

1 **Candidatus Ornithobacterium hominis sp. nov.: insights gained from draft genomes obtained from**
2 **nasopharyngeal swabs.**

3

4 Susannah J Salter ^{1*}

5 Paul Scott ¹

6 Andrew J Page ^{1#a}

7 Alan Tracey ¹

8 Marcus C de Goffau ¹

9 Bernardo Ochoa-Montaña ^{2#b}

10 Clare L Ling ^{3,4}

11 Jiraporn Tangmanakit ³

12 Paul Turner ^{4,5}

13 Julian Parkhill ¹

14

15 ¹ Pathogen Genomics, Wellcome Sanger Institute, Hinxton, United Kingdom

16 ² Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom

17 ³ Shoklo Malaria Research Unit, Mahidol-Oxford Tropical Medicine Research Unit, Faculty of Tropical
18 Medicine, Mahidol University, Mae Sot, Thailand

19 ⁴ Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of
20 Oxford, Oxford, UK.

21 ⁵ Cambodia-Oxford Medical Research Unit, Angkor Hospital for Children, Siem Reap, Cambodia

22 #a Current Address: Quadram Institute Bioscience, Norwich, United Kingdom

23 #b Current Address: Illumina Cambridge Ltd, Little Chesterford, United Kingdom

24

25 * Corresponding author

26 Email: sb18@sanger.ac.uk (SJS)

27 ABSTRACT

28 *Candidatus Ornithobacterium hominis* sp. nov. represents a new member of the Flavobacteriaceae
29 detected in 16S rRNA gene surveys from Southeast Asia, Africa and Australia. It frequently colonises
30 the infant nasopharynx at high proportional abundance, and we demonstrate its presence in 42% of
31 nasopharyngeal swabs from 12 month old children in the Maela refugee camp in Thailand. The
32 species, a Gram negative bacillus, has not yet been cultured but the cells can be identified in mixed
33 samples by fluorescent hybridisation. Here we report seven genomes assembled from metagenomic
34 data, two to improved draft standard. The genomes are approximately 1.9Mb, sharing 62% average
35 amino acid identity with the only other member of the genus, the bird pathogen *Ornithobacterium*
36 *rhinotracheale*. The draft genomes encode multiple antibiotic resistance genes, competition factors,
37 *Flavobacterium johnsoniae*-like gliding motility genes and a homolog of the *Pasteurella multocida*
38 mitogenic toxin. Intra- and inter-host genome comparison suggests that colonisation with this
39 bacterium is both persistent and strain exclusive.

40

41 AUTHOR SUMMARY

42 The nasopharynx is part of the respiratory tract and hosts a unique microbial community that is
43 established during infancy, changing throughout life. The nasopharyngeal microbiome is important
44 to study as it includes bacteria that can cause diseases such as otitis media or pneumonia, as well as
45 non-pathogenic species. In Maela, a refugee camp in Thailand, we identified a prevalent bacterial
46 species colonising children under the age of two years and occasionally their mothers. We were not
47 able to culture it from frozen swabs, but could visualise the cells microscopically using a fluorescent
48 probe. Its genetic signature can be seen in published data from several countries suggesting that the
49 species may be widespread. From analysis of the genome we confirm it is highly divergent from its
50 closest characterised relative, the respiratory pathogen *Ornithobacterium rhinotracheale*, which
51 infects turkeys, chickens and other birds. We propose the name *Candidatus Ornithobacterium*
52 *hominis* sp. nov.

53

54 INTRODUCTION

55 During previous work on the nasopharyngeal microbiota of children in the MaeLa refugee camp in
56 Thailand, an abundant unclassified taxon was discovered through 16S rRNA gene sequencing [1]. It
57 was >99% identical to other unclassified sequences reported in nasopharyngeal samples from the
58 Gambia [2, 3], Kenya [4], and Australia [5], and the gene shared 93% nucleotide identity with that of
59 the avian respiratory pathogen *Ornithobacterium rhinotracheale* (ORT). On the basis of 16S rRNA
60 gene similarity the taxon was presumed to represent a new species of Flavobacteriaceae, closely
61 related to the genus *Ornithobacterium*. The taxon is of interest because it was ubiquitous in the
62 study group of 21 children, appearing to be a persistent coloniser and at a proportional abundance
63 up to 71%. The 16S rRNA gene sequences could be divided into three oligotypes [6]: each appeared
64 to be carried persistently and exclusively by their host [1].

65 As the bacterium could not be cultured from swabs, ten DNA samples from the initial study were
66 selected for metagenomic sequencing to maximise recovery of the genome of interest while
67 representing a range of children, ages, and 16S rRNA gene oligotypes: seven were successfully
68 sequenced. The extracted genomes were then used to design a PCR-based prevalence screen for
69 samples from the MaeLa cohort and a fluorescent probe to visualise the cells in mixed samples. On
70 the basis of this genomic analysis, we propose the unclassified taxon as *Candidatus*
71 *Ornithobacterium hominis* sp. nov. (OH).

72

73 RESULTS

74

75 Genomes of OH were assembled from metagenomic data generated on an Illumina Miseq. Despite
76 significant loss of sequence coverage to human and other bacterial genomes, two samples
77 assembled into 9 and 15 contigs from OH, yielding draft genomes predicted to be nearly complete
78 based on the detection of all ribosomal protein genes and by inter-sample comparison. A further five

79 samples assembled into larger numbers of smaller contigs, which were aligned to the draft genomes
 80 and found to cover most of the expected genome (Table 1). The OH genome is approximately
 81 1.9Mb, 20% smaller than its closest relative ORT.

82

83 **Table 1: Sequenced sample information; well-assembled draft genomes are highlighted in green.**

Child	Age (months)	Sample	Date	Genome ID (accession #)	No. extracted contigs	Total length	Max contig length	Proportional abundance in sample [1]
ARI0073	11	08B08946	Nov 2008	OH-22797 (ERS2321173)	302	1.81Mb	32kb	56%
	15	09B02844	Mar 2009	OH-22872 (ERS2321176)	41	1.93Mb	509kb	38%
	21	09B09452	Sep 2009	OH-22767 (ERS2321172)	15	1.93Mb	611kb	31%
ARI0106	20	09B08231	Aug 2009	OH-22298 (ERS2321170)	183	1.80Mb	68kb	67%
ARI0218	14	09B00559	Jan 2009	OH-22819 (ERS2321175)	172	1.42Mb	51kb	49%
	19	09B06140	Jun 2009	OH-22803 (ERS2321174)	9	1.87Mb	684kb	62%
ARI0484	15	09B08221	Aug 2009	OH-22763 (ERS2321171)	205	1.88Mb	57kb	51%

84

85

86 Prevalence

87 An OH specific real-time PCR detection protocol for V2-V5 of the 16S rRNA gene was designed using
 88 full length gene sequences and tested on archived nasopharyngeal swabs (NPS) in STGG (skim milk,
 89 tryptone, glucose and glycerol storage medium) and on metagenomic DNA of known bacterial
 90 composition. This PCR screen was then applied directly to STGG from archived NPS of 100 randomly
 91 selected 12-month old infants in Maela, and concurrent swabs from their mothers. A second PCR
 92 screen was developed targeting the toxin gene *toxA* and was also performed on the archived NPS.
 93 The two PCR targets were concordant in infant samples, resulting in 42 positive and 58 negative
 94 results, giving an estimated carriage prevalence among 12 month old infants in Maela of 42% (95%
 95 CI: 32.3-51.7). From the mothers 2 samples were positive, 93 negative, and 5 were either equivocal
 96 or nonconcordant (Table 2). The cycle threshold (Ct) values were higher in maternal samples than
 97 infants, which may indicate a lower bacterial load.

98

99 **Table 2: OH prevalence in mother and infant samples**

100

Infant or mother	No. samples	OH PCR: 16S target	OH PCR: toxin target
Infant	42	+	+
Infant	58	-	-
Total	100		
Mother	2	+	+
Mother	93	-	-
Mother	5	+/equivocal	-
Total	100		

101

102

103 Similarity to *O. rhinotracheale*

104 The draft genomes appear to be very distantly related to ORT by several measures, summarised in

105 Table 3. The average nucleotide identity (ANI) between OH-22767, OH-22803 and ORT UMN-88 [7]

106 can only be calculated from a small fraction of the genome. The two-way average amino acid

107 identity (AAI) between the two drafts and UMN-88 is approximately 62% based on three quarters of

108 predicted proteins. Another measure, the reciprocal percentage of conserved proteins (POCP) [8],

109 may be used to gauge the relatedness of two genomes at the genus level. To be considered

110 conserved for this measure a gene must share >40% amino acid identity over >50% of its length: two

111 members of the same genus are expected to have at least half of their proteins in common. The

112 POCP between UMN-88 and the draft genomes is approximately 58%. Although these figures are

113 based on incomplete genomes, as 50.7% of UMN-88 proteins are conserved in OH from these data,

114 they are likely to be very distantly related members of the same genus.

115

116 **Table 3: comparison of OH draft genomes with ORT genome UMN-88**

Draft genome	Average nucleotide identity (ANI)		Average amino acid identity (AAI)		Percentage conserved proteins (POCP)
OH-22767	81.7%	7.2% genome aligned	61.8%	75.0% proteins aligned	57.6% 50.7% ORT conserved 66.7% OH conserved
OH-22803	77.7%	5.8% genome aligned	61.5%	77.7% proteins aligned	58.1% 50.7% ORT conserved 68.2% OH conserved

117

118 Detection in swab material using fluorescent hybridisation

119 A fluorescent probe targeting the V5 region of the 16S ribosomal RNA was designed for OH using
120 full-length sequences of the 16S rRNA gene. It was used to visualise OH cells directly from
121 nasopharyngeal swab samples. OH is a Gram negative bacillus, often observed in pairs and
122 occasionally longer chains, similar to the morphology of ORT (Figure 1). ORT was tested in parallel
123 and could not be visualised with the same fluorescent probe.

124

125 **Figure 1: OH cells visualised with fluorescent ribosomal probe**

126

127 Core genome

128 The core genes shared between the draft genomes and lower quality assemblies, along with the
129 accessory genes unique to each, were calculated using Roary [9]. Due to the low quality assemblies
130 containing gaps, just 935kb or approximately 50% of the draft genome size was identified as “core
131 genome” using this analysis. A core genome phylogeny was generated using >13,000 SNPs identified
132 in this shared sequence [10]. Samples taken from the same child at different dates are very similar
133 (Figure 2), adding to the initial 16S rRNA gene oligotype data that inferred long-term carriage of the
134 same or closely related strains in this cohort. Features of interest that are present in the core
135 genome include a large toxin gene *toxA* and gliding motility-associated genes.

136

137 **Figure 2: Unrooted phylogenetic tree built from 935kb “core genome” sequence from high and low**
138 **quality assemblies using RAxML. Boxes represent groups based on accessory genome discussed**
139 **below.**

140

141 The 3.8kb gene *toxA* is present in all sequenced samples and was detected by PCR in all 16S-positive
142 swabs from children. It is predicted to produce a secreted toxin similar to the *Pasteurella* mitogenic

143 toxin (PMT). Although OH ToxA and PMT share only 35% amino acid identity overall, there is greater
144 conservation around the predicted active sites, and the modelled structure of the C-terminal domain
145 is extremely similar (Figure 3).

146

147 **Figure 3: Modelled structure of *Pasteurella* mitogenic toxin (PMT) C-terminal domain, left, and of**
148 **equivalent region of OH ToxA, right. The PMT model is coloured on a rainbow spectrum to indicate**
149 **position. The OH model is coloured according to amino acid identity to PMT with identical residues**
150 **in blue, non-identical in purple, and insertions in grey. In PMT the C1 region is responsible for**
151 **plasma membrane localisation, the C3 region forms an active pocket and is responsible for**
152 **mitogenic activity in mammalian cells.**

153

154 The core genome includes a full complement of 14 *gld* genes that are homologs of those required for
155 gliding motility by *Flavobacterium johnsoniae*. This mechanism involves the movement of an adhesin
156 around the cell membrane in a helical path, thereby pulling the bacterium rapidly along a substrate
157 [11]. Most of these genes in *F. johnsoniae* are also components of the Bacteroidetes type IX
158 secretion system (T9SS) [12, 13], which is responsible for secretion of *F. johnsoniae* gliding adhesins
159 SprB and RemA. However *sprB* has not been identified in OH.

160

161 Accessory genome

162 Around half of the accessory genome is made up of hypothetical protein genes, most of which are
163 similar to those of other members of the Flavobacteriaceae such as ORT, *Capnocytophaga*,
164 *Chryseobacterium*, *Elizabethkingia*, *Flavobacterium*, *Riemerella*, or *Weeksella*. It also includes
165 evidence of mobile elements and associated drug resistance genes, Rhs (rearrangement hotspot)
166 genes, and two distinct lipopolysaccharide production clusters.

167

168 A portion of the hypothetical protein genes encode the *Fibrobacter succinogenes* major paralogous
169 domain (PF09603). Up to 17 of these genes are present in each genome. A quarter of them are also
170 predicted to possess an immunoglobulin-like fold (IPR013783). No two samples have a complement
171 of completely identical predicted genes, but more are shared between samples that group together
172 based on the core genome (Figure 2).

173

174 The genomes contain different complements of bacteriophage associated genes, Type I, II and III
175 restriction modification systems, diverse variants of DNA degradation genes *dndCDE* [14], abortive
176 infection systems, transfer and mobilisation genes. In some cases these are present in small regions
177 of <10kb and lack any clear structure, while in others they are part of a defined element such as a
178 30kb tailed bacteriophage (22803_00899-00942), or a 19kb element (22803_01685-01710)
179 containing efflux genes and flanked by 350bp imperfect direct repeats. All sequenced samples
180 possess efflux transporters of the MATE and RND families, but genomes from cluster A (Figure 2)
181 have an additional B1 metallo-beta lactamase (22767_01182), streptogramin lyase (22767_01181)
182 and *ampC* gene (22767_01179) within a partially-assembled mobile element. Genomes from cluster
183 B possess an extended spectrum beta lactamase (ESBL) gene *per1* on the well-characterised
184 transposon Tn4555 [15].

185

186 The accessory genome of strains from cluster A encodes a number of elements containing the
187 conserved RHS repeat-associated core domain (TIGR03696) with extremely variable C termini and
188 unique hypothetical protein genes immediately downstream. There is one large Rhs gene of >9kb,
189 22767_01758, that encodes a protein sharing signatures with the *Salmonella* plasmid virulence
190 protein SpvB (IPR003284) and the bacterial insecticide toxin TcdB (IPR022045), while the other
191 smaller ORFs may be the dissociated tips generated from lateral acquisition of variable C termini as
192 seen in *Serratia marcescens* [16]. This large Rhs gene with displaced tips is only found in samples
193 from two children, ARI0106 and ARI0073: some but not all of the Rhs gene tips differ between them.

194 These samples also possess a further Rhs gene with no displaced tips, that has two predicted
195 phospholipase D domains (IPR001736). Some Rhs proteins have been shown to act as competition
196 factors [17], consistent with OH being a persistent member of the nasopharyngeal microbiota.

197

198 Although all genomes possess a 30kb lipopolysaccharide (LPS) production gene cluster, 11kb of this
199 region differs between samples from clusters A and B, containing a different complement of
200 transferases and synthases, and suggesting the presence of at least two serotypes in the species. The
201 A variant (22767_00486-00496) includes an ABC transporter and several transferase and synthase
202 genes 45-55% identical to those of ORT, while the B variant (22803_00491-00502) has a Wzx flippase
203 and genes that are more similar to *Chryseobacterium* or other Flavobacteriaceae.

204

205 DISCUSSION

206

207 Previously there has been little evidence available for OH colonising adults in Thailand or other
208 countries, as most nasopharyngeal microbiota studies target young children. By screening 100 pairs
209 of mother/child samples (Table 2), 2% of samples from mothers and 42% from infants were
210 unequivocally PCR-positive for OH. The prevalence at 12 months of age could be estimated as 32.3-
211 51.7% (95% CI) using this screen but the maternal prevalence could not be accurately measured
212 from this number of samples. The Ct values for mothers may indicate a lower bacterial load, leading
213 to a modest under-detection of colonisation by PCR. It contrasts with other bacteria such as
214 *Streptococcus pneumoniae* which is commonly carried in both adults and infants in MaeLa: among
215 these 100 pairs of swabs 31% from mothers and 79% from infants were culture positive for *S.*
216 *pneumoniae*.

217

218 Preliminary efforts have been made to culture the bacterium from nasopharyngeal swabs expected
219 to contain a large proportion of this species, following protocols appropriate for ORT or various

220 nasopharyngeal bacteria. OH has not yet been successfully cultured from any archived sample.
221 Furthermore, the frequency of fluorescent cells observed during ribosomal probing of mixed
222 samples was not as high as expected from the 16S analysis. This difficulty culturing the bacterium
223 may be due to unknown metabolic requirements, or it may be that the cells have not survived
224 storage due to environmental stresses or lysis.

225

226 In earlier work [1] it was noted that each child was colonised with only one of the three detected 16S
227 rRNA gene oligotypes representing OH. Colonisation was persistent, i.e. constituting >5% of the
228 proportional abundance of taxa for at least 5 consecutive months, in 13 out of 21 children but was
229 detected in all children at some point during the study. Here we describe multiple similar OH
230 genomes taken from time-points that are 4-10 months apart in two children, adding evidence to the
231 hypothesis that long-term colonisation is restricted to a particular strain for each host. Given the
232 high prevalence of OH in Maela (42% of 12 month old children) and the genetic diversity observed
233 between contemporaneous samples from only four hosts, this exclusion of diversity from the host
234 may be explained by microbial competition systems such as the SdpABC-like toxin system identified
235 in OH-22298 [18] or Rhs proteins [17]. Due to the high frequency of clinical pneumonia in Maela
236 (0.73 episodes per child year [19]), the children and their microbiota are frequently exposed to beta
237 lactam antibiotics. In all sequenced samples we found evidence of horizontally acquired drug
238 resistance genes, which may also aid persistent colonisation.

239

240 Bacterial LPS may confer advantages in adhesion and avoidance of complement mediated cell lysis,
241 although it is also a key target for the host immune system [20]. The gene content of LPS cluster
242 variant A is somewhat similar to that of the ORT serotype A (Figure 4), the most common of the 18
243 known serotypes [21] and the only ORT LPS type currently represented in public sequence
244 databases. Despite individual gene similarity to LPS clusters of several *Chryseobacterium* species

245 including *C. gallinarum* and *C. senegalense*, OH variant B in its entirety does not resemble that from
246 any known genomes.

247

248 **Figure 4: tblastx alignment of OH LPS variant A (OH-22767, top) and ORT serotype A (type strain**
249 **DSM 15997, bottom).**

250

251 Gliding motility is often associated with firm dry surfaces [22] though it is also found among oral
252 bacteria such as those from the genera *Cytophaga* and *Capnocytophaga*. The ability to move
253 independently in the environment is advantageous for scavenging nutrients, for complex biofilm
254 formation, and to bring about contact with other bacterial or host cells. The *gld* genes that are
255 required for gliding motility among the Flavobacteriia, Cytophagia and Sphingobacteriia overlap with
256 those of the T9SS, so a subset of these genes are also found in non-motile relatives [23]. Despite
257 possessing all 14 *gld* genes and further required genes *sprAET*, these OH genomes do not include an
258 SprB-like adhesin and putative gliding motility must be confirmed phenotypically.

259

260 PMT is a toxin produced by some serovars of *Pasteurella multocida* that causes a range of host
261 pathologies including nasal bone resorption [24], lower respiratory tract disease [25, 26], and
262 dermonecrotic wound infections [27], and has also been shown experimentally to affect the heart
263 [28], liver [29], and bladder [30]. It acts by deamidating the α subunits of several heterotrimeric G
264 proteins, activating mitogenic signalling pathways [31, 32]. In the Maela cohort there are no reports
265 of PMT-like toxin mediated disease, strongly suggesting that despite structural similarities the
266 expression or function of OH ToxA is different to that of *P. multocida*.

267

268 In conclusion, we have assembled seven genomes representing a new species of nasopharyngeal
269 bacteria, proposed as *Candidatus Ornithobacterium hominis* sp. nov., from nasopharyngeal swabs
270 collected from a cohort of children in the Maela refugee camp in Thailand. Although phenotypic

271 characterisation is not yet possible due to undetermined storage or culture requirements, several
272 points of interest have been identified for further investigation. These include the predicted gliding
273 motility phenotype, two lipopolysaccharide variants, and the production of a protein similar to the
274 *Pasteurella* mitogenic toxin. The prevalence of OH colonisation appears to be approximately 42% of
275 12 month old children in Maela refugee camp and needs to be estimated in other populations.

276

277 MATERIAL AND METHODS

278

279 Samples

280 Between 2007 and 2010, a cohort of 955 infants born in the Maela refugee camp on the Thailand-
281 Myanmar border were followed from birth until 24 months of age in a study of pneumococcal
282 colonisation and pneumonia epidemiology [19, 33]. Nasopharyngeal samples were collected from
283 each infant at monthly intervals using Dacron tipped swabs (Medical Wire & Equipment, Corsham,
284 UK). Immediately following collection, the NPS were placed into STGG transport medium and frozen
285 at -80°C within 8 hours. An additional nasopharyngeal swab was taken if the infant was diagnosed
286 with pneumonia according to WHO clinical criteria [34].

287

288 Ethics statement

289 Infants were eligible for inclusion in the cohort study if informed written consent had been obtained
290 from the mother during the antenatal period. Participants and their data were anonymised using a
291 4-digit code prefixed by "ARI-". The study protocol was reviewed and approved by the ethics
292 committees of the Faculty of Tropical Medicine, Mahidol University, Thailand (MUTM-2009-306) and
293 University of Oxford, UK (OXTREC-031-06).

294

295 Prevalence screen

296 Prevalence of carriage in the infant population of Maela was estimated using a qPCR screen direct
297 from NPS storage medium. The 12-month routine samples from 100 randomly selected infants
298 (excluding twins) were used. The age of child at time of sampling ranged from 359-377 days, median
299 365 days. Concurrent swabs were also acquired from the mothers and screened to assess maternal
300 carriage in relation to infant carriage. qPCR was performed on an Applied Biosystems 7500 RT PCR
301 machine using PowerUp SYBR Green mastermix (Applied Biosystems) in 20 μ l reaction volumes. The
302 16S rRNA gene screen targets the V2-V5 region with forward primer CTTATCGGGAGGATAGCCCG and
303 reverse GAAGTTCTTCACCCCGAAAACG, yielding a 700bp product under the conditions: 94°C for 5
304 minutes (cell lysis), then 40 cycles of 94°C for 30 seconds, 53°C for 30 seconds, 68°C for 1 minute,
305 ending with a melt curve. A positive result was a Ct<40 and peak melting temperature (T_m) of 80-
306 86°C. The ToxA gene screen with forward primer TATCTCTCACAGAGCTAGGCTTGAGCGTGG and
307 reverse TGCTATATTTGGGAAAGGCGCATGAATACC yields a 1.95kb product under the conditions: 94°C
308 for 5 minutes (cell lysis), then 40 cycles of 94°C for 30 seconds, 58°C for 2.5 minutes, 68°C for 2.5
309 minutes, ending with a melt curve. A positive result was Ct<40 and peak T_m of 77-79°C.
310 A positive result for both targets was interpreted as carriage-positive, a negative result for both
311 targets was interpreted as carriage-negative. Non-concordant results (positive/negative or
312 negative/positive) were treated as a separate group. This assessment of carriage prevalence may be
313 affected by several factors: recent antibiotic consumption, low microbial biomass, age under 6
314 months (as inferred from previous work [1]), or technical error during swab collection may lead to a
315 lower estimate. Presence of dead bacterial cells in the nasopharynx or cross-contamination during
316 sample handling may lead to false positives. The sample size of 100 was selected as adequate to
317 encompass a predicted prevalence of 10-90% with a precision of 5% and 95% CI [35]. This sample
318 size is approximately one tenth of the total population being estimated, i.e. all 12 month old children
319 born in Maela between 2007-2010.

320

321 Microscopy

322 A suspension of NPS STGG sample was fixed overnight at 4°C in 3% paraformaldehyde, and
323 dehydrated in suspension with 96% ethanol. Fluorescent hybridisation with an Alexa546 labelled
324 probe (Invitrogen) was performed on fixed sample in buffered suspension (20mM Tris-HCl, 0.9M
325 NaCl, 0.1% SDS) for 2 hours at 55°C, and washed with 20mM Tris-HCl, 0.9M NaCl for 5 minutes at
326 55°C. The samples were then suspended in water and applied to standard microscope slides, dried,
327 and a coverslip applied with Vectashield mounting medium (Vector Laboratories, California, USA).
328 The probe with sequence GUUCUUCACCCCGAAAACG targets the V5 region of the 16S ribosomal
329 RNA, corresponding approximately to position 822-840 of the 16S rRNA gene of *E. coli*. It was not
330 found to bind to an ORT sample when processed in parallel. A Zeiss Axiovert 200M fluorescence
331 microscope with Zeiss filter set 43 (Ex 545/25, FT 570, Em 605/70) was used to visualise probed cells.
332 A Sony DXC-390p colour video camera was used for imaging, which was recorded using Debut video
333 capture (v.4.04, NCH Software). Images were collected from captured video files using Camtasia
334 video processing software (v.8.6, TechSmith).

335

336 DNA extraction and sequencing

337 Preliminary work was performed on DNA extracted from swabs with FastDNA Spin Kit For Soil (MP
338 Biomedicals, Ohio, USA) which was then amplified using multiple displacement amplification (MDA).
339 MDA reagents were filtered at 0.2µm, endonuclease digested with φ29 enzyme, and UV irradiated
340 (254nm) prior to use to remove any exogenous DNA from the subsequent amplification reaction
341 [36]. 3µl of sample was heat denatured with 2µl Heat Denaturation Buffer (20mM Tris HCl pH 8.0,
342 2mM EDTA, and 400µM PTO random hexamers (Eurofins Genomics)) at 95°C for 3 minutes. 15µl of
343 Reaction Master Mix (1x RepliPHI reaction buffer, RepliPHI φ29 enzyme (6.7 U/µl) (Epicentre),
344 0.5mM dNTP, 50µM PTO random hexamers (Eurofins MWG), 5% DMSO, 10mM DTT) was then
345 added, samples were incubated at 30°C for 16 hours, and the reaction halted by final incubation at
346 65°C for 20 minutes. Eight separate reaction volumes were processed per sample, these parallel

347 reactions were pooled before sequencing using the Miseq 250bp paired end protocol with a 450bp
348 library fragment size.

349

350 Genome analysis

351 Samples were selected to maximise recovery of the genome of interest by choosing those with a
352 high proportional abundance of the bacterium from previous 16S rRNA gene sequencing data [1].
353 Raw reads from MDA-amplified samples were first classified using Kraken v. 0.10.6 [37] and parsed
354 to remove any reads classified as mammal, *Moraxella*, *Haemophilus* or *Streptococcus*. The remaining
355 reads were then assembled using SPAdes v. 3.10.0 [38]. Contigs shorter than 500bp or with average
356 coverage below 4X were discarded. For the two well-assembled samples, a BLAST+ v. 2.7.0 [39]
357 screen of all contigs against the nr database was used to discard those that closely matched other
358 known nasopharyngeal bacteria. The contigs first brought forward for the draft genomes were those
359 that had consistent, low-identity matches to ORT. Samples were then reciprocally compared using
360 BLAST+ to find further contigs present in all runs. These curated contig sets were manually improved
361 using Gap5 v. 1.2.14 [40] and targeted PCR for gap closure, resulting in two syntenic draft genomes.
362 The other assemblies with large numbers of short contigs were screened by comparison against the
363 draft genomes using BLAST+ and extracting all contigs with >10% length hit. Automated annotation
364 of curated contigs was performed using Prokka v. 1.11 [41] and the RefSeq database [42]. ANI and
365 AAI were calculated using the enveomics calculators [43, 44]. ANI used a minimum length 700bp and
366 minimum identity 70%, with 1000bp window size and 200bp step. AAI used a 20% identity cut-off.
367 POCP was calculated as described by Qin *et al* [8]. The core and accessory genomes were calculated
368 using Roary v. 3.11.3 [9]. Phylogenetic trees were built using RAxML v. 8.2.8 [45].

369

370 Protein modelling

371 The OH ToxA protein sequence was searched using the program FUGUE [46] against a database of all
372 chains of the Protein Data Bank (PDB) as of June 2017 [47]. Significant similarity with 30% sequence

373 identity was found for residues 554-1269 to chain X of PDB ID: 2EBF [48], with corresponds to the C-
374 terminal region of the *Pasteurella* mitogenic toxin. The matched region was aligned to the chain
375 sequence using FUGUE, and models were generated with MODELLER v.9.15 using "very slow"
376 refinement [49]. Visualization of the resulting models was performed on PyMOL Molecular Graphics
377 System v1.8.

378

379 Culture methods

380 NPS STGG samples were streaked out on blood or chocolate agar and incubated at 37°C in aerobic,
381 enriched CO₂ or anaerobic conditions for 48 hours. 25ml brain heart infusion (BHI) broths, with and
382 without 1µg/ml ampicillin, were inoculated with 5ul of STGG and incubated either static or shaking
383 at 37°C for up to one week. Other nasopharyngeal species were recovered from NPS STGG following
384 these methods, but OH was not.

385

386 Accession numbers

387 Accession numbers for the 7 samples are listed in Table 1, and are deposited under project accession
388 ERP107699 (ERS2321170-6). The ORT genomes used for comparison are UMN-88, accession
389 CP006828.1 and DSM 15997, accession CP003283.1.

390

391 ACKNOWLEDGEMENTS

392 The authors acknowledge the generosity of Paul Wigley and Rachel Gilroy at the University of
393 Liverpool for provision of an ORT strain as a negative control for fluorescent probe testing, the
394 support of the SMRU laboratory and clinical teams, the Sanger Institute Pathogen Informatics team
395 and the core sequencing and informatics teams.

396

397

- 398 1. Salter SJ, Turner C, Watthanaworawit W, de Goffau MC, Wagner J, Parkhill J, et al. A
399 longitudinal study of the infant nasopharyngeal microbiota: The effects of age, illness and antibiotic
400 use in a cohort of South East Asian children. *PLoS Neglected Tropical Diseases*.
401 2017;11(10):e0005975.
- 402 2. Kwambana BA, Mohammed NI, Jeffries D, Barrer M, Adegbola RA, Antonio M. Differential
403 effects of frozen storage on the molecular detection of bacterial taxa that inhabit the nasopharynx.
404 *BMC Clinical Pathology*. 2011;11:2.
- 405 3. Kwambana BA, Hanson B, Worwui A, Agbla S, Foster-Nyarko E, Ceesay F, et al. Rapid
406 replacement by non-vaccine pneumococcal serotypes may mitigate the impact of the pneumococcal
407 conjugate vaccine on nasopharyngeal bacterial ecology. *Scientific Reports*. 2017;7:8127.
- 408 4. Feazel LM, Santorico SA, Robertson CE, Bashraheil M, Scott JA, Frank DN, et al. Effects of
409 Vaccination with 10-Valent Pneumococcal Non-Typeable Haemophilus influenza Protein D Conjugate
410 Vaccine (PHiD-CV) on the Nasopharyngeal Microbiome of Kenyan Toddlers. *PLoS One*.
411 2015;10(6):e0128064.
- 412 5. Marsh RL, Kaestli M, Chang AB, Binks MJ, Pope CE, Hoffman LR, et al. The microbiota in
413 bronchoalveolar lavage from young children with chronic lung disease includes taxa present in both
414 the oropharynx and nasopharynx. *Microbiome*. 2016;4:37.
- 415 6. Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML. Minimum entropy
416 decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene
417 sequences. *The ISME Journal*. 2015;9:968-79.
- 418 7. Zehr ES, Bayles DO, Boatwright WD, Tabatabai LB, Register KB. Complete genome sequence
419 of *Ornithobacterium rhinotracheale* strain ORT-UMN 88. *Standards in Genomic Sciences*. 2014;9:16.
- 420 8. Qin QL, Xie BB, Zhang XY, Chen XL, Zhou BC, Zhou J, et al. A proposed genus boundary for the
421 prokaryotes based on genomic insights. *Journal of Bacteriology*. 2014;196(12):2210-5.
- 422 9. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-
423 scale prokaryote pan genome analysis. *Bioinformatics*. 2015;31(22):3691-3.
- 424 10. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites: rapid efficient
425 extraction of SNPs from multi-FASTA alignments. *Microbial Genomics*. 2016;2(4):e000056.
- 426 11. Nakane D, Sato K, Wada H, McBride MJ, Nakayama K. Helical flow of surface protein
427 required for bacterial gliding motility. *Proceedings of the National Academy of Sciences of the*
428 *United States of America*. 2013;110(27):11145-50.
- 429 12. Lasica AM, Ksiazek M, Madej M, Potempa J. The Type IX Secretion System (T9SS): Highlights
430 and Recent Insights into Its Structure and Function. *Frontiers in Cellular and Infection Microbiology*.
431 2017;7:215.
- 432 13. Johnston JJ, Shrivastava A, McBride MJ. Untangling *Flavobacterium johnsoniae* Gliding
433 Motility and Protein Secretion. *Journal of Bacteriology*. 2017;200(2):e00362-17.
- 434 14. Barbier P, Lunazzi A, Fujiwara-Nagata E, Avendaño-Herrera R, Bernardet J-F, Touchon M, et
435 al. From the *Flavobacterium* genus to the phylum *Bacteroidetes*: genomic analysis of *dnd* gene
436 clusters. *FEMS Microbiology Letters*. 2013;348(1):26-35.
- 437 15. Tribble GD, Parker AC, Smith CJ. The *Bacteroides* mobilizable transposon Tn4555 integrates
438 by a site-specific recombination mechanism similar to that of the gram-positive bacterial element
439 Tn916. *Journal of Bacteriology*. 1997;179(8):2731-9.
- 440 16. Jackson AP, Thomas GH, Parkhill J, Thomson NR. Evolutionary diversification of an ancient
441 gene family (*rhs*) through C-terminal displacement. *BMC Genomics*. 2009;10:584.
- 442 17. Koskiniemi S, Lamoureux JG, Nikolakakis KC, t'Kint de Roodenbeke C, Kaplan MD, Low DA, et
443 al. *Rhs* proteins from diverse bacteria mediate intercellular competition. *PNAS*. 2013;110(17):7032-7.
- 444 18. Pérez Morales TG, Ho TD, Liu WT, Dorrestein PC, Ellermeier CD. Production of the
445 cannibalism toxin SDP is a multistep process that requires SdpA and SdpB. *Journal of Bacteriology*.
446 2013;195(14):3244-51.

- 447 19. Turner C, Turner P, Carrara V, Burgoine K, Tha Ler Htoo S, Watthanaworawit W, et al. High
448 rates of pneumonia in children under two years of age in a South East Asian refugee population.
449 PLoS One. 2013;8(1):e54026.
- 450 20. Lerouge I, Vanderleyden J. O-antigen structural variation: mechanisms and possible roles in
451 animal/plant–microbe interactions. FEMS Microbiology Reviews. 2006;26(1):17-47.
- 452 21. Thieme S, Mühldorfer K, Lüschow D, Hafez HM. Molecular Characterization of the Recently
453 Emerged Poultry Pathogen *Ornithobacterium rhinotracheale* by Multilocus Sequence Typing. PLoS
454 One. 2016;11(2):e0148158.
- 455 22. Nan B, Zusman DR. Novel mechanisms power bacterial gliding motility. Molecular
456 Microbiology. 2016;101(2):186-93.
- 457 23. McBride MJ, Zhu Y. Gliding Motility and Por Secretion System Genes Are Widespread among
458 Members of the Phylum Bacteroidetes. Journal of Bacteriology. 2013;195(2):270-8.
- 459 24. Chakraborty S, Kloos B, Harre U, Schett G, Kubatzky KF. *Pasteurella multocida* toxin triggers
460 RANKL-independent osteoclastogenesis. Frontiers in Immunology. 2017;8:185.
- 461 25. Klein NC, Cunha BA. *Pasteurella multocida* pneumonia. Seminars in Respiratory Infections.
462 1997;12(1):54-6.
- 463 26. Pijoan C, Lastra A, Ramirez C, Leman AD. Isolation of toxigenic strains of *Pasteurella*
464 *multocida* from lungs of pneumonic swine. Journal of the American Veterinary Medical Association.
465 1984;185(5):522-3.
- 466 27. Wilson BA, Ho M. *Pasteurella multocida* toxin interaction with host cells: entry and cellular
467 effects. Current Topics in Microbiology and Immunology. 2012;361:93-111.
- 468 28. Weise M, Vettel C, Spiger K, Gilsbach R, Hein L, Lorenz K, et al. A systemic *Pasteurella*
469 *multocida* toxin aggravates cardiac hypertrophy and fibrosis in mice. Cellular microbiology.
470 2015;17(9):1320-31.
- 471 29. Chevillon NF, Rimler RB. A protein toxin from *Pasteurella multocida* type D causes acute and
472 chronic hepatic toxicity in rats. Veterinary Pathology. 1989;26(2):148-57.
- 473 30. Hoskins IC, Thomas LH, Lax AJ. Nasal infection with *Pasteurella multocida* causes
474 proliferation of bladder epithelium in gnotobiotic pigs. The Veterinary Record. 1997;140(1):22.
- 475 31. Orth JH, Aktories K. Molecular biology of *Pasteurella multocida* toxin. Current Topics in
476 Microbiology and Immunology. 2012;361:73-92.
- 477 32. Orth JH, Preuss I, Fester I, Schlosser A, Wilson BA, Aktories K. *Pasteurella multocida* toxin
478 activation of heterotrimeric G proteins by deamidation. PNAS. 2009;106(17):7179-84.
- 479 33. Turner P, Turner C, Jankhot A, Helen N, Lee SJ, Day NP, et al. A Longitudinal Study of
480 *Streptococcus pneumoniae* Carriage in a Cohort of Infants and Their Mothers on the Thailand-
481 Myanmar Border. PLoS One. 2012;7(5):e38271.
- 482 34. WHO. Cough or difficulty breathing. Pocket book of hospital care for children: guidelines for
483 the management of common illnesses with limited resources. 2nd Edition ed. Organisation WH,
484 editor2013.
- 485 35. Pourhoseingholi MA, Vahedi M, Rahimzadeh M. Sample size calculation in medical studies.
486 Gastroenterology and Hepatology from Bed to Bench. 2013;6(1):14-7.
- 487 36. Scott P, Walker A. Whole Genome Amplification of Single Bacterial Cells. In: McGenity TJ,
488 Timmis KN, Nogales B, editors. Hydrocarbon and Lipid Microbiology Protocols: Single-Cell and Single-
489 Molecule Methods: Springer Nature; 2016. p. 29-41.
- 490 37. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact
491 alignments. Genome Biology. 2014;15:R46.
- 492 38. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A new
493 genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational
494 Biology. 2012;19(5):455-77.
- 495 39. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+:
496 Architecture and applications. BMC Bioinformatics. 2008;10:421.

- 497 40. Bonfield JK, Whitwham A. Gap5 - editing the billion fragment sequence assembly.
498 Bioinformatics. 2010;26(14):1699-703.
- 499 41. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics.
500 2014;30(14):2068-9.
- 501 42. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current
502 status, new features and genome annotation policy. Nucleic Acids Research. 2012;40(D):130-5.
- 503 43. Rodriguez-R LM, Konstantinidis KT. The enveomics collection: a toolbox for specialized
504 analyses of microbial genomes and metagenomes. PeerJ Preprints. 2016;4:e1900v1.
- 505 44. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. DNA-DNA
506 hybridization values and their relationship to whole-genome sequence similarities. International
507 Journal of Systematic and Evolutionary Microbiology. 2007;57:81-91.
- 508 45. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
509 phylogenies. Bioinformatics. 2014;30(9):1312-3.
- 510 46. Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence-structure homology recognition using
511 environment-specific substitution tables and structure-dependent gap penalties. Journal of
512 Molecular Biology. 2001;310(1):243-57.
- 513 47. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data
514 Bank. Nucleic Acids Research. 2000;28:235-42.
- 515 48. Kitadokoro K, Kamitani S, Miyazawa M, Hanajima-Ozawa M, Fukui A, Miyake M, et al. Crystal
516 structures reveal a thiol protease-like catalytic triad in the C-terminal region of Pasteurella multocida
517 toxin. PNAS. 2007;104(12):5139-44.
- 518 49. Webb B, Sali A. Comparative protein structure modeling using MODELLER. Current Protocols
519 in Bioinformatics. 2016;54:5.6.1-5.6.37.

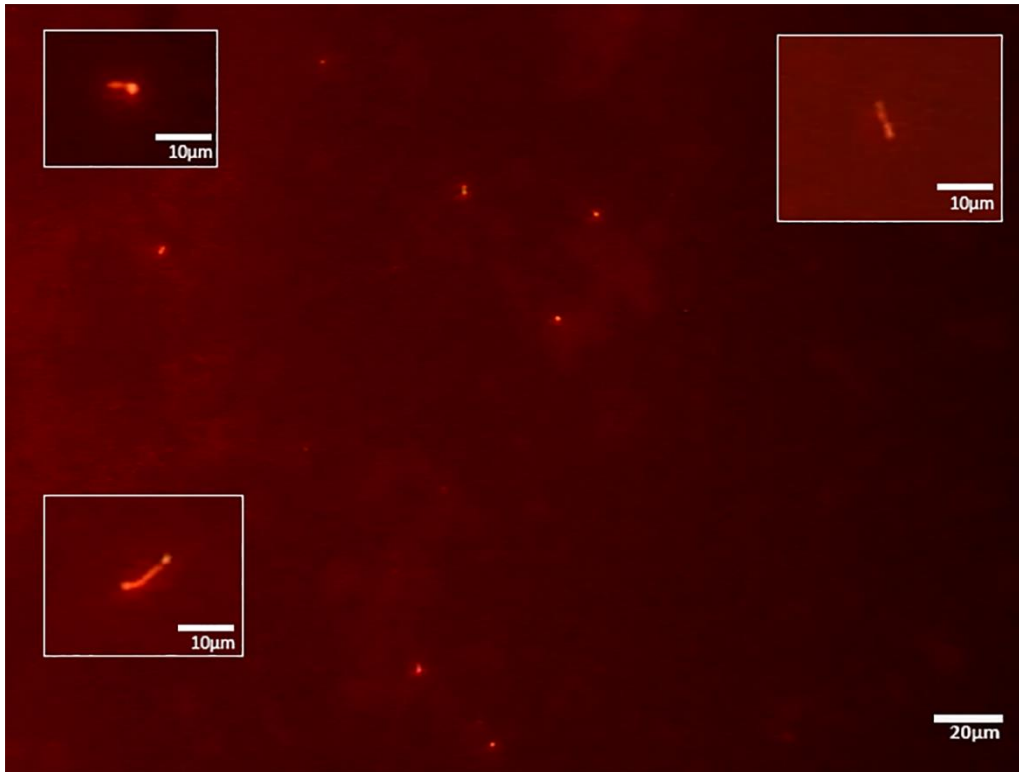
520

521

522 FIGURES

523

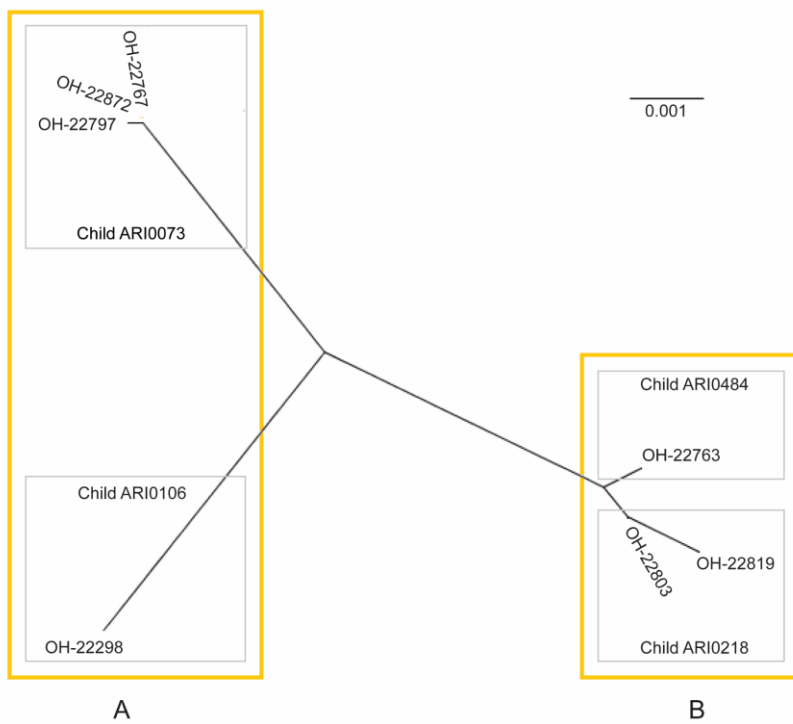
524 **Figure 1**



525

526

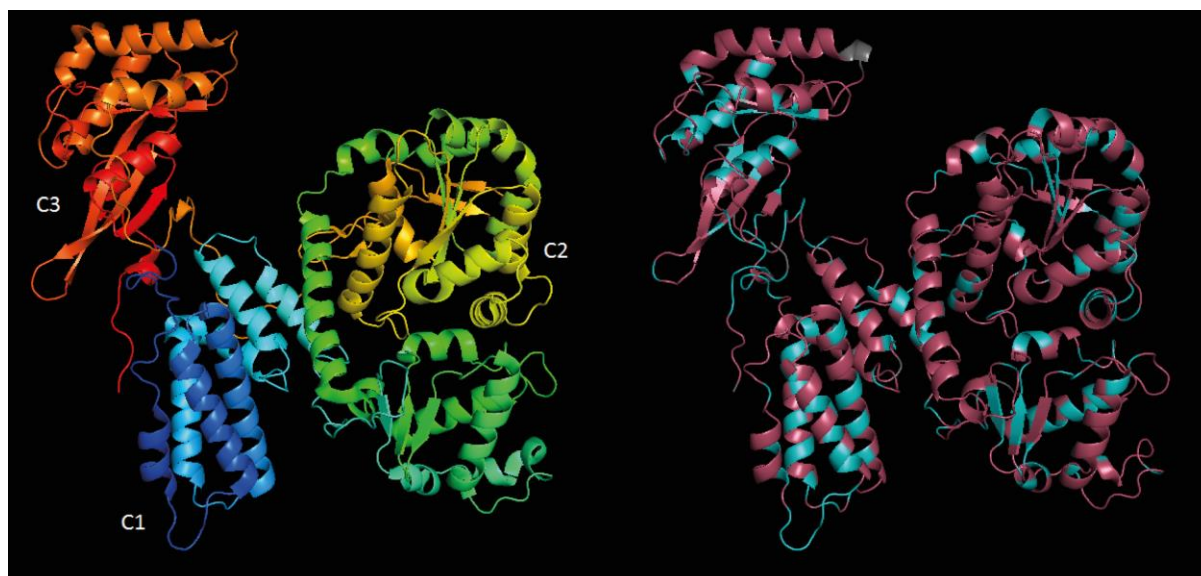
527 **Figure 2**



528

529

530 **Figure 3**



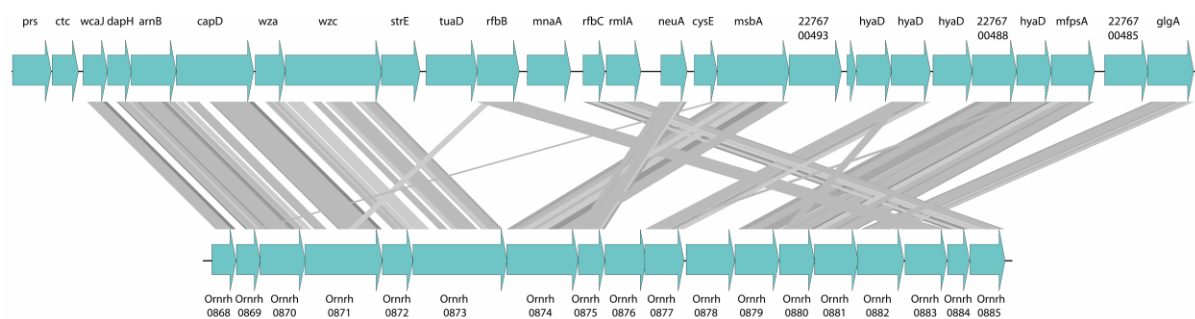
531

532

533

534

535 **Figure 4**



536