



HAL
open science

Ffp1, an ancestral *Porphyromonas* spp. fimbrillin

Luis A Acuña-Amador, Frédérique Barloy-Hubler

► **To cite this version:**

Luis A Acuña-Amador, Frédérique Barloy-Hubler. Ffp1, an ancestral *Porphyromonas* spp. fimbrillin. 2023. hal-04334115

HAL Id: hal-04334115

<https://hal.science/hal-04334115>

Preprint submitted on 10 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Ffp1, an ancestral *Porphyromonas* spp. fimbrillin**

2 Luis Acuña-Amador^{a,*} and Frédérique Barloy-Hubler^{b,*}

3

4 ^aLaboratorio de Investigación en Bacteriología Anaerobia, Centro de Investigación en
5 Enfermedades Tropicales, Facultad de Microbiología, Universidad de Costa Rica, San
6 José, Costa Rica

7 ^bUniversité de Rennes 1, CNRS, UMR 6553 ECOBIO (Écosystèmes, Biodiversité,
8 Évolution), 35042 Rennes, France

9 * luisalberto.acuna@ucr.ac.cr

10

11 **Keywords:** *Porphyromonas*, fimbriae, bioinformatic, phylogenomic, 3D protein
12 modeling

13

14 **Abstract**

15 **Background:** Little is known about fimbriae in the *Porphyromonas* genus. Besides *fim* and
16 *mfa*, a third *Porphyromonas gingivalis* adhesin called Ffp1 has been described, and seems to
17 be capital for outer membrane vesicle (OMV) production. **Objective:** We aimed to
18 investigate the distribution and diversity of type V fibrillin, particularly Ffp1, in the
19 *Porphyromonas* genus. **Methods:** A bioinformatic phylogenomic analysis was conducted
20 using all accessible *Porphyromonas* genomes in order to generate a domain search for
21 fimbriae, using HMM profiles. **Results:** Ffp1 was found as the sole fimbrillin in all the
22 analyzed genomes. After manual biocuration and 3D modeling, this protein was determined
23 to be a type V fimbrillin, with a closer structural resemblance to a *Bacteroides ovatus*
24 fimbrillin than to FimA or Mfa1 from *P. gingivalis*. **Conclusion:** It appears that Ffp1
25 represents ancestral fimbriae present in all *Porphyromonas* species. Additional investigations

26 are necessary to elucidate the biogenesis of Ffp1 fimbriae and his potential role in OMV
27 production and niche adaptation.

28

29 **Introduction**

30 Fimbriae (fibrillae or pili) are adhesins consisting of protein polymers forming
31 filamentous appendages that protrude from the bacterial cell surface. Unlike motility flagella,
32 fimbriae have adhesive properties to attach to surfaces. In Gram-negative bacteria, fimbriae
33 are classified according to their assembly pathways, including the chaperone-usher (CU)
34 pilus system, the type IV pilus, and the conjugative type IV secretion pilus (1,2).

35 In 2016, a new prevalent type V pilus was discovered within the human gut microbiome
36 (3) and was described as a new donor strand-mediated system restricted to the Bacteroidia
37 class (2). This system resembles the CU type, but requires the lipoprotein sorting pathway,
38 and outer membrane proteinases (4).

39 Type V fimbriae have been mainly studied in *P. gingivalis* which classically produces
40 two distinct adhesins, termed FimA (described in 1984 (5)), and Mfa1 (described in 1996
41 (6)), according to the names of stalk subunits (7). Both stalk proteins must be processed and
42 matured. They possess long leader peptides (8) that facilitate their transport to the periplasm
43 via the Sec system. Subsequently, they undergo lipid modification and are cleaved by type II
44 signal peptidase (9), followed by a proteolytic maturation achieved by RgpA, RgpB and Kgp
45 proteinases called gingipains (10). Finally, mature fibrillin monomers polymerize (11). The
46 genetic loci for both fimbriae are distinct but organized into two clusters: *fimA-E* and *mfa1-5*
47 (12).

48 In 2017, a third *P. gingivalis* adhesin was described (PGN_1808 in the ATCC 33277
49 strain or PG1881 in the W83 strain) and termed Ffp1 for filament-forming protein 1 (13). It
50 corresponds to filaments 200 to 400 nm in length and 2 to 3 nm in diameter, that can be

51 degraded, unlike FimA or Mfa1, by detergents and temperature into 50 kDa monomers (14).
52 Ffp1 is among the exclusive repertoire of proteins within the order Bacteroidales and is
53 conserved across *Porphyromonas* and *Bacteroides* (15,16). This protein was identified among
54 the outer membrane proteins and especially the O-glycoproteome of *P. gingivalis* (17) and
55 was described as essential in the production of outer membrane vesicles (OMVs), as the Ffp1
56 null-mutants exhibited a 30% reduction in OMVs production compared to the wild-type
57 strain (13). Moreover, a recent study indicates a connection between Ffp1 and the production
58 of sphingolipids (SL). In the absence of SL, *P. gingivalis* generates OMVs without Ffp1,
59 whereas OMVs containing SLs exhibit an enrichment of Ffp1. Interestingly, these SL-
60 containing OMVs limit host inflammation (18).

61 Ffp1 C-terminal region is homologous to type IV fimbriae from *Bacillus* spp. (15) and
62 its sequence bears a significant similarity to the adhesion protein BACOVA_01548 (PDB ID:
63 4rfj) from *Bacteroides ovatus* (3). Its structural modeling suggests a donor strand-mediated
64 assembly mechanism (14), which would classify Ffp1 as a new type V pilin (13). However,
65 unlike FimA or Mfa1, no accessory component has yet been identified for Ffp1 despite its
66 apparent co-expression as an operon with three upstream genes, annotated as a Cys-RNAt
67 ligase, a patatin (lipase) and a glycosyl transferase. This co-transcription suggests the
68 involvement of these four proteins in the same biochemical pathway or utilization of the same
69 substrates/transporters, albeit without physical interaction (14).

70 To date, Ffp1 has been the subject of few works limited to *P. gingivalis*, only on two
71 reference strains ATCC33277 and W83, and no information is available for the other 21
72 *Porphyromonas* species. At the genus level, knowledge for *non-P. gingivalis* Ffp1 or other
73 fimbriae is scarce, with the exception of description of FimA-like and Mfa1-like fimbriae in
74 *P. gulae*, a closely related species to *P. gingivalis* (19,20), and reports indicating fimbriation

75 in *P. circumdentaria*, *P. macacae* and *P. asaccharolytica* (21–23), without further
76 characterizations.

77 In this context, the aim of this study is to complete this knowledge gap and to
78 investigate the distribution and diversity of type V fibrillin, particularly Ffp1, in the
79 *Porphyromonas* genus. To do so, we performed an *in silico* analysis of type V fimbriin
80 locus in all 144 available genomes of *Porphyromonas*, investigating their presence/absence
81 and then focus on Ffp1 diversity, and 3D predicted structure.

82

83 **Material and Methods**

84 ***Porphyromonas* taxogenomics**

85 All 144 *Porphyromonas* genomes (Table S1) were automatically downloaded from
86 the NCBI RefSeq database using the `ncbi-genome-download` script*. Unannotated
87 Metagenome-Assembled Genomes (MAGs) with inconsistent taxonomic labels were not
88 considered. To categorize all genomes into reliable groups, genomic data-driven taxonomic
89 confirmation and/or assignment were performed. To confirm the assignment of genomes with
90 a species name, we conducted a comparison of three metrics : i) the 16S rRNA gene
91 percentage identity (when annotated), evaluated using a threshold of 98.65% (24); ii) the
92 digital DNA-DNA hybridization distance (DDH) using the GGDC v2.1 (25) and `ggdc-robot`
93 script†, with the default threshold of 70% using formula 2 (25–27); and iii) the whole genome
94 Average Nucleotide Identity (gANI), calculated using FastANI with a threshold of 96% for
95 species demarcation (28). In case of a disagreement between these three metrics, we
96 combined alignment fraction values (AF) with gANI using 60% and 96.5 % as threshold
97 values respectively, to assign a genome pair to the same species (29). Additionally, when

* <https://github.com/kbclin/ncbi-genome-download>

† <https://github.com/andrewfrank/ggdc-robot>

98 needed, we also used OrthoANI[‡] to measure and visualize the overall similarity between
99 some *Porphyromonas* species.

100 For the genomes without a specified species name (*Porphyromonas* sp.), as the
101 majority of them originated from environmental samples (human- or animal-associated
102 habitats) and are often highly fragmented, it was crucial to ensure that they were not
103 contaminated and do not correspond to genome assemblies containing a mixture of different
104 species. This genomic homogeneity was evaluated with Kraken2 (30) using the non-
105 redundant nucleic database (updated April 22). Only assemblies that consisted of over 80% of
106 *Porphyromonas* content and/or larger than 80% of the expected average genome size (2.5
107 Mb) were retained for our analysis. Their affiliation to the *Porphyromonas* genus was first
108 confirmed using fDBAC server[§] (31) and their position within the *Porphyromonas*
109 taxonomy was validated using an Orthofinder rooted species tree (32). This tree was
110 constructed using all *Porphyromonas* sp. (*P.* sp.) and one reference genome per
111 *Porphyromonas* species (refer to Table S1) and was visualized using FigTree^{**}. For each
112 branch, one or several *P. sp* were associated to a *Porphyromonas* species through ANI and
113 DDH, employing the same thresholds as previously described.

114

115 ***Porphyromonas* fimbriae identification and classification**

116 **1. Dataset construction:** Sequences from type V fimbriae (FimABCDE, Mfa12345 and
117 Ffp1) were manually extracted from the 59 *P. gingivalis* genomes, and were used as queries
118 to identify homologous sequences all in the genomes of other *Porphyromonas* spp. using

[‡] <https://www.ezbiocloud.net/tools/orthoani>

[§] <http://fbac.dmicrobe.cn/tools/>

^{**} <http://tree.bio.ed.ac.uk/software/figtree/>

119 BlastP (identity $\geq 30\%$; query coverage $\geq 60\%$; e-value $< 10e^{-5}$). All sequences were grouped
120 as dataset 1.

121 **2. Functional domain-based screening:** Dataset 1 was subjected to analysis using
122 InterProScan to identify all protein domains associated with those sequences. The resulting
123 domains were searched in the complete orfeomes of *Porphyromonas* downloaded from
124 PATRIC 3.6.6 (33), using HMMsearch from HMMER v3.3.1 (34) and the hidden Markov
125 models (HMMs) from Pfam 33.1 (May 2020) database (35). Sequences harboring the
126 targeted domains with an e-value $< 10e^{-06}$ were retained and grouped into dataset 2.

127 **3. Protein clustering, biocuration and HMM profile construction:** Dataset 2 was clustered
128 with MMseqs2 (36) via the easy-cluster command. Each cluster obtained underwent manual
129 biocuration after multiple alignment using Clustal Omega (37) and any missing genes were
130 annotated. Subsequently, for each cluster, the multiple alignments were converted from
131 FASTA format to Stockholm format with 'sreformat' command and HMM profiles were
132 generated using the 'hmmbuild' command with default settings. Clustering and HMM
133 profiles creation was first performed on raw data and then refined on biocurated data.

134 **4. Final classification:** The obtained HMM profiles (refer to Supplementary material) were
135 used to identify and classify all fimbrillins within the *Porphyromonas* orfeomes, downloaded
136 from PATRIC 3.6.6 database, using 'hmmsearch' command from HMMER package.

137 **5. In silico analysis of *Porphyromonas* fimbrillins:** Geneious Prime (38) was used to
138 visualize the genomic context of each identified fimbrillin. Biocuration for start codon were
139 proposed, based on sequence homology, to optimize the prediction of SPII signal peptide and
140 the cleavage site positions. N-terminal region was identified using *charge* (window size=3)
141 from EMBOSS 6.6.0 (39), the H hydrophobic region was characterized with Kyte-Doolittle
142 hydropathy plot made with ProtScale (40) (window size=3), and the cleavage site was
143 confirmed by SignalP 6.0 (41) and LipoP (42). Palmitoylation in the lipobox cysteine residue

144 was verified using CSS-Palm (43). Protein sizes were represented using violin plots
145 (`geom_violin`) and/or boxplot (`geom_box`), both functions from the `ggplot2` package (44).

146 For each fimbrillin family, a multiple alignment was performed using MAFFT (L-
147 INS-I algorithm and BLOSUM62 matrix; gap open penalty and offset value by default) (45).
148 This alignment was visualized in two dimensions using Alignmentviewer v1.1^{††} which
149 employs the UMAP algorithm (46) and Hamming distance to cluster aligned sequences.
150 Phylogenetic trees were calculated using FastTree 2.1.11 (47), PhyML 3.3 (48), and RaxML
151 (49) with default parameters.

152 The taxonomic distribution of fimbrillin genes was analyzed across a phylogenetic
153 tree constructed using OrthoFinder based on the pangenomes of all confirmed
154 *Porphyromonas* species groups (32) and visualize using FigTree. The phylogenetic
155 reconstruction was performed both using native and mature proteins (i.e excluding their
156 signal peptides) using RaxML (evolution model GAMMA LG and 100 bootstrap). Robinson-
157 Foulds, Nye Similarity and Jaccard Robinson Foulds distances between the phylogenetic
158 trees were calculated using TreeDist^{‡‡} R library and tanglegrams were created with the R
159 package phytools^{§§} (scripts TREE.R and Tanglegram.R).

160 **6. 3D modeling:** Secondary protein structure was predicted with PSIPRED in Phyre2 (50).
161 3D structures of Ffp1 mature proteins were modeled, based on homology modeling, using
162 Robetta (51) and the RoseTTAFold method, as well as Phyre2. The quality of all five 3D
163 models generated by Robetta for each Ffp1 protein was assessed and validated using two
164 quality calculation tools: ERRAT (52) and Verify3D (53). The most accurate predicted
165 structure was chosen and superposed to the best model target, found by VAST+ (54), Phyre2

^{††} <https://github.com/sanderlab/alignmentviewer>

^{‡‡} <https://cran.rstudio.com/web/packages/TreeDist/index.html>

^{§§} <https://cran.r-project.org/web/packages/phytools/index.html>

166 and iPBA (55). The RMSD value (56) as well as the number and percentage of aligned
167 residues were retrieved and compared to Phyre2 results. RMSD < 3Å were considered
168 significant between Ffp1 predicted structure and 3D models (57).

169

170 **Results**

171 **1. *Porphyromonas* taxonomic assignment**

172 The 144 *Porphyromonas* genomes studied in this work (Table S1) were
173 predominantly in draft form (85% of the genomes), with only 6 out of the 17 analyzed
174 species possessing at least one complete genome.

175 The taxonomic assignment for the genomes classified into the 17 *Porphyromonas*
176 species was verified (Table S1). Species *P. loveana* and *P. pasteri* have only one
177 representative genome and therefore cannot be verified intra-specifically. For the other
178 species, intra-specific analysis combining ANI, 16s rRNA and DDH comparison (Figure 1A)
179 showed no anomalies for taxonomic placements, except for *P. uenonis*, *P. somerae*, and *P.*
180 *canoris*.

181 Firstly, for *P. uenonis*, the differences in metrics reflect a significant distance between
182 strain 60-3 and the two other strains (Figure 1A and 1B). In fact, strain 60-3 rRNA operon is
183 found within a single contig (6019 nt) that does not contain any other genes. Given that *P.*
184 *uenonis* strain 60-3 was isolated from a human metagenome (vagina) and exhibits a highly
185 fragmented genome, its 250 contigs underwent analysis using Kraken2 (Figure S1). Of the
186 238 contigs classified by Kraken2 (95.2% of total contigs), 78% corresponds to
187 *Porphyromonas*, totaling 2.1 Mb, which is approximately 85% of the expected average size.
188 It was concluded that *P. uenonis* 60-3 belongs to the *Porphyromonas* genus but its
189 classification within the *P. uenonis* species appears to be incorrect based on ANI/DDH

190 analysis. This genome has been retained for the study but as an unclassified *Porphyromonas*,
191 denoted as PSP_60-3 (Table S1).

192 Secondly, in the case of *P. somerae* KA00683, all indicators suggest a taxonomic
193 assignment inaccuracy. BlastN analysis of the 16S gene fragment (852 nt) reveals a 100%
194 identity with *P. pasteri*, in accordance with ANI (95.7) and DDH (64.53) values, even though
195 the latter two values are slightly below the established threshold (Figure 1B). However, the
196 Kraken2 analysis indicates a genomic mixture, with only 32.5% of the reads being attributed
197 to *Porphyromonas*. Consequently, we have opted not to include *P. somerae* KA00683 in our
198 study, leaving only two strains within this species (Table S1).

199 Finally, regarding *P. canoris* (2 genomes), the difference in the 16S rDNA sequences
200 was associated with an additional 173-nt fragment in strain OH1224, resulting in a longer
201 gene (Figure 1C). However, it's worth noting that these genomes are in draft and fragmented
202 into 14 and 21 contigs. As a result, it is impossible to determine whether this difference
203 represents genuine genomic diversity or a sequencing error. Nevertheless, since all other
204 indicators (ANI, DDH and orthology) confirmed the uniformity of this species, we
205 disregarded this 16S rRNA disparity and consider both genomes as belonging to *P. canoris*.

206 Furthermore, 28 *Porphyromonas* genomes lacked a species label, necessitating a
207 multi-stage analysis. Initially, the genomic contents of these strains were examined using
208 Kraken2, and genomes with less than 80% of *Porphyromonas* reads and/or that reconstructed
209 less than 80% of *Porphyromonas* average genome size (2.5 Mb) were excluded from the
210 study (Figure S1 and Table S1). Consequently, 17 strains were omitted from this study (Table
211 S1). Among the 11 remaining *Porphyromonas* sp., their placement in the Orthofinder species
212 tree based on ANI/DDH metrics (Figure 2) allowed to assign *P. sp.* OH4946 to species *P.*
213 *gulae*; *P. sp.* MGYG-HGUT-04267 to species *P. asaccharolytica*; *P. sp.* UMGS1452 to
214 species *P. uenonis* and *P. sp.* OH1349 and OH2963 to species *P. canoris* (Table S1). Finally,

215 there were 6 *P. sp.* genomes that could not be assigned to any specific group and were
216 individually examined (unassigned, Table S1). This examination further supported the
217 reclassification of *P. uenonis* 60-3 as PSP_60-3 (Figure 2 and Table S1).

218 After completing this taxogenomic biocurated analysis, our study retains a total of
219 126 *Porphyromonas* genomes clustered into 24 groups (comprising 17 species and 7 *P. sp.*
220 singletons), unequally distributed between the genus, ranging from 59 genomes for *P.*
221 *gingivalis* (almost half of all available genomes in the genus) to just one genome for *P.*
222 *loveana*, *P. pasteri* and each *Porphyromonas sp.* (PSP).

223

224 **2. Ffp1 is the only fimbrillin common to all *Porphyromonas***

225 Screening and clustering fimbrillin genes from *Porphyromonas* genomes resulted in
226 the definition of 12 HHM profiles, one for each gene in either FimABCDE or Mfa12345, and
227 two for Ffp1. Searching for sequence similarity in each *Porphyromonas* orfeome, using each
228 of the 12 HHM profiles, enabled the identification and classification of these three fimbriae
229 systems in all *Porphyromonas* genomes (Figure 3).

230

231 **2.1. *fimABCDE* locus**

232 For the FimABCDE proteins (Figure S2A), an expect value (E-value) calibration was
233 performed and set to a minimum threshold of e^{-100} for each of the 5 profiles. Using this
234 threshold, the detection of the locus *fimABCDE* exhibited both sensitivity and specificity,
235 perfectly correlating with presence/absence of each gene.

236 In each genome, these genes are colocalized and organized into operons, with an
237 average size of 7.3 kb. Out of all the genomes, two stand out as outliers: *P. gingivalis* A7436
238 due to an IS5 family transposase ISPg8 insertion in *fimC*, and *P. uenonis* UMGS1452 for
239 which the locus remains incomplete because located at the end of a contig.

240 It is noteworthy that all *P. macacae* strains possess two complete *fimABCDE* loci, a
241 unique feature in *Porphyromonas*. This duplication raises questions about the redundancy or
242 functional complementarity of both *loci*, especially as *P. macacae* JCM15984 has a
243 pseudogenized *fimE* in locus 1 and a pseudogenized *fimD* in locus 2.

244 The utilization of HMM profiles in our search strategy allows for the rapid and
245 unambiguous identification and classification of fimbrial genes, even in cases with low mean
246 amino acid percentage identities: 52.3% (FimA), 63.7 % (FimB), 56.7% (FimC), 48.2%
247 (FimD) and 49.8% (FimE). Additionally, the annotations of FimABCDE proteins are
248 inconsistent, with the majority being labeled as hypothetical proteins or simply categorized as
249 fimbrial proteins without any additional characterization (Figure S2B). As such ontology
250 searches are almost impossible.

251 Moreover, the establishment an E-value threshold facilitates pinpointing
252 abnormalities.

253 For instance, in *P. gingivalis*, for the FimB HMM profile, the E-value is greater than the
254 established threshold due to a nonsense mutation in *fimB* for the ATCC33277 strain (58), this
255 gene is annotated as two genes (PGN_0181, e-value = $2.8e^{-63}$ and PGN_0182, e-value = $1.4e^{-$
256 55). The same case occurs in *P. uenonis*, for the FimE HMM profile, due to the
257 incompleteness of this gene (at the end of contig) for the UMGS1452 strain.

258 In every analyzed *Porphyromonas* genome, the *fimABCDE* locus is consistently
259 present, with only nine groups lacking this operon: *P. asaccharolytica*, *P. bennonis*, *P.*
260 *catoniae*, *P. circumdentaria*, *P. gingivicanis*, *P. pasteri*, *P. somerae*, *P. sp.* OH3588 and *P.*
261 *sp.* UMGS907.

262

263 2.2. *mfa12345* locus

264 Significant E-values ranging from e^{-200} and e^{-100} were observed for each of the 5
265 *Mfa12345* profiles (Figure S2C). Specifically, regarding the *Mfa1* HMM profile, three
266 distinct situations were evident: i. *Mfa1* was recovered, with low E-values, in four species (*P.*
267 *gingivalis*, *P. gulae*, *P. loveana* and *P. macacae*); ii. in 14 groups, *Mfa1* was identified with
268 higher E-values; and iii. in six species (*P. bennonis*, *P. canoris*, *P. cationae*, *P. cangingivalis*
269 and *P. pasteri*, as well as PSP_OH3588), no *Mfa1* was detected. The *Mfa2* HMM profile
270 produces identical results, yielding the same three groups.

271 The *Mfa3* HMM profile successfully identified this protein in the same four species
272 (*P. gingivalis*, *P. gulae*, *P. loveana* and *P. macacae*) and additionally in *P. endodontalis* that
273 contains an *Mfa3*-like protein. Finally, both the *Mfa4* and *Mfa5* HMM profiles exclusively
274 detected these proteins in *P. gingivalis*, *P. gulae*, and *P. loveana* and in three of the six strains
275 of *P. macacae*: JCM15984 and NCTC11632 (isolated from the oral cavity of cats) and
276 OH2859 (isolated from a canine oral cavity). In OH2859, the *mfa12345* operon locus is
277 intact, while in the cases of JCM15984 and NCTC11632, we observed two distinct loci: the
278 first one contains genes encoding *Mfa123* proteins, followed by two genes encoding proteins
279 similar to *FimD* and *FimE* (referred to as *mfa123_fimDE*) and the second comprises genes
280 encoding *Mfa2345* proteins preceded by a non-characterized fimbriin gene that shares
281 similarity with *Ffp1*, indicated by low E-values of $7.3e^{-58}$ for *Ffp1* profile A and $7.8e^{-41}$ for
282 *Ffp1* profile B (referred to as *ffp1-like_mfa2345*). It is worth noting that three strains of *P.*
283 *macacae*, specifically OH2631 (isolated from the canine oral cavity), as well as NCTC13100
284 and DSM20710/JCM13914 (isolated from the macaque oral cavity), exhibit two tandemly
285 organized *mfa123_fimDE* loci. Remarkably, these loci are not identical, displaying an
286 average sequence identity of 53%. In the case of OH2631, these loci are separated by less

287 than 2 kb, while in NCTC13100 and DSM20710, they are separated by a 3 kb region that
288 includes an IS4 pseudogene. None of these three strains harbor the *ffp1-like_mfa2345* locus.

289 *P. endodontalis* features an additional alternative locus comprising six genes,
290 including Mfa1-like, Mfa2, Mfa3-like, followed by two genes encoding lipoproteins and one
291 gene encoding a von Willebrand factor type A (VWA) domain-containing protein.
292 Interestingly, several other species, such as *P. asaccharolytica*, *P. circumdentaria*, *P.*
293 *crevioricanis*, *P. gingivicanis* and *P. uenonis*, also exhibit alternative *loci*, which likely
294 correspond to novel fimbriin systems. These systems require in-depth dedicated future
295 studies for thorough characterization.

296 In conclusion, when considering only the complete *mfa12345* locus as a reference, we
297 identified its presence in four species: *P. gingivalis*, *P. gulae*, *P. loveana*, and *P. macacae*
298 strain OH2859. We also illustrate the effectiveness of HMM profiles in distinguish true *mfa*
299 *loci* from alternative *loci*. As for FimABCDE, the descriptions found in the annotations of
300 Mfa12345 proteins are uninformative, often annotated as hypothetical or fimbria. This
301 labeling makes it nearly impossible to conduct meaningful ontology searches (Figure S2D).

302

303 **2.3. *ffp1***

304 MMseqs2 clustering reveals the separation of Ffp1 orthologs in two distinct groups
305 which resulted in two distinct HMM profiles termed Ffp1_A and Ffp1_B (Figure S2E).
306 Ffp1_A mature amino acid sequences, excluding the signal peptide, share a 57.4% identity,
307 while Ffp1_B sequences exhibit only a 37% identity, primarily due to divergence in *P.*
308 *bennonis*. The identity between the two groups decreases to 24%.

309 Ffp1_A HMM profile retrieves genes from all *Porphyromonas* species except for *P.*
310 *bennonis*, *P. canoris*, *P. cangingivalis*, *P. levii*, and *P. somerae*, which are recovered with
311 Ffp1_B HMM profile. So, remarkably, fimbriin Ffp1 is indeed present in all

312 *Porphyromonas* spp., contrary to FimABCDE and Mfa12345 (except for *P.* sp. UMGS1085
313 where a 186 nt fragment of a gene (at the start of a contig) is identified by Ffp1_A HMM
314 profile with an E-value at 6.7×10^{-22} (Figure S2E). This higher E-value is the result of being
315 obtained for only 61 amino acids instead of about 500 for a Ffp1_A protein.

316 As shown in figure Figure S2F, approximately 70% of the identified Ffp1 proteins are
317 annotated as hypothetical or uncharacterized, 22% as fimbrillin/fimbriae (with half linked to
318 the PGN_1808 protein, described as Ffp1 in the *P. gingivalis* ATCC33277 reference strain)
319 and 8% are described as lipoproteins.

320 Using HMMsearch with both Ffp1_A and Ffp1_B profiles, using an E-value
321 threshold at e^{-100} , in Ensembl Genome Bacteria (taxid:2) database, only *Porphyromonas*
322 proteins are retrieved. As a conclusion, Ffp1 fimbrillins are the sole fimbriae proteins
323 conserved across all *Porphyromonas* species, making them unique to the genus.

324

325 **3. Characterization of *Porphyromonas* FFp1 fimbriae**

326 Ffp1 exhibits variable pre-cleavage sizes among *Porphyromonas* species, in both
327 subclasses. For the Ffp1_A group, protein sizes range from 439 aa (*P. circumdentaria* DSM
328 103022) to 553 aa (*P. asaccharolytica* PR426713P-I), and for the Ffp2_B, from 483 aa (*P.*
329 *somerae* DSM 23387) to 527 aa (*P. canoris*) (Figure 4A). Size is well conserved within
330 *Porphyromonas* species with the exception of the *P. asaccharolytica*, *P. circumdentaria*, *P.*
331 *macacae* and *P. uenonis* for Ffp1_A, and *P. bennonis* for Ffp1_B (Figure 4A).

332 The observed differences for *P. asaccharolytica* are due to the presence of 33
333 additional nucleotides in strain PR426713P-I (at position 88-120), absent in strain DSM
334 20707. For *P. circumdentaria*, it is a 175 nt shorter annotation in strain DSM 103022
335 (compared to strain ATCC 51356). For *P. macacae* these are due to the gene encoding
336 Ffp1_A being at the end of the contig and truncated at the 5' end, in strain *P. macacae* JCM

337 15984. For *P. uenonis*, it is also the choice of an alternative start codon for the UMGS1452
338 strain, 34 amino acids upstream of those chosen for the DSM23387 and JCM13868 strains.
339 Finally, for *P. bennonis*, at position 1410 in the DSM 23058 strain, a C base, absent from the
340 JCM 16335 strain, leads to a frameshift. This frameshift leads to a shorter C-terminal
341 sequence compared to DSM 23058 strain. Note that for *P. somerae*, the sizes are similar but
342 the annotated sequences are "shifted" and proteins different on the N-terminal (20 aa longer
343 in DSM_23387 compared to St14) and C-terminal (21 aa shorter in DSM_23387 due to a
344 partial CDS at the end of the contig).

345 Accurate annotation of the N-terminus of proteins, which predicts their cellular
346 localization, is crucial and deserves the attention of annotators. For this purpose, we re-
347 annotated the start codons of Ffp1, when needed, to optimize both the SPII cleavage
348 prediction score and the presence of charged residues at the N-terminus, followed by
349 hydrophobic amino acids. The resulting re-annotations and their implications for cell
350 localization predictions are listed in Table S2.

351 In the absence of comprehensive biocuration, a substantial part of Ffp1 proteins are
352 predicted as cytoplasmic (*P. asaccharolytica*, *P. catoniae*, *P. circumdentaria* DSM 103022,
353 *P. somerae* St14) or having localization predictions classified as indeterminate (PSP
354 UMGS107, PSP UMGS166, PSP UMGS907, *P. uenonis* DSM 23387, *P. uenonis* JCM
355 13868). Some proteins are predicted to be cleaved by SPII, but biocuration enhances both the
356 signal peptide prediction score and the likelihood of cleavage by SPII. As a result of this
357 reannotation work, all Ffp1 proteins are predicted as lipoproteins, with a signal peptide of
358 about 20 amino acids (15 to 25 aa), consistent with the requirements cited previously: 2 to 4
359 positively charged amino acids followed by a hydrophobic region of 10 to 15 aa (Figure S3)
360 and a lipobox [ASG] \downarrow C positions -1 to 1 (Figure 4B). *In silico* predictions also confirm the
361 predicted palmitoylation (addition of acyl chains) of the cysteine residue.

362 These biocurated peptide signals exhibit a high degree of intra-species conservation,
363 while demonstrating significant inter-species variability, with only a 25% pairwise identity
364 when considering all species collectively (min. 5% - max. 100%, Figure 4C). However, two
365 groups characterized by similar signal peptide sequences can be discerned: a first one formed
366 by *P. gingivalis* and *P. gulae* (ca. 86% identity) and a second more consistent, composed of
367 species *P. asaccharolytica*, *P. uenonis*, PSP_60-3, PSP_UMGS907, PSP_UMGS18 and
368 PSP_UMGS166 (66.7 to 100% identity, Figure 4C). The same groups were observed when
369 examining the lipobox motif.

370 As shown in Figure 4A (second panel), Ffp1 signal peptides biocuration not only results
371 in more consistent predictions of their cellular localization, but also leads to a
372 homogenization of their size, both within and across species, with the exception of *P.*
373 *bennonis* (since the frameshift occurs in the 3' region of the gene). This size homogenization
374 becomes even more pronounced following signal peptide cleavage (Figure 4A, third frame).
375 It can be seen that mature Ffp1s in group B are larger than those in group A by about 20 aa.

376 As shown in Figure 4D, the average intra-specific identity of the Ffp1_A subclass is very
377 high and ranges from 100% to 94.8% depending on the species. The most divergent species
378 are *P. macacae*, *P. gulae*, and *P. uenonis*. In the first two cases, this divergence can be
379 attributed to the coexistence of two distinct homology groups within the same species.
380 However, regrettably, the available metadata does not provide sufficient information to
381 elucidate the underlying reasons for these discrepancies. For *P. uenonis*, the strain
382 UMGS1452 that derived from a metagenome is different from the two other strains. As
383 previously noted, the conservation of interspecific Ffp1_A sequences is small (57.5%) with
384 only 4.5% of identical sites between all of them. When examining the Ffp1_B group, it is
385 worth noting that the average intra-specific identity is elevated, oscillating between 98.8 and
386 91.1% (Figure 4D). *P. bennonis* is the most divergent because the two strains have proteins

387 with the last 75 aa that differ. It is noteworthy that Ffp1_B is less homogeneous than Ffp1_A
388 with only an average inter-specific identity of 36.4% and 4.3% identical sites. The number of
389 conserved sites decreases to 0.7% if we compare both groups, Ffp1_A and Ffp1_B.

390

391 **4. 3D structures confirm that *Porphyromonas Ffp1* are fimbrillins**

392 As the signal peptide is absent in the mature protein, it was excised prior to structure
393 prediction for all Ffp1 proteins. PSIPred predict 30% to 44% residues as strand (mean = 36.6,
394 SD = 3.4) and 2% to 7% residues as helix (mean = 5.4, SD = 1.4) for Ffp1_A group. For
395 group Ffp1_B, predictions concern 24% to 42% amino acids in strand (mean = 29.6, SD =
396 6.5) and 4% to 9% in helix (mean = 7, SD = 2.1).

397 The optimal structures for all *Porphyromonas* Ffp1 representatives, as predicted by
398 Robetta and assessed by ERRAT and Verify3D, are depicted Figure 5. These structures were
399 subjected to comparison with existing models and, the best hit, obtained either via VAST+ or
400 Phyre2, correspond to *Bacteroides ovatus* cell adhesion protein (BACOVA_01548,
401 4JRF.pdb) for all *Porphyromonas* Ffp1 proteins, irrespective of species or Ffp1_class.

402 According to Phyre2 and iBPA results (Figure S4), more than 82% of Ffp1 sequences
403 were modeled with 100.0% confidence against 4JRF.pdb. Superposition of *Porphyromonas*
404 Ffp1 and BACOVA_01548 3D structures were performed by iBPA and all evaluation values
405 (RMSD, GDT_TS), reflect good overall similarity. For all overlapping morphologies, the
406 aligned fraction is about 50% of the protein sequence, with mean reported RMSDs of 2.26 Å
407 [range 2.09 to 2.53 Å] and mean GDT-TS distance scores of 32 [range 32 -37.3] (Figure S4).
408 For *Porphyromonas gingivalis*, the structures of FimA (4Q98.pdb) and Mfa1(5NF2.pdb) are
409 available and comparisons by superposition between Ffp1 and these two other fimbrillins
410 (Figure S5) confirm that Ffp1 is indeed a new distinct *Porphyromonas* fimbrillin family.

411

412 5. *Porphyromonas* Ffp1 are ancestral orthologs but not syntelogs

413 In *P. gingivalis*, *ffp1* is the fourth gene in an operon-like structure comprising a gene
414 encoding a cysteinyl-tRNA synthetase, a second gene encoding a patatin-like protein, and a
415 third gene encoding a group 2 glycosyltransferase. An identical *locus* is found in all *P. gulae*
416 genomes, while is absent in any other *Porphyromonas* (Figure S6). The *P. asaccharolytica*,
417 *P. uenonis*, *P. sp.* UMGS18 and *P. sp.* UMGS107 group, mentioned earlier in this article,
418 show a syntenic pattern upstream of *ffp1*, characterized by the presence of two conserved
419 genes encoding dihydroorotate dehydrogenases, crucial enzymes involved in *de novo*
420 pyrimidine biosynthesis in prokaryotic cells. However, there is no direct association with the
421 fimbriin function. Furthermore, the intergenic space spanning approximately 300
422 nucleotides is sufficiently substantial to preclude any functional linkage between these genes.
423 The second group, comprising *P. sp.* UMGS166 and *P. sp.* UMGS907, also previously
424 mentioned, shows synteny downstream of *ffp1* with a gene encoding a nitronate
425 monooxygenase (degradation of propionate-3-nitronate) and another encoding a 4-hydroxy-
426 tetrahydrodipicolinate synthase (involved in lysine biosynthesis). Similar to the previous
427 group, no discernible functional relationship appears to exist, and the intergenic space of
428 about 200 nt seems to confirm this hypothesis. For the other *Porphyromonas*, each species
429 exhibits a distinct gene organization arrangement surrounding *ffp1* (Figure S6).

430 In conclusion, with the exception of phylogenetically closely related species, we find no
431 preserved synteny in the *ffp1* locus, which would reflect the absence of co-localization
432 constraints for co-functional genes. Nevertheless, as demonstrated by the tanglegram
433 juxtaposing the orfeome tree and the Ffp1 tree (Figure 6), the remarkable congruence
434 between these two trees provides compelling evidence that Ffp1 is an ancestral protein of
435 *Porphyromonas*, and its evolution would have closely paralleled the evolutionary trajectory
436 of the entire genus. This observation also holds true for the differentiation between the two

437 Ffp1 classes (Figure 6). The absence of genes conservation in close chromosomal proximity
438 to *ffp1*, along with the presence of a significant 5' intergenic space (Figure S6), not only
439 signifies the absence of selection pressure around this gene but also strongly suggests that
440 *ffp1* functions as an independent transcriptional unit.

441

442 **Discussion**

443 Our analysis of fimbriin *loci* within the *Porphyromonas* genus initiated with the
444 retrieval of genomes from the NCBI RefSeq database. The first step encompassed the
445 validation of genus-level assignment for each genome retrieved, followed, when feasible, by
446 species-level confirmation. The Overall Genome Relatedness Indices (OGRI), namely: digital
447 DNA-DNA hybridization distance (DDH) and genome Average Nucleotide Identity (gANI)
448 were used to classify genomes into monophyletic groups. These OGRIs are increasingly used
449 in taxogenomic studies and serve as a valuable tool for validating the taxonomic
450 classification of isolates of interest (59). Likewise, in accordance with prior research, we
451 employed more conventional methodologies for species-level genomes grouping, such as
452 evaluating the percentage identity of the gene encoding 16S rRNA (when annotated) (60). Our
453 study underscores the critical necessity of rigorously confirming the taxonomic classifications
454 of genomes before embarking on any comparative genomics analysis to ensure their
455 accuracy. Moreover, this checking step enables the possibility of taxonomic reassignment
456 when warranted, aligning with findings from previous studies (61–63). In this investigation,
457 we have identified genomes erroneously labeled as *Porphyromonas* (i.e strain 31_2, which is
458 a *Parabacteroides*), misassignment of *Porphyromonas* to species (i.e strain 60.3, which does
459 not belong to *P. uenonis*) as well as metagenomic mixture such as strain KA000683
460 imperfectly assigned to *P. somerae*.

461 Our study also raises questions about genomes assigned to *Porphyromonas* without
462 any species assignment (28 out of 144 genomes, i.e., 19.5%). They all correspond to
463 incomplete draft genomes which introduces bias into studies that rely on them (64). We can
464 specifically mention the presence of gaps, local assembly errors, chimeras and contamination
465 by fragments from other genomes (65,66). This contamination, defined as the presence of
466 foreign sequences within a genome, can lead to incorrect functional inferences such as higher
467 rates of horizontal gene transfer (HGT) and errors in phylogenomic studies. Such errors can
468 be propagated throughout the scientific community and have been documented to exist in
469 databases (67). To mitigate these types of errors, several studies, including the present one,
470 advocate the practice of data biocuration throughout the study. To identify potential
471 contamination in draft genomes, we employed Kraken2 software and assessed the cumulative
472 contig size of incomplete genomes. By applying specific inclusion criteria, we were able to
473 disqualify 17 draft genomes, corresponding to metagenomic mixtures and inaccurately
474 labeled as *Porphyromonas*. Furthermore, among the 11 remaining draft genomes, our
475 taxonomic approach led to the reclassification of 5 genomes into four previously described
476 species (*P. gulae*, *P. asaccharolytica*, *P. uenonis* and 2 genomes in *P. canoris*). The
477 remaining 6 genomes that cannot be assigned to already described species may potentially
478 represent novel, yet undescribed species, akin to hypotheses proposed in other bacterial
479 genera (63,67). This suggests that the genus *Porphyromonas* may encompass a greater degree
480 of species diversity than previously recognized.

481 Thus, in this study, we retained 126 *Porphyromonas* genomes (24 clades comprising
482 17 species and 7 singletons) to describe fimbriae loci. To accomplish our research objectives,
483 distant homology between proteins must be detected and is fundamental for enabling
484 comparative and evolutionary investigations, shedding light on protein families, and
485 providing insights into their molecular structures and functions (68).

486 Current orthology detection methods include Position Specific Scoring Matrix
487 (PSSM) techniques, like PSI-BLAST (Position-specific iterated BLAST, (69), which
488 generate substitution score profiles by accounting for residue variability within homologous
489 sequence families (70). An even more effective approach involves Hidden Markov Model
490 (HMM) profiles, which incorporate emission and transition state probabilities at each protein
491 sequence position, making them a superior choice for identifying distant homology (70,71).

492 Using ontology as a protein search strategy search is ineffective, as most fimbrillin
493 genes are poorly annotated or annotated as "hypothetical protein" (between 21.1% and 88.6%
494 of annotated genes). Specifically, stem and anchor proteins (FimAB or Mfa12) are better
495 annotated with deficient annotation rates ranging from 21.1% to 50.7%. In contrast, accessory
496 proteins (FimCDE or Mfa345) suffer from particularly poor annotations with error
497 percentages ranging from 58.9% to 88.6%. These annotation errors are present within the
498 databases and, without biocuration and correction, are likely to persist, potentially
499 perpetuating inconsistencies, inaccuracies, and errors in subsequent genome annotations (72).
500 For example, for a gene family, nearly 20% of sequences may exhibit significant errors such
501 as inaccuracies in gene names, partial sequences or initiation codon misassignments (73). In
502 the context of less extensively researched bacterial species, as is the case in this study, the
503 prevalence of erroneous or uninformative annotations are much higher, reaching 77.1% of
504 sequences identified as Ffp1 where the annotation was "hypothetical protein" or
505 "lipoprotein".

506 In this study, we leveraged 12 HMM profiles developed from *P. gingivalis* genomes,
507 which were further refined through a strategy involving functional domain screening,
508 clustering and biocuration. This approach enabled a comprehensive exploration of the
509 *Porphyromonas* orfeomes, revealing variations in the three fimbriae loci across all species
510 within this genus.

511 The *fimABCDE* locus is present in 9 (of 24 groups, or 37.5% of *Porphyromonas*
512 species) with two distinct *fim* loci present in all *P. macacae* genomes. The *mfa12345* locus is
513 present only in three closely related species (*P. gingivalis*, *P. gulae* and *P. loveana*). For this
514 locus, hybrid *fim/mfa* or *ffp1/mfa* loci are present in two species (*P. endodontalis* and *P.*
515 *macacae*): *mfa123_fimDE* and *ffp1-like_mfa2345* in *P. macacae*; and a distinctive six-gene
516 locus in *P. endodontalis*. This locus encompasses genes encoding Mfa1-like, Mfa2, and
517 Mfa3-like proteins, along with two genes responsible for lipoproteins and a gene encoding a
518 protein featuring a von Willebrand factor type A (VWA) domain. Interestingly, for the gene
519 encoding Mfa5, the prevailing description is rather nondescript, simply stating it as a "protein
520 containing a VWA domain". This description, however, falls short in conveying the
521 functional significance of this gene. It's worth emphasizing that proteins featuring VWA
522 domains play pivotal roles in diverse biological processes, including but not limited to cell
523 adhesion and defense mechanisms. Thus, a more detailed annotation is warranted to better
524 appreciate the functional implications of Mfa5 (74).

525 Finally, other species (i.e. *P. asaccharolytica*, *P. circumdentaria*, *P. crevioricanis*, *P.*
526 *gingivicanis* and *P. uenonis*) have fimbriin genes identified through HMM profiles that
527 remain uncharacterized. These two loci, *fimABDCE* and *mfa12345*, have been described in
528 other closely related species, for example an Mfa system (with only *mfa1* and *mfa2*) in
529 *Bacteroides thetaiotaomicron* (75), and a cluster with *fimABCDE*-like genes and genes
530 similar to either *mfa1/mfa2* or *mfa4/mfa2* with either *mfa1* or *mfa4* encoding the fimbriae
531 stem and *mfa2* as an anchor in *Parabacteroides distasonis* (76). The *fim* and *mfa* loci in
532 *Porphyromonas* spp. will be the main subject of an ulterior publication.

533 Concerning Ffp1 fimbriae (77.1% of all *ffp1* genes were deficiently annotated), this
534 protein was most recently described in *P. gingivalis* (13,14). The encoding gene has two
535 variants, denoted as A and B in our study. Ffp1_A is the predominant variant found in 19

536 *Porphyromonas* species/groups, whereas Ffp1_B is restricted to only 5 species (*P. bennonis*,
537 *P. canoris*, *P. cangingivalis*, *P. levii*, and *P. somerae*). Furthermore, this study demonstrates
538 that the utilization of HMM profiles reveals that *ffp1* is confined to the *Porphyromonas* genus
539 and is absent in closely related genera like *Bacteroides* or *Prevotella*. This finding contrasts
540 with approaches employing BLASTp and psiBLAST (16).

541 The presence of multiple fimbriae loci within genomes is a common phenomenon
542 observed in other bacterial models. These loci are often associated with general niche
543 colonization abilities or the adhesion to more specific substrates (77,78). Further
544 investigations are needed on species more closely related to *Porphyromonas* and within this
545 bacterial genus. These studies can shed light on aspects such as host specificity and their
546 association with species-related pathologies (79).

547 Given that the majority of *in silico* coding sequence (CDS) annotators tend to
548 prioritize the prediction of the longest possible Open Reading Frame (ORF) by favoring the
549 initiation codon (ATG) over alternative codons (TTG and GTG) (80,81), and considering the
550 variable size of proteins across *Porphyromonas* species, we conducted a thorough
551 examination of the annotated initiation codons for each predicted Ffp1. Given that fimbrellins
552 are lipoproteins (9), their N-terminal region is expected to feature a signal peptide starting
553 with positively charged amino acids, followed by hydrophobic amino acids, and concluding
554 with a cysteine-terminated lipobox, which serves as the cleavage site for signal peptidase II.
555 The biocuration of start codons led to a more consistent protein size post-signal peptide
556 cleavage. Additionally, the extracellular prediction of mature lipoproteins was confirmed,
557 characterized by the presence of charged and hydrophobic residues, the lipobox, and a
558 palmitoylation site. These features align with the ancestral nature of FFp1.

559 In addition, Ffp1 3D modeling of the mature protein was performed with several
560 software packages, and the predictions were evaluated with classical metrics (82,83). In all

561 cases, the generated models were compared with existing 3D structures, and the most
562 significant match was found with the cell adhesion protein BACOVA_01548 from
563 *Bacteroides ovatus* (3). This *B. ovatus* protein has not been extensively studied, but was
564 classified by the authors as the stem of a type V pilus, sharing common features with type V
565 fimbriae. These characteristics include export to the periplasm as a lipoprotein (prepilin),
566 subsequent delivery to the outer membrane, translocation to the cell surface and cleavage by
567 Rgp (Arg-gingipain) (4,84).

568 Moreover, this new fimbrillin, Ffp1 exhibits notable distinctions from both FimA and
569 Mfa1, as evident from the obtained metrics when superimposing the 3D structures of these
570 proteins available for *P. gingivalis*. Furthermore, the gene arrangement of *ffp1* differs from
571 the *fim* and *mfa* operons as the gene encoding Ffp1 does not appear to be in an operon
572 structure.

573

574 **Concluding remarks**

575 HMM profiles are potent tools for detecting distant homologies and facilitating
576 phylogenetic studies. For conducting these investigations, meticulous manual biocuration is
577 essential, as with any *in silico* research. In this article, these HMM profiles make it possible
578 to discriminate, without ambiguity, three *Porphyromonas* fimbriae and to describe their
579 distribution: *mfa12345*, limited to the three closely related species (*P. gingivalis*, *P. gulae* and
580 *P. loveana*), *fimABCDE* present in nearly 40% of the *Porphyromonas* species and *ffp1*,
581 present in all *Porphyromonas* but restricted to this bacterial genus. Our study predicts that
582 Ffp1 is a new fimbrillin, distinct from FimA and Mfa1. It is closely related to another type V
583 fimbrillin protein, BACOVA_01548, as evidenced by manual start codon curation and 3D
584 modeling. Given the ancestral nature of Ffp1, as elucidated by our study, and its presence in
585 all studied *Porphyromonas* genomes, in contrast to the fimbrillins Fim and Mfa, the question

586 of its function becomes paramount. Especially in absence of co-localization of accessory
587 genes ensuring its stability, assembly, and anchorage to the cell surface. What role does it
588 play in the production and cargo of Outer Membrane Vesicles (OMVs), a phenomenon
589 observed in numerous studies? Further wet-lab investigations are necessary to address these
590 pending inquiries.
591

592 Legend to figures

593 Figure 1. Validation of the taxogenomic assignment of *Porphyromonas* genomes. **A.**

594 Intra-species homogeneity was checked by calculating intra-species distances using gANI,
595 rRNA 16S identity and DDH. **B.** Checking *P. uenonis* and *P. somerae* genomes homogeneity
596 using OrthoANI. **C.** Difference in 16S rRNA sequences of two strains of *P. canoris* with 173
597 nt insertion in strain OH1224. 3-letter code acronyms correspond to ASA: *P.*
598 *asaccharolytica*; BEN: *P. bennonis*; CAN: *P. canoris*; CAT: *P. catoniae*; CGI: *P.*
599 *cangingivalis*; CIR: *P. circumdentaria*; CRE: *P. crevioricanis*; END: *P. endodontalis*; GGI:
600 *P. gingivicanis*; GIN: *P. gingivalis*; GUL: *P. gulae*; LEV: *P. levii*; LOV: *P. loveana*; MAC:
601 *P. macacae*; PAS: *P. pasteri*; SOM: *P. somerae*; and UEN: *P. uenonis*.

602

603 Figure 2. Phylogenetic species tree derived from OrthoFinder analysis. This tree was

604 used to place some *Porphyromonas* spp. into UEN (UMGS1452), ASA (MGYG-HGUT-
605 0467), CAN (OH2963 and OH1349) and GUL (OH4946) genus, after confirmation via
606 OrthoANI.

607

608 Figure 3. Heatmap depicting the presence/absence of fimbrillins. The heatmap scale color

609 indicates whether fimbriae systems (FimABCDE, Mfa12345 or Ffp1_A or B) were detected:
610 white (absence), dark purple (presence as one *locus*) and light purple (presence as two *loci*).

611

612 Figure 4. **A.** Violin plots of Ffp1_A and Ffp1_B amino acid lengths. From left to right:

613 sizes as initially annotated in Genbank files (no curation), sizes after signal peptide (SP)
614 biocuration prior to cleavage, and sizes after SP cleavage by signal peptidase II (SPII). In the
615 box plot associated with each violin plot, the middle line represents the median and the
616 whiskers indicate the interquartile range. **B.** Multiple sequence alignment and sequence

617 **logo of Ffp1 lipobox.** Boxes represent groups of identical sequences. **C. Heat map**
618 **illustrating the percent nucleotide identity** of Ffp1 signal peptides. **D. Circular**
619 **phylogram of *Porphyromonas* Ffp1 proteins distance tree.** The Ffp1_A proteins are
620 depicted in warm colors, while Ffp1_B are shown in various shades of blue. The boxes
621 indicate the minimum, average, and maximum intraspecific identity values. If only one value
622 is displayed, it represents the average identity percentage.

623

624 **Figure 5.** Predicted tertiary structure for mature proteins of *Porphyromonas* reference strains
625 (one per genus). These structures correspond to predictions made by Robetta and evaluated
626 by ERRAT and Verify3D. Only the best prediction is represented. Ffp1_A proteins are in red,
627 Ffp1_B in blue. BACOVA_01548 was also predicted using Robetta.

628

629 **Figure 6.** Tanglegram comparing the tree constructed from the primary sequences of the Ffp1
630 proteins in representative strains of *Porphyromonas* (on the left) with the species tree based
631 on the orthology of the orfeomes.

632

633 **Legend to supplementary figures**

634 **Figure S1. Sankey diagrams representing Kraken 2 report results for each**
635 ***Porphyromonas* sp. genome.** Each diagram illustrates the percentage of the genome
636 classified under the genus *Porphyromonas* and the cumulative size of the accurately assigned
637 fragments for each strain.

638

639 **Figure S2. A-C-E:** Box plot representing hmmsearch E-value results for all *Porphyromonas*
640 genus groups and all HMM profiles used in this study. The dotted lines represent the

641 thresholds used. **B-D-F:** Frequencies of the terms employed to describe the genes of the 3
642 fimbriae systems of *Porphyromonas*, as originally annotated in the Genbank files.

643

644 **Figure S3. Ffp1 signal peptide prediction by the SignalP-6.0.** For each *Porphyromonas*
645 group, a reference sequence was chosen (named at the top of the SignalP6 graphs). For each
646 group, we present the intra-specific consensus sequence of the signal peptides (after
647 biocuration when required) in the form of a logo. The blue stars represent the charged amino
648 acids (predicted by EMBOSS charge prediction) and the orange curve represents the
649 prediction of hydrophobicity (ProtScale, Amino acid Hydrophobicity using Kyte and Doolittle
650 method).

651

652 **Figure S4. Evaluation of 3D models of Ffp1 proteins predicted *in silico*.** Assessments
653 were performed using iBPA and Phyre2 based on template 4JRF (BACOVA_01548). Right
654 column shows superposition of predicted structure (green) with model structure (in red) using
655 iPBA.

656

657 **Figure S5. Comparison of Robetta structures prediction for *P. gingivalis* FimA, Mfa1
658 and Ffp1.** The boxes represent the results of the superposition in Ffp1 and either
659 BACOVA_01548, GIN_FimA or GIN_Mfa1.

660

661 **Figure S6. Diagram of synteny in the vicinity of the *Ffp1* gene in the different species of
662 *Porphyromonas*.** Ffp1 genes are shown in light blue and surrounding genes in dark blue
663 using Geneious Prime. The yellow boxes correspond to contig extremities.

664

665 **Supplementary tables**

666 **Table S1. List of *Porphyromonas* genomes used in this study.** Genomes were grouped into
667 clades following genomic data-driven taxonomic clustering. Complete genomes are in green,
668 *Porphyromonas* sp. grouped in a clade are in blue, a mislabelled *P. somerae* is indicated in
669 yellow, *P.* sp. genomes that could not be grouped with others are in dark grey and two
670 mislabelled “*Porphyromonas*” genomes are in purple. All accession numbers are indicated as
671 well as the strain.

672

673 **Table S2. List and details of all *ffp1* genes identified during this study.** Information
674 includes HMM group membership, old and new locus_tags, items related to start codon
675 reannotation (when needed), cell localization predictions, peptide signal, and before and after
676 cleavage sizes.

677

678 **Supplementary material**

679 **HMM profiles.** This archive contains the 12 HMM profiles that were generated and
680 employed in this study for the detection and the classification of *Porphyromonas* fimbriae.

681

682 **References**

- 683 1. Lukaszczyk M, Pradhan B, Remaut H. The biosynthesis and structures of bacterial pili.
684 In: *Bacterial Cell Walls and Membranes*. Springer; 2019. p. 369–413.
- 685 2. Hospenthal MK, Costa TRD, Waksman G. A comprehensive guide to pilus biogenesis
686 in Gram-negative bacteria. *Nat Rev Microbiol*. 2017;15(6):365–79.
- 687 3. Xu Q, Shoji M, Shibata S, Naito M, Sato K, Elsliger M-AA, et al. A distinct type of
688 pilus from the human microbiome. *Cell*. 2016;165(3):690–703.
- 689 4. Shibata S, Shoji M, Okada K, Matsunami H, Matthews MM, Imada K, et al. Structure
690 of polymerized type V pilin reveals assembly mechanism involving protease-mediated
691 strand exchange. *Nat Microbiol*. 2020 Apr 13;5(6):830–7.
- 692 5. Yoshimura F, Takahashi K, Nodasaka Y, Suzuki T. Purification and characterization
693 of a novel type of fimbriae from the oral anaerobe *Bacteroides gingivalis*. *J Bacteriol*.
694 1984/12/01. 1984;160(3):949–57.
- 695 6. Hamada N, Sojar HT, Cho MI, Genco RJ. Isolation and characterization of a minor
696 fimbria from *Porphyromonas gingivalis*. *Infect Immun*. 1996/11/01.
697 1996;64(11):4788–94.
- 698 7. Fujiwara-Takahashi K, Watanabe T, Shimogishi M, Shibasaki M, Umeda M, Izumi Y,
699 et al. Phylogenetic diversity in *fim* and *mfa* gene clusters between *Porphyromonas*
700 *gingivalis* and *Porphyromonas gulae*, as a potential cause of host specificity. *J Oral*
701 *Microbiol*. 2020 Jan 1;12(1):1775333.
- 702 8. Onoe T, Hoover CI, Nakayama K, Ideka T, Nakamura H, Yoshimura F. Identification
703 of *Porphyromonas gingivalis* pefimbrilin possessing a long leader peptide: possible
704 involvement of trypsin-like protease in fimbrilin maturation. *Microb Pathog*.
705 1995/11/01. 1995;19(5):351–64.
- 706 9. Shoji M, Naito M, Yukitake H, Sato K, Sakai E, Ohara N, et al. The major structural

- 707 components of two cell surface filaments of *Porphyromonas gingivalis* are matured
708 through lipoprotein precursors. *Mol Microbiol.* 2004/05/29. 2004 Jun;52(5):1513–25.
- 709 10. Kuboniwa M, Amano A, Hashino E, Yamamoto Y, Inaba H, Hamada N, et al. Distinct
710 roles of long/short fimbriae and gingipains in homotypic biofilm development by
711 *porphyromonas gingivalis*. *BMC Microbiol.* 2009;9:1–13.
- 712 11. Shoji M, Yoshimura A, Yoshioka H, Takade A, Takuma Y, Yukitake H, et al.
713 Recombinant *Porphyromonas gingivalis* FimA preproprotein expressed in *Escherichia*
714 *coli* is lipidated and the mature or processed recombinant FimA protein forms a short
715 filament in vitro. *Can J Microbiol.* 2010/11/16. 2010;56(11):959–67.
- 716 12. Yoshimura F, Murakami Y, Nishikawa K, Hasegawa Y, Kawaminami S. Surface
717 components of *Porphyromonas gingivalis*. *J Periodontal Res.* 2008/11/01.
718 2009;44(1):1–12.
- 719 13. Gui MJ. Characterization of the *Porphyromonas gingivalis* protein PG1881 and its
720 roles in outer membrane vesicle biogenesis and biofilm formation. University of
721 Melbourne; 2016.
- 722 14. Nagano K, Hasegawa Y, Yoshida Y, Yoshimura F. Novel fimbriilin PGN_1808 in
723 *Porphyromonas gingivalis*. *PLoS One.* 2017/03/16. 2017;12(3):e0173541.
- 724 15. Hasegawa Y, Iijima Y, Persson K, Nagano K, Yoshida Y, Lamont RJ, et al. Role of
725 Mfa5 in expression of Mfa1 fimbriae in *porphyromonas gingivalis*. *J Dent Res.*
726 2016/06/22. 2016;95(11):1291–7.
- 727 16. Gupta RS, Lorenzini E. Phylogeny and molecular signatures (conserved proteins and
728 indels) that are specific for the Bacteroidetes and Chlorobi species. *BMC Evol Biol.*
729 2007;7(1):71.
- 730 17. Veith PD, Shoji M, Scott NE, Reynolds EC. Characterization of the O-Glycoproteome
731 of *Porphyromonas gingivalis*. Avci FY, editor. *Microbiol Spectr.* 2022 Feb 23;10(1).

- 732 18. Rocha FG, Ottenberg G, Eure ZG, Davey ME, Gibson FC. Sphingolipid-Containing
733 Outer Membrane Vesicles Serve as a Delivery Vehicle To Limit Macrophage Immune
734 Response to *Porphyromonas gingivalis*. Whiteley M, editor. *Infect Immun*. 2021 Mar
735 17;89(4).
- 736 19. Iwashita N, Nomura R, Shirai M, Kato Y, Murakami M, Matayoshi S, et al.
737 Identification and molecular characterization of *Porphyromonas gulae* fimA types
738 among cat isolates. *Vet Microbiol*. 2019 Feb;229:100–9.
- 739 20. Oishi Y, Watanabe K, Kumada H, Ishikawa E, Hamada N. Purification and
740 characterization of a novel secondary fimbrial protein from *Porphyromonas gulae*. *J*
741 *Oral Microbiol*. 2012/09/25. 2012;4.
- 742 21. Collings S, Love DN. Further studies on some physical and biochemical characteristics
743 of asaccharolytic pigmented *Bacteroides* of feline origin. *J Appl Bacteriol*. 1992/06/01.
744 1992 Jun;72(6):529–35.
- 745 22. Love DN, Bailey GD, Collings S, Briscoe DA. Description of *Porphyromonas*
746 *circumdentaria* sp. nov. and reassignment of *bacteroides salivus* (Love, Johnson,
747 Jones, and Calverley 1987) as *Porphyromonas* (Shah and Collins 1988) *salivosa* comb.
748 nov. *Int J Syst Bacteriol*. 1992/07/01. 1992;42(3):434–8.
- 749 23. KOYATA Y, WATANABE K, TOYAMA T, SASAKI H, HAMADA N. Purification
750 and characterization of a fimbrial protein from *Porphyromonas salivosa* ATCC 49407.
751 *J Vet Med Sci*. 2019;81(6):916–23.
- 752 24. Kim M, Oh H-S, Park S-C, Chun J. Towards a taxonomic coherence between average
753 nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of
754 prokaryotes. *Int J Syst Evol Microbiol*. 2014 Feb;64(Pt 2):346–51.
- 755 25. Meier-Kolthoff JP, Auch AF, Klenk HP, Goker M. Genome sequence-based species
756 delimitation with confidence intervals and improved distance functions. *BMC*

- 757 Bioinformatics. 2013/02/26. 2013;14:60.
- 758 26. Wayne LG. International committee on systematic bacteriology: Announcement of the
759 report of the ad hoc committee on reconciliation of approaches to bacterial
760 systematics. Zentralblatt für Bakteriologie Mikrobiologie und Hygiene Ser A Med Microbiol Infect
761 Dis Virol Parasitol. 1988 Jun;268(4):433–4.
- 762 27. STACKEBRANDT E, GOEBEL BM. Taxonomic Note: A Place for DNA-DNA
763 Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in
764 Bacteriology. Int J Syst Evol Microbiol. 1994 Oct 1;44(4):846–9.
- 765 28. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput
766 ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat
767 Commun. 2018 Nov;9(1):5114.
- 768 29. Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides
769 NC, et al. Microbial species delineation using whole genome sequences. Nucleic Acids
770 Res. 2015 Aug;43(14):6761–71.
- 771 30. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2.
772 Genome Biol. 2019 Nov 28;20(1):257.
- 773 31. Liang Q, Liu C, Xu R, Song M, Zhou Z, Li H, et al. fIDBAC: A Platform for Fast
774 Bacterial Genome Identification and Typing. Front Microbiol. 2021;12:723577.
- 775 32. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative
776 genomics. Genome Biol. 2019 Nov;20(1):238.
- 777 33. Davis JJ, Wattam AR, Aziz RK, Brettin T, Butler R, Butler RM, et al. The PATRIC
778 Bioinformatics Resource Center: expanding data and analysis capabilities. Nucleic
779 Acids Res. 2019 Oct 31;
- 780 34. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search:
781 HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 2013

- 782 Jul 1;41(12):e121–e121.
- 783 35. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et
784 al. Pfam: The protein families database in 2021. *Nucleic Acids Res.* 2021 Jan
785 8;49(D1):D412–9.
- 786 36. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the
787 analysis of massive data sets. *Nat Biotechnol.* 2017 Nov 16;35(11):1026–8.
- 788 37. Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many
789 protein sequences. *Protein Sci.* 2018 Jan;27(1):135–45.
- 790 38. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious
791 Basic: an integrated and extendable desktop software platform for the organization and
792 analysis of sequence data. *Bioinformatics.* 2012/05/01. 2012;28(12):1647–9.
- 793 39. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open
794 Software Suite. *Trends Genet.* 2000 Jun;16(6):276–7.
- 795 40. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, et al. Protein
796 Identification and Analysis Tools on the ExPASy Server. In: Walker JM, editor. *The*
797 *Proteomics Protocols Handbook*. Totowa, NJ: Humana Press; 2005. p. 571–607.
- 798 41. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak
799 S, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks.
800 *Nat Biotechnol.* 2019 Apr 18;37(4):420–3.
- 801 42. Juncker AS, Willenbrock H, von Heijne G, Brunak S, Nielsen H, Krogh A. Prediction
802 of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci.* 2003/07/24. 2003
803 Aug;12(8):1652–62.
- 804 43. Ren J, Wen L, Gao X, Jin C, Xue Y, Yao X. CSS-Palm 2.0: an updated software for
805 palmitoylation sites prediction. *Protein Eng Des Sel.* 2008 Aug 27;21(11):639–44.
- 806 44. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York;

- 807 2016.
- 808 45. Kato K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple
809 sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 2002/07/24.
810 2002;30(14):3059–66.
- 811 46. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation
812 and Projection. *J Open Source Softw.* 2018 Sep 2;3(29):861.
- 813 47. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood
814 Trees for Large Alignments. Poon AFY, editor. *PLoS One.* 2010 Mar 10;5(3):e9490.
- 815 48. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New
816 algorithms and methods to estimate maximum-likelihood phylogenies: assessing the
817 performance of PhyML 3.0. *Syst Biol.* 2010 May;59(3):307–21.
- 818 49. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
819 large phylogenies. *Bioinformatics.* 2014 May 1;30(9):1312–3.
- 820 50. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal
821 for protein modeling, prediction and analysis. *Nat Protoc.* 2015 Jun;10(6):845–58.
- 822 51. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al.
823 Accurate prediction of protein structures and interactions using a three-track neural
824 network. *Science (80-).* 2021 Aug 20;373(6557):871–6.
- 825 52. Colovos C, Yeates TO. Verification of protein structures: patterns of nonbonded
826 atomic interactions. *Protein Sci.* 1993 Sep;2(9):1511–9.
- 827 53. Bowie JU, Lüthy R, Eisenberg D. A method to identify protein sequences that fold into
828 a known three-dimensional structure. *Science.* 1991 Jul 12;253(5016):164–70.
- 829 54. Madej T, Marchler-Bauer A, Lanczycki C, Zhang D, Bryant SH. Biological Assembly
830 Comparison with VAST+. In 2020. p. 175–86.
- 831 55. Gelly J-C, Joseph AP, Srinivasan N, de Brevern AG. iPBA: a tool for protein structure

- 832 comparison using sequence alignment strategies. *Nucleic Acids Res.* 2011 Jul
833 1;39(suppl_2):W18–23.
- 834 56. Carugo O, Pongor S. A normalized root-mean-square distance for comparing protein
835 three-dimensional structures. *Protein Sci.* 2001 Jul;10(7):1470–3.
- 836 57. Reva BA, Finkelstein A V, Skolnick J. What is the probability of a chance prediction
837 of a protein structure with an rmsd of 6 Å? *Fold Des.* 1998;3(2):141–7.
- 838 58. Dashper SG, Mitchell HL, Seers CA, Gladman SL, Seemann T, Bulach DM, et al.
839 *Porphyromonas gingivalis* Uses Specific Domain Rearrangements and Allelic
840 Exchange to Generate Diversity in Surface Virulence Factors. *Front Microbiol.*
841 2017/02/12. 2017;8:48.
- 842 59. Kirdat K, Tiwarekar B, Sathe S, Yadav A. From sequences to species: Charting the
843 phytoplasma classification and taxonomy in the era of taxogenomics. *Front Microbiol.*
844 2023 Mar 9;14.
- 845 60. Wagner J, Coupland P, Browne HP, Lawley TD, Francis SC, Parkhill J. Evaluation of
846 PacBio sequencing for full-length bacterial 16S rRNA gene classification. *BMC*
847 *Microbiol.* 2016 Dec 14;16(1):274.
- 848 61. Liu Y, Du J, Pei T, Du H, Feng G-D, Zhu H. Genome-based taxonomic classification
849 of the closest-to-Comamonadaceae group supports a new family Sphaerotilaceae fam.
850 nov. and taxonomic revisions. *Syst Appl Microbiol.* 2022 Nov;45(6):126352.
- 851 62. Khoder M, Osman M, Kassem II, Rafei R, Shahin A, Fournier PE, et al. Whole
852 Genome Analyses Accurately Identify *Neisseria* spp. and Limit Taxonomic
853 Ambiguity. *Int J Mol Sci.* 2022 Nov 3;23(21):13456.
- 854 63. Tambong JT. Taxogenomics and Systematics of the Genus *Pantoea*. *Front Microbiol.*
855 2019 Oct 30;10.
- 856 64. Sousa T de J, Parise D, Profeta R, Parise MTD, Gomide ACP, Kato RB, et al. Re-

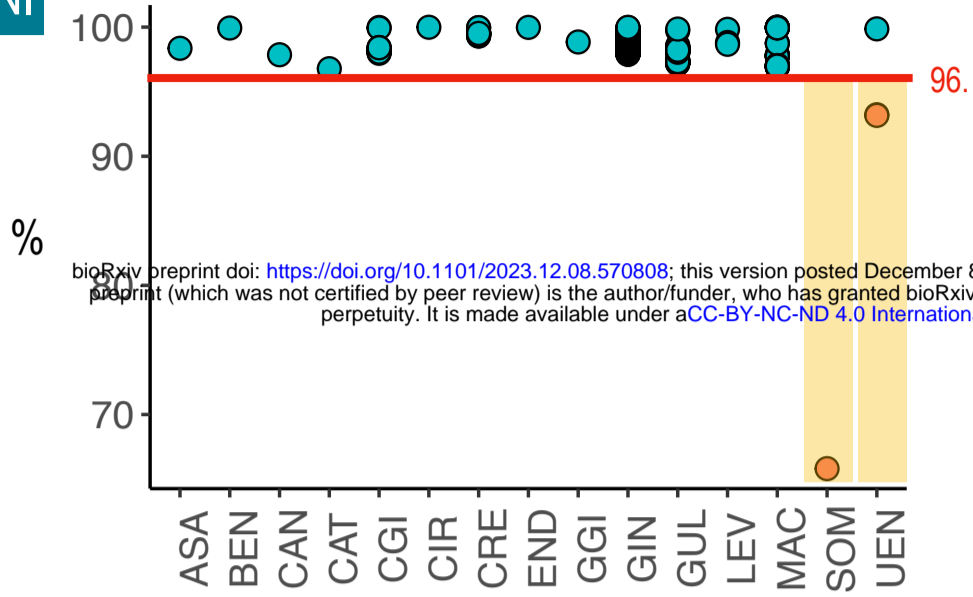
- 857 sequencing and optical mapping reveals misassemblies and real inversions on
858 *Corynebacterium pseudotuberculosis* genomes. *Sci Rep.* 2019 Nov 8;9(1):16387.
- 859 65. Chen L-X, Anantharaman K, Shaiber A, Eren AM, Banfield JF. Accurate and
860 complete genomes from metagenomes. *Genome Res.* 2020 Mar;30(3):315–33.
- 861 66. De Simone G, Pasquadibisceglie A, Proietto R, Polticelli F, Aime S, J.M. Op den
862 Camp H, et al. Contaminations in (meta)genome data: An open issue for the scientific
863 community. *IUBMB Life.* 2020 Apr 23;72(4):698–705.
- 864 67. Colston SM, Fullmer MS, Beka L, Lamy B, Gogarten JP, Graf J. Bioinformatic
865 Genome Comparisons for Taxonomic and Phylogenetic Assignments Using
866 *Aeromonas* as a Test Case. Ruby EG, editor. *MBio.* 2014 Dec 31;5(6).
- 867 68. Trachana K, Forslund K, Larsson T, Powell S, Doerks T, von Mering C, et al. A
868 Phylogeny-Based Benchmarking Test for Orthology Inference Reveals the Limitations
869 of Function-Based Validation. Anisimova M, editor. *PLoS One.* 2014 Nov
870 4;9(11):e111122.
- 871 69. Altschul S. Gapped BLAST and PSI-BLAST: a new generation of protein database
872 search programs. *Nucleic Acids Res.* 1997 Sep 1;25(17):3389–402.
- 873 70. Kumar G, Srinivasan N, Sandhya S. Profiles of Natural and Designed Protein-Like
874 Sequences Effectively Bridge Protein Sequence Gaps: Implications in Distant
875 Homology Detection. In 2022. p. 149–67.
- 876 71. Jin X, Liao Q, Liu B. PL-search: a profile-link-based search method for protein remote
877 homology detection. *Brief Bioinform.* 2021 May 20;22(3).
- 878 72. de Crécy-lagard V, Amorin de Hegedus R, Arighi C, Babor J, Bateman A, Blaby I, et
879 al. A roadmap for the functional annotation of protein families: a community
880 perspective. *Database.* 2022 Aug 12;2022.
- 881 73. Ruiz J. Analysis of the presence of erroneous Qnr sequences in GenBank. *J*

- 882 Antimicrob Chemother. 2018 May 1;73(5):1213–6.
- 883 74. Colombatti A, Bonaldo P. The superfamily of proteins with von Willebrand factor type
884 A-like domains: one theme common to components of extracellular matrix,
885 hemostasis, cellular adhesion, and defense mechanisms. *Blood*. 1991 Jun
886 1;77(11):2305–15.
- 887 75. Mihajlovic J, Bechon N, Ivanova C, Chain F, Almeida A, Langella P, et al. A Putative
888 Type V Pilus Contributes to *Bacteroides thetaiotaomicron* Biofilm Formation
889 Capacity. O’Toole G, editor. *J Bacteriol*. 2019 Sep 15;201(18).
- 890 76. Chamarande J, Cunat L, Alauzet C, Cailliez-Grimal C. In Silico Study of Cell Surface
891 Structures of *Parabacteroides distasonis* Involved in Its Maintenance within the Gut
892 Microbiota. *Int J Mol Sci*. 2022 Aug 20;23(16):9411.
- 893 77. González-Montalvo MA, Tavares-Carreón F, González GM, Villanueva-Lozano H,
894 García-Romero I, Zomosa-Signoret VC, et al. Defining chaperone-usher fimbriae
895 repertoire in *Serratia marcescens*. *Microb Pathog*. 2021 May;154:104857.
- 896 78. Khater F, Balestrino D, Charbonnel N, Dufayard JF, Brisse S, Forestier C. In Silico
897 Analysis of Usher Encoding Genes in *Klebsiella pneumoniae* and Characterization of
898 Their Role in Adhesion and Colonization. Bengoechea JA, editor. *PLoS One*. 2015
899 Mar 9;10(3):e0116215.
- 900 79. Acuña-Amador L, Barloy-Hubler F. *Porphyromonas* spp. have an extensive host range
901 in ill and healthy individuals and an unexpected environmental distribution: A
902 systematic review and meta-analysis. *Anaerobe*. 2020 Dec;66:102280.
- 903 80. Villegas A, Kropinski AM. An analysis of initiation codon utilization in the Domain
904 Bacteria – concerns about the quality of bacterial genome annotation. *Microbiology*.
905 2008 Sep 1;154(9):2559–661.
- 906 81. Nielsen P, Krogh A. Large-scale prokaryotic gene prediction and comparison to

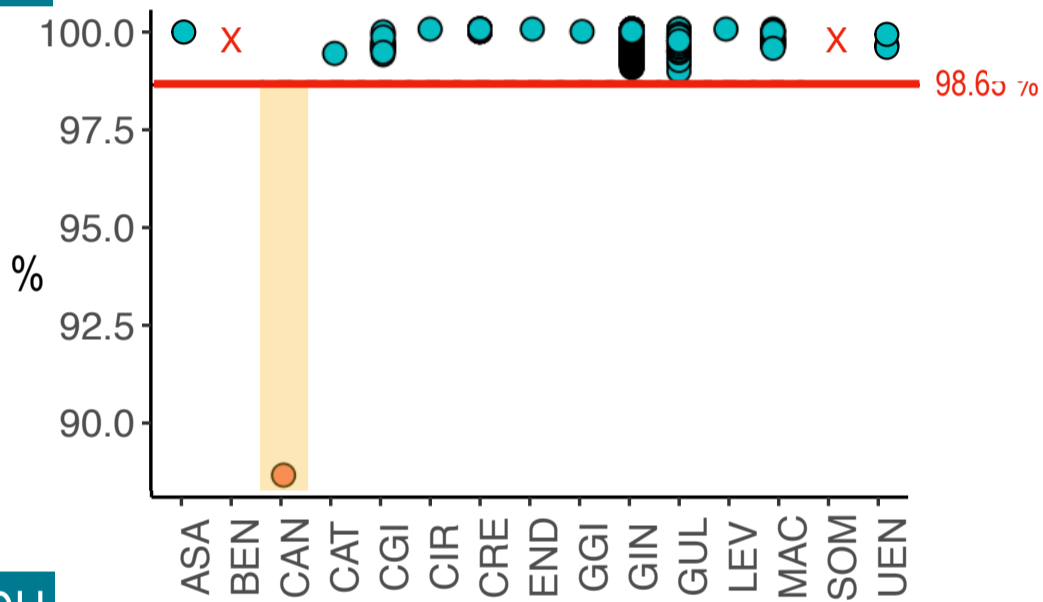
- 907 genome annotation. *Bioinformatics*. 2005 Dec 15;21(24):4322–9.
- 908 82. Xueyi Wang, Snoeyink J. Defining and Computing Optimum RMSD for Gapped and
909 Weighted Multiple-Structure Alignment. *IEEE/ACM Trans Comput Biol Bioinforma*.
910 2008 Oct;5(4):525–33.
- 911 83. Shibuya T. Efficient Substructure RMSD Query Algorithms. *J Comput Biol*. 2007
912 Nov;14(9):1201–7.
- 913 84. Burrows LL. Heads or tails for type V pilus assembly. *Nat Microbiol*. 2020 May
914 28;5(6):782–4.
- 915

A

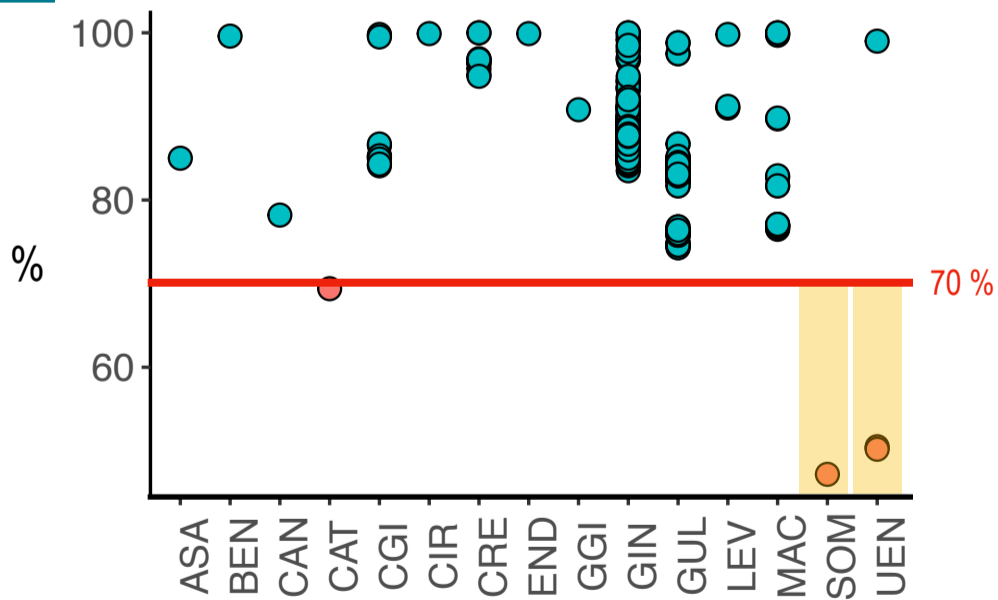
ANI



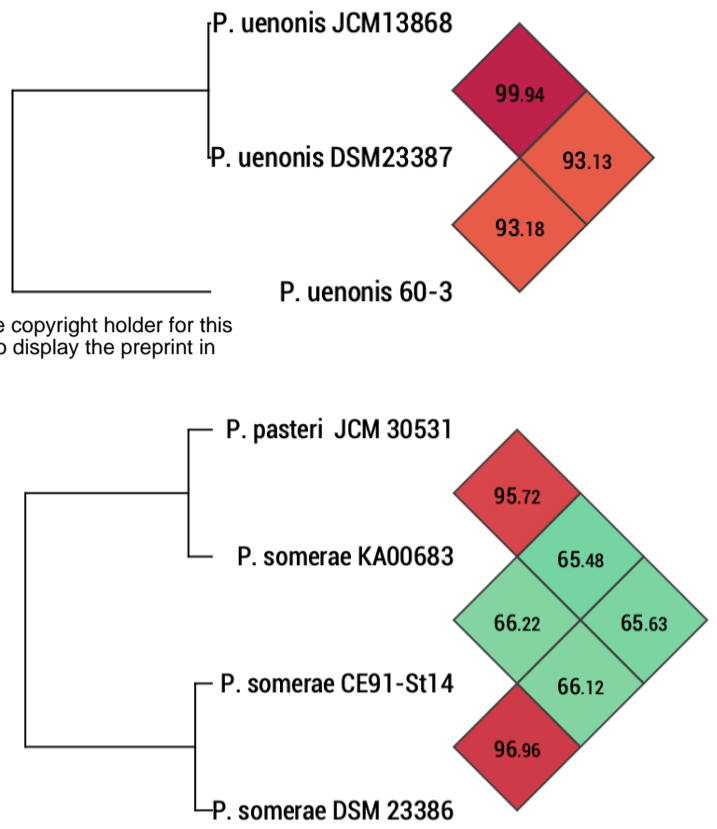
16S



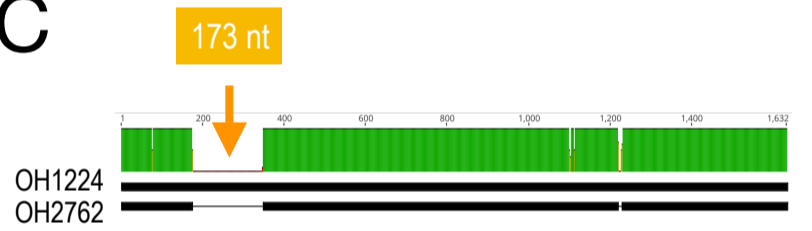
DDH



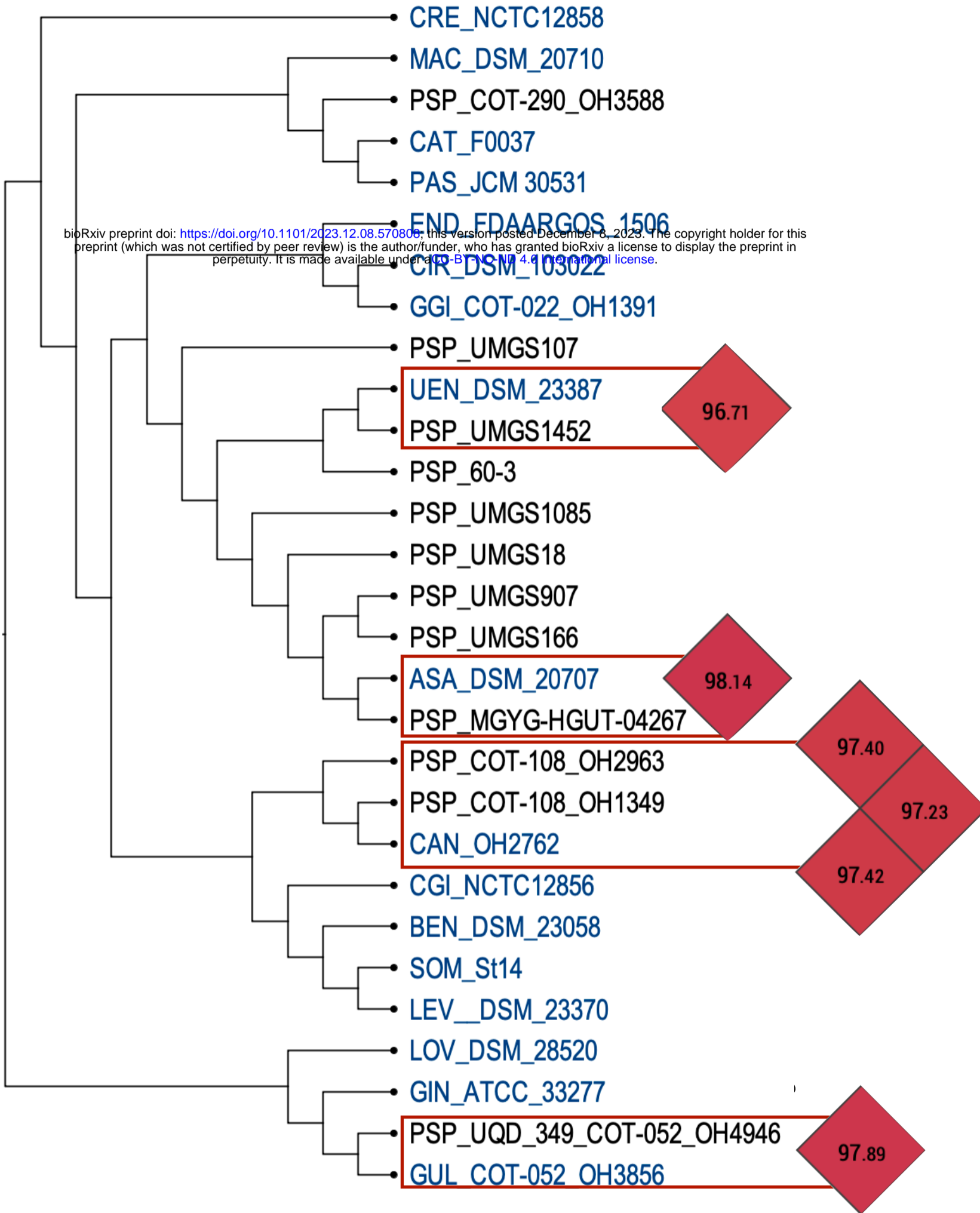
B



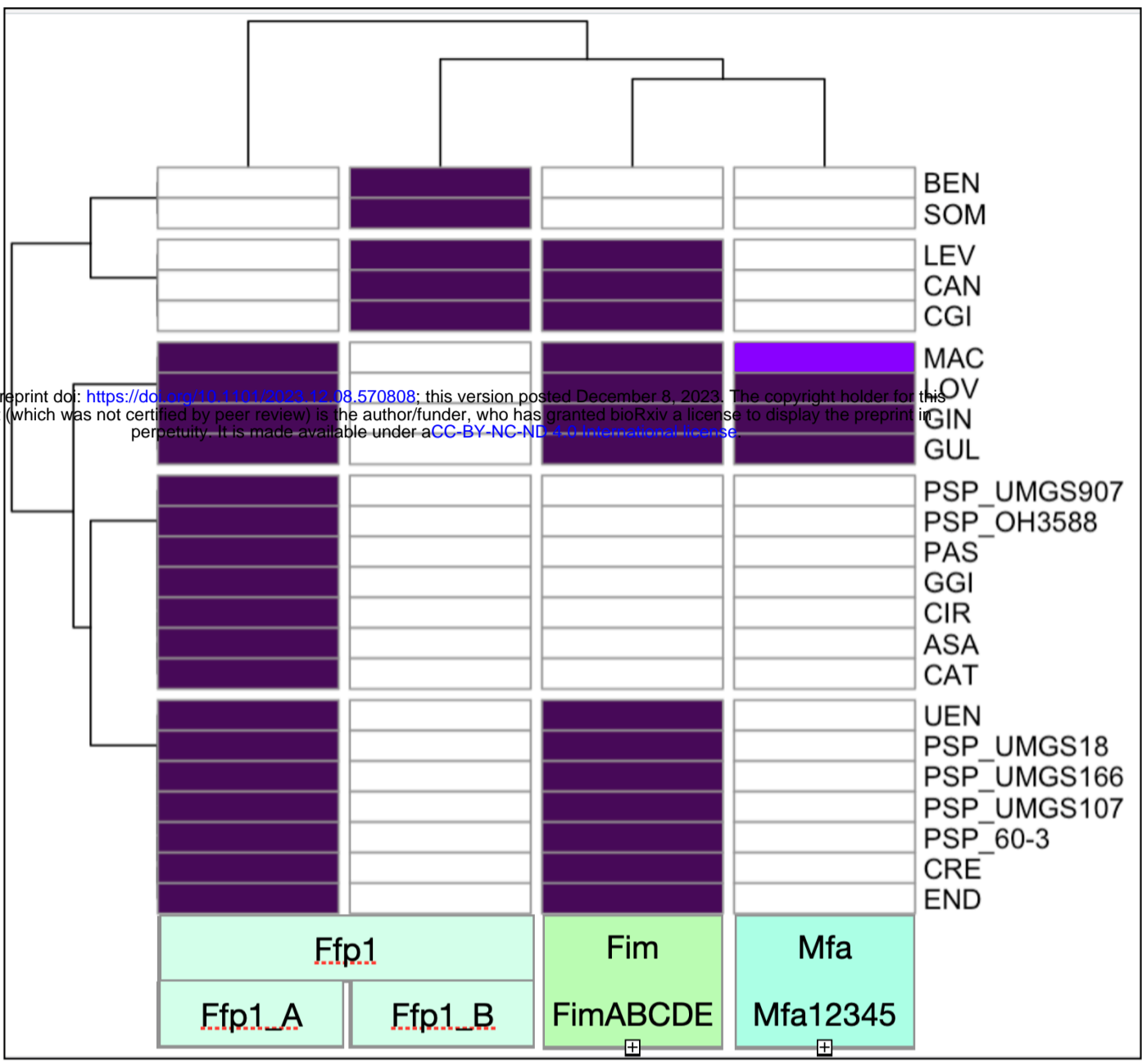
C

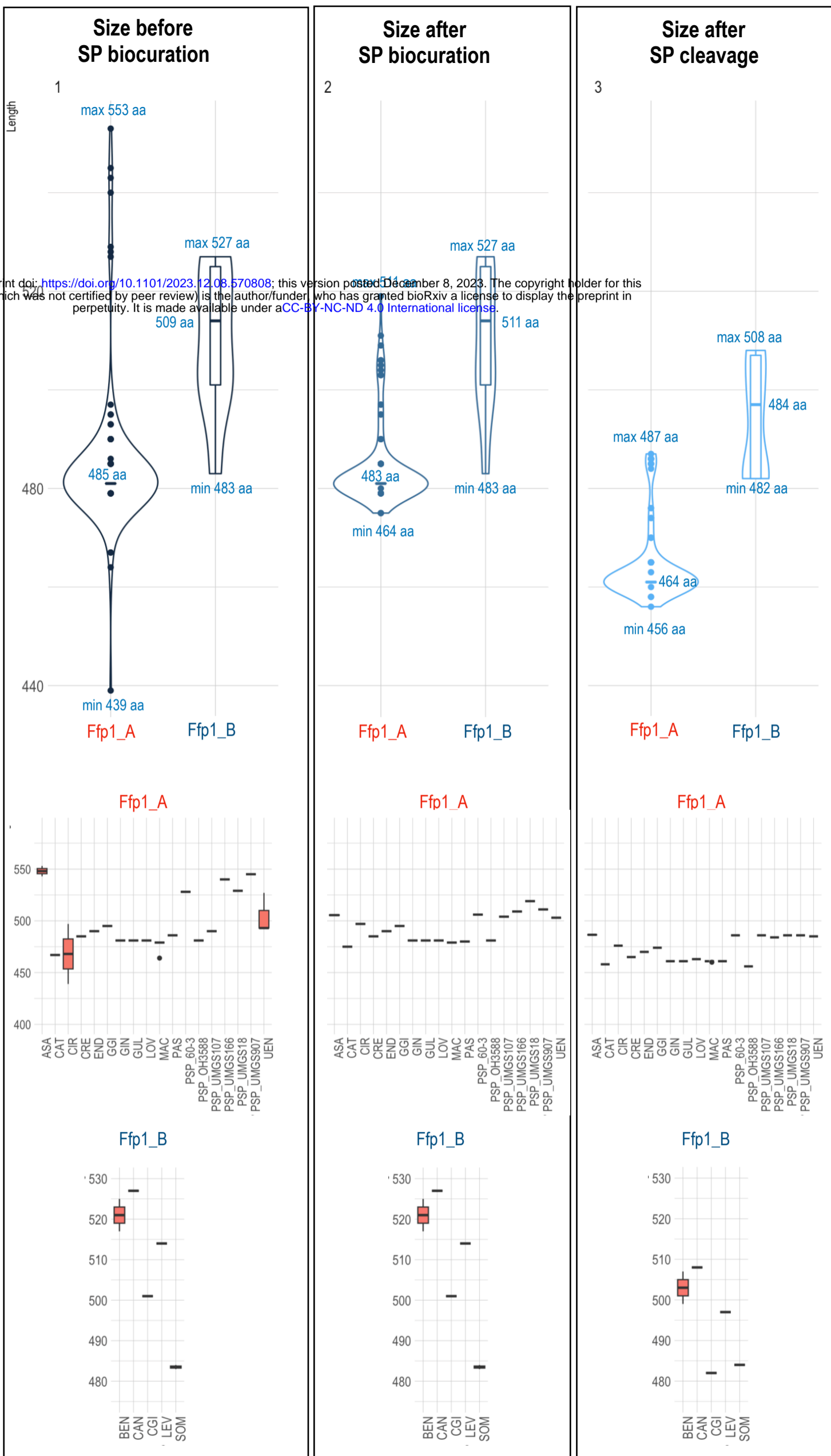


bioRxiv preprint doi: <https://doi.org/10.1101/2023.12.08.570806>; this version posted December 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



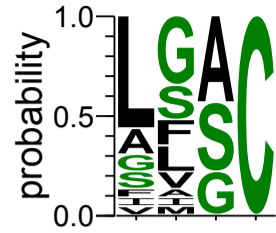
bioRxiv preprint doi: <https://doi.org/10.1101/2023.12.08.570808>; this version posted December 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



A

bioRxiv preprint doi: <https://doi.org/10.1101/2023.12.08.570808>; this version posted December 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

B



bioRxiv preprint doi: <https://doi.org/10.1101/2023.12.08.570800>; this version posted December 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

| | | | | |
|----------------|---|---|---|---|
| 3_GIN | L | L | A | C |
| 4_GUL | G | F | S | C |
| 5_END | L | L | A | C |
| 6_MAC | I | S | S | C |
| 7_LEV | L | L | G | C |
| 8_PAS | L | A | G | C |
| 9_SOM | V | S | S | C |
| 10_PSP_UMGS107 | L | G | A | C |
| 11_PSP_UMGS907 | L | G | A | C |
| 12_ASA | L | G | S | C |
| 13_PSP_UMGS166 | L | G | A | C |
| 14_PSP_60-3 | L | G | A | C |
| 15_PSP_UMGS18 | L | G | A | C |
| 16_UEN | L | G | A | C |
| 17_CAT | F | S | S | C |
| 18_GGI | L | G | S | C |
| 19_CRE | A | M | G | C |
| 20_BEN | A | V | G | C |
| 21_CAN | S | S | G | C |
| 22_CGI | L | V | A | C |
| 23_CIR | L | L | S | C |

C

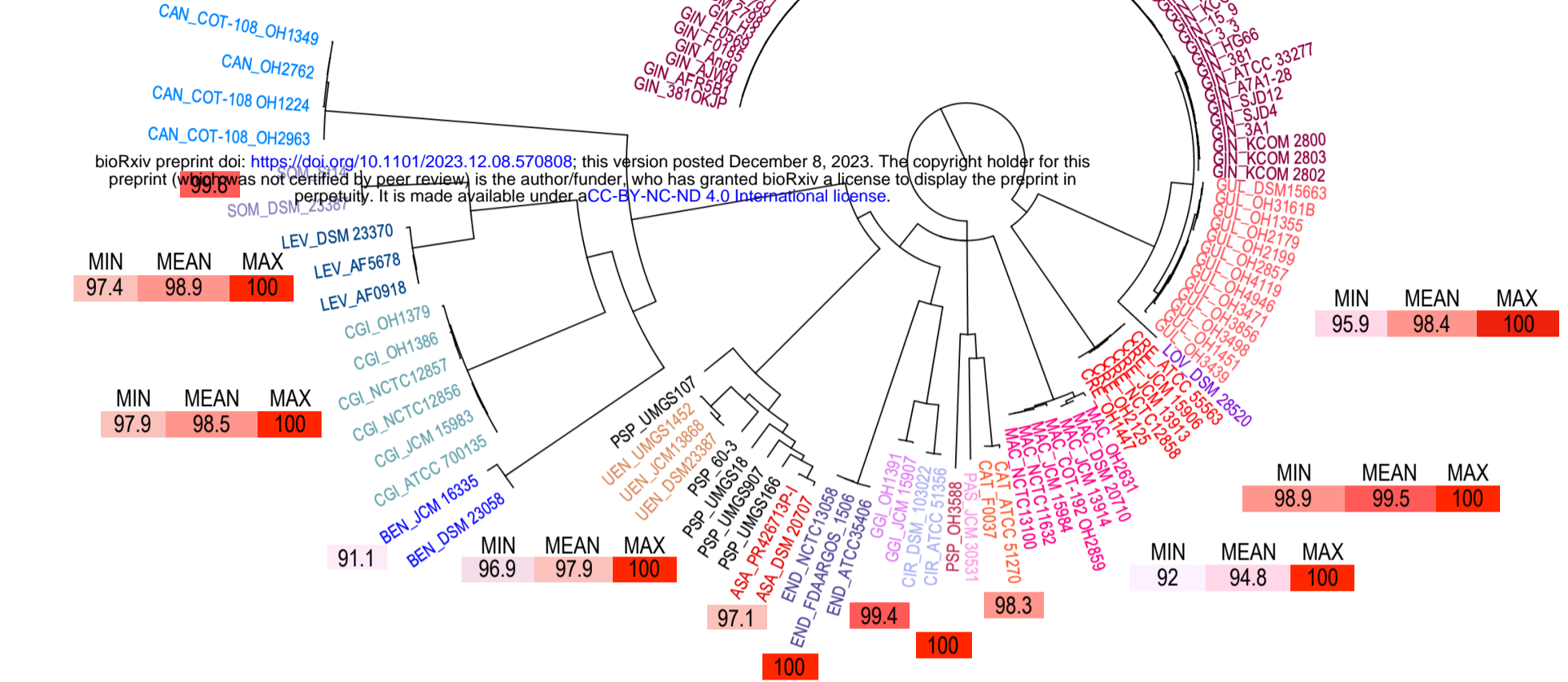
| | PSP_OH3588 | LOV | GIN | GUL | END | MAC | LEV | PAS | SOM | PSP_UMGS107 | PSP_UMGS907 | ASA | PSP_UMGS166 | PSP_60-3 | PSP_UMGS18 | UEN | CAT | GGI | CRE | BEN | CAN | CGI | CIR |
|-------------|------------|-----|-----|-----|-----|-----|-----|-----|-----|-------------|-------------|-----|-------------|----------|------------|-----|-----|-----|-----|-----|-----|-----|-----|
| LOV | 20 | 100 | 40 | 35 | 35 | 30 | 26 | 17 | 17 | 25 | 15 | 10 | 20 | 15 | 18 | 20 | 16 | 11 | 22 | 25 | 25 | 30 | 5 |
| GIN | 25 | 40 | 100 | 86 | 23 | 20 | 21 | 17 | 39 | 10 | 18 | 24 | 18 | 18 | 18 | 15 | 32 | 11 | 22 | 30 | 24 | 19 | 16 |
| GUL | 25 | 35 | 86 | 100 | 23 | 25 | 16 | 17 | 28 | 10 | 18 | 24 | 18 | 18 | 18 | 15 | 32 | 16 | 22 | 30 | 19 | 19 | 21 |
| END | 30 | 35 | 23 | 23 | 100 | 15 | 26 | 17 | 11 | 20 | 23 | 19 | 18 | 23 | 29 | 25 | 11 | 21 | 17 | 10 | 10 | 19 | 16 |
| MAC | 10 | 30 | 20 | 25 | 15 | 100 | 26 | 11 | 28 | 15 | 10 | 15 | 20 | 10 | 12 | 15 | 37 | 16 | 11 | 15 | 20 | 15 | 16 |
| LEV | 21 | 26 | 21 | 16 | 26 | 26 | 100 | 33 | 33 | 11 | 21 | 21 | 26 | 26 | 24 | 21 | 26 | 11 | 17 | 16 | 21 | 32 | 21 |
| PAS | 22 | 17 | 17 | 17 | 17 | 11 | 33 | 100 | 44 | 28 | 39 | 33 | 39 | 39 | 41 | 39 | 11 | 28 | 33 | 33 | 17 | 22 | 33 |
| SOM | 17 | 17 | 39 | 28 | 11 | 28 | 33 | 44 | 100 | 11 | 17 | 22 | 17 | 17 | 18 | 17 | 44 | 33 | 11 | 17 | 28 | 28 | 22 |
| PSP_UMGS107 | 20 | 25 | 10 | 10 | 20 | 15 | 11 | 28 | 11 | 100 | 50 | 40 | 55 | 50 | 59 | 50 | 11 | 21 | 11 | 20 | 15 | 30 | 26 |
| PSP_UMGS907 | 20 | 15 | 18 | 18 | 23 | 10 | 21 | 39 | 17 | 50 | 100 | 67 | 67 | 59 | 82 | 75 | 11 | 21 | 17 | 30 | 14 | 19 | 26 |
| ASA | 20 | 10 | 24 | 24 | 19 | 15 | 21 | 33 | 22 | 40 | 67 | 100 | 71 | 71 | 88 | 80 | 16 | 26 | 11 | 30 | 24 | 14 | 26 |
| PSP_UMGS166 | 30 | 20 | 18 | 18 | 18 | 20 | 26 | 39 | 17 | 55 | 67 | 71 | 100 | 70 | 94 | 85 | 16 | 16 | 17 | 30 | 14 | 19 | 21 |
| PSP_60-3 | 30 | 15 | 18 | 18 | 23 | 10 | 26 | 39 | 17 | 50 | 59 | 71 | 70 | 100 | 94 | 80 | 16 | 16 | 17 | 30 | 19 | 19 | 21 |
| PSP_UMGS18 | 29 | 18 | 18 | 18 | 29 | 12 | 24 | 41 | 18 | 59 | 82 | 88 | 94 | 94 | 100 | 100 | 12 | 18 | 18 | 35 | 24 | 24 | 24 |
| UEN | 25 | 20 | 15 | 15 | 25 | 15 | 21 | 39 | 17 | 50 | 75 | 80 | 85 | 80 | 100 | 100 | 11 | 16 | 17 | 40 | 20 | 20 | 21 |
| CAT | 16 | 16 | 32 | 32 | 11 | 37 | 26 | 11 | 44 | 11 | 11 | 16 | 16 | 16 | 12 | 11 | 100 | 16 | 17 | 21 | 32 | 16 | 26 |
| GGI | 11 | 11 | 11 | 16 | 21 | 16 | 11 | 28 | 33 | 21 | 21 | 26 | 16 | 16 | 18 | 16 | 16 | 100 | 11 | 5 | 16 | 21 | 37 |
| CRE | 22 | 22 | 22 | 22 | 17 | 11 | 17 | 33 | 11 | 11 | 17 | 11 | 17 | 17 | 18 | 17 | 17 | 11 | 100 | 39 | 17 | 6 | 28 |
| BEN | 25 | 25 | 30 | 30 | 10 | 15 | 16 | 33 | 17 | 20 | 30 | 30 | 30 | 30 | 35 | 40 | 21 | 5 | 39 | 100 | 35 | 15 | 21 |
| CAN | 10 | 25 | 24 | 19 | 10 | 20 | 21 | 17 | 28 | 15 | 14 | 24 | 14 | 19 | 24 | 20 | 32 | 16 | 17 | 35 | 100 | 14 | 16 |
| CGI | 20 | 30 | 19 | 19 | 19 | 15 | 32 | 22 | 28 | 30 | 19 | 14 | 19 | 19 | 24 | 20 | 16 | 21 | 6 | 15 | 14 | 100 | 11 |
| CIR | 5 | 5 | 16 | 21 | 16 | 16 | 21 | 33 | 22 | 26 | 26 | 26 | 21 | 21 | 24 | 21 | 26 | 37 | 28 | 21 | 16 | 11 | 100 |

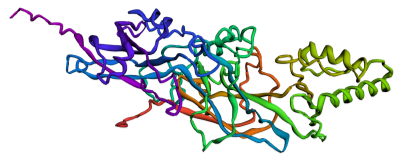
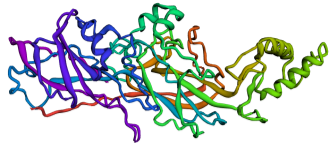
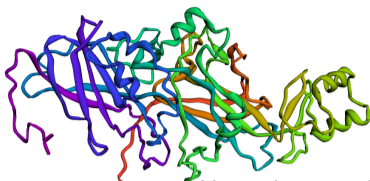
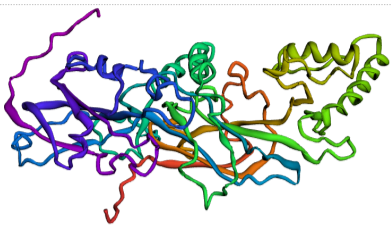
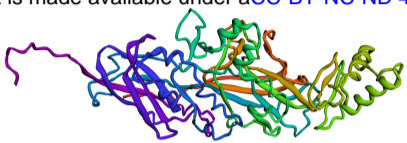
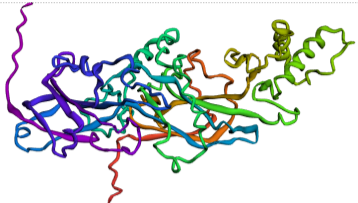
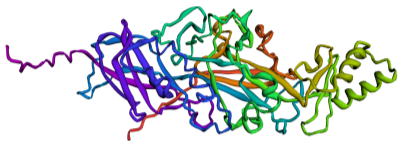
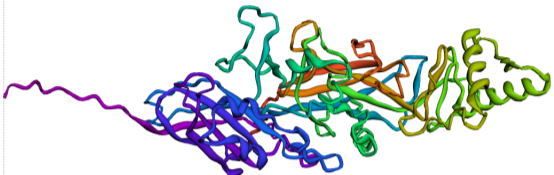
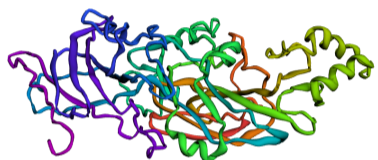
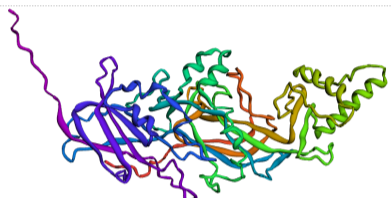
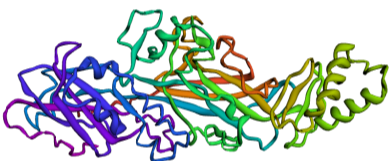
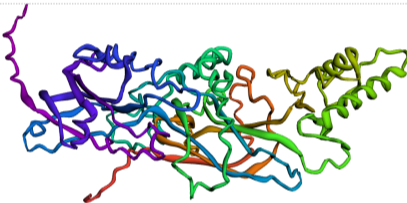
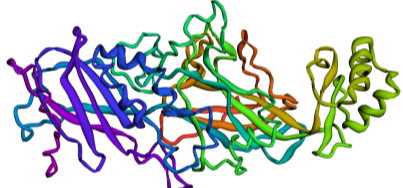
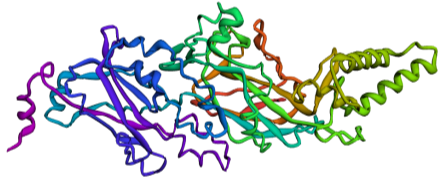
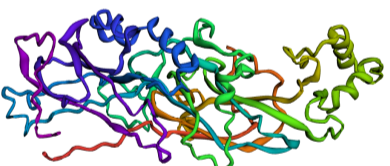
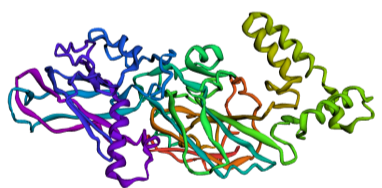
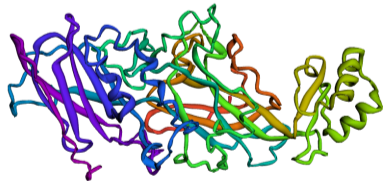
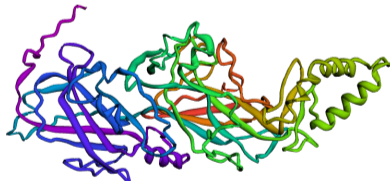
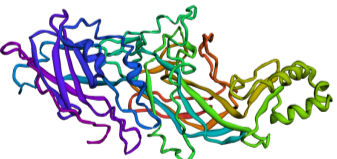
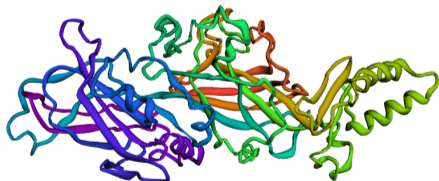
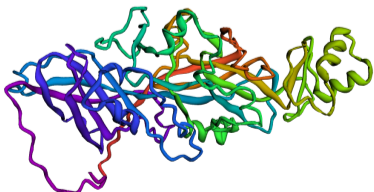
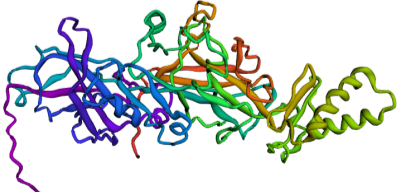
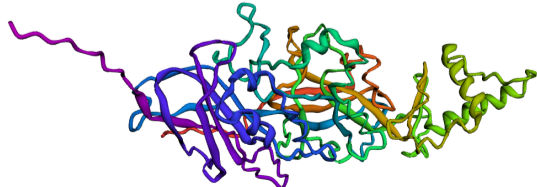
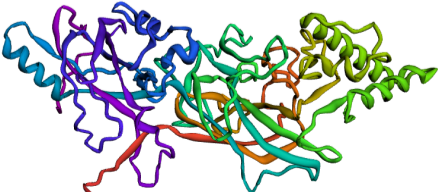
D

| MIN | MEAN | MAX |
|------|------|------|
| 98.4 | 98.8 | 99.4 |

| MIN | MEAN | MAX |
|------|------|-----|
| 97.5 | 99.1 | 100 |

bioRxiv preprint doi: <https://doi.org/10.1101/2023.12.08.570808>; this version posted December 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

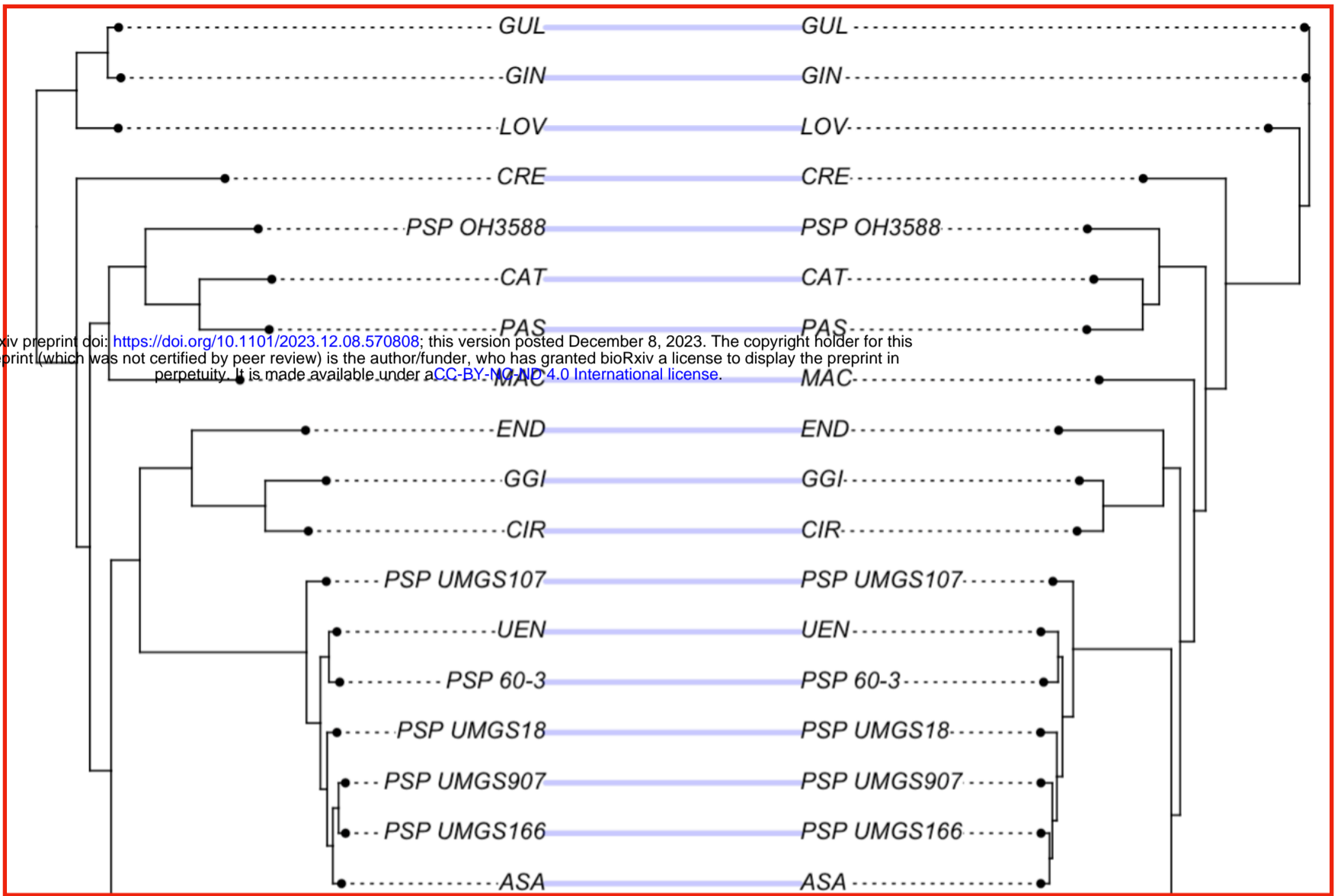


| | | | |
|----------|--|-------------|---|
| ASA |  | PSP_OH3588 |  |
| CAT |  | PSP_UMGS18 |  |
| CIR |  | PSP_UMGS107 |  |
| CRE |  | PSP_UMGS166 |  |
| END |  | PSP_UMGS907 |  |
| GGI |  | UEN |  |
| GIN |  | BEN |  |
| GUL |  | CAN |  |
| LOV |  | CGI |  |
| MAC |  | LEV |  |
| PAS |  | SOM |  |
| PSP_60-3 |  | BACOVA |  |

bioRxiv preprint doi: <https://doi.org/10.1101/2023.12.08.570808>; this version posted December 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Ffp1_A

bioRxiv preprint doi: <https://doi.org/10.1101/2023.12.08.570808>; this version posted December 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Ffp1_B

