

1 **Title:**
2 **Misclassification of a whole genome sequence reference defined by the Human Microbiome**

3 **Project: a detrimental carryover effect to microbiome studies**

4
5

6 **Authors:**

7 DJ Darwin R. Bandy^{1,2} B Carol Huang, and Bart C. Weimer^{1*}

8
9

10 ¹University of California Davis, School of Veterinary Medicine, 100 K Pathogen Genome Project,
11 Davis, CA 95616, USA; ²Department of Veterinary Paraclinical Sciences, College of Veterinary
12 Medicine, University of the Philippines Los Baños, Laguna 4031 Philippines

13

14 *corresponding author: bcweimer@ucdavis.edu; (01) 530-760-9550

15

16 **Abstract**

17 Taxonomic classification is an essential step in the analysis of microbiome data that depends
18 on a reference database of whole genome sequences. Taxonomic classifiers are built on
19 established reference species, such as the Human Microbiome Project database, that is growing
20 rapidly. While constructing a population wide pangenome of the bacterium *Hungatella*, we
21 discovered that the Human Microbiome Project reference species *Hungatella hathewayi* (WAL
22 18680) was significantly different to other members of this genus. Specifically, the reference
23 lacked the core genome as compared to the other members. Further analysis, using average
24 nucleotide identity (ANI) and 16s rRNA comparisons, indicated that WAL18680 was
25 misclassified as *Hungatella*. The error in classification is being amplified in the taxonomic
26 classifiers and will have a compounding effect as microbiome analyses are done, resulting in
27 inaccurate assignment of community members and will lead to fallacious conclusions and
28 possibly treatment. As automated genome homology assessment expands for microbiome
29 analysis, outbreak detection, and public health reliance on whole genomes increases this issue
30 will likely occur at an increasing rate. These observations highlight the need for developing
31 reference free methods for epidemiological investigation using whole genome sequences and the
32 criticality of accurate reference databases.

33 34 **Background**

35 Clostridia are a very diverse group of organisms. The taxonomy is in constant revision in
36 light of new whole genome sequence production and genomic flux¹. While organism
37 classification can be reassigned, the identified isolates within the same species retain their
38 relatedness. In the analysis of 13,151 microbial genomes, the misclassification (18%) was
39 determined by binning into cliques and singletons with ANI data using the Bron-Kerbosch
40 algorithm, which resulted in the misclassification of 31 out of the 445 type strains². The different

41 causes of the type strain misclassification include poor DNA-DNA hybridization (e.g. high
42 genomic diversity), low DNA-DNA hybridization values, naming without referencing to another
43 type strain, and lack of 16s rRNA data. *Hungatella hathewayi*, or its prior designation
44 *Clostridium hathewayi*, was not included in the previous as there were very few *Hungatella*
45 genomes in the time of that publication. As more metagenomes are published increasing claims
46 of finding new organisms are mounting. To this point, Almeida et al. reported an increase of
47 1952 uncultured organisms that are not represented in well-studied human populations, where
48 they presented data to support that rare species will be difficult to accurately identify and do not
49 match existing references³.

50 Public repositories of genomic data have experienced tremendous expansion beyond human
51 curatorial capacities, which is an ever increasing issue with the high rate of WGS production^{4,5}.
52 Recently, it was estimated that ~18% of the organisms are misclassified in microbial genome
53 databases². This high rate of error led to investigation of misclassification of specific organisms,
54 including *Aeromonas*⁶ *Fusobacterium*⁷, and ultimately entire reference databases². These studies
55 found misclassified type strains, which calls into question the foundation of the taxonomy and
56 inferred relatedness when population genomes are being used for epidemiological purposes,
57 especially with rare organisms that are not well represented in the reference database. The work
58 presented here uniquely identified a misclassified reference species and found propagation of
59 incorrectly labelled genomes in several highly cited microbiome studies^{8,9,10,11}.

60

61 **Observation**

62 Based on this species delineation notion, we discovered that the Human Microbiome Project
63 reference genome for *Hungatella hathewayi* (WAL18680) was misidentified while building a
64 phylogeny of *Hungatella* species using a population of whole genome sequences¹². Both 16s

65 rRNA and average nucleotide identity (ANI²) analysis indicated that WAL18680 was not a
66 member of the *Hungatella* genus based on genome assessment (Table 1). Population genome
67 comparison analysis was instrumental in discovering that WAL18680 was misclassified and the
68 impact for genomic epidemiology purposes would be important.

69 The misclassified *H. hathewayi* WAL18680 has been used to generate phylogenomic
70 analysis, reference WGS for metagenome analysis, and web server identification platforms
71 utilizing the metagenomic classifiers^{10,13,14}. Epidemiologically, association with clinical disease
72 will be discordant with genomic data and result in inaccurate conclusions on the microbiome
73 ecology or therapies based on the microbiome membership to mitigate disease leading to the
74 wrong causal relationship to be concluded⁹. As more microbiome studies are linking rare
75 microbes to biological outcomes, a need exists to quickly identify inaccurate assignment when
76 only a few WGS of individual organisms are available for use as a reference. This creates an
77 issue with low sampling of the genome space for rare organisms and may result in mis-naming
78 based on a small set of phenotypic assays that do not represent the genome content or flux¹⁵.

79 *H. hathewayi* was first described as an isolate was from human feces¹⁶ and was subsequently
80 reported in a patient with acute cholecystitis, hepatic abscess, and bacteremia^{17,18}. It was also
81 later reported in a case of appendicitis¹⁹. *H. hathewayi* is (WAL18680) one of the designated
82 reference strains in Human Microbiome Project and is used extensively for binning and
83 classification of microbiome related studies, which confounds analysis of the genus *Hungatella*.
84 This organism can be isolated from the microbiome depending on the enrichment conditions⁹.
85 Having a reference species misclassified is detrimental to microbiome research and in
86 epidemiological investigations. To solve this issue, we developed a heuristic to minimizing

87 misclassification for rare reference species as a result of cross-validation of the genomic
88 information for name assignment.

89 The standard procedure of the 100K Pathogen Genome Sequencing Project^{4,5,20-22} determines
90 the identity of bacterial pathogen isolates in clinical samples using WGS and the genome
91 distance (ANI^{23,24}) before proceeding with additional comparisons. This analysis was done with
92 a group of isolates from suspected *Clostridioides difficile* infection cases. We identified a species
93 of *H. hathewayi* using genome distance using the entire genome sequence that was implemented
94 for high dimensional comparison using MASH²⁵ (with the maximum sketch size). This was
95 coupled to comparison of all of the available WGS to represent the entire genome diversity to
96 build a whole genome phylogeny¹² to determine the naming accuracy of the clinical isolates .
97 Unexpectedly, one particular sequence was well beyond the species ANI threshold for
98 *C. difficile*. We found that based on ANI, is a putative new species of *Hungatella* (strain
99 2789STDY5834916). Weis et al.^{26,27} used this method with *Campylobacter* species to
100 demonstrate that genome distance accurately estimates host-specific genotypes, zoonotic
101 genotypes, and disease within livestock disease with validated reference genomes. While ANI
102 was the first estimate to raise questions for the accurate identification of this organism, we
103 proceeded with a cross-validation strategy to verify the potential misclassification of the
104 reference species.

105 We advanced with the initial mis-identification by determining the pangenome analysis with
106 the hypothesis that outbreak isolates would cluster together based on the isolate origin (i.e. an
107 individual or location)¹² as well as contain the same core genome. We found that WAL18680 did
108 not contain any of the core genome relative to all of the other *Hungatella* genomes (Figure 1).
109 Together, these genomic metrics prove that this reference genome was misclassified, which has

110 extensive implications as reference sequences are commonly used for genomic identity for
111 outbreak investigations. Additionally, metagenome studies require reference genome databases
112 to identify bacterial community members. This result indicates that if the epidemiological
113 workflow did not include specific whole genome alignment, inaccurate conclusions and
114 misleading deductions will be made – as was observed by Kaufman et al.¹⁵ – where they found
115 that genome diversity is unexpectedly large and expands based on a power law with each new
116 WGS that is added to the database. Combining the fact that this is a reference genome from a
117 rare organism from a very diverse group, that the genome evolution rate is a power law, and that
118 this is a reference genome from the Human Genome Project the implications for the mis-
119 identification have far reaching implications.

120 Conflicts of taxonomic classification based on traditional methods, such as phenotypic
121 assays, metabolism, with genomic based parameters will likely increase as more genomes are
122 produced and use of the entire genetic potential (i.e. the entire genome). The need for heuristical
123 indicators of misclassification are needed as is the need to expand WGS that adequately
124 represent bacterial diversity among and within taxonomy to represent the genetic diversity of any
125 single organism.

126

127 **Genome sequence availability:**

128 The WGS for each genome is via the NCBI with Biosample numbers of SAMD00008809,
129 SAMN02463855, SAMN02596771, SAMEA3545258, SAMEA3545379, SAMN09074768. The
130 WGS sequence for BCW8888 is available via the 100K Project BioProject at the NCBI
131 (PRJNA186441) as Biosample SAMN12055167.

132

133 **Figure legends:**

- 134 1. Pangenome of *Hungatella*. WAL18680 was originally identified as *Clostridium*
135 *hathewayi*. After a recent taxonomic reclassification it was renamed as *Hungatella*
136 *hathewayi*. **(WAL 18680) does not have the core genome of other *Hungatella* species**
137 **(*hathewayi* or *efluvii*)** and possess very few core genes common to the other *Hungatella*
138 species. The bulk of its genome is not found in other *Hungatella* species, indicating it
139 belongs to another genus. Strain 2789STDY5834916 is a novel *Hungatella* species.
- 140 2. Phylogenomic of all *Hungatella* relatedness estimated using genome distance.

141

142 **Tables:**

- 143 1. Average nucleotide identity (ANI) of *Hungatella* isolates using the WGS. The reference
144 WGS from WAL18680 (isolated in Canada in 2011) was classified as a different genus
145 using the ANI criteria as compared to the other isolates examined. Strains
146 2789STDY5834916 (isolated in the UK in 2015) and BCW8888 (isolated in Mexico in
147 2015) would be considered new and novel species using ANI.

148

149 **References:**

- 150 1 Yutin, N. & Galperin, M. Y. A genomic update on clostridial phylogeny: Gram-negative
151 spore formers and other misplaced clostridia. *Environ Microbiol* **15**, 2631-2641,
152 doi:10.1111/1462-2920.12173 (2013).
- 153 2 Varghese NJ, M. S., Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, Pati.
154 Microbial species delineation using whole genome sequences. *Nucleic Acids Res* **Aug**
155 **18;43(14):6761-71**, doi:10.1093/nar/gkv657 (2015).

- 156 3 Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**,
157 499-504, doi:10.1038/s41586-019-0965-1 (2019).
- 158 4 Weimer, B. C. 100K Pathogen Genome Project. *Genome Announcements*,
159 genomeA.00594-00517, doi:DOI: 10.1128/genomeA.00594-17 (2017).
- 160 5 Kong, N. *et al.* Draft Genome Sequences of 1,183 Salmonella Strains from the 100K
161 Pathogen Genome Project. *Genome Announc* **5**, e00518-00537,
162 doi:10.1128/genomeA.00518-17 (2017).
- 163 6 Awan, F. *et al.* Comparative genome analysis provides deep insights into *Aeromonas*
164 *hydrophila* taxonomy and virulence-related factors. *BMC Genomics* **19**, 712,
165 doi:10.1186/s12864-018-5100-4 (2018).
- 166 7 Kook, J., Park, SN., Lim, Y.K. Genome-Based Reclassification of *Fusobacterium*
167 *nucleatum* Subspecies at the Species Level. *Current Microbiology* **74**, 1137-1147,
168 doi:<https://doi.org/10.1007/s00284-017-1296-9> (29 June 2017).
- 169 8 I, R. D. a. T. Comparative Genomic Analysis of the Human Gut Microbiome Reveals a
170 Broad Distribution of Metabolic Pathways for the Degradation of Host-Synthesized
171 Mucin Glycans and Utilization of Mucin-Derived Monosaccharides. *Front. Genet.* **8**,
172 doi:10.3389/fgene.2017.00111 (2017).
- 173 9 Atarashi, K. *et al.* Treg induction by a rationally selected mixture of *Clostridia* strains
174 from the human microbiota. *Nature* **500**, 232-236, doi:10.1038/nature12331 (2013).
- 175 10 Sabag-Daigle A, W. J., Borton MA., Sengupta A, G. V., Wrighton KC, & Wysocki VH,
176 A. B. Identification of bacterial species that can utilize fructoseasparagine. *Appl Environ*
177 *Microbiol* **84**:e01957-17, doi: 10.1128/AEM.01957-17 (2018).

- 178 11 Yu, L. *et al.* Grammar of protein domain architectures. *Proc Natl Acad Sci U S A* **116**,
179 3636-3645, doi:10.1073/pnas.1814684116 (2019).
- 180 12 Bandoy, D. Pangenome guided pharmacophore modelling of enterohemorrhagic
181 *Escherichia coli* sdiA. *F1000Research*
182 doi:<https://doi.org/10.12688/f1000research.17620.1> (2019).
- 183 13 Davis, M. P., van Dongen, S., Abreu-Goodger, C., Bartonicek, N. & Enright, A. J.
184 Kraken: a set of tools for quality control and analysis of high-throughput sequence data.
185 *Methods* **63**, 41-49, doi:10.1016/j.ymeth.2013.06.027 (2013).
- 186 14 Carrico, J. A., Rossi, M., Moran-Gilad, J., Van Domselaar, G. & Ramirez, M. A primer
187 on microbial bioinformatics for nonbioinformaticians. *Clin Microbiol Infect* **24**, 342-349,
188 doi:10.1016/j.cmi.2017.12.015 (2018).
- 189 15 Kaufman, J. H., Christopher A. Elkins, Matthew Davis, Allison M Weis, Bihua C.
190 Huang, Mark K Mammel, Isha R. Patel, Kristen L. Beck, Stefan Edlund, David
191 Chambliss, Simone Bianco, Mark Kunitomi, Bart C. Weimer. Microbiogeography and
192 microbial genome evolution. *arXiv*, 1703.07454 (2017).
193 <<https://arxiv.org/abs/1703.07454>>.
- 194 16 Steer, T., Collins, M. D., Gibson, G. R., Hippe, H. & Lawson, P. A. *Clostridium*
195 *hathewayi* sp. nov., from human faeces. *Syst Appl Microbiol* **24**, 353-357,
196 doi:10.1078/0723-2020-00044 (2001).
- 197 17 Kaur, S., Yawar, M., Kumar, P. A. & Suresh, K. *Hungatella effluvii* gen. nov., sp. nov.,
198 an obligately anaerobic bacterium isolated from an effluent treatment plant, and
199 reclassification of *Clostridium hathewayi* as *Hungatella hathewayi* gen. nov., comb. nov.
200 *Int J Syst Evol Microbiol* **64**, 710-718, doi:10.1099/ijs.0.056986-0 (2014).

- 201 18 Elsayed, S. & Zhang, K. Human infection caused by *Clostridium hathewayi*. *Emerg*
202 *Infect Dis* **10**, 1950-1952, doi:10.3201/eid1011.040006 (2004).
- 203 19 Woo, P. C. *et al.* Bacteremia due to *Clostridium hathewayi* in a patient with acute
204 appendicitis. *J Clin Microbiol* **42**, 5947-5949, doi:10.1128/JCM.42.12.5947-5949.2004
205 (2004).
- 206 20 Weis, A. M. *et al.* Large-Scale Release of *Campylobacter* Draft Genomes: Resources for
207 Food Safety and Public Health from the 100K Pathogen Genome Project. *Genome*
208 *Announc* **5**, e00925-00916, doi:10.1128/genomeA.00925-16 (2017).
- 209 21 Weis, A. M., Bihua C. Huang, Dylan B. Storey, Nguyet Kong, Poyin Chen, Narine
210 Arabyan, Brent Gilpin, Carl Mason, Andrea K. Townsend, Woutrina A. Miller, Barbara
211 A. Byrne, Conor C. Taff, Bart C. Weimer. Large-scale release of *Campylobacter* draft
212 genomes; resources for food safety and public health from the 100K Pathogen Genome
213 Project. *Genome Announcements* **5**, e00925-00916 (2016).
- 214 22 Chen, P. *et al.* 100K Pathogen Genome Project: 306 *Listeria* Draft Genome Sequences
215 for Food Safety and Public Health. *Genome Announc* **5**, e00967-00916,
216 doi:10.1128/genomeA.00967-16 (2017).
- 217 23 Auch, A. F., von Jan, M., Klenk, H. P. & Goker, M. Digital DNA-DNA hybridization for
218 microbial species delineation by means of genome-to-genome sequence comparison.
219 *Stand Genomic Sci* **2**, 117-134, doi:10.4056/sigs.531120 (2010).
- 220 24 Auch, A. F., Klenk, H. P. & Goker, M. Standard operating procedure for calculating
221 genome-to-genome distances based on high-scoring segment pairs. *Stand Genomic Sci* **2**,
222 142-148, doi:10.4056/sigs.541628 (2010).

- 223 25 Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using
224 MinHash. *Genome Biol* **17**, 132, doi:10.1186/s13059-016-0997-x (2016).
- 225 26 Weis, A. M. *et al.* Genomic Comparisons and Zoonotic Potential of Campylobacter
226 Between Birds, Primates, and Livestock. *Applied and environmental microbiology*, 7165-
227 7175, doi:10.1128/AEM.01746-16 (2016).
- 228 27 Lawton, S. J. *et al.* Comparative analysis of Campylobacter isolates from wild birds and
229 chickens using MALDI-TOF MS, biochemical testing, and DNA sequencing. *J Vet*
230 *Diagn Invest* **30**, 354-361, doi:10.1177/1040638718762562 (2018).
- 231

232 **Table 1.**

	<i>H. hathewayi</i> BCW8888	<i>H. effluvii</i> DSM24995	<i>H. hathewayi</i> WAL18680	<i>H. hathewayi</i> VE202-11	<i>H. hathewayi</i> 2789STDY5834916	<i>H. hathewayi</i> 2789STDY5608850	<i>H. hathewayi</i> 12489931
<i>H. hathewayi</i> BCW8888	100	94.3	70.9	96.7	85.4	98.1	96.8
<i>H. effluvii</i> DSM24995	94.4	100	70.5	94.5	85.5	94.4	94.9
<i>H. hathewayi</i> WAL18680	71.2	70.8	100	71.0	72.0	71.1	71.1
<i>H. hathewayi</i> VE202-11	96.6	94.5	70.7	100	85.2	96.5	98.73
<i>H. hathewayi</i> 2789STDY5834916	85.6	85.6	72.0	85.4	100	85.7	85.8
<i>H. hathewayi</i> 2789STDY5608850	98.1	94.3	70.7	96.4	85.4	100	96.5
<i>H. hathewayi</i> 12489931	96.8	94.9	70.8	98.8	85.7	96.7	100

233

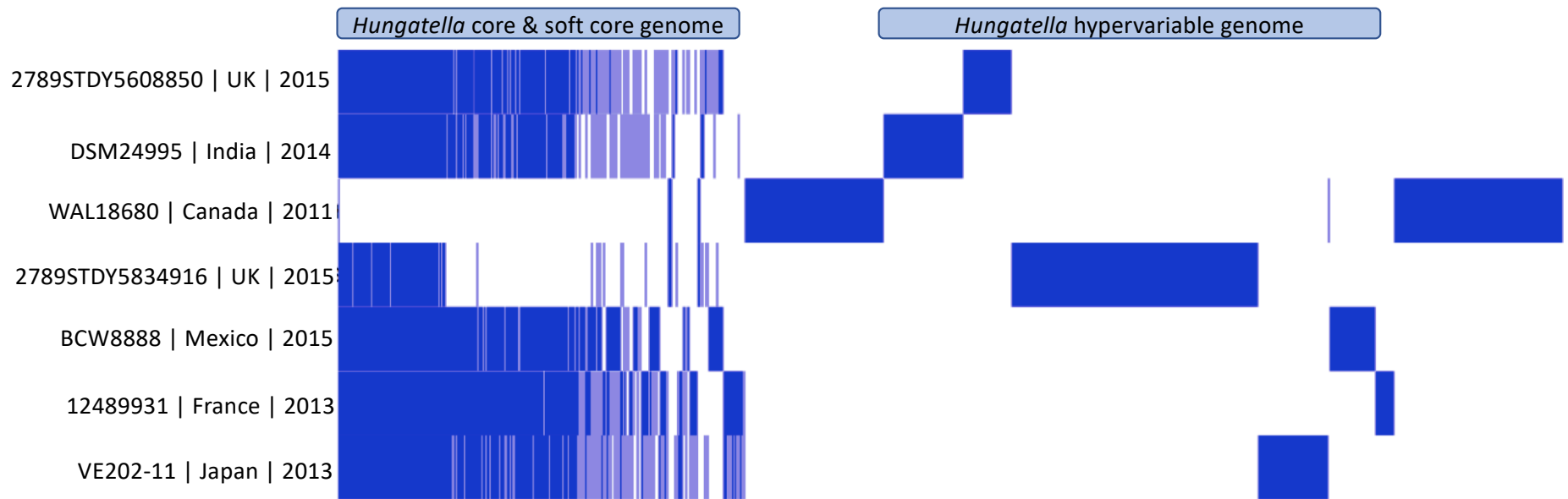


Fig 1

0.1

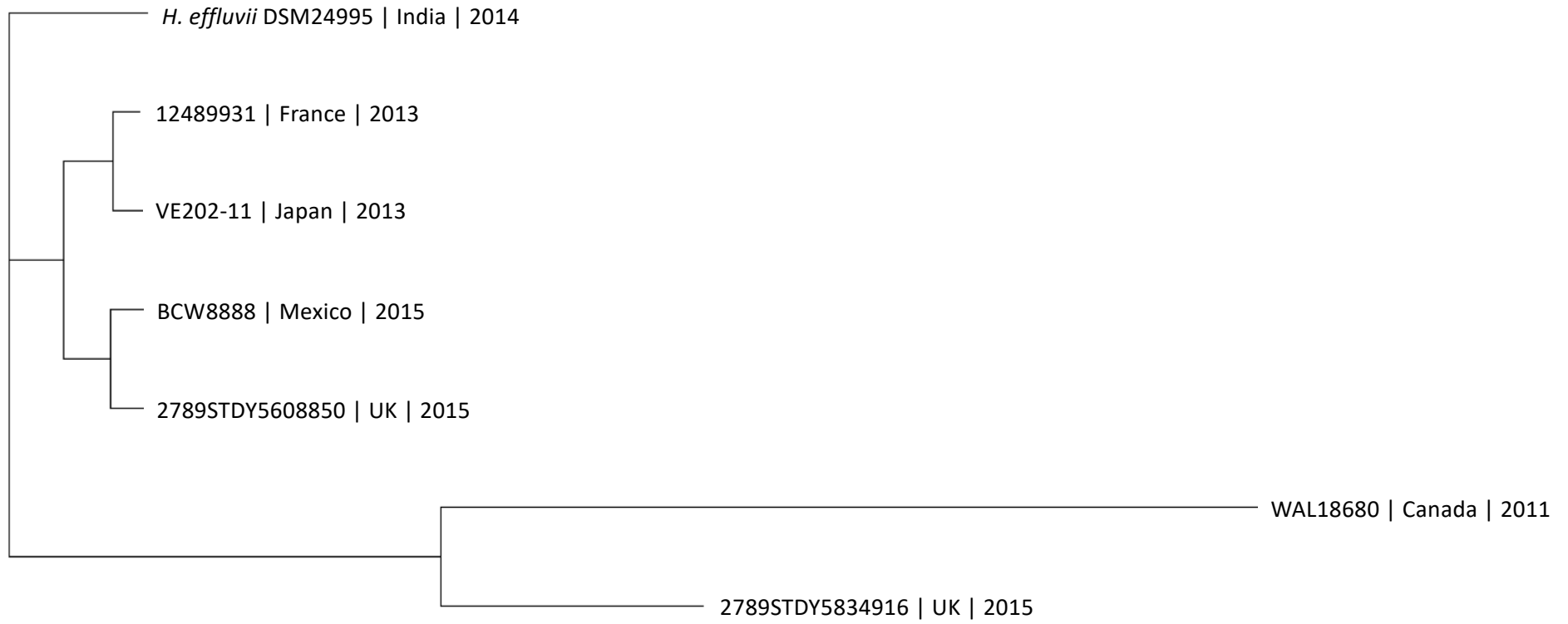


Fig 2