In silico analysis of Ffp1, an ancestral Porphyromonas spp. fimbrillin, shows differences with Fim and Mfa

Luis A Acuña-Amador, Frédérique Barloy-Hubler

ACCESS MICROBIOLOGY
an open research platform

RESEARCH ARTICLE
Acuña-Amador and Barloy-Hubler, *Access Microbiology* 2024;6:000771.v3
DOI 10.1099/acmi.0.000771.v3

MICROBIOLOGY SOCIETY
OPEN DATA OPEN ACCESS

In silico analysis of Ffp1, an ancestral *Porphyromonas* spp. fimbrillin, shows differences with Fim and Mfa

Luis Acuña-Amador[1,*] and Frederique Barloy-Hubler[2]

Abstract

Background. Scant information is available regarding fimbrillins within the genus *Porphyromonas*, with the notable exception of those belonging to *Porphyromonas gingivalis*, which have been extensively researched for several years. Besides *fim* and *mfa*, a third *P. gingivalis* adhesin called filament-forming protein 1 (Ffp1) has recently been described and seems to be pivotal for outer membrane vesicle (OMV) production.

Objective. We aimed to investigate the distribution and diversity of type V fimbrillin, particularly Ffp1, in the genus *Porphyromonas*.

Methods. A bioinformatics phylogenomic analysis was conducted using all accessible *Porphyromonas* genomes to generate a domain search for fimbriae, using hidden Markov model profiles.

Results. Ffp1 was identified as the sole fimbrillin present in all analysed genomes. After manual verification (i.e. biocuration) of both structural and functional annotations and 3D modelling, this protein was determined to be a type V fimbrillin, with a closer structural resemblance to a *Bacteroides ovatus* fimbrillin than to FimA or Mfa1 from *P. gingivalis*.

Conclusion. It appears that Ffp1 is an ancestral fimbria, transmitted through vertical inheritance and present across all *Porphyromonas* species. Additional investigations are necessary to elucidate the biogenesis of Ffp1 fimbriae and their potential role in OMV production and niche adaptation.

Impact Statement

Three distinct fimbriae have been described in *Porphyromonas gingivalis*. Hidden Markov model profiles were used to search genes from these three fimbriae in all the *Porphyromonas* genomes, and it was found that they were differentially present within the genus. Unlike Fim or Mfa, Ffp1 is the only fimbriae common to all *Porphyromonas* monophyletic groups. This gene codes for a stem protein distinct from FimA and Mfa1, and similar to BACOVA_01548. Ffp1 is not present in other bacteria, and seems to be ancestral in *Porphyromonas* spp. As such, studying this gene might help our understanding of niche adaptation and pathogenicity, and other biological process such as outer membrane vesicle production. Characterization of this novel fimbrillin in terms of biogenesis and its involvement in bacterial fitness is lacking and should be addressed.

Access Microbiology is an open research platform. Pre-prints, peer review reports, and editorial decisions can be found with the online version of this article.
Received 16 January 2024; Accepted 08 May 2024; Published 11 July 2024
Author affiliations: [1]Laboratorio de Investigación en Bacteriología Anaerobia, Centro de Investigación en Enfermedades Tropicales, Facultad de Microbiología, Universidad de Costa Rica, San José, Costa Rica; [2]Université de Rennes 1, CNRS, UMR 6553 ECOBIO (Écosystèmes, Biodiversité, Évolution), 35042 Rennes, France.
***Correspondence:** Luis Acuña-Amador, luisalberto.acuna@ucr.ac.cr
Keywords: 3D protein modelling; bioinformatics; fimbriae; phylogenomics; *Porphyromonas*.
Abbreviations: AF, alignment fraction; CDS, coding sequence; CU, chaperone-usher; DDH, DNA–DNA hybridization; gANI, genome average nucelotide identity; HGT, horizontal gene transfer; HMM, hidden Markov model; MAG, metagenome-assembled genome; OGRIs, overall genome relatedness indices; OMV, outer membrane vesicle; PSI-BLAST, position-specific iterated BLAST; PSSM, position-specific scoring matrix; SL, sphingolipid; SPII, signal peptidase II; VWA, von Willebrand factor type A.
Seven supplementary figures and two supplementary tables are available with the online version of this article.

DATA SUMMARY

External code and software were used as stated in the Methods section, and appropriate literature and/or URLs are provided for all. The authors confirm all supporting data, code and protocols have been provided within the article or through supplementary data files. The hidden Markov model profiles used for this work are publicly available and can be found using DOI: 10.5281/zenodo.10519420[1].

INTRODUCTION

Fimbriae (fibrillae or pili) are adhesins consisting of protein polymers forming filamentous appendages that protrude from the bacterial cell surface. Unlike motility flagella, fimbriae have adhesive properties to attach to surfaces. In Gram-negative bacteria, fimbriae are classified according to their assembly pathways, including the chaperone-usher (CU) pilus system, the type IV pilus and the conjugative type IV secretion pilus [2, 3].

In 2016, a new prevalent type V pilus was discovered within the human gut microbiome [4] and was described as a new donor strand-mediated system restricted to the class *Bacteroidia* [3]. This system resembles the CU type, but requires the lipoprotein sorting pathway, and outer membrane proteinases [5].

Type V fimbriae have been mainly studied in *Porphyromonas gingivalis* which classically produces two distinct adhesins, termed FimA (described in 1984 [6]) and Mfa1 (described in 1996 [7]), according to the names of stalk subunits [8]. Both stalk proteins must be processed and matured. They possess long leader peptides [9] that facilitate their transport to the periplasm via the Sec system. Subsequently, they undergo lipid modification and are cleaved by type II signal peptidase [10], followed by a proteolytic maturation achieved by RgpA, RgpB and Kgp proteinases called gingipains [11]. Finally, mature fibrillin monomers polymerize [12]. The genetic loci for both fimbriae are distinct but organized into two clusters: *fimA-E* and *mfa1-5* [13].

In 2017, a third *P. gingivalis* adhesin was described (PGN_1808 in the ATCC 33277 strain or PG1881 in the W83 strain) and termed Ffp1 for filament-forming protein 1 [14]. It corresponds to filaments 200–400 nm in length and 2–3 nm in diameter that can be degraded, unlike FimA or Mfa1, by detergents and temperature into 50 kDa monomers [15]. Ffp1 is among the exclusive repertoire of proteins within the order *Bacteroidales* and was described as conserved across *Porphyromonas* and *Bacteroides* [16, 17]. This protein was identified among the outer membrane proteins and especially the O-glycoproteome of *P. gingivalis* [18] and was described as essential in the production of outer membrane vesicles (OMVs), as the Ffp1 null-mutants exhibited a 55% reduction in OMV production compared to the wild-type strain [14]. Moreover, a recent study indicated a connection between Ffp1 and the production of sphingolipids (SLs). In the absence of SLs, *P. gingivalis* generates OMVs without Ffp1, whereas OMVs containing SLs exhibit an enrichment of Ffp1. Interestingly, these SL-containing OMVs limit host inflammation [19].

The Ffp1 C-terminal region is homologous to type IV fimbriae from *Bacillus* spp. [16], and its sequence bears a significant similarity to the adhesion protein BACOVA_01548 (PDB ID: 4rfj) from *Bacteroides ovatus* [4]. Structural modelling suggests a donor strand-mediated assembly mechanism [15], which would classify Ffp1 as a new type V pilin [14]. However, unlike FimA or Mfa1, no accessory component has yet been identified for Ffp1 despite its apparent co-expression as an operon with three upstream genes, annotated as a Cys-RNAt ligase, a patatin (lipase) and a glycosyl transferase. This co-transcription suggests the involvement of these four proteins in the same biochemical pathway or utilization of the same substrates/transporters, albeit without physical interaction [15].

To date, Ffp1 has been the subject of few studies limited to *P. gingivalis*, only on two reference strains, ATCC 33277 and W83, and no information is available for the other 21 *Porphyromonas* species. At the genus level, knowledge for non-*P. gingivalis* Ffp1 or other fimbriae is scarce, except for description of FimA-like and Mfa1-like fimbriae in *P. gulae*, a closely related species to *P. gingivalis* [20, 21], and reports indicating fimbriation in *P. circumdentaria*, *P. macacae* and *P. asaccharolytica* [22–24], without further characterizations.

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

METHODS

Porphyromonas taxogenomics

All 144 *Porphyromonas* genomes (Table S1, available in the online version of this article) were automatically downloaded from the NCBI RefSeq database [25] (release 217, 8 March 2023) using the ncbi-genome-download script v0.2.12 [26]. Unannotated metagenome-assembled genomes (MAGs) with inconsistent taxonomic labels were not considered. To categorize all genomes into reliable groups, genomic data-driven taxonomic confirmation and/or assignment were performed. To confirm the assignment of genomes with a species name, we conducted a comparison of three metrics: (i) the 16S rRNA gene percentage identity (when annotated), evaluated using a threshold of 98.65% [27]; (ii) the digital DNA–DNA hybridization distance (DDH) using the GGDC v2.1 [28] and ggdc-robot script v0.04 [29], with the default threshold of 70% using formula 2 [28, 30, 31]; and (iii) the whole genome

average nucleotide identity (gANI), calculated using FastANI v1.34 [32] with a threshold of 96% for species demarcation [33]. In case of a disagreement between these three metrics, we combined alignment fraction values (AFs) with gANI using 60 and 96.5% as threshold values respectively, to assign a genome pair to the same species [34]. Additionally, when needed, we also used OrthoANI v0.93.1 [35] to measure and visualize the overall similarity between some *Porphyromonas* species.

For the genomes without a specified species name (i.e. *Porphyromonas* sp.), as most of them originated from environmental samples (human- or animal-associated habitats) and are often highly fragmented, it was crucial to ensure that they were not contaminated and do not correspond to genome assemblies containing a mixture of different species. This genomic homogeneity was evaluated with Kraken2 v2.1.3 [36] using the non-redundant nucleic database (updated April 22). Only assemblies that consisted of over 80% of *Porphyromonas* content and/or larger than 80% of the expected average genome size (2.5 Mb) were retained for our analysis. Their affiliation to the genus *Porphyromonas* was first confirmed using the fIDBAC server[37, 38] and their position within the *Porphyromonas* taxonomy was validated using an OrthoFinder v2.5.5 [39] rooted species tree [40]. This tree was reconstructed using all *Porphyromonas* sp. (*P.* sp.) and one reference genome per *Porphyromonas* species (see Table S1) and was visualized using FigTree v1.4.4 [41]. For each branch, one or several *P.* sp. were associated with a *Porphyromonas* species through ANI and DDH, employing the same thresholds as previously described.

Porphyromonas fimbriae identification and classification

(1) **Dataset construction:** Individual sequences from type V fimbriae (FimABCDE, Mfa12345 and Ffp1) were manually extracted from the 59 *P. gingivalis* genomes and each one was used as query to identify homologous sequences all in the genomes of other *Porphyromonas* spp. using BlastP [42] (identity ≥30%; query coverage ≥60%; e-value <10e^{-5}). All sequences were grouped as dataset 1.

(2) **Functional domain-based screening:** Dataset 1 was subjected to analysis using InterProScan [43] to identify all protein domains associated with those sequences. The resulting domains were searched in the complete orfeomes of *Porphyromonas* downloaded from PATRIC v3.6.6[44, 45], using 'hmmsearch' from HMMER v3.3.1 [46, 47] and the hidden Markov models (HMMs) from the Pfam v33.1 database [48] (May 2020). Sequences harbouring the targeted domains with an e-value <10e^{-06} were retained and grouped into dataset 2.

(3) **Protein clustering, biocuration and HMM profile construction:** Dataset 2 was clustered with MMseqs2 v15-6f452 [49] via the 'easy-cluster' command. Each cluster obtained underwent manual biocuration after multiple alignment using Clustal Omega[50, 51] and any missing genes were annotated. Subsequently, for each cluster, using HMMER, the multiple alignments were converted from FASTA format to Stockholm format with 'esl-reformat' command and HMM profiles were generated using the 'hmmbuild' command with default settings. Clustering and HMM profile creation was first performed on raw data and then refined on biocurated data.

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

(5) ***In silico* analysis of *Porphyromonas* fimbrillins:** Geneious Prime v2023 [52] was used to visualize the genomic context of each identified fimbrillin. Biocuration for start codons was proposed, based on sequence homology, to optimize the prediction of signal peptidase II (SPII) signal peptide and the cleavage site positions. The N-terminal region was identified using *charge* (window size=3) from EMBOSS v6.6.0 [53], the H hydrophobic region was characterized with a Kyte-Doolittle hydropathy plot made with ProtScale [54, 55] (window size=3), and the cleavage site was confirmed by SignalP v6.0 [56] and LipoP v1.0 [57]. Palmitoylation in the lipobox cysteine residue was verified using CSS-Palm v4.0 [58]. Protein sizes were represented using violin plots (geom_violin) and/or boxplots (geom_box), both functions from the ggplot2 package [59].

For each fimbrillin family, a multiple alignment was performed using MAFFT v7.490 (L-INS-I algorithm and BLOSUM62 matrix; gap open penalty and offset value by default) [60]. This alignment was visualized in two dimensions using Alignmentviewer v1.1 [61] which employs the UMAP algorithm [62] and Hamming distance to cluster aligned sequences. Phylogenetic trees were calculated using FastTree v2.1.11 [63], PhyML v3.3 [64] and RaxML v4.0 [65] with default parameters.

The taxonomic distribution of fimbrillin genes was analysed across a phylogenetic tree reconstructed using OrthoFinder based on the pangenomes of all confirmed *Porphyromonas* species groups and visualized using FigTree. The phylogenetic reconstruction was performed both using native and mature proteins (i.e. excluding their signal peptides) using RaxML (evolution model GAMMA LG and 100 bootstraps). Robinson-Foulds, Nye Similarity and Jaccard Robinson Foulds distances between the phylogenetic trees were calculated using the TreeDist [66] R library and tanglegrams were created with the R package phytools [67] (scripts TREE.R and Tanglegram.R).

6. 3D modelling: Secondary protein structure was predicted with PSIPRED v4.0 [68, 69] and Phyre2 v2.0[70, 71]. 3D structures of Ffp1 mature proteins were modelled, based on homology modelling, using Robetta [72, 73] and the RoseTTAFold method, as well as Phyre2. The quality of all five 3D models generated by Robetta for each Ffp1 protein was assessed and validated using two quality calculation tools: ERRAT [74, 75] and Verify3D[76]. The most accurate predicted structure was chosen and superposed to the best model target, found by VAST+[77, 78], Phyre2 and iPBA [79, 80]. The RMSD value [81] as well as the number and percentage of

aligned residues were retrieved and compared to Phyre2 results. RMSD values of <3 Å were considered significant between Ffp1 predicted structure and 3D models [82].

RESULTS

Porphyromonas taxogenomic assignment

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

The taxogenomic assignment for the genomes classified into the 17 *Porphyromonas* species was verified (Table S1). The *Porphyromonas* species *P. loveana* and *P. pasteri* have only one representative genome and therefore cannot be verified intra-specifically. For the other species, intra-specific analysis combining ANI, 16S rRNA and DDH comparisons (Fig. S1A) showed no anomalies for taxonomic placements, except for *P. uenonis*, *P. somerae* and *P. canoris*.

Firstly, for *P. uenonis*, the differences in metrics reflect a significant distance between strain 60-3 and the two other strains (Fig. S1A and S1B). Strain 60-3 was analysed using Kraken2 and it was concluded that *P. uenonis* 60-3 belongs to the genus *Porphyromonas* but not to *P. uenonis* (Fig. S2). This genome has been retained for the study but as an unclassified *Porphyromonas*, denoted as PSP_60-3 (Table S1).

Secondly, in the case of *P. somerae* KA00683, Kraken2 analysis indicates a genomic mixture, and our taxonomic analysis separates this strain from the other two within the species (Fig. S1B). Consequently, we have opted not to include *P. somerae* KA00683 in our study (Table S1).

Finally, regarding *P. canoris* (two genomes), the difference in the 16S rRNA gene sequences was associated with a longer gene in one strain (Fig. S1C). It is impossible to determine whether this difference represents genuine genomic diversity or a sequencing error; we consider both genomes as belonging to *P. canoris* (Table S1).

Furthermore, 28 *Porphyromonas* genomes lacked a species label. All genomes were examined using Kraken2, and genomes with less than 80% of *Porphyromonas* reads and/or that reconstructed less than 80% of *Porphyromonas* average genome size (2.5 Mb) were excluded from the study (Fig. S2 and Table S1). Consequently, 17 strains were omitted from this study (Table S1). Among the 11 remaining *Porphyromonas* sp., their placement in the OrthoFinder species tree based on ANI/DDH metrics (Fig. 1) allowed us to assign genomes to: *P. gulae*, *P. asaccharolytica*, *P. uenonis*, and two genomes to *P. canoris* (Fig. 1 and Table S1). Finally, there were six *P*. sp. genomes that could not be assigned to any specific group and were individually examined (unassigned, Table S1).

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

FFp1 is the only fimbrillin common to all *Porphyromonas*

Screening and clustering fimbrillin genes from *Porphyromonas* genomes resulted in the definition of 12 HHM profiles, one for each gene in either FimABCDE or Mfa12345, and two for Ffp1. Searching for sequence similarity in each *Porphyromonas* orfeome, using each of the 12 HHM profiles, enabled the identification and classification of these three fimbriae systems in all *Porphyromonas* genomes (Fig. 2).

fimABCDE locus

For the FimABCDE proteins (Fig. S3A), an expected value (E-value) calibration was performed and set to a minimum threshold of e^{-100} for each of the five profiles. Using this threshold, the detection of the locus *fimABCDE* exhibited both sensitivity and specificity, perfectly correlating with presence/absence of each gene.

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

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

The utilization of HMM profiles in our search strategy allows for the rapid and unambiguous identification and classification of fimbrial genes, even in cases with low mean amino acid percentage identities: 52.3% (FimA), 63.7% (FimB), 56.7% (FimC), 48.2% (FimD) and 49.8% (FimE). Additionally, the annotations of FimABCDE proteins are inconsistent, with the majority being labelled as hypothetical proteins or simply categorized as fimbrial proteins without any additional characterization (Fig. S3B). As such, ontology searches are almost impossible.

Moreover, the establishment an E-value threshold facilitates pinpointing abnormalities. For instance, in *P. gingivalis*, for the FimB HMM profile, the E-value is greater than the established threshold due to a nonsense mutation in *fimB* for the ATCC 33277 strain [83], and this gene is annotated as two genes (PGN_0181, e-value=$2.8e^{-63}$ and PGN_0182, e-value=$1.4e^{-55}$). The same case occurs in *P. uenonis*, for the FimE HMM profile, due to the incompleteness of this gene (at the end of contig) for the UMGS1452 strain.

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

mfa12345 locus

Significant E-values ranging from e^{-200} and e^{-100} were observed for each of the five Mfa12345 profiles (Fig. S3C). Specifically, regarding the Mfa1 HMM profile, three distinct situations were evident: (i) Mfa1 was recovered, with low E-values, in four species (*P. gingivalis*, *P. gulae*, *P. loveana* and *P. macacae*); (ii) in 14 groups, Mfa1 was identified with higher E-values; and (iii) in six species

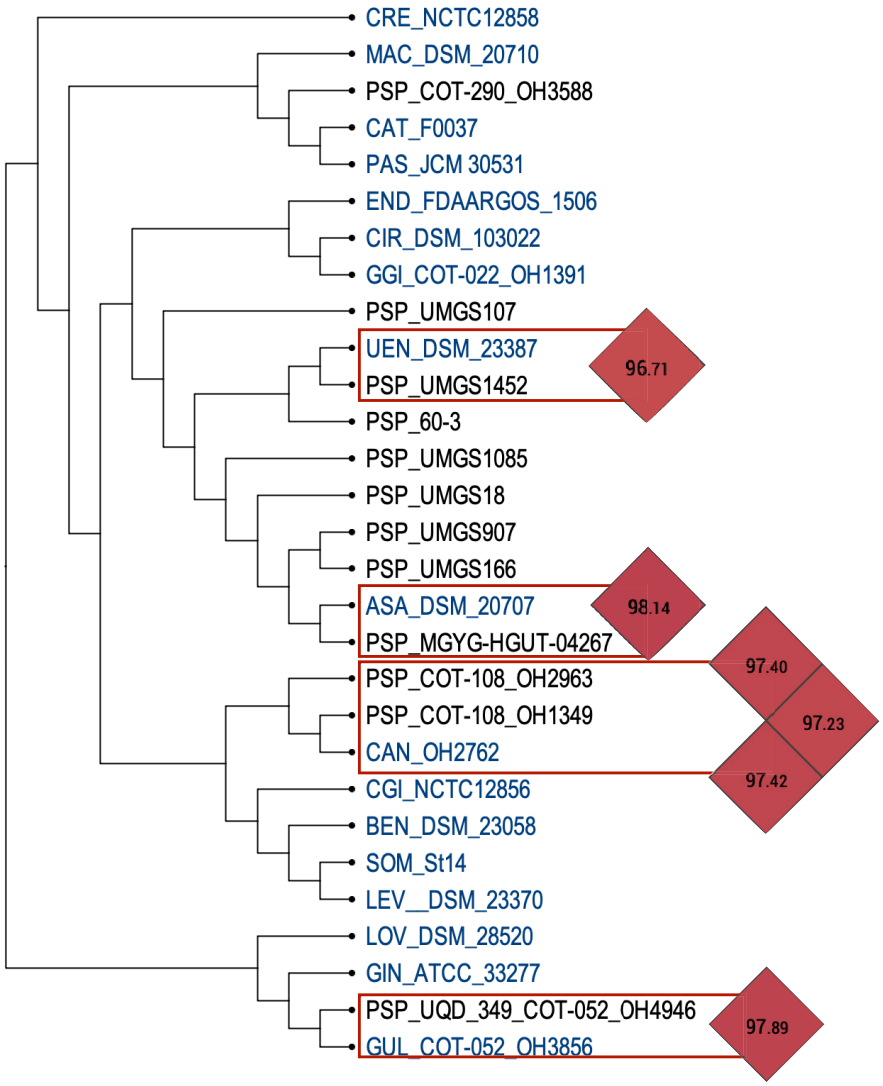

Fig. 1. Phylogenetic species tree derived from OrthoFinder analysis. This tree was used to place some *Porphyromonas* spp. within UEN (UMGS1452), ASA (MGYG-HGUT-0467), CAN (OH2963 and OH1349) and GUL (OH4946), after confirmation via OrthoANI. Three-letter code acronyms correspond to ASA: *P. asaccharolytica*; BEN: *P. bennonis*; CAN: *P. canoris*; CAT: *P. catoniae*; CGI: *P. cangingivalis*; CIR: *P. circumdentaria*; CRE: *P. crevioricanis*; END: *P. endodontalis*; GGI: *P. gingivicanis*; GIN: *P. gingivalis*; GUL: *P. gulae*; LEV: *P. levii*; LOV: *P. loveana*; MAC: *P. macacae*; PAS: *P. pasteri*; SOM: *P. somerae*; and UEN: *P. uenonis*.

(*P. bennonis*, *P. canoris*, *P. catoniae, P. cangingivalis* and *P. pasteri*, as well as PSP_OH3588), no Mfa1 was detected. The Mfa2 HMM profile produces identical results, yielding the same three groups.

The Mfa3 HMM profile successfully identified this protein in the same four species (*P. gingivalis*, *P. gulae*, *P. loveana* and *P. macacae*) and additionally in *P. endodontalis* that contains an Mfa3-like protein. Finally, both the Mfa4 and Mfa5 HMM profiles exclusively detected these proteins in *P. gingivalis*, *P. gulae* and *P. loveana* and in three of the six strains of *P. macacae*: JCM15984 and NCTC11632 (isolated from the oral cavity of cats) and OH2859 (isolated from a canine oral cavity). In OH2859, the *mfa12345* operon locus is intact, while in JCM15984 and NCTC11632, we observed two distinct loci: the first one contains genes encoding Mfa123 proteins, followed by two genes encoding proteins similar to FimD and FimE (referred to as *mfa123_fimDE*), and the second comprises genes encoding Mfa2345 proteins preceded by a non-characterized fimbrilin gene that shares similarity with Ffp1, indicated by low E-values of $7.3e^{-58}$ for Ffp1 profile A and $7.8e^{-41}$ for Ffp1 profile B (referred to as *ffp1-like_mfa2345*). It is worth noting that three strains of *P. macacae*, specifically OH2631 (isolated from the canine oral cavity), as well as NCTC13100 and DSM20710/JCM13914 (isolated from the macaque oral cavity), exhibit two tandemly organized *mfa123_fimDE* loci. Remarkably, these loci are not identical, displaying an average sequence identity of 53%. In OH2631, these loci are separated by less than 2 kb, while in NCTC13100 and DSM20710, they are separated by a 3 kb region that includes an IS4 pseudogene. None of these three strains harbour the *ffp1-like_mfa2345* locus.

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

In conclusion, when considering only the complete *mfa12345* locus as a reference, we identified its presence in four species: *P. gingivalis*, *P. gulae*, *P. loveana* and *P. macacae* strain OH2859. We also illustrate the effectiveness of HMM profiles in distinguishing true *mfa* loci from alternative loci. As for FimABCDE, the descriptions found in the annotations of Mfa12345 proteins are uninformative, often annotated as hypothetical or fimbria. This labelling makes it nearly impossible to conduct meaningful ontology searches (Fig. S3D).

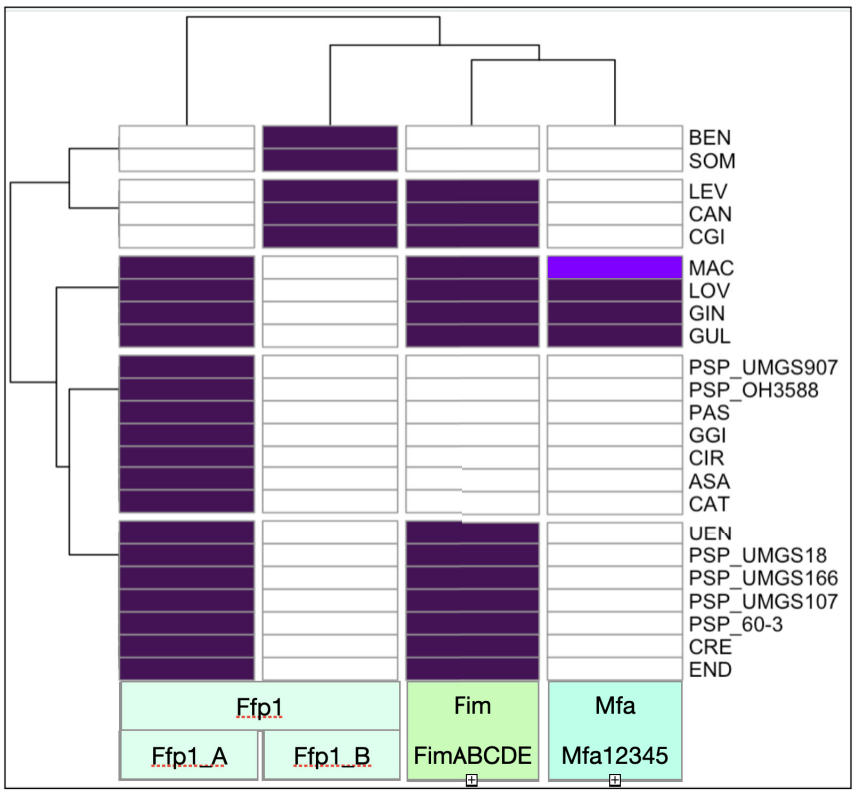

Fig. 2. Heatmap depicting the presence/absence of fimbrillins. The heatmap scale colour indicates whether fimbriae systems (FimABCDE, Mfa12345, or Ffp1_A or B) were detected: white (absence), dark purple (presence as one *locus*) and light purple (presence as two *loci*).

ffp1

MMseqs2 clustering reveals the separation of Ffp1 orthologues in two distinct groups which resulted in two distinct HMM profiles termed Ffp1_A and Ffp1_B (Fig. S3E). Ffp1_A mature amino acid sequences, excluding the signal peptide, share 57.4% identity and 23 conserved amino acids (logo in Fig. S3E), while Ffp1_B sequences exhibit only 37% identity primarily due to divergence in *P. bennonis,* but with 30 conserved amino acids (logo in Fig. S3E). The identity between the two groups decreases to 24% with only eight conserved amino acids (represented with an asterisk in the logos Fig. S3E).

The Ffp1_A HMM profile retrieves genes from all *Porphyromonas* species except *P. bennonis*, *P. canoris*, *P. cangingivalis*, *P. levii* and *P. somerae*, which are recovered with the Ffp1_B HMM profile. So, remarkably, fimbrillin Ffp1 is indeed present in all *Porphyromonas* spp., contrary to FimABCDE and Mfa12345 [except for *P*. sp. UMGS1085 where a 186 nt fragment of a gene (at the start of a contig) is identified by the Ffp1_A HMM profile with an E-value at 6.7 e-22 (Fig. S3E)]. This higher E-value is the result of being obtained for only 61 amino acids instead of about 500 for an Ffp1_A protein.

As shown in figure Fig. S3F, approximately 70% of the identified Ffp1 proteins are annotated as hypothetical or uncharacterized, 22% as fimbrillin/fimbriae (with half linked to the PGN_1808 protein, described as Ffp1 in the *P. gingivalis* ATCC 33277 reference strain) and 8% as lipoproteins.

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

CHARACTERIZATION OF *PORPHYROMONAS* FFP1 FIMBRIAE

Ffp1 exhibits variable pre-cleavage sizes among *Porphyromonas* species, in both subclasses. For the Ffp1_A group, protein sizes range from 439 aa (*P. circumdentaria* DSM 103022) to 553 aa (*P. asaccharolytica* PR426713P-I), and for the Ffp2_B group, from 483 aa (*P. somerae* DSM 23387) to 527 aa (*P. canoris*) (Fig. 3a). Size is well conserved within *Porphyromonas* species except for *P. asaccharolytica*, *P. circumdentaria*, *P. macacae* and *P. uenonis* for Ffp1_A, and *P. bennonis* for Ffp1_B (Fig. 3a).

The observed differences for *P. asaccharolytica* are due to the presence of 33 additional nucleotides in strain PR426713P-I (at position 88–120), absent in strain DSM 20707. For *P. circumdentaria*, it is a 175 nt shorter annotation in strain DSM 103022 (compared to strain ATCC 51356). For *P. macacae* these are due to the gene encoding Ffp1_A being at the end of the contig and truncated at the 5′ end, in strain *P. macacae* JCM 15984. For *P. uenonis*, it is also the choice of an alternative start codon for the UMGS1452 strain, 34 aa upstream of those chosen for the DSM 23387 and JCM 13868 strains. Finally, for *P. bennonis*, at position 1410 in the DSM 23058 strain, a C base, absent from the JCM 16335 strain, leads to a frameshift. This frameshift leads to a shorter C-terminal sequence compared to DSM 23058. Note that for *P. somerae*, the sizes are similar, but the annotated sequences are 'shifted' and proteins different on the N-terminal (20 aa longer in DSM 23387 compared to St14) and C-terminal [21 aa shorter in DSM 23387 due to a partial coding sequence (CDS) at the end of the contig].

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

In the absence of thorough human biocuration for structural annotation, particularly regarding the selection of start codons, a significant portion of Ffp1 proteins are predicted to be cytoplasmic (*P. asaccharolytica*, *P. catoniae*, *P. circumdentaria* DSM 103022, *P. somerae* St14) or having localization predictions classified as indeterminate (PSP UMGS107, PSP UMGS166, PSP UMGS907, *P. uenonis* DSM 23387, *P. uenonis* JCM 13868). Some proteins are predicted to be cleaved by SPII, but biocuration enhances both the signal peptide prediction score and the likelihood of cleavage by SPII. As a result of this reannotation work, all Ffp1 proteins are predicted as lipoproteins, with a signal peptide of about 20 aa (15–25 aa), consistent with the requirements cited previously: two to four positively charged amino acids followed by a hydrophobic region of 10–15 aa (Fig. S4) and a lipobox [ASG]↓C positions −1 to 1 (Fig. 3b). *In silico* predictions also confirm the predicted palmitoylation (addition of acyl chains) of the cysteine residue.

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

As shown in Fig. 3a (second panel), Ffp1 signal peptide biocuration not only results in more consistent predictions of their cellular localization, but also leads to a homogenization of their size, both within and across species, except for *P. bennonis*

(since the frameshift occurs in the 3′ region of the gene). This size homogenization becomes even more pronounced following signal peptide cleavage (Fig. 3a, third frame). Mature Ffp1s in group B are larger than those in group A by about 20 aa.

As shown in Fig. 3d, the average intra-specific identity of the Ffp1_A subclass is very high and ranges from 100% to 94.8% depending on the species. The most divergent species are *P. macacae*, *P. gulae* and *P. uenonis*. In the first two cases, this divergence can be attributed to the coexistence of two distinct homology groups within the same species. However, regrettably, the available metadata do not provide sufficient information to elucidate the underlying reasons for these discrepancies. For *P. uenonis*, strain UMGS1452 derived from a metagenome is different from the two other strains. As previously noted, the conservation of interspecific Ffp1_A sequences is low (57.5%) with only 4.5% of identical sites between all of them. When examining the Ffp1_B group, it is worth noting that the average intra-specific identity is elevated, oscillating between 98.8 and 91.1% (Fig. 3d). *P. bennonis* is the most divergent because the two strains have proteins with the last 75 aa that differ. It is noteworthy that Ffp1_B is less homogeneous than Ffp1_A with an average inter-specific identity of only 36.4% and 4.3% identical sites. The number of conserved sites decreases to 0.7% if we compare both groups, Ffp1_A and Ffp1_B.

3D STRUCTURES CONFIRM THAT *PORPHYROMONAS FFP1* ARE FIMBRILLINS

As the signal peptide is absent in the mature protein, it was excised prior to structure prediction for all Ffp1 proteins. PSIPRED predicts 30–44% residues as strand (mean=36.6, SD=3.4) and 2–7% residues as helix (mean=5.4, SD=1.4) for the Ffp1_A

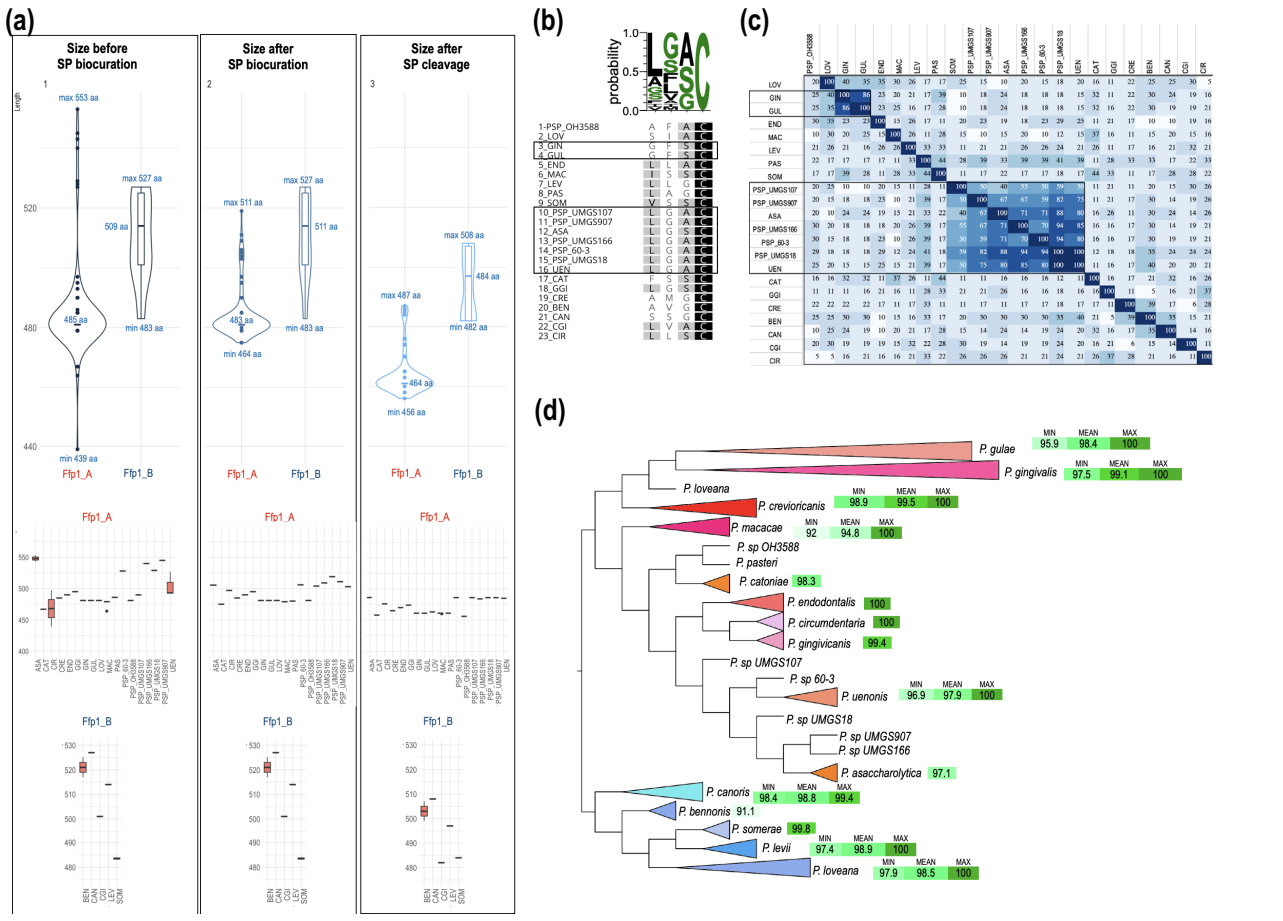

Fig. 3. (a) Violin plots of Ffp1_A and Ffp1_B amino acid lengths. From left to right: sizes as initially annotated in GenBank files (no curation), sizes after signal peptide (SP) biocuration prior to cleavage and sizes after SP cleavage by signal peptidase II (SPII). In the box plot associated with each violin plot, the middle line represents the median and the whiskers indicate the interquartile range. (b) Multiple sequence alignment and sequence logo of Ffp1 lipobox. Boxes represent groups of identical sequences. (c) Heat map illustrating the percentage nucleotide identity of Ffp1 signal peptides. (d) Phylogram of *Porphyromonas* Ffp1 proteins distance tree. The Ffp1_A proteins are depicted in warm colours, while Ffp1_B proteins are shown in various shades of blue. The boxes indicate the minimum, average and maximum intraspecific identity values. If only one value is displayed, it represents the average identity percentage.

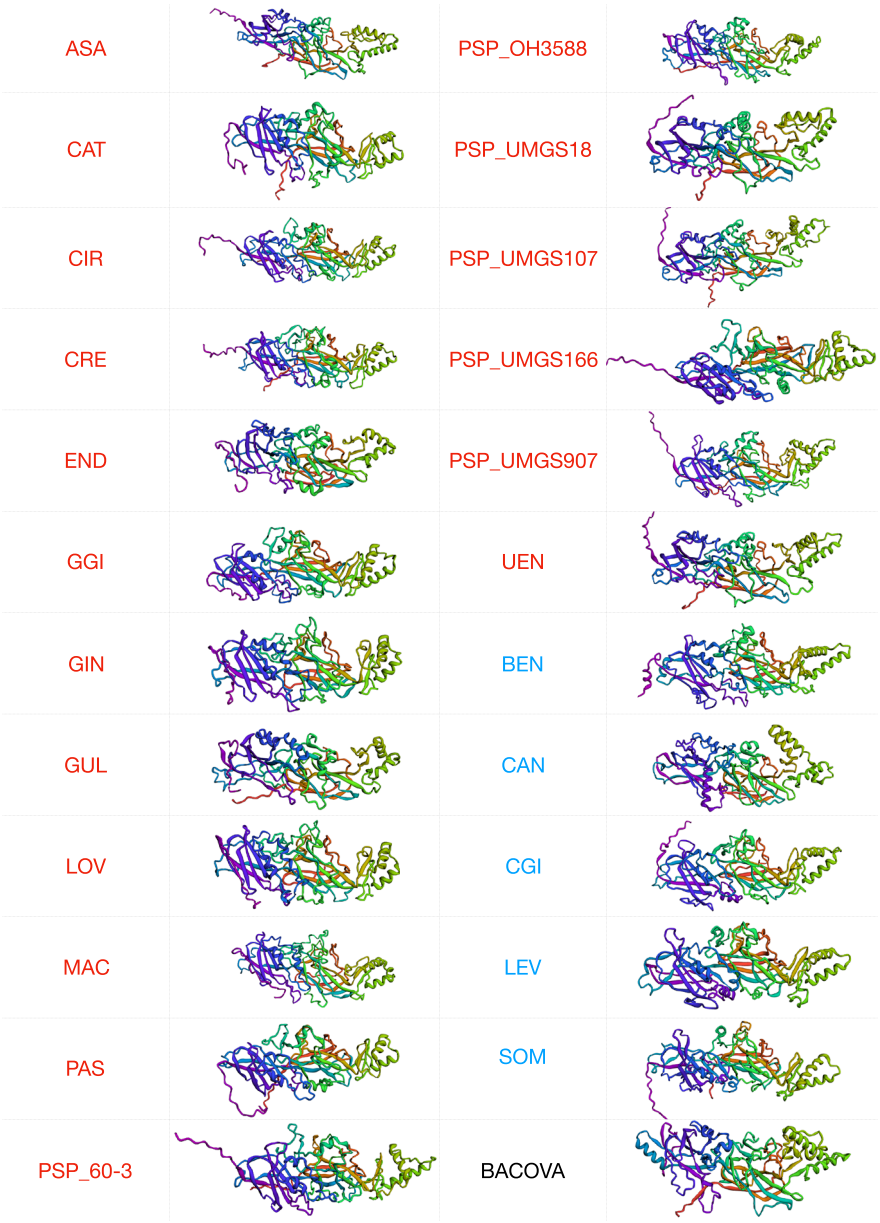

Fig. 4. Predicted tertiary structure for mature proteins of *Porphyromonas* reference strains (one per genus). These structures correspond to predictions made by Robetta and evaluated by ERRAT and Verify3D. Only the best prediction is represented. Ffp1_A proteins are in red and Ffp1_B in blue. BACOVA_01548 was also predicted using Robetta.

group. For the Ffp1_B group, predictions concern 24%–42% amino acids as strand (mean=29.6, SD=6.5) and 4%–9% as helix (mean=7, SD=2.1).

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

According to Phyre2 and iBPA results (Fig. S5), more than 82% of Ffp1 sequences were modelled with 100.0% confidence against 4JRF.pdb. Superposition of *Porphyromonas* Ffp1 and BACOVA_01548 3D structures were performed by iBPA and all evaluation values (RMSD, GDT_TS) reflect good overall similarity. For all overlapping morphologies, the aligned fraction is about 50% of the protein sequence, with mean reported RMSDs of 2.26 Å (range 2.09–2.53 Å) and mean GDT-TS distance scores of 32

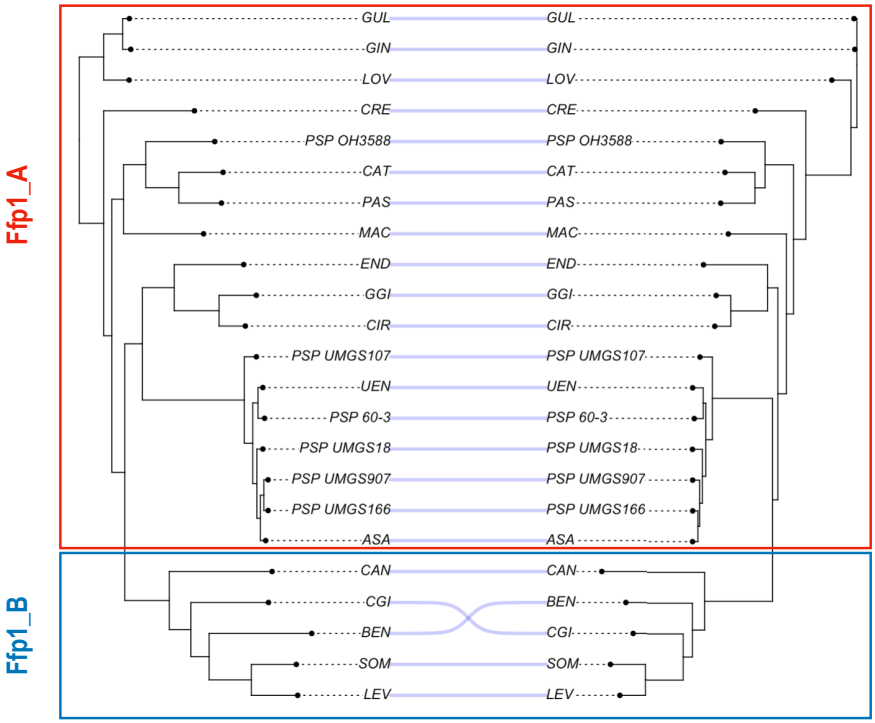

Fig. 5. Tanglegram comparing the tree reconstructed from the primary sequences of the Ffp1 proteins in representative strains of *Porphyromonas* (on the left) with the species tree based on the orthology of the orfeomes.

(range 32–37.3 Å) (Fig. S5). For *P. gingivalis*, the structures of FimA (4Q98.pdb) and Mfa1(5NF2.pdb) are available and comparisons by superposition between Ffp1 and these two other fimbrillins (Fig. S6) confirm that Ffp1 is indeed a new distinct *Porphyromonas* fimbrillin family.

PORPHYROMONAS FFP1 ARE ANCESTRAL ORTHOLOGUES BUT NOT SYNTELOGUES

In *P. gingivalis*, *ffp1* is the fourth gene in an operon-like structure comprising a gene encoding a cysteinyl-tRNA synthetase, a second gene encoding a patatin-like protein, and a third gene encoding a group 2 glycosyltransferase. An identical locus is found in all *P. gulae* genomes, while t is absent in all other *Porphyromonas* (Fig. S7). The *P. asaccharolytica*, *P. uenonis*, *P.* sp. UMGS18 and *P.* sp. UMGS107 group, mentioned above, show a syntenic pattern upstream of *ffp1*, characterized by the presence of two conserved genes encoding dihydroorotate dehydrogenases, crucial enzymes involved in *de novo* pyrimidine biosynthesis in prokaryotic cells. *P. uenonis* and *P.* sp. UMGS18 even extend this 5′ synteny with the gene *uvrA* encoding excinuclease ABC subunit A. *P. uenonis, P.* sp. UMGS18 and *P.* sp. 60-3 also share a *ffp1* downstream gene encoding a potassium/proton antiporter. *P. asaccharolytica*, *P.* sp. UMGS166 and *P.* sp. UMGS907 show three syntenic genes downstream of *ffp1*, one encoding a nitronate monooxygenase (degradation of propionate-3-nitronate), another encoding a 4-hydroxy-tetrahydrodipicolinate synthase (involved in lysine biosynthesis) and *recF*, involved in DNA replication and repair. Finally, *P. catoniae* and *P. pasteri* also share three conserved genes, upstream of *ffp1*, encoding respectively a serC phosphoserine transaminase, an NAD(P)-binding domain-containing protein and a protein with the DUF1015 domain (Fig. S7).

However, linking all these surrounding gene functions with *ffp1* fimbrillin is challenging, if not impossible, without functional experimentation. Furthermore, the intergenic spaces, often spanning several hundred nucleotides, suggest separate regulatory mechanisms and rule out any functional correlation between these genes. For the other *Porphyromonas* species, each exhibits a distinct gene organization arrangement surrounding *ffp1* (Fig. S7).

In conclusion, except for phylogenetically closely related species, we find no preserved synteny in the *ffp1* locus, which would reflect the absence of co-localization constraints for co-functional genes. Nevertheless, as demonstrated by the tanglegram juxtaposing the orfeome tree and the Ffp1 tree (Fig. 5), the remarkable congruence between these two trees provides compelling evidence that Ffp1 is an ancestral protein of *Porphyromonas,* and its evolution would have closely paralleled the evolutionary trajectory of the entire genus. This observation also holds true for the differentiation between the two Ffp1 classes (Fig. 5). The absence of gene conservation in close chromosomal proximity to *ffp1,* along with the presence of a significant 5′ intergenic space (Fig. S7), not only signifies the absence of

selection pressure around this gene but also strongly suggests that *ffp1* functions as an independent transcriptional unit. These results suggest strict vertical inheritance of *ffp1* in the genus *Porphyromonas* over a long period of evolutionary time, demonstrating that this fimbrillin is part of the *Porphyromonas* pangenome and is not an accessory gene. However, it should be noted that this locus appears to have evolved differently in each species or related group of species, as no strict synteny is observed.

DISCUSSION

Our analysis of fimbrillin loci within the genus *Porphyromonas* was initiated with the retrieval of genomes from the NCBI RefSeq database. The first step encompassed the validation of genus-level assignment for each genome retrieved, followed, when feasible, by species-level confirmation. The overall genome relatedness indices (OGRIs), namely digital DDH distance and gANI, were used to classify genomes into monophyletic groups. These OGRIs are increasingly used in taxogenomic studies and serve as a valuable tool for validating the taxonomic classification of isolates of interest [84]. Likewise, in accordance with prior research, we employed more conventional methodologies for species-level genome grouping, such as evaluating the percentage identity of the gene encoding 16S rRNA (when annotated) [85].Our study underscores the critical necessity of rigorously confirming the taxonomic classifications of genomes before embarking on any comparative genomics analysis to ensure their accuracy. Moreover, this checking step enables the possibility of taxonomic reassignment when warranted, aligning with findings from previous studies [86–88]. In this investigation, we have identified genomes erroneously labelled as *Porphyromonas* (i.e strain 31_2, which is a *Parabacteroides*), misassignment of *Porphyromonas* to species (i.e strain 60.3, which does not belong to *P. uenonis*) as well as metagenomic mixture such as strain KA000683 imperfectly assigned to *P. somerae*.

Our study also raises questions about genomes assigned to *Porphyromonas* without any species assignment (28 out of 144 genomes, i.e. 19.5%). They all correspond to incomplete draft genomes which introduces bias into studies that rely on them [89]. We specifically note the presence of gaps, local assembly errors, chimeras and contamination by fragments from other genomes [90, 91]. This contamination, defined as the presence of foreign sequences within a genome, can lead to incorrect functional inferences such as higher rates of horizontal gene transfer (HGT) and errors in phylogenomic studies. Such errors can be propagated throughout the scientific community and have been documented to exist in databases [91]. To mitigate these types of errors, several studies, including the present one, advocate the practice of data biocuration throughout the study. To identify potential contamination in draft genomes, we employed Kraken2 software and assessed the cumulative contig size of incomplete genomes. By applying specific inclusion criteria, we were able to disqualify 17 draft genomes, corresponding to metagenomic mixtures and inaccurately labelled as *Porphyromonas*. Furthermore, among the 11 remaining draft genomes, our taxogenomic approach led to the reclassification of five genomes into four previously described species (*P. gulae*, *P. asaccharolytica*, *P. uenonis* and two genomes in *P. canoris*). The remaining six genomes that cannot be assigned to already described species may potentially represent novel, yet undescribed species, akin to hypotheses proposed in other bacterial genera [88, 92]. This suggests that the genus *Porphyromonas* may encompass a greater degree of species diversity than previously recognized.

Thus, in this study, we retained 126 *Porphyromonas* genomes (24 clades comprising 17 species and seven singletons) to describe fimbriae loci. To accomplish our research objectives, distant homology between proteins must be detected and is fundamental for enabling comparative and evolutionary investigations, shedding light on protein families, and providing insights into their molecular structures and functions [93].

Current orthology detection methods include position-specific scoring matrix (PSSM) techniques, like PSI-BLAST (position-specific iterated BLAST [94], which generate substitution score profiles by accounting for residue variability within homologous sequence families [95]. An even more effective approach involves HMM profiles, which incorporate emission and transition state probabilities at each protein sequence position, making them a superior choice for identifying distant homology [95, 96].

Using ontology as a protein search strategy search is ineffective, as most fimbrillin genes are poorly annotated or annotated as 'hypothetical protein' (between 21.1 and 88.6% of annotated genes). Specifically, stem and anchor proteins (FimAB or Mfa12) are better annotated with deficient annotation rates ranging from 21.1 to 50.7%. In contrast, accessory proteins (FimCDE or Mfa345) suffer from particularly poor annotations with error percentages ranging from 58.9 to 88.6%. These annotation errors are present within the databases and, without biocuration and correction, are likely to persist, potentially perpetuating inconsistencies, inaccuracies and errors in subsequent genome annotations [97]. For example, for a gene family, nearly 20% of sequences may exhibit significant errors such as inaccuracies in gene names, partial sequences or initiation codon misassignments [98]. In the context of less extensively researched bacterial species, as is the case in this study, the prevalence of erroneous or uninformative annotations are much higher, reaching 77.1% of sequences identified as Ffp1 where the annotation was 'hypothetical protein' or 'lipoprotein'.

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

The *fimABCDE* locus is present in nine (of 24 groups, or 37.5% of *Porphyromonas* species) with two distinct *fim* loci present in all *P. macacae* genomes. The *mfa12345* locus is present only in three closely related species (*P. gingivalis*, *P. gulae* and *P. loveana*). For this locus, hybrid *fim/mfa* or *ffp1/mfa* loci are present in two species (*P. endodontalis* and *P. macacae*): *mfa123_fimDE* and *ffp1-like_mfa2345* in *P. macacae*; and a distinctive six-gene locus in *P. endodontalis*. This locus encompasses genes encoding Mfa1-like, Mfa2 and Mfa3-like proteins, along with two genes responsible for lipoproteins and a gene encoding a protein featuring a VWA domain. Interestingly, for the gene encoding Mfa5, the prevailing description is rather nondescript, simply stating it as a 'protein containing a VWA domain'. This description, however, falls short in conveying the functional significance of this gene. It is worth emphasizing that proteins featuring VWA domains play pivotal roles in diverse biological processes, including but not limited to cell adhesion and defence mechanisms. Thus, a more detailed annotation is warranted to better appreciate the functional implications of Mfa5 [99].

Finally, other species (i.e. *P. asaccharolytica*, *P. circumdentaria*, *P. crevioricanis*, *P. gingivicanis* and *P. uenonis*) have fimbrilin genes identified through HMM profiles that remain uncharacterized. These two loci, *fimABDCE* and *mfa12345*, have been described in other closely related species, for example an Mfa system (with only *mfa1* and *mfa2*) in *Bacteroides thetaiotaomicron* [100], and a cluster with *fimABCDE*-like genes and genes similar to either *mfa1*/*mfa2* or *mfa4*/*mfa2* with either *mfa1* or *mfa4* encoding the fimbriae stem and *mfa2* as an anchor in *Parabacteroides distasonis* [101]. The *fim* and *mfa loci* in *Porphyromonas* spp. will be the main subject of an further publication.

Concerning Ffp1 fimbriae (77.1% of all *ffp1* genes were deficiently annotated), this protein was most recently described in *P. gingivalis* [14, 15]. The encoding gene has two variants, denoted as A and B in our study. Ffp1_A is the predominant variant found in 19 *Porphyromonas* species/groups, whereas Ffp1_B is restricted to only five species (*P. bennonis*, *P. canoris*, *P. cangingivalis*, *P. levii* and *P. somerae*). Furthermore, this study demonstrates that the utilization of HMM profiles reveals that *ffp1* is confined to the genus *Porphyromonas* and is absent in closely related genera such as *Bacteroides* or *Prevotella*. This finding contrasts with approaches employing BLASTp and PSI-BLAST [17].

In the future, for certain *Porphyromonas* species with limited genomic data, it will be important to revalidate the absence of *fim* and/or *mfa* fimbrillin loci through extensive genome sequencing efforts, particularly targeting several strains. In comparative genomics, the inherent incompleteness of draft genomes demands careful consideration to nuance results, particularly when the conclusions concern gene absence. Nevertheless, in our investigation, absence of the *fim* and/or *mfa* locus was substantiated by the non-detection of all ten corresponding genes, which, we believe, accentuate the robustness of our findings. In a broader sense, it is regrettable that raw reads are not available for download for most draft genomes. Having access to these raw reads could potentially allow for the confirmation of gene absence when needed.

The presence of multiple fimbriae loci within genomes is a common phenomenon observed in other bacterial models. These loci are often associated with general niche colonization abilities or the adhesion to more specific substrates [102, 103]. Further investigations are needed on species more closely related to *Porphyromonas* and within this bacterial genus. These studies can shed light on aspects such as host specificity and their association with species-related pathologies [104].

Given that the majority of *in silico* CDS annotators tend to prioritize the prediction of the longest possible ORF by favouring the initiation codon (ATG) over alternative codons (TTG and GTG) [105, 106], and considering the variable size of proteins across *Porphyromonas* species, we conducted a thorough examination of the annotated initiation codons for each predicted Ffp1. Given that fimbrillins are lipoproteins [10], their N-terminal region is expected to feature a signal peptide starting with positively charged amino acids, followed by hydrophobic amino acids, and concluding with a cysteine-terminated lipobox, which serves as the cleavage site for SPII. The biocuration of start codons led to a more consistent protein size post-signal peptide cleavage. Additionally, the extracellular prediction of mature lipoproteins was confirmed, characterized by the presence of charged and hydrophobic residues, the lipobox, and a palmitoylation site. These features align with the ancestral nature of FFp1.

In addition, Ffp1 3D modelling of the mature protein was performed with several software packages, and the predictions were evaluated with classical metrics [107, 108]. In all cases, the generated models were compared with existing 3D structures, and the most significant match was found with the cell adhesion protein BACOVA_01548 from *Bacteroides ovatus* [4]. This *B. ovatus* protein has not been extensively studied but was classified by the authors as the stem of a type V pilus, sharing common features with type V fimbriae. These characteristics include export to the periplasm as a lipoprotein (prepilin), subsequent delivery to the outer membrane, translocation to the cell surface and cleavage by Rgp (Arg-gingipain) [5, 109].

Moreover, this new fimbrillin, Ffp1, exhibits notable distinctions from both FimA and Mfa1, as evident from the obtained metrics when superimposing the 3D structures of these proteins available for *P. gingivalis*. Furthermore, the gene arrangement of *ffp1* differs from the *fim* and *mfa* operons as the gene encoding Ffp1 does not appear to be in an operon structure. The strict vertical inheritance of *ffp1* in *Porphyromonas* suggests a vital role for this fimbrillin, as Ffp1 plays a role in responding to environmental signals, such as acid stress, and in polymicrobial biofilm production [14]. Moreover, it has been identified as enriched in sphingolipid-containing OMVs [18, 19]. Thus, Ffp1 appears to serve crucial and diverse functions, facilitating *Porphyromonas* host colonization by promoting stress

adaptation, biofilm formation and OMV production. The significance of these functions probably explains its vertical transmission and conservation within the genus *Porphyromonas*.

CONCLUDING REMARKS

HMM profiles are potent tools for detecting distant homologies and facilitating phylogenetic studies. For conducting these investigations, meticulous manual biocuration is essential, as with any *in silico* research. In this article, these HMM profiles make it possible to discriminate, without ambiguity, three *Porphyromonas* fimbriae and to describe their distribution: *mfa12345*, limited to the three closely related species (*P. gingivalis*, *P. gulae* and *P. loveana*); *fimABCDE* present in nearly 40% of the *Porphyromonas* species; and *ffp1*, present in all *Porphyromonas* but restricted to this bacterial genus. Our study predicts that Ffp1 is a new fimbrillin, distinct from FimA and Mfa1. It is closely related to another type V fimbrillin protein, BACOVA_01548, as evidenced by manual start codon curation and 3D modelling. Given the ancestral nature of Ffp1, as elucidated by our study, and its presence in all studied *Porphyromonas* genomes, in contrast to the fimbrillins Fim and Mfa, the question of its function becomes paramount, especially in the absence of co-localization of accessory genes ensuring its stability, assembly and anchorage to the cell surface. What role does it play in the production and cargo of OMVs, a phenomenon observed in numerous studies? Further wet-lab investigations are necessary to address these pending inquiries.

Funding information
This work received no specific grant from any funding agency.

Author contributions
L.A.-A.: conceptualization, validation, writing – original draft, writing – review and editing, visualization; F.B.H.: conceptualization, methodology, software, data curation, writing – review and editing, supervision, project administration.

Conflicts of interest
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Acuña-Amador L, Barloy-Hubler F. HMM profiles used in "Ffp1, an ancestral Porphyromonas spp. fimbrillin." Zenodo; 2024. https://zenodo.org/records/10519421?token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6IjQ5MjE4NWEzLWZiYTQtNDAyYi1iZDgxLWM0NmYxNjA3ZTBmNCIsImRhdGEiOnt9LCJyYW5kb20iOiI1ZWFmNDdkODhiMDQwNDc3MDYxNWM3MjcyYmViZmRlYiJ9.4w7roQaNawNkxfWaPZTW2KgidmqyDQGiqvLHm-tlONnN9d9xqQfp5biGaCXpnqUANGHulMouVDaaeak1lPwU2g
2. Lukaszczyk M, Pradhan B, Remaut H. The biosynthesis and structures of bacterial Pili. In: *Bacterial Cell Walls and Membranes*. Springer, 2019. pp. 369–413.
3. Hospenthal MK, Costa TRD, Waksman G. A comprehensive guide to pilus biogenesis in Gram-negative bacteria. *Nat Rev Microbiol* 2017;15:365–379.
4. Xu Q, Shoji M, Shibata S, Naito M, Sato K, *et al*. A distinct type of pilus from the human microbiome. *Cell* 2016;165:690–703.
5. Shibata S, Shoji M, Okada K, Matsunami H, Matthews MM, *et al*. Structure of polymerized type V pilin reveals assembly mechanism involving protease-mediated strand exchange. *Nat Microbiol* 2020;5:830–837.
6. Yoshimura F, Takahashi K, Nodasaka Y, Suzuki T. Purification and characterization of a novel type of fimbriae from the oral anaerobe *Bacteroides gingivalis*. *J Bacteriol* 1984;160:949–957.
7. Hamada N, Sojar HT, Cho MI, Genco RJ. Isolation and characterization of a minor fimbria from *Porphyromonas gingivalis*. *Infect Immun* 1996;64:4788–4794.
8. Fujiwara-Takahashi K, Watanabe T, Shimogishi M, Shibasaki M, Umeda M, *et al*. Phylogenetic diversity in Fim and Mfa gene clusters between *Porphyromonas Gingivalis* and *Porphyromonas Gulae*, as a potential cause of host specificity. *J Oral Microbiol* 2020;12:1775333.
9. Onoe T, Hoover CI, Nakayama K, Ideka T, Nakamura H, *et al*. Identification of *Porphyromonas gingivalis* prefimbrilin possessing a long leader peptide: possible involvement of trypsin-like protease in fimbrilin maturation. *Microb Pathog* 1995;19:351–364.
10. Shoji M, Naito M, Yukitake H, Sato K, Sakai E, *et al*. The major structural components of two cell surface filaments of *Porphyromonas gingivalis* are matured through lipoprotein precursors. *Mol Microbiol* 2004;52:1513–1525.
11. Kuboniwa M, Amano A, Hashino E, Yamamoto Y, Inaba H, *et al*. Distinct roles of long/short fimbriae and gingipains in homotypic biofilm development by *Porphyromonas gingivalis*. *BMC Microbiol* 2009;9:1–13.
12. Shoji M, Yoshimura A, Yoshioka H, Takade A, Takuma Y, *et al*. Recombinant *Porphyromonas gingivalis* FimA preproprotein expressed in *Escherichia coli* is lipidated and the mature or processed recombinant FimA protein forms a short filament in vitro. *Can J Microbiol* 2010;56:959–967.
13. Yoshimura F, Murakami Y, Nishikawa K, Hasegawa Y, Kawaminami S. Surface components of *Porphyromonas gingivalis*. *J Periodontal Res* 2009;44:1–12.
14. Gui MJ. Characterization of the Porphyromonas Gingivalis Protein PG1881 and Its Roles in Outer Membrane Vesicle Biogenesis and Biofilm Formation. PhD Thesis, University of Melbourne 2016.
15. Nagano K, Hasegawa Y, Yoshida Y, Yoshimura F. Novel fimbrilin PGN_1808 in *Porphyromonas gingivalis*. *PLoS One* 2017;12:e0173541.
16. Hasegawa Y, Iijima Y, Persson K, Nagano K, Yoshida Y, *et al*. Role of Mfa5 in expression of Mfa1 fimbriae in *Porphyromonas gingivalis*. *J Dent Res* 2016;95:1291–1297.
17. Gupta RS, Lorenzini E. Phylogeny and molecular signatures (conserved proteins and indels) that are specific for the *Bacteroidetes* and *Chlorobi* species. *BMC Evol Biol* 2007;7:71.
18. Veith PD, Shoji M, Scott NE, Reynolds EC. Characterization of the O-glycoproteome of *Porphyromonas gingivalis*. *Microbiol Spectr* 2022;10:e0150221.
19. Rocha FG, Ottenberg G, Eure ZG, Davey ME, Gibson FC. Sphingolipid-containing outer membrane vesicles serve as a delivery vehicle to limit macrophage immune response to *Porphyromonas gingivalis*. *Infect Immun* 2021;89:e00614-20.
20. Iwashita N, Nomura R, Shirai M, Kato Y, Murakami M, *et al*. Identification and molecular characterization of *Porphyromonas gulae* fimA types among cat isolates. *Vet Microbiol* 2019;229:100–109.

21. Oishi Y, Watanabe K, Kumada H, Ishikawa E, Hamada N. Purification and characterization of a novel secondary fimbrial protein from *Porphyromonas gulae*. *J Oral Microbiol* 2012;4.
22. Collings S, Love DN. Further studies on some physical and biochemical characteristics of asaccharolytic pigmented Bacteroides of feline origin. *J Appl Bacteriol* 1992;72:529–535.
23. Love DN, Bailey GD, Collings S, Briscoe DA. Description of *Porphyromonas circumdentaria* sp. nov. and reassignment of *Bacteroides salivosus* (Love, Johnson, Jones, and Calverley 1987) as *Porphyromonas* (Shah and Collins 1988) *salivosa* comb. nov. *Int J Syst Bacteriol* 1992;42:434–438.
24. Koyata Y, Watanabe K, Toyama T, Sasaki H, Hamada N. Purification and characterization of a fimbrial protein from *Porphyromonas salivosa* ATCC 49407. *J Vet Med Sci* 2019;81:916–923.
25. NCBI. NCBI RefSeq Database; 2024. https://www.ncbi.nlm.nih.gov/refseq/
26. Blin K. NCBI-genome-download v0.2.12; 2023. https://github.com/kblin/ncbi-genome-download
27. Kim M, Oh HS, Park SC, Chun J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int J Syst Evol Microbiol* 2014;64:346–351.
28. Meier-Kolthoff JP, Auch AF, Klenk H-P, Göker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinform* 2013;14:60.
29. Frank A. GGDC-robot v0.04; 2018. https://github.com/andrewfrank/ggdc-robot
30. Wayne LG. International Committee on Systematic Bacteriology: announcement of the report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Zentralbl Bakteriol Mikrobiol Hyg A* 1988;268:433–434.
31. Stackebrandt E, Goebel BM. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int J Syst Evol Microbiol* 1994;44:846–849.
32. Scofield DG. FastANI v1.34; 2023. https://github.com/ParBLiSS/FastANI
33. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 2018;9:5114.
34. Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, *et al*. Microbial species delineation using whole genome sequences. *Nucleic Acids Res* 2015;43:6761–6771.
35. Lee I, Kim YO, Park SC, Chun J. OrthoANI v0.93.1; 2015. https://www.ezbiocloud.net/tools/orthoani
36. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol* 2019;20:257.
37. Liang Q, Liu C, Xu R, Song M, Zhou Z, *et al*. fIDBAC: a platform for fast bacterial genome identification and typing. *Front Microbiol* 2021;12:723577.
38. Hangzhou Digital-Micro Biotech. fIDBAC; 2021. http://fbac.dmicrobe.cn
39. Emms D. OrthoFinder v2.5.5; 2023. https://github.com/davidemms/OrthoFinder
40. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 2019;20:238.
41. Rimbaut A. FigTree v1.4.4; 2018. http://tree.bio.ed.ac.uk/software/figtree/
42. NCBI. Basic Local Alignment Search Tool; 2024. https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins
43. EMBL-EBI. InterPro: Classification of protein families; 2024. https://www.ebi.ac.uk/interpro/search/sequence/
44. Davis JJ, Wattam AR, Aziz RK, Brettin T, Butler R, *et al*. The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. *Nucleic Acids Res* 2020;48:D606–D612.
45. BV-BRC. PATRIC v3.6.6; 2024. https://www.bv-brc.org
46. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res* 2013;41:e121.
47. HMMER. HMMER: biosequence analysis using profile hidden Markov models; 2023. http://hmmer.org
48. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, *et al*. Pfam: the protein families database in 2021. *Nucleic Acids Res* 2021;49:D412–D419.
49. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;35:1026–1028.
50. Sievers F, Higgins DG. Clustal omega for making accurate alignments of many protein sequences. *Protein Sci* 2018;27:135–145.
51. EMBL-EBI. Clustal Omega: Multiple Sequence Alignment (MSA) [Internet]. The EMBL-EBI Job Dispatcher sequence analysis tools framework in 2024; 2024. https://www.ebi.ac.uk/jdispatcher/msa/clustalo
52. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, *et al*. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 2012;28:1647–1649.
53. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 2000;16:276–277.
54. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, *et al*. Protein identification and analysis tools on the ExPASy server. In: Walker JM (ed). *The Proteomics Protocols Handbook*. Totowa, NJ: Humana Press; 2005. pp. 571–607.
55. Swiss Institute of Bioinformatics. Swiss-Prot protein knowledgebase; 2024. https://web.expasy.org/protscale/
56. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, *et al*. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol* 2019;37:420–423.
57. Juncker AS, Willenbrock H, Von Heijne G, Brunak S, Nielsen H, *et al*. Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci* 2003;12:1652–1662.
58. Ren J, Wen L, Gao X, Jin C, Xue Y, *et al*. CSS-Palm 2.0: an updated software for palmitoylation sites prediction. *Protein Eng Des Sel* 2008;21:639–644.
59. Wickham H. ggplot2. In: *Ggplot2: Elegant Graphics for Data Analysis*. Cham: Springer-Verlag New York, 2016.
60. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 2002;30:3059–3066.
61. Reguant R. Alignmentviewer v1.1; 2020. https://github.com/sanderlab/alignmentviewer
62. McInnes L, Healy J, Saul N, Großberger L. UMAP: uniform manifold approximation and projection. *J Open Source Softw* 2018;3:861.
63. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* 2010;5:e9490.
64. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, *et al*. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 2010;59:307–321.
65. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;30:1312–1313.
66. Smith MR. TreeDist: Distances between Phylogenetic Trees. R package version 2.7.0. Comprehensive R Archive Network; 2020. https://cran.rstudio.com/web/packages/TreeDist/index.html
67. Revell LJ. phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things); 2024. https://cran.r-project.org/web/packages/phytools/index.html
68. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. *Bioinformatics* 2000;16:404–405.

69. UCL Department of computer Science. PSIPRED v4.0; 2024. http://bioinf.cs.ucl.ac.uk/psipred/
70. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* 2015;10:845–858.
71. Structural Bioinformatics Group at Imperial College in London. Phyre2 v2.0; 2024. http://www.sbg.bio.ic.ac.uk/~phyre2/html/page.cgi?id=index
72. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, *et al*. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 2021;373:871–876.
73. Baker Lab at University of Washington. Robetta; 2024. http://robetta.bakerlab.org
74. Colovos C, Yeates TO. Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci* 1993;2:1511–1519.
75. UCLA & Institute for Genomics and Proteomics US Department of Energy Office of Science. ERRAT & Verify3D; 2024. https://saves.mbi.ucla.edu
76. Bowie JU, Lüthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 1991;253:164–170.
77. Madej T, Lanczycki CJ, Zhang D, Thiessen PA, Geer RC, *et al*. MMDB and VAST+: tracking structural similarities between macromolecular complexes. *Nucleic Acids Res* 2014;42:D297–D303.
78. NCBI. VAST+; 2024. https://structure.ncbi.nlm.nih.gov/Structure/vastplus/vastplus.cgi
79. Gelly J-C, Joseph AP, Srinivasan N, de Brevern AG. iPBA: a tool for protein structure comparison using sequence alignment strategies. *Nucleic Acids Res* 2011;39:W18–W23.
80. INSERM-Université Paris Cité-UMR_S U1134. iPBA; 2024. https://www.dsimb.inserm.fr/dsimb_tools/ipba/index.php
81. Carugo O, Pongor S. A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Sci* 2001;10:1470–1473.
82. Reva BA, Finkelstein AV, Skolnick J. What is the probability of a chance prediction of a protein structure with an rmsd of 6 A? *Fold Des* 1998;3:141–147.
83. Dashper SG, Mitchell HL, Seers CA, Gladman SL, Seemann T, *et al*. *Porphyromonas gingivalis* uses specific domain rearrangements and allelic exchange to generate diversity in surface virulence factors. *Front Microbiol* 2017;8:48.
84. Kirdat K, Tiwarekar B, Sathe S, Yadav A. From sequences to species: charting the phytoplasma classification and taxonomy in the era of taxogenomics. *Front Microbiol* 2023;14:1123783.
85. Wagner J, Coupland P, Browne HP, Lawley TD, Francis SC, *et al*. Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification. *BMC Microbiol* 2016;16:274.
86. Liu Y, Du J, Pei T, Du H, Feng G-D, *et al*. Genome-based taxonomic classification of the closest-to-Comamonadaceae group supports a new family *Sphaerotilaceae* fam. nov. and taxonomic revisions. *Syst Appl Microbiol* 2022;45:126352.
87. Khoder M, Osman M, Kassem II, Rafei R, Shahin A, *et al*. Whole genome analyses accurately identify *Neisseria* spp. and limit taxonomic ambiguity. *Int J Mol Sci* 2022;23:13456.
88. Tambong JT. Taxogenomics and systematics of the genus *Pantoea*. *Front Microbiol* 2019;10.
89. Sousa T de J, Parise D, Profeta R, Parise MTD, Gomide ACP, *et al*. Re-sequencing and optical mapping reveals misassemblies and real inversions on *Corynebacterium pseudotuberculosis* genomes. *Sci Rep* 2019;9:16387.
90. Chen LX, Anantharaman K, Shaiber A, Eren AM, Banfield JF. Accurate and complete genomes from metagenomes. *Genome Res* 2020;30:315–333.
91. De Simone G, Pasquadibisceglie A, Proietto R, Polticelli F, Aime S, *et al*. Contaminations in (meta)genome data: an open issue for the scientific community. *IUBMB Life* 2020;72:698–705.
92. Colston SM, Fullmer MS, Beka L, Lamy B, Gogarten JP, *et al*. Bioinformatic genome comparisons for taxonomic and phylogenetic assignments using *Aeromonas* as a test case. *mBio* 2014;5:e02136.
93. Trachana K, Forslund K, Larsson T, Powell S, Doerks T, *et al*. A phylogeny-based benchmarking test for orthology inference reveals the limitations of function-based validation. *PLoS One* 2014;9:e111122.
94. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–3402.
95. Kumar G, Srinivasan N, Sandhya S. Profiles of natural and designed protein-like sequences effectively bridge protein sequence gaps: implications in distant homology detection. In: *Data Mining Techniques for the Life Sciences*. New York: Springer US, 2022. pp. 149–167.
96. Jin X, Liao Q, Liu B. PL-search: a profile-link-based search method for protein remote homology detection. *Brief Bioinform* 2021;22:bbaa051.
97. de Crécy-Lagard V, Amorin de Hegedus R, Arighi C, Babor J, Bateman A, *et al*. A roadmap for the functional annotation of protein families: a community perspective. *Database* 2022;2022:baac062.
98. Ruiz J. Analysis of the presence of erroneous Qnr sequences in GenBank. *J Antimicrob Chemother* 2018;73:1213–1216.
99. Colombatti A, Bonaldo P. The superfamily of proteins with von Willebrand factor type A-like domains: one theme common to components of extracellular matrix, hemostasis, cellular adhesion, and defense mechanisms. *Blood* 1991;77:2305–2315.
100. Mihajlovic J, Bechon N, Ivanova C, Chain F, Almeida A, *et al*. A putative type V pilus contributes to bacteroides thetaiotaomicron biofilm formation capacity. *J Bacteriol* 2019;201:e00650-18.
101. Chamarande J, Cunat L, Alauzet C, Cailliez-Grimal C. In silico study of cell surface structures of *Parabacteroides distasonis* involved in its maintenance within the gut microbiota. *Int J Mol Sci* 2022;23:9411.
102. González-Montalvo MA, Tavares-Carreón F, González GM, Villanueva-Lozano H, García-Romero I, *et al*. Defining chaperone-usher fimbriae repertoire in *Serratia marcescens*. *Microb Pathog* 2021;154:104857.
103. Khater F, Balestrino D, Charbonnel N, Dufayard JF, Brisse S, *et al*. In silico analysis of usher encoding genes in *Klebsiella pneumoniae* and characterization of their role in adhesion and colonization. *PLoS One* 2015;10:e0116215.
104. Acuña-Amador L, Barloy-Hubler F. *Porphyromonas* spp. have an extensive host range in ill and healthy individuals and an unexpected environmental distribution: a systematic review and meta-analysis. *Anaerobe* 2020;66:102280.
105. Villegas A, Kropinski AM. An analysis of initiation codon utilization in the domain bacteria - concerns about the quality of bacterial genome annotation. *Microbiology* 2008;154:2559–2661.
106. Nielsen P, Krogh A. Large-scale prokaryotic gene prediction and comparison to genome annotation. *Bioinformatics* 2005;21:4322–4329.
107. Wang X, Snoeyink J. Defining and computing optimum RMSD for gapped and weighted multiple-structure alignment. *IEEE/ACM Trans Comput Biol Bioinform* 2008;5:525–533.
108. Shibuya T. Efficient substructure RMSD query algorithms. *J Comput Biol* 2007;14:1201–1207.
109. Burrows LL. Heads or tails for type V pilus assembly. *Nat Microbiol* 2020;5:782–784.