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

Andrew H. Buultjens[1,2], Kyra Y. L. Chua[3], Sarah L. Baines[1], Jason Kwong[1,2,3], Wei Gao[1], Zoe Cutcher[4,5], Stuart Adcock[4], Susan Ballard[3], Mark B. Schultz[3], Takehiro Tomita[3], Nela Subasinghe[3], Glen P. Carter[1,2], Sacha J. Pidot[1], Lucinda Franklin[4], Torsten Seemann[3,6], Anders Gonçalves Da Silva[2,3], Benjamin P. Howden[1,2,3,*], Timothy P. Stinear[1,2,*]

Affiliations:

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

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

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

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

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

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

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

23 **Running title:** *L. pneumophila* source tracking using genomics

Abstract

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

Importance

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

Introduction

Legionellae are Gram-negative bacteria that replicate within free-living aquatic amoebae and are present in aquatic environments worldwide. These bacteria can proliferate in man-made water systems and cause large outbreaks of pneumonia known as Legionnaires' disease when contaminated water is aerosolized and inhaled (1). The majority of human infections are caused by *Legionella pneumophila* serogroup 1 (2). Public health investigations of Legionnaires' disease outbreaks are typically supported by molecular typing methods to establish the likely source of the bacteria and the extent of the outbreak. Investigations usually proceed with the assumption that a single *Legionella* genotype is responsible for an environmental point source reservoir (3). Traditional molecular typing methods described for fingerprinting Legionellae include pulsed-field gel electrophoresis (PFGE) and sequence-based typing (SBT) (4). Increasingly, whole genome sequencing (WGS) is being employed to investigate individual *Legionella* outbreaks and the insights obtained from these high-resolution comparisons are challenging our expectations regarding common-source outbreaks, which usually are characterized by a single strain or genotype (5-9). It is becoming evident that outbreaks can be caused by multiple co-circulating *L. pneumophila* genotypes (5, 10) and that *L. pneumophila* core genomes can be surprisingly conserved across space and time (8, 11-13).

Melbourne is in the state of Victoria and it is the second largest city in Australia with a population approaching five million inhabitants, and considered the ninth largest city in the Southern Hemisphere. Legionellosis has been a notifiable disease in Victoria since 1979 and there are 50-100 cases reported each year, most occurring in the greater metropolitan region of Melbourne (14). The Microbiological Diagnostic Unit Public Health Laboratory (MDU PHL) is Victoria's State Reference Laboratory for the characterization and typing of *Legionella* spp. The laboratory's collection includes isolates from a particularly noteworthy outbreak at the Melbourne Aquarium in April 2000. This was the largest single episode of Legionellosis reported in Australia (15), approximately three months after the aquarium was opened to visitors, with construction of the site completed in December 1999. It resulted in 125 confirmed cases, with positive cultures obtained from 11 patients. Our isolate collection also spanned 28 other potential legionellosis outbreaks or infection clusters, for which at least one culture isolate had been obtained.

In this study, we used comparative genomics to explore the population structure of 234 *Legionella pneumophila* isolates recovered from human and environmental sources submitted to the MDU PHL in Melbourne over a 21-year period. This collection included 11 clinical and 14 environmental isolates from the Aquarium outbreak and 42 clinical and 50 environmental isolates from 28 other likely point source case clusters. We also assessed genomic data from a recently described investigation of Legionnaires' cases at a

89 UK hospital (8). The aim of this project was to develop a robust genomic approach that would surmount the
90 unusual population structure of
91 *L. pneumophila* and assist identification of case clusters and source tracking efforts during Legionnaires'
92 disease outbreak investigations.

93

94 Materials and Methods

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

108

109 **Sequence based typing.** This was performed as previously described according to the European
110 Legionnaires' Disease Surveillance Network (ELDSNet) method
111 (http://bioinformatics.phe.org.uk/legionella/legionella_sbt/php/sbt_homepage.php, accessed 31 August
112 2015) (16).

113

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

120 Biosciences) using P6-C4 chemistry. All sequence reads and the completed genome are available (GenBank
121 BioProject ID: PRJEB13594)

122

123 ***Legionella pneumophila* serogroup 1 isolate Lpm7613 assembly and closure.** A high quality finished ST 30
124 reference genome was established for *L. pneumophila* serogroup 1 clinical isolate Lpm7613 using the
125 SMRT® Analysis System v2.3.0.140936 (Pacific Biosciences). Raw sequence data were *de novo* assembled
126 using the HGAP v3 protocol with a genome size of 4 Mb. Polished contigs were error corrected using Quiver
127 v1. The resulting assembly was then checked using BridgeMapper v1 in the SMRT® Analysis System, and the
128 consensus sequence corrected with short-read Illumina data, using the program Snippy
129 (https://github.com/tseemann/snippy). Whole genome annotation was performed using Prokka (17),
130 preferentially using the *L. pneumophila* Paris strain annotation (18). BRIG was used to visualize BLASTn
131 DNA:DNA comparisons of *L. pneumophila* Lpm7613 against other *L. pneumophila* genomes (19).
132 Nomenclature of the genomic islands demonstrated in *L. pneumophila* Lpm7613 was based on previously
133 described islands (20). CRISPR databases were used to search for CRISPR sequences
134 (http://crispi.genouest.org and http://crispr.u-psud.fr/Server/, accessed 14 February 2016).

135

136 **Variant detection and phylogenetic analysis.** The genomes of ten publicly available complete *L.*
137 *pneumophila* genomes (Table S1) were shredded to generate short *in silico* sequence reads of 250bp and all
138 244 *L. pneumophila* reads sets were mapped against the Lpm7613 reference genome using Snippy v3.2. An
139 alignment file from pairwise comparisons of core genome SNPs (with inferred recombining sites removed)
140 was used as input to FastTree v2.1.8 with double precision (21) to infer a maximum likelihood phylogenetic
141 tree using the general time reversible model of nucleotide substitution. Branch support was estimated
142 using 1,000 bootstrap replicates. Resulting trees were visualized in FigTree v1.4.2
143 (http://tree.bio.ed.ac.uk/software/figtree/). Single nucleotide polymorphism (SNP) differences between
144 isolates were tabulated and visualized using a custom R-script (https://github.com/MDU-
145 PHL/pairwise_snp_differences). The core genome SNPs were also used as the input into a Bayesian analysis
146 of population structure (BAPS) using iterative clustering to a depth of 10 levels and a pre-specified
147 maximum of 20 clusters (22).

148

149 **Recombination and molecular clock analysis.** Recombination detection was performed using
150 ClonalFrameML (23), taking as input a full genome alignment (included invariant sites) prepared using
151 Snippy as above and the ML phylogeny as a guide tree with polytomies removed from the FastTree tree
152 using a custom python script (https://github.com/kwongj/nw_multi2bifurcation). Results were visualized

153 using a custom Python script to render separate and superposable images of extant and ancestral inferred
154 recombination regions (https://github.com/kwongj/cfml-maskrc). Molecular clock-likeness of the ML tree
155 with ClonalFrameML-adjusted branch lengths was assessed using TempEst v1.5
156 (http://tree.bio.ed.ac.uk/software/tempest/).

157

158 **Phylogeographic analysis.** Variant detection for the 64 environmental genomes was undertaken by running
159 snippy-core. Core SNPs were used to reconstruct a phylogenomic tree with FastTree that was overlaid upon
160 a base map in GenGIS (24). Victorian population mesh data was downloaded from the Australia Bureau of
161 Statistics webpage
162 (http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.001July%202016?OpenDocument)
163 and Local Government Area data was downloaded from the Victorian Government Data Directory webpage
164 (https://www.data.vic.gov.au/data/dataset/lga-geographical-profiles-2014-beta/resource/f6c49074-0679-
165 4c79-a0db-04dac8eda364).

166

167 **DAPC model building using core SNPs.** DAPC was implemented in the R package *adegenet* v2.0.1 (25). For
168 input, we used a matrix of single nucleotide polymorphisms (SNP) for all genomes originating from disease
169 clusters that possessed at least one environmental and clinical representative (Table S1). SNP detection was
170 undertaken by running Snippy and sites that were recombinogenic and or invariant among the
171 environmental genomes were discarded. The input matrix was used to generate a *genind* (data structure
172 class) object in *adegenet* that was subsequently separated into two subset *geninds*. The first subset
173 consisted of environmental isolates (hereon referred to as the training set) and the second *genind*
174 consisted of the clinical isolates from groups in common with validation set groups (hereon referred to as
175 the validation set). The training set was used to develop a DAPC model using the training set disease
176 clusters (Table S1) to define discriminant functions (26). The ability of the model to predict the source
177 attribution of the validation set genomes was simulated across the first to the 20th principal components,
178 allowing an optimal number of principal components to be identified. The optimized model was then used
179 to predict the groupings of validation set clinical isolate genomes.

180

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

186 genomes and used to predict the groupings of the validation set clinical isolate genomes.

187

188 ## Results

189 **Isolates and epidemiology.** There were 234 *Legionella pneumophila* serogroup 1 (Lpn-SG 1) isolates
190 obtained across a 21-year period between 1994 and 2014. Initial MLST analysis indicated that 180 isolates
191 (77%) belonged to ST30. The collection comprised 180 clinical isolates of respiratory origin (sputum or
192 bronchoscopy specimens) and 64 environmental isolates recovered from cooling tower water samples. All
193 isolates were collected in the state of Victoria with the exception of six isolates from patients who were
194 exposed elsewhere. Further information for each isolate is available in Table S1, including NCBI SRA
195 accession numbers. One hundred and ten of the 234 isolates were epidemiologically associated with 29
196 formally investigated case clusters or outbreaks, designated as outbreaks A-AC (Table S1). The majority of
197 these cases occurred within a 42 km radius of Melbourne city center and over a 16-year period. Outbreak
198 A, the Melbourne Aquarium outbreak, was the largest (15).

199

200 **Complete genome sequence of *Legionella pneumophila* serogroup 1 isolate Lpm7613.** Before this study,
201 there were no closed, fully assembled ST30 *L. pneumophila* genomes. Thus, to ensure identification of
202 maximum genetic variation among this dominant ST in our collection, we first established a ST30 reference
203 genome sequence, selecting a clinical isolate from the Melbourne Aquarium outbreak (Lpm7613). The
204 finished genome consisted of a single circular 3,261,562 bp chromosome (38.3% GC) and a 129,875 bp
205 circular plasmid (pLpm7613) (Fig. S1). Although the chromosome indicated this genome belonged to the
206 same lineage as *L. pneumophila* Philadelphia (Fig. 1A), the plasmid shared 100% nucleotide identity with
207 pLPP reported in *L. pneumophila* Paris, but 2kb shorter in length (18). A total of 2,891 chromosomal
208 protein-coding sequences (CDS), 43 tRNA genes and nine rRNA loci were predicted using Prokka (17).
209 CRISPR-Cas regions were not detected (27).

210

211 **Assessment of *L. pneumophila* population structure.** Sequence reads from the other 233 genomes and the
212 ten selected publicly available completed genomes were mapped to the chromosome of reference strain
213 Lpm7613. Approximately 90% of the Lpm7613 genome was present in all genomes (*i.e.*, core), with 188,049
214 variable core nucleotide positions identified. Population structure analyses using an unsupervised Bayesian
215 clustering approach revealed six distinct groups (BAPS groups) (Fig. 1A). Comparison of intra- and inter-
216 BAPS group pairwise SNP distances confirmed the validity of these clusters and highlighted the extensive
217 genetic variation among this Lpn-SG1 population (Fig. 1C). The exceptions were BAPS groups 3 and 4, which
218 classified isolates across two clades, and is likely explained by recombination. Most striking, however, was

219 the lack of diversity within the 186 genomes comprising BAPS group 5 (hereafter referred to as BAPS-5),
220 with a median core SNP distance of only 5 SNPs (IQR 3 – 7). Isolates dispersed across time and space
221 (including isolates from England, New South Wales, South Australia and Tasmania) were scattered
222 throughout the entire phylogeny. All 180 ST30 isolates were encompassed by BAPS-5, as was ST37 *L.*
223 *pneumophila* Philadelphia (Philadelphia, USA), ST211 *L. pneumophila* ATCC 43290 (Denver, USA) and ST733
224 *L. pneumophila* Thunder Bay (Ontario, Canada) (Fig. 1A,B). The median inter-BAPS group distances ranged
225 between 27,506 to 63,136 SNPs (Fig. 1C), highlighting that there is also substantial genetic diversity among
226 Lpn-SG1 isolates circulating in Melbourne.

227

228 A rooted maximum likelihood phylogeny of the population was then inferred using the 181,633 non-
229 recombining core SNP loci. The phylogenomic tree reflected the BAPS clusters with BAPS-5 forming a
230 distinct, well-supported lineage (Fig. 1A). The separation of the three North American reference isolates
231 from the Melbourne ST30 isolates is suggestive of contemporaneous global dispersal of this BAPS-5 lineage
232 (Fig. 1A,B). All BAPS groups displayed monophyletic origins with the exception of BAPS-3 and BAPS-4. BAPS-
233 3 had a single isolate of paraphyletic origin that shared a most recent common ancestor (MRCA) with BAPS-
234 2 while BAPS-4 contained two paraphyletic sub-clades, one of which shared a MRCA with the majority of
235 BAPS-3 isolates.

236

237 **Impact of recombination.** Recombination is a driving force in the evolution of the Legionellae (5, 7, 28-31).
238 Therefore, to further understand the structure and evolution of this Lpn-SG1 population we assessed the
239 impact of DNA exchange. There was evidence of extensive recombination among isolates across BAPS
240 groups 1-4, and 6 with approximately 3% of variable nucleotide sites impacted relative to the Lpm7613
241 reference chromosome. The detection of two paraphyletic groups (BAPS-3 and BAPS-4) is likely explained
242 by ancestral recombination among the component sub-clades. In comparison, there was little
243 recombination evident among BAPS-5 isolates (Fig. S2), in accord with the core SNP phylogeny described
244 above, and suggesting the relatively recent emergence of this *L. pneumophila* lineage. After removal of
245 putative sequences affected by recombination, the ClonalFrameML-adjusted tree branch lengths showed
246 no correlation with isolation dates (r^2=0.116). This observation indicates that nucleotide substitutions in the
247 population have not been evolving under a molecular clock model, thus limiting estimates for dates of
248 emergence for particular lineages.

249

250 **Genomic molecular epidemiology of local outbreaks.** We next compared only the 180 ST30 genomes to
251 our Lpm7613 reference genome, and again confirmed the very restricted genomic diversity within this

252 lineage (median core SNP distance was 6 SNPs (IQR 4 – 9), with five outlier genomes, impacted by
253 recombination. (Fig. 1B,C Fig. S2). Within this reconstructed core-genome ST30-specific phylogeny, many
254 but not all epidemiologically-related isolates formed distinct, well-supported, monophyletic clades. In some
255 instances, epidemiologically-associated isolates spanned multiple clades (outbreaks A, B, C, D and K) (Fig.
256 1B). In addition, Outbreak A (the Melbourne aquarium outbreak), which was previously considered to
257 represent infections caused by a single clone (Table S1) (15), actually contained five distinct genotypes (A1-
258 A5) (Fig 1B,C).

259

260 The analysis of environmental surveillance isolates provided an ideal means to gain insights into the
261 diversity within potential reservoirs of Lpn-SG1 - diversity that might enable prospective source tracking. A
262 phylogeographic analysis was therefore undertaken to assess the relationship between 64 environmental
263 Melbourne metropolitan isolates against their 11 cooling tower sampling locations. Based on variation in
264 core SNPs, striking geographical structure was observed, with the majority of isolates from common cooling
265 towers tightly clustering in the phylogeny (Fig. 2A). Comparisons of pairwise core SNPs depicted smaller
266 within group diversity and larger between location group diversity, further indicating the existence of
267 geographical population structure (Fig. 2B). This structure among the environmental Lpn-SG1 isolates
268 suggested it might be possible to use the genome data to build models predictive of environmental source
269 to assist epidemiological efforts during outbreak investigations.

270

271 **A multivariate statistical model for source attribution.** To enhance resolution and try to detect outbreak-
272 specific genomic signals, a supervised statistical learning approach called Discriminant Analysis of Principal
273 Components (DAPC) (26) was employed. DAPC is a linear discriminant analyses (LDA) that accommodates
274 discrete genetic-based predictors by first transforming the genetic data into continuous Principal
275 Components (PC) and builds predictive classification models. The PCs are used to build discriminant
276 functions (DF) under the constraint that they must minimize within group variance, and maximize variance
277 between groups. Infection clusters are specified *a priori* from the epidemiological findings, and training
278 isolates are used to define the discriminant functions. The model can then be used to estimate the
279 posterior probability of membership for unknown (*e.g.* clinical) isolates to each pre-specified infection
280 cluster given the training data. Here, we used 43 of the 64 environmental isolates as our training set
281 (cooling tower isolates originating from epidemiologically defined infection clusters that possessed at least
282 one environmental and clinical representative), under the assumption that each outbreak was caused by
283 exposure to a point source of Lpn-SG1. We used core genome SNPs from only environmental Lpn-SG1
284 genomes to build the classifier, in order to maximize our discriminatory power (32).

285

286 Outbreak-associated environmental Lpn-SG1 were grouped *a priori* into training set groups based on the
287 origin of the cooling towers from which they were isolated (see model building details in methods). The DFs
288 were then used to classify 15 clinical isolates that had been independently assigned based on
289 epidemiological data to the training set groups, hereon referred to as the validation genomes (Table S1).
290 The input matrix for DAPC was an alignment of 714 non-recombinogenic SNPs variable among the 43
291 environmental genomes. Plots depicting the separation of isolates according to the first two discriminant
292 functions and the amount of variation explained is shown (Fig. 3). A model was trained using the first four
293 principal components (PC), as this was found to be optimal (see methods). We next classified our clinical
294 validation genomes using the model and found a 93% match between our model's assignment and that
295 proposed by the epidemiological data (Fig. 4A,B). These data show that despite the high level of genome
296 conservation, it is possible to utilize signature differences in core genome SNPs to build predictive
297 probabilistic classification models. The single discrepancy between model predictions and epidemiological
298 groupings was an infection cluster C genome that was predicted as originating from the Aquarium.
299 Interestingly, cluster C was located closest to the Melbourne Aquarium at a distance of approximately 500
300 metres. Given the proximity of clusters A and C, these data may indicate cooling towers seeded from a
301 common *L. pneumophila* source. In order to appraise the utility of this method beyond a large urban setting
302 and the ST30 genotype, we built a sister model using 31 ST1 environmental *L. pneumophila* genomes from
303 a previously published hospital investigation in Essex, UK, and used it to predict the origins of seven
304 nosocomial clinical isolates (Fig. 3A, Table S1) (8). Here, the model was trained using the first 15 PCs, as this
305 was found to be optimal. As with the Melbourne disease clusters, the model performed very well. For 86%
306 of the clinical isolates there was a match between the model's ward assignment and the origin suggested
307 by epidemiology (Fig. 4A,B). Again, a single discrepancy occurred with a ward G genome predicted to
308 originate from ward A. Wards A and G were co-located on the same corner and level of a common building,
309 again suggesting a common *L. pneumophila* source (8). As before, isolates from a common source would be
310 miss-assigned by the model, owing to the lack of location-specific genomic variants.

311

312 **Core genome multilocus sequence typing (cgMLST) has reduced discrimination.** In order to evaluate the
313 utility of the recently described cgMLST scheme for source tracking (33, 34), we trained new DAPC models
314 for both the Melbourne and Essex hospital datasets using a matrix of allelic integers derived from SNP
315 profiles of the 1,529 cgMLST loci (Fig. 3B). When using the first one and seven PCs, we observed only 60%
316 and 71% concordance between our model's assignment and that predicted by the epidemiological data for
317 the Melbourne and Essex hospital datasets, respectively (Fig. 4A,C).

318

319 ## Discussion

320 In this study, we have retrospectively examined a large collection of 234 clinical and environmental isolates
321 Lpn-SG1 isolates spanning 29 defined outbreaks. Isolates were collected over wide temporal and spatial
322 scales and detailed genomic comparisons revealed wide extremes Lpn-SG1 genetic diversity among distinct
323 genomic populations; a phenomenon not fully appreciated from previous genomic investigations that have
324 sampled less extensively and focused on single outbreaks (5, 6, 35). Most striking in our collection was the
325 high sequence conservation and dominance of a single genotype (BAPS-5, ST30), shared by 77% of isolates
326 with a median core SNP distance of only 5 SNPs across 21 years. In agreement with our findings, two recent
327 population genomic investigations of Lpn-SG1 also describe the unusual restriction in core genome diversity
328 (8, 12).

329

330 Based on our previous experience with other bacterial pathogens (36) and reports in the literature of
331 Legionnaires' disease outbreak investigations using genomics (35, 37) - we expected to be able to use Lpn-
332 SG1 genomic comparisons and develop genetic rule-in or rule-out criteria to guide outbreak assessment
333 and source attribution. For example, we recently proposed a 'traffic-light' system for *Listeria*
334 *monocytogenes* based on SNP difference cutoffs of 'likely related', 'possibly-related' and 'not-related' (36).
335 This approach has also been proposed for *L. pneumophila* (33). A comparison of genotyping approaches
336 using 335 Lpn isolates, including 106 from the European Society for Clinical Microbiology Study Group's
337 *Legionella Typing Panel*, proposed an escalating, hierarchical approach to genotyping, beginning with an
338 extended 50-gene MLST scheme up to a 1529-gene cgMLST (33, 34).

339

340 The analysis of the population structure of Lpn-SG1 presented here indicates that SNP-based typing with
341 threshold cut-offs, whether they are based on seven genes, 50 genes, 1500 genes or whole genomes will
342 not necessarily provide sufficient discriminative power. These genotyping approaches will be confounded
343 by the presence of (i) indistinguishable Lpn-SG1 genotypes present in unrelated cases and (ii) polyclonal
344 outbreaks. Our retrospective analysis of the Melbourne Aquarium outbreak illustrates clearly both these
345 issues, where five distinct subtypes were recovered from 25 clinical and environmental isolates (Fig. 1B).
346 There is a growing awareness of single source, polyphyletic Lpn-SG1 outbreaks (8, 10, 13, 38, 39). These
347 data all point to the need for a different approach in order to use molecular epidemiology and genomics in
348 support of *Legionella* outbreak investigations.

349

350 We address this issue by exploiting all core genome information to train probabilistic classification models.
351 Our DAPC analysis demonstrates that it is possible to build predictive models based on Lpn-SG1
352 environmentally derived genomes that help in identifying the source of clinical isolates during complex

353 outbreak investigations in both the community and hospital environments (Fig. 4). By including all core SNP
354 variation, DAPC was able to identify outbreak-specific genotypes, even when the source of the outbreak
355 was polyclonal. This enabled us to build robust models that assigned validation set genomes, with known
356 provenance, back to their original groupings with high concordance. The fact that this model was built
357 purely from environmental surveillance isolates demonstrates that such approaches can be developed
358 prospectively and be preexisting, ready to deploy at the onset of outbreaks.

359

360 In contrast to the high performance of the DAPC model developed from core genome SNPs, the model built
361 using variants identified by cgMLST scheme had a lower matching rate when assigning validation genomes
362 back to their putative epidemiological groupings (Fig. 4C). Despite cgMLST being a useful tool for broad
363 Lpn-SG1 population structure assessment, our analysis suggests it may have insufficient resolution and thus
364 predictive capacity for outbreak investigations.

365

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

382

383 The modeling approach, is not intended to be used in isolation, but rather employed as an adjunct to
384 traditional epidemiological investigations. In this way, insights gained through epidemiological
385 investigations can be informed by microbiological evidence from our predictive models. A limitation of our

386 current models are the relatively small sample sizes. Performance measures for validation sets this small
387 are often sensitive to slight perturbations in the data and may be influenced by small features of the data.
388 However, as a proof-of-concept implementation of our approach, we have built two models from
389 independent datasets, and both demonstrate high predictive capacity. More robust appraisals of model
390 performance will require validation with larger datasets, collected prospectively.

391

392 From a biological perspective, the lack of genetic diversity in Lpn-SG1 over such coarse temporal and spatial
393 scales is potentially explained by a reservoir of latent-state bacteria intermittently seeding warm water
394 sources in the greater Melbourne region and is supported by the frequently-reported and widespread
395 presence of *Legionella* species in drinking water supply systems (DWSS) (40-42). Independent studies
396 propose similar hypotheses to explain the surprisingly high sequence conservation among some *L.*
397 *pneumophila* genomes (8, 12).

398

399 This study is, to our knowledge, the largest genomic investigation of environmental and clinical Legionella
400 reported to date from a single jurisdiction and confirms that Lpn-SG1 is an unusual 'edge case' in the
401 application of genomics in public health microbiology. In the absence of a deep understanding of local *L.*
402 *pneumophila* population structure (both clinical and environmental) the combination of extreme genomic
403 monomorphism combined with outbreaks caused by mixed pathogen populations could easily lead to
404 erroneous conclusions regarding source attribution. Thus, we require new approaches that can better
405 utilize the genomic information available, and harmoniously combine it with epidemiological evidence, in
406 order to provide public health officials with useful and timely information.

407

408 ## References

409 1. **Fields BS, Benson RF, Besser RE.** 2002. *Legionella* and Legionnaires' disease: 25 years of
410 investigation. Clin Microbiol Rev **15:**506-526.

411 2. **Yu VL, Plouffe JF, Pastoris MC, Stout JE, Schousboe M, Widmer A, Summersgill J, File T, Heath CM,**
412 **Paterson DL, Chereshsky A.** 2002. Distribution of *Legionella* species and serogroups isolated by
413 culture in patients with sporadic community-acquired legionellosis: an international collaborative
414 survey. J Infect Dis **186:**127-128.

415 3. **Luck C, Fry NK, Helbig JH, Jarraud S, Harrison TG.** 2013. Typing methods for *Legionella*. Methods
416 Mol Biol **954:**119-148.

417 4. **Mercante JW, Winchell JM.** 2015. Current and emerging *Legionella* diagnostics for laboratory and
418 outbreak investigations. Clin Microbiol Rev **28:**95-133.

419 5. **McAdam PR, Vander Broek CW, Lindsay DS, Ward MJ, Hanson MF, Gillies M, Watson M, Stevens**
420 **JM, Edwards GF, Fitzgerald JR.** 2014. Gene flow in environmental *Legionella pneumophila* leads to
421 genetic and pathogenic heterogeneity within a Legionnaires' disease outbreak. Genome Biol
422 **15:**504.

423 6. **Reuter S, Harrison TG, Koser CU, Ellington MJ, Smith GP, Parkhill J, Peacock SJ, Bentley SD, Torok**
424 **ME.** 2013. A pilot study of rapid whole-genome sequencing for the investigation of a *Legionella*
425 outbreak. BMJ Open **3**.

426 7. **Sanchez-Buso L, Comas I, Jorques G, Gonzalez-Candelas F.** 2014. Recombination drives genome
427 evolution in outbreak-related *Legionella pneumophila* isolates. Nat Genet **46:**1205-1211.

428 8. **David S, Afshar B, Mentasti M, Ginevra C, Podglajen I, Harris SR, Chalker VJ, Jarraud S, Harrison**
429 **TG, Parkhill J.** 2017. Seeding and Establishment of *Legionella pneumophila* in Hospitals:
430 Implications for Genomic Investigations of Nosocomial Legionnaires' Disease. Clin Infect Dis
431 **64:**1251-1259.

432 9. **Weiss D, Boyd C, Rakeman JL, Greene SK, Fitzhenry R, McProud T, Musser K, Huang L, Kornblum J,**
433 **Nazarian EJ, Fine AD, Braunstein SL, Kass D, Landman K, Lapierre P, Hughes S, Tran A, Taylor J,**
434 **Baker D, Jones L, Kornstein L, Liu B, Perez R, Lucero DE, Peterson E, Benowitz I, Lee KF, Ngai S,**
435 **Stripling M, Varma JK, South Bronx Legionnaires' Disease Investigation T.** 2017. A Large
436 Community Outbreak of Legionnaires' Disease Associated With a Cooling Tower in New York City,
437 2015. Public Health Rep **132:**241-250.

438 10. **Sanchez-Buso L, Guiral S, Crespi S, Moya V, Camaro ML, Olmos MP, Adrian F, Morera V, Gonzalez-**
439 **Moran F, Vanaclocha H, Gonzalez-Candelas F.** 2015. Genomic Investigation of a Legionellosis
440 Outbreak in a Persistently Colonized Hotel. Front Microbiol **6:**1556.

441 11. **Underwood AP, Jones G, Mentasti M, Fry NK, Harrison TG.** 2013. Comparison of the *Legionella*
442 *pneumophila* population structure as determined by sequence-based typing and whole genome
443 sequencing. BMC Microbiol **13:**302.

444 12. **David S, Rusniok C, Mentasti M, Gomez-Valero L, Harris SR, Lechat P, Lees, J., Ginerva C, Glaser C,**
445 **Ma L, Bouchier C, Underwood A, Jarraud S, Harrison TG, Parkhill J, Buchrieser C.** 2016. Multiple
446 major disease-associated clones of *Legionella pneumophila* have emerged recently and
447 independently. Genome Res **In press**.

448 13. **Schjorring S, Stegger M, Kjelso C, Lilje B, Bangsborg JM, Petersen RF, David S, Uldum SA,**
449 **Infections ESGfL.** 2017. Genomic investigation of a suspected outbreak of *Legionella pneumophila*
450 ST82 reveals undetected heterogeneity by the present gold-standard methods, Denmark, July to
451 November 2014. Euro Surveill **22**.

452 14. **Anon.** 2014. Communicable Disease Surveillance, Victoria, Oct-Dec 2014. Victorian Infectious
453 Diseases Bulletin **17:**22-23.

454 15. **Greig JE, Carnie JA, Tallis GF, Ryan NJ, Tan AG, Gordon IR, Zwolak B, Leydon JA, Guest CS, Hart**
455 **WG.** 2004. An outbreak of Legionnaires' disease at the Melbourne Aquarium, April 2000:
456 investigation and case-control studies. Med J Aust **180:**566-572.

457 16. **Gaia V, Fry NK, Afshar B, Luck PC, Meugnier H, Etienne J, Peduzzi R, Harrison TG.** 2005. Consensus
458 sequence-based scheme for epidemiological typing of clinical and environmental isolates of
459 *Legionella pneumophila*. J Clin Microbiol **43:**2047-2052.

460 17. **Seemann T.** 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics **30:**2068-2069.

461 18. **Cazalet C, Rusniok C, Bruggemann H, Zidane N, Magnier A, Ma L, Tichit M, Jarraud S, Bouchier C,**
462 **Vandenesch F, Kunst F, Etienne J, Glaser P, Buchrieser C.** 2004. Evidence in the *Legionella*
463 *pneumophila* genome for exploitation of host cell functions and high genome plasticity. Nat Genet
464 **36:**1165-1173.

465 19. **Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA.** 2011. BLAST Ring Image Generator (BRIG):
466 simple prokaryote genome comparisons. BMC Genomics **12:**402.

467 20. **D'Auria G, Jimenez-Hernandez N, Peris-Bondia F, Moya A, Latorre A.** 2010. *Legionella*
468 *pneumophila* pangenome reveals strain-specific virulence factors. BMC Genomics **11:**181.

469 21. **Price MN, Dehal PS, Arkin AP.** 2010. FastTree 2--approximately maximum-likelihood trees for large
470 alignments. PLoS One **5:**e9490.

471 22. **Cheng L, Connor TR, Siren J, Aanensen DM, Corander J.** 2013. Hierarchical and spatially explicit
472 clustering of DNA sequences with BAPS software. Mol Biol Evol **30:**1224-1228.

473 23. **Didelot X, Wilson DJ.** 2015. ClonalFrameML: efficient inference of recombination in whole bacterial
474 genomes. PLoS Comput Biol **11:**e1004041.

475 24. **Parks DH, Mankowski T, Zangooei S, Porter MS, Armanini DG, Baird DJ, Langille MG, Beiko RG.**
476 2013. GenGIS 2: geospatial analysis of traditional and genetic biodiversity, with new gradient
477 algorithms and an extensible plugin framework. PLoS One **8:**e69885.

478 25. **Jombart T, Ahmed I.** 2011. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data.
479 Bioinformatics **27:**3070-3071.

480 26. **Jombart T, Devillard S, Balloux F.** 2010. Discriminant analysis of principal components: a new
481 method for the analysis of genetically structured populations. BMC Genet **11:**94.

482 27. **Rao C, Guyard C, Pelaz C, Wasserscheid J, Bondy-Denomy J, Dewar K, Ensminger AW.** 2016. Active
483 and Adaptive *Legionella* CRISPR-Cas reveals a recurrent challenge to the pathogen. Cell Microbiol
484 doi:10.1111/cmi.12586.

485 28. **Coscolla M, Gonzalez-Candelas F.** 2007. Population structure and recombination in environmental
486 isolates of *Legionella pneumophila*. Environ Microbiol **9:**643-656.

487 29. **Coscolla M, Gonzalez-Candelas F.** 2009. Comparison of clinical and environmental samples of
488 *Legionella pneumophila* at the nucleotide sequence level. Infect Genet Evol **9:**882-888.

489 30. **Gomez-Valero L, Rusniok C, Jarraud S, Vacherie B, Rouy Z, Barbe V, Medigue C, Etienne J,**
490 **Buchrieser C.** 2011. Extensive recombination events and horizontal gene transfer shaped the
491 *Legionella pneumophila* genomes. BMC Genomics **12:**536.

492 31. **Costa J, Teixeira PG, d'Avo AF, Junior CS, Verissimo A.** 2014. Intragenic recombination has a critical
493 role on the evolution of *Legionella pneumophila* virulence-related effector sidJ. PLoS One
494 **9:**e109840.

495 32. **McNally A, Oren Y, Kelly D, Pascoe B, Dunn S, Sreecharan T, Vehkala M, Valimaki N, Prentice MB,**
496 **Ashour A, Avram O, Pupko T, Dobrindt U, Literak I, Guenther S, Schaufler K, Wieler LH, Zhiyong Z,**
497 **Sheppard SK, McInerney JO, Corander J.** 2016. Combined Analysis of Variation in Core, Accessory
498 and Regulatory Genome Regions Provides a Super-Resolution View into the Evolution of Bacterial
499 Populations. PLoS Genet **12:**e1006280.

500 33. **David S, Mentasti M, Tewolde R, Aslett M, Harris SR, Afshar B, Underwood A, Fry NK, Parkhill J,**
501 **Harrison TG.** 2016. Evaluation of an optimal epidemiologic typing scheme for *Legionella*
502 *pneumophila* with whole genome sequence data using validation guidelines. J Clin Microbiol
503 doi:10.1128/JCM.00432-16.

504 34. **Moran-Gilad J, Prior K, Yakunin E, Harrison TG, Underwood A, Lazarovitch T, Valinsky L, Luck C,**
505 **Krux F, Agmon V, Grotto I, Harmsen D.** 2015. Design and application of a core genome multilocus
506 sequence typing scheme for investigation of Legionnaires' disease incidents. Euro Surveill **20**.

507 35. **Bartley PB, Ben Zakour NL, Stanton-Cook M, Muguli R, Prado L, Garnys V, Taylor K, Barnett TC,**
508 **Pinna G, Robson J, Paterson DL, Walker MJ, Schembri MA, Beatson SA.** 2016. Hospital-wide

509 Eradication of a Nosocomial *Legionella pneumophila* Serogroup 1 Outbreak. Clin Infect Dis **62:**273-
510 279.

511 36. **Kwong JC, Mercoulia K, Tomita T, Easton M, Li HY, Bulach DM, Stinear TP, Seemann T, Howden**
512 **BP.** 2016. Prospective Whole-Genome Sequencing Enhances National Surveillance of *Listeria*
513 *monocytogenes*. J Clin Microbiol **54:**333-342.

514 37. **Graham RM, Doyle CJ, Jennison AV.** 2014. Real-time investigation of a *Legionella pneumophila*
515 outbreak using whole genome sequencing. Epidemiol Infect **142:**2347-2351.

516 38. **Sanchez-Buso L, Olmos MP, Camaro ML, Adrian F, Calafat JM, Gonzalez-Candelas F.** 2015.
517 Phylogenetic analysis of environmental *Legionella pneumophila* isolates from an endemic area
518 (Alcoy, Spain). Infect Genet Evol **30:**45-54.

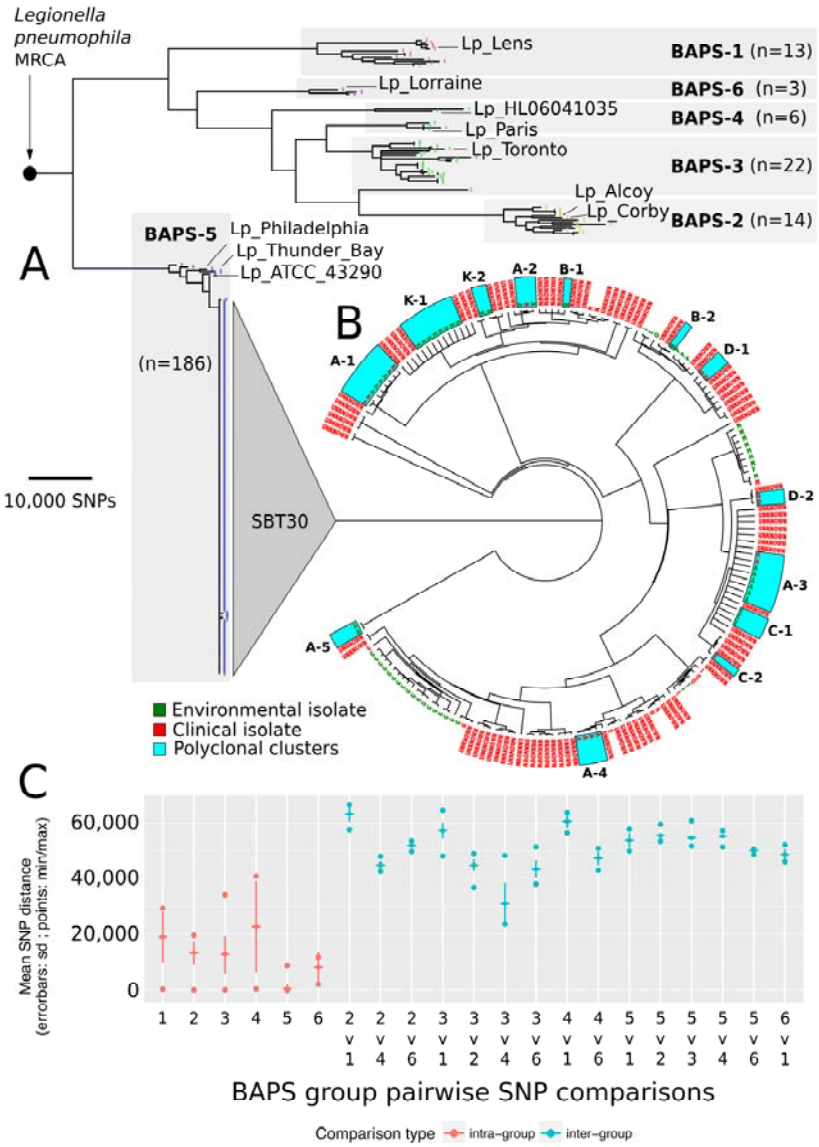
519 39. **Coscolla M, Fernandez C, Colomina J, Sanchez-Buso L, Gonzalez-Candelas F.** 2014. Mixed infection
520 by *Legionella pneumophila* in outbreak patients. Int J Med Microbiol **304:**307-313.

521 40. **Diederen BM, de Jong CM, Aarts I, Peeters MF, van der Zee A.** 2007. Molecular evidence for the
522 ubiquitous presence of *Legionella* species in Dutch tap water installations. J Water Health **5:**375-
523 383.

524 41. **Rivera JM, Aguilar L, Granizo JJ, Vos-Arenilla A, Gimenez MJ, Aguiar JM, Prieto J.** 2007. Isolation of
525 *Legionella* species/serogroups from water cooling systems compared with potable water systems in
526 Spanish healthcare facilities. J Hosp Infect **67:**360-366.

527 42. **Storey MV, Langmark J, Ashbolt NJ, Stenstrom TA.** 2004. The fate of legionellae within distribution
528 pipe biofilms: measurement of their persistence, inactivation and detachment. Water Sci Technol
529 **49:**269-275.

530



Figures

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

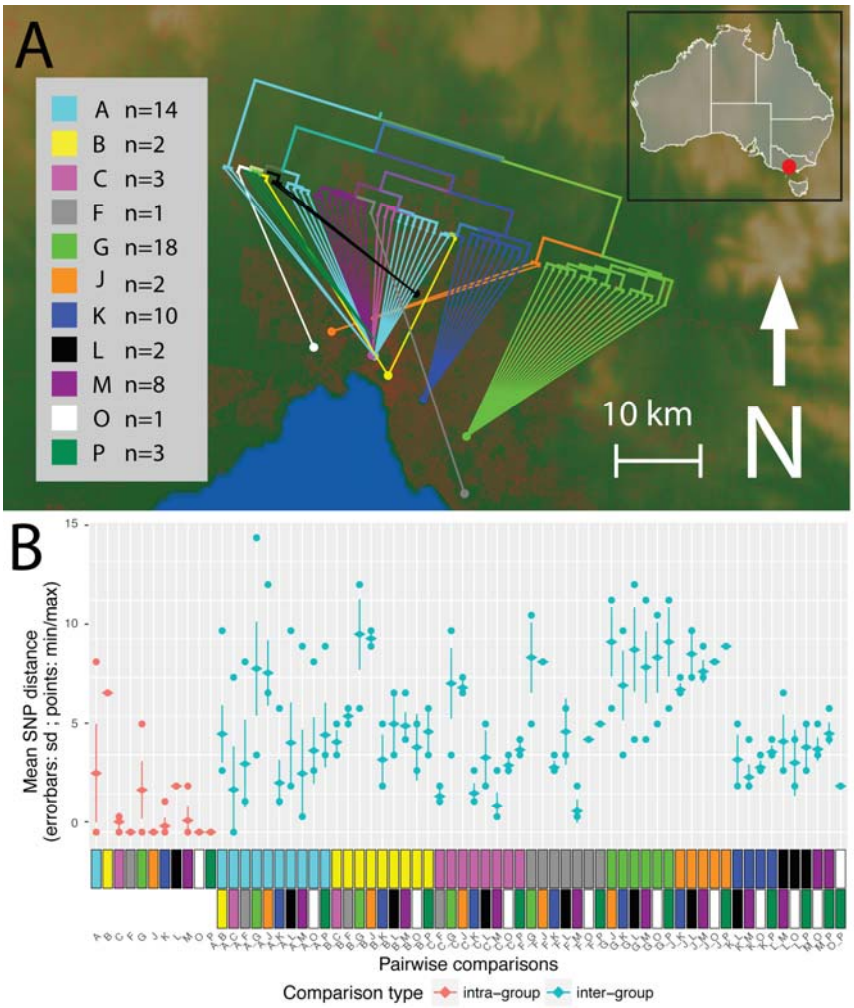


Fig. 2: **Phylogeography of 64 Lpn-SG1 environmental isolate genomes.** (A) Map of the greater Melbourne area, showing the location of the 11 cooling towers assessed during Legionellosis outbreaks, designated by colored circles. 'A' (light blue) represents the location Melbourne aquarium outbreak and is close to the centre of Melbourne. Inset shows the location of Melbourne (red circle) within the State of Victoria in south east Australia. Overall the phylogeny aligns closely with the geography of originating cooling towers. For several outbreak codes polyclonality is apparent, as some common origins have connecting lines drawn from different sub-clades of the phylogeny. Red coloration on the base map represents population density within the greater Melbourne region. The branch lengths of the trees have been transformed and are proportional to the number of nodes under each parent node. (B) Core genome pairwise SNP comparisons of within, and between, cooling tower isolate groups. Comparisons of specific epidemiologically defined groups are indicated with color codes as defined in the key. All groups had smaller within group diversity than between groups.

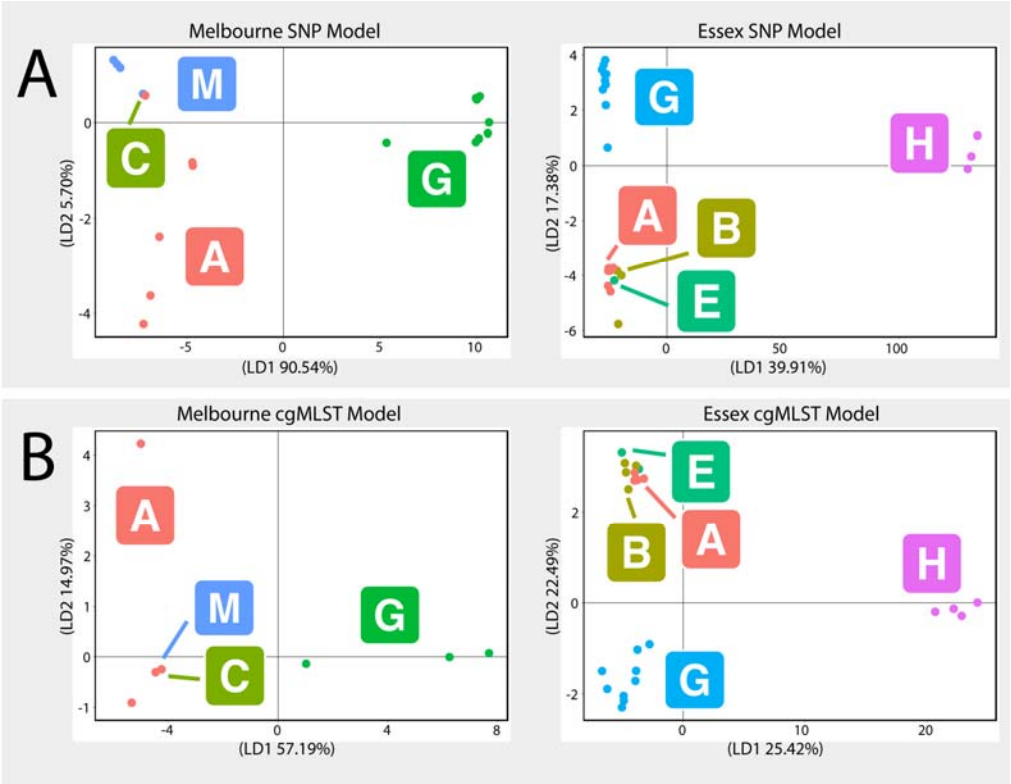


Fig. 3. Scatterplots resulting from discriminant analysis of principal components (DAPC). (A): Core genome single nucleotide polymorphisms (SNP) based models of the Melbourne (top left) and Essex (top right) datasets and; (B): core genome MLST (cgMLST) based models of the Melbourne (bottom left) and Essex (bottom right) datasets. The membership of each point within an epidemiologically defined group is indicated by the colored circles and the corresponding labeled squares. The amount of variation explained by the first and second discriminant functions are specified on the axes of each plot.

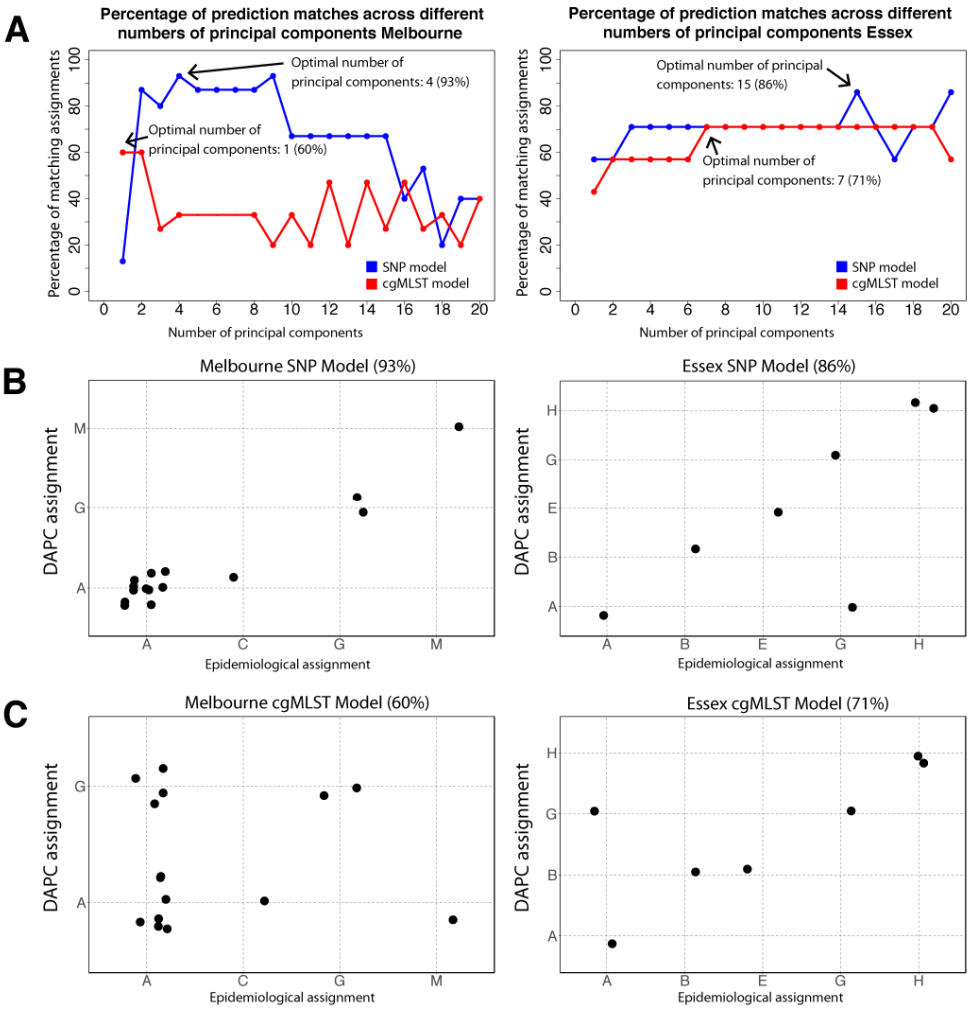


Fig. 4: **Discriminatory analysis of principal components (DAPC) modeling of Lpn-SG1 genomic data.** (A) Model comparison plots depicting the percentage of matches between the predicted and epidemiologically determined groupings of the validation set genomes across a range of 1-20 principal components for single nucleotide polymorphisms (SNP) and core genome MLST (cgMLST) DAPC models for the Melbourne and Essex datasets. The retention of four and one principal components was found to be optimal for the SNP (93% match) and cgMLST (60% match) models in Melbourne, respectively, while 15 and seven principal components were found to be optimal for the SNP (86% match) and cgMLST (71% match) models in the Essex hospital, respectively. (B) Assignment plots depicting the ability of the SNP models to predict the source attribution of the validation set clinical isolate genomes for the Melbourne and Essex hospital datasets. (C) Assignment plots depicting the ability of the cgMLST models to predict the source attribution of the validation set clinical isolate genomes for the Melbourne and Essex hospital datasets.