting all core genome information to train probabilistic classification models.
343 Our DAPC analysis demonstrates that it is possible to build predictive models based on Lpn-SG1
344 environmentally derived genomes that help in identifying the source of clinical isolates during complex
345 outbreak investigations in both the community and hospital environments (Fig. 3). By including all core SNP
346 variation, DAPC was able to identify outbreak-specific genotypes, even when the source of the outbreak
347 was polyclonal. This enabled us to build robust models that assigned validation set genomes, with known
348 provenance, back to their original groupings with high concordance. The fact that this model was built
349 purely from environmental surveillance isolates demonstrates that such approaches can be developed
350 prospectively and be preexisting, ready to deploy at the onset of outbreaks.

351

352 In contrast to the high performance of the DAPC model developed from core genome SNPs, the model built
353 using variants identified by cgMLST scheme had a lower matching rate when assigning validation genomes
354 back to their putative epidemiological groupings (Fig. 3). Despite cgMLST being a useful tool for broad Lpn-
355 SG1 population structure assessment, our analysis suggests it has insufficient resolution and thus predictive
356 capacity for outbreak investigations.

357

358 The DAPC approach however, while promising, does not permit discrimination among isolates that do not
359 belong to defined clusters. This is because the model assumes that the world is composed of only the k
360 groups used to train it, and therefore assigns unknown isolates to one of these groups, even if the isolate is
361 known not to be part of any of the groups. One way to address this issue would be to create a single group
362 classifier, which is trained with environmental samples. Isolates with low probability of membership to this
363 single large group would then be excluded before being analyzed with the multi-group model. Future
364 models could be further improved by adding epidemiological evidence (e.g. patient zip codes), and assess
365 how that improves our assignment of a clinical isolate to a particular location. An advantage of a
366 classification-based model is that its output could be distilled down to a zip code (or group of zip codes) and
367 a probability that a clinical isolate is associated with the zip code (indicating uncertainty about the
368 classification). This would obviate the need to interpret, and explain phylogenetic trees. Interpreting trees
369 is often not intuitive and trees may fail to communicate what action is required from a public health
370 perspective. Crucial for such a classification approach to work however, is an extensive database of
371 environmental Lpn-SG1 genotypes. We are currently investigating how to implement such models.

372

373 From a biological perspective, the lack of genetic diversity in Lpn-SG1 over such coarse temporal and spatial
374 scales is potentially explained by a reservoir of latent-state bacteria intermittently seeding warm water
375 sources in the greater Melbourne region and is supported by the frequently-reported and widespread
376 presence of *Legionella* species in drinking water supply systems (DWSS)³⁹⁻⁴¹. Independent studies propose
377 similar hypotheses to explain the surprisingly high sequence conservation among some *L. pneumophila*
378 genomes^{8,12}.

379

380 This study is, to our knowledge, the largest genomic investigation of environmental and clinical *Legionella*
381 reported to date from a single jurisdiction and confirms that Lpn-SG1 is an unusual 'edge case' in the
382 application of genomics in public health microbiology. In the absence of a deep understanding of local *L.*
383 *pneumophila* population structure (both clinical and environmental) the combination of extreme genomic
384 monomorphism combined with outbreaks caused by mixed pathogen populations could easily lead to

385 erroneous conclusions regarding source attribution. Thus, we require new approaches that can better
386 utilize the genomic information available, and harmoniously combine it with epidemiological evidence, in
387 order to provide public health officials with useful and timely information.

388

389 References

- 390 1 Fields, B. S., Benson, R. F. & Besser, R. E. *Legionella* and Legionnaires' disease: 25 years of
391 investigation. *Clin Microbiol Rev* **15**, 506-526 (2002).
- 392 2 Yu, V. L. *et al.* Distribution of *Legionella* species and serogroups isolated by culture in patients with
393 sporadic community-acquired legionellosis: an international collaborative survey. *J Infect Dis* **186**,
394 127-128, doi:10.1086/341087 (2002).
- 395 3 Luck, C., Fry, N. K., Helbig, J. H., Jarraud, S. & Harrison, T. G. Typing methods for *Legionella*.
396 *Methods Mol Biol* **954**, 119-148, doi:10.1007/978-1-62703-161-5_6 (2013).
- 397 4 Mercante, J. W. & Winchell, J. M. Current and emerging *Legionella* diagnostics for laboratory and
398 outbreak investigations. *Clin Microbiol Rev* **28**, 95-133, doi:10.1128/CMR.00029-14 (2015).
- 399 5 McAdam, P. R. *et al.* Gene flow in environmental *Legionella pneumophila* leads to genetic and
400 pathogenic heterogeneity within a Legionnaires' disease outbreak. *Genome Biol* **15**, 504,
401 doi:10.1186/PREACCEPT-1675723368141690 (2014).
- 402 6 Reuter, S. *et al.* A pilot study of rapid whole-genome sequencing for the investigation of a
403 *Legionella* outbreak. *BMJ Open* **3**, doi:10.1136/bmjopen-2012-002175 (2013).
- 404 7 Sanchez-Buso, L., Comas, I., Jorques, G. & Gonzalez-Candelas, F. Recombination drives genome
405 evolution in outbreak-related *Legionella pneumophila* isolates. *Nat Genet* **46**, 1205-1211,
406 doi:10.1038/ng.3114 (2014).
- 407 8 David, S. *et al.* Seeding and Establishment of *Legionella pneumophila* in Hospitals: Implications for
408 Genomic Investigations of Nosocomial Legionnaires' Disease. *Clin Infect Dis* **64**, 1251-1259,
409 doi:10.1093/cid/cix153 (2017).
- 410 9 Weiss, D. *et al.* A Large Community Outbreak of Legionnaires' Disease Associated With a Cooling
411 Tower in New York City, 2015. *Public Health Rep* **132**, 241-250, doi:10.1177/0033354916689620
412 (2017).
- 413 10 Sanchez-Buso, L. *et al.* Genomic Investigation of a Legionellosis Outbreak in a Persistently Colonized
414 Hotel. *Front Microbiol* **6**, 1556, doi:10.3389/fmicb.2015.01556 (2015).

- 415 11 Underwood, A. P., Jones, G., Mentasti, M., Fry, N. K. & Harrison, T. G. Comparison of the *Legionella*
416 *pneumophila* population structure as determined by sequence-based typing and whole genome
417 sequencing. *BMC Microbiol* **13**, 302, doi:10.1186/1471-2180-13-302 (2013).
- 418 12 David, S. *et al.* Multiple major disease-associated clones of *Legionella pneumophila* have emerged
419 recently and independently. *Genome Res In press*, doi:10.1101/gr.209536.116 (2016).
- 420 13 Anon. Communicable Disease Surveillance, Victoria, Oct-Dec 2014. *Vic Infect Dis Bull* **17**, 22-23
421 (2014).
- 422 14 Greig, J. E. *et al.* An outbreak of Legionnaires' disease at the Melbourne Aquarium, April 2000:
423 investigation and case-control studies. *Med J Aust* **180**, 566-572 (2004).
- 424 15 Gaia, V. *et al.* Consensus sequence-based scheme for epidemiological typing of clinical and
425 environmental isolates of *Legionella pneumophila*. *J Clin Microbiol* **43**, 2047-2052,
426 doi:10.1128/JCM.43.5.2047-2052.2005 (2005).
- 427 16 Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069,
428 doi:10.1093/bioinformatics/btu153 (2014).
- 429 17 Cazalet, C. *et al.* Evidence in the *Legionella pneumophila* genome for exploitation of host cell
430 functions and high genome plasticity. *Nat Genet* **36**, 1165-1173, doi:10.1038/ng1447 (2004).
- 431 18 Alikhan, N. F., Petty, N. K., Ben Zakour, N. L. & Beatson, S. A. BLAST Ring Image Generator (BRIG):
432 simple prokaryote genome comparisons. *BMC Genomics* **12**, 402, doi:10.1186/1471-2164-12-402
433 (2011).
- 434 19 D'Auria, G., Jimenez-Hernandez, N., Peris-Bondia, F., Moya, A. & Latorre, A. *Legionella pneumophila*
435 pangenome reveals strain-specific virulence factors. *BMC Genomics* **11**, 181, doi:10.1186/1471-
436 2164-11-181 (2010).
- 437 20 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for
438 large alignments. *PLoS One* **5**, e9490, doi:10.1371/journal.pone.0009490 (2010).
- 439 21 Cheng, L., Connor, T. R., Siren, J., Aanensen, D. M. & Corander, J. Hierarchical and spatially explicit
440 clustering of DNA sequences with BAPS software. *Mol Biol Evol* **30**, 1224-1228,
441 doi:10.1093/molbev/mst028 (2013).
- 442 22 Didelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in whole bacterial
443 genomes. *PLoS Comput Biol* **11**, e1004041, doi:10.1371/journal.pcbi.1004041 (2015).

- 444 23 Parks, D. H. *et al.* GenGIS 2: geospatial analysis of traditional and genetic biodiversity, with new
445 gradient algorithms and an extensible plugin framework. *PLoS One* **8**, e69885,
446 doi:10.1371/journal.pone.0069885 (2013).
- 447 24 Jombart, T. & Ahmed, I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data.
448 *Bioinformatics* **27**, 3070-3071, doi:10.1093/bioinformatics/btr521 (2011).
- 449 25 Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: a new method
450 for the analysis of genetically structured populations. *BMC Genet* **11**, 94, doi:10.1186/1471-2156-
451 11-94 (2010).
- 452 26 Rao, C. *et al.* Active and Adaptive *Legionella* CRISPR-Cas reveals a recurrent challenge to the
453 pathogen. *Cell Microbiol*, doi:10.1111/cmi.12586 (2016).
- 454 27 Coscolla, M. & Gonzalez-Candelas, F. Population structure and recombination in environmental
455 isolates of *Legionella pneumophila*. *Environ Microbiol* **9**, 643-656, doi:10.1111/j.1462-
456 2920.2006.01184.x (2007).
- 457 28 Coscolla, M. & Gonzalez-Candelas, F. Comparison of clinical and environmental samples of
458 *Legionella pneumophila* at the nucleotide sequence level. *Infect Genet Evol* **9**, 882-888,
459 doi:10.1016/j.meegid.2009.05.013 (2009).
- 460 29 Gomez-Valero, L. *et al.* Extensive recombination events and horizontal gene transfer shaped the
461 *Legionella pneumophila* genomes. *BMC Genomics* **12**, 536, doi:10.1186/1471-2164-12-536 (2011).
- 462 30 Costa, J., Teixeira, P. G., d'Avo, A. F., Junior, C. S. & Verissimo, A. Intragenic recombination has a
463 critical role on the evolution of *Legionella pneumophila* virulence-related effector sidJ. *PLoS One* **9**,
464 e109840, doi:10.1371/journal.pone.0109840 (2014).
- 465 31 McNally, A. *et al.* Combined Analysis of Variation in Core, Accessory and Regulatory Genome
466 Regions Provides a Super-Resolution View into the Evolution of Bacterial Populations. *PLoS Genet*
467 **12**, e1006280, doi:10.1371/journal.pgen.1006280 (2016).
- 468 32 David, S. *et al.* Evaluation of an optimal epidemiologic typing scheme for *Legionella pneumophila*
469 with whole genome sequence data using validation guidelines. *J Clin Microbiol*,
470 doi:10.1128/JCM.00432-16 (2016).
- 471 33 Moran-Gilad, J. *et al.* Design and application of a core genome multilocus sequence typing scheme
472 for investigation of Legionnaires' disease incidents. *Euro Surveill* **20** (2015).
- 473 34 Bartley, P. B. *et al.* Hospital-wide Eradication of a Nosocomial *Legionella pneumophila* Serogroup 1
474 Outbreak. *Clin Infect Dis* **62**, 273-279, doi:10.1093/cid/civ870 (2016).

- 475 35 Kwong, J. C. *et al.* Prospective Whole-Genome Sequencing Enhances National Surveillance of
476 *Listeria monocytogenes*. *J Clin Microbiol* **54**, 333-342, doi:10.1128/JCM.02344-15 (2016).
- 477 36 Graham, R. M., Doyle, C. J. & Jennison, A. V. Real-time investigation of a *Legionella pneumophila*
478 outbreak using whole genome sequencing. *Epidemiol Infect* **142**, 2347-2351,
479 doi:10.1017/S0950268814000375 (2014).
- 480 37 Sanchez-Buso, L. *et al.* Phylogenetic analysis of environmental *Legionella pneumophila* isolates from
481 an endemic area (Alcoy, Spain). *Infect Genet Evol* **30**, 45-54, doi:10.1016/j.meegid.2014.12.008
482 (2015).
- 483 38 Coscolla, M., Fernandez, C., Colomina, J., Sanchez-Buso, L. & Gonzalez-Candelas, F. Mixed infection
484 by *Legionella pneumophila* in outbreak patients. *Int J Med Microbiol* **304**, 307-313,
485 doi:10.1016/j.ijmm.2013.11.002 (2014).
- 486 39 Diederren, B. M., de Jong, C. M., Aarts, I., Peeters, M. F. & van der Zee, A. Molecular evidence for
487 the ubiquitous presence of *Legionella* species in Dutch tap water installations. *J Water Health* **5**,
488 375-383, doi:10.2166/wh.2007.033 (2007).
- 489 40 Rivera, J. M. *et al.* Isolation of *Legionella* species/serogroups from water cooling systems compared
490 with potable water systems in Spanish healthcare facilities. *J Hosp Infect* **67**, 360-366,
491 doi:10.1016/j.jhin.2007.07.022 (2007).
- 492 41 Storey, M. V., Langmark, J., Ashbolt, N. J. & Stenstrom, T. A. The fate of legionellae within
493 distribution pipe biofilms: measurement of their persistence, inactivation and detachment. *Water*
494 *Sci Technol* **49**, 269-275 (2004).
- 495

496 **Figure Legends**

497

498 **Fig. 1: Global *Legionella pneumophila* population clustering, phylogenomics and genomic molecular**
499 **epidemiology of local outbreaks.** (A) Core genome phylogeny estimated using maximum likelihood
500 corresponds with six BAPS groups. Branches with less than 70% bootstrap support were collapsed and scale
501 indicates the number of core SNPs. The locations of the ten international genomes are labeled. (B) ST30
502 core genome phylogeny. Tree tips are labeled with outbreak codes. Environmental and clinical isolates are
503 colored according to the key. Polyclonal outbreaks/case clusters are highlighted with blue boxes. Branch
504 lengths have been transformed and are proportional to the number of nodes under each parent node.

505

506 **Fig. 2: Phylogeography of 64 Lpn-SG1 environmental isolate genomes.** Map of the greater Melbourne
507 area, showing the location of the 11 Legionellosis outbreaks, designated by colored circles. 'A' represents
508 the location Melbourne aquarium outbreak and is close to the centre of Melbourne. Inset shows the
509 location of Melbourne (red circle) within the State of Victoria in south east Australia. Overall the phylogeny
510 aligns closely with the geography of originating cooling towers. For several outbreak codes polyclonality is
511 apparent, as some common origins have connecting lines drawn from different sub-clades of the
512 phylogeny. Red coloration on the base map represents population density within the greater Melbourne
513 region. The branch lengths of the trees have been transformed and are proportional to the number of
514 nodes under each parent node.

515

516 **Fig. 3: Discriminatory analysis of principal components (DAPC) modeling of Lpn-SG1 genomic data.** (A)
517 Model comparison plots depicting the percentage of matches between the predicted and epidemiologically
518 determined groupings of the validation set genomes across a range of 1-20 principal components for single
519 nucleotide polymorphisms (SNP) and core genome MLST (cgMLST) DAPC models for the Melbourne and
520 Essex datasets. The retention of four and one principal components was found to be optimal for the SNP
521 (93% match) and cgMLST (60% match) models in Melbourne, respectively, while 15 and seven principal
522 components were found to be optimal for the SNP (86% match) and cgMLST (71% match) models in the
523 Essex hospital, respectively. (B) Assignment plots depicting the ability of the SNP models to predict the
524 source attribution of the validation set clinical isolate genomes for the Melbourne and Essex hospital
525 datasets. (C) Assignment plots depicting the ability of the cgMLST models to predict the source attribution
526 of the validation set clinical isolate genomes for the Melbourne and Essex hospital datasets.

527

Legionella pneumophila
MRCA

A

BAPS-5 Lp_Philadelphia
Lp_Thunder_Bay
Lp_ATCC_43290

Lp_Lens

BAPS-1

Lp_Lorraine

BAPS-6

Lp_HL06041035

BAPS-4

Lp_Paris

Lp_Toronto

BAPS-3

Lp_Alcoy

Lp_Corby

BAPS-2

10,000 SNPs

SBT30

B

K-2 A-2 B-1

K-1

A-1

B-2

D-1

D-2

A-3

C-1

C-2

A-5

A-4

■ Environmental isolate
■ Clinical isolate
■ Polyclonal clusters





