Bioinformatic process for the identification and characterization of bacterial repeat-in-toxin adhesins.

Thomas Hansen[1], Laurie A. Graham[1], Blake P. Soares[1], Daniel Lee[1], Justin R. Gagnon[1], Trina Dykstra-MacPherson[1], Shuaiqi Guo[2], Peter L. Davies[1*]

[1]Department of Biomedical and Molecular Sciences, Queen's University, Kingston, Ontario

[2]Department of Anatomy and Cell Biology, McGill University, Montreal, Quebec

*Corresponding author:

E-mail address: daviesp@queensu.ca

Abstract

Bacteria bind to host surfaces with ligand-binding domains (LBDs) at the distal tips of fibrillar adhesins. Blocking these initial binding interactions could potentially stop colonization that would otherwise lead to biofilm formation and infections. To develop such preventative strategies, it will be essential to know the full range of adhesins and their LBDs that are present in the various strains of a particular bacterial pathogen. The bioinformatic process we have developed to find RTX adhesins in Gram-negative bacteria starts with the collation of all predicted proteins from long-read genome sequences for the target species. Proteins are clustered by similarity of their C-terminal regions then scanned through InterPro to detect domains characteristic of RTX adhesins. Candidates are then modeled using AlphaFold3 and classified as RTX adhesins if they contain both an N-terminal retention domain and a C-terminal Type I Secretion Signal. Many species have several distinct RTX adhesins, most of which have LBDs, including putative carbohydrate-binding modules and von Willebrand Factor A-like domains, projecting from split domains. Similar adhesin structures can be shared by several species, possibly by descent or by acquisition though DNA uptake. Some species have an RTX adhesin of uncertain function that has no obvious LBDs.

198 words

Abbreviations

BIg, Bacterial Immunoglobulin-like

CBM, Carbohydrate-Binding Module

FA, Fibrillar Adhesins

FrhA, Flagellar regulated hemagglutinin A

LapA, Long adhesin protein A

LBR, Ligand-Binding Region

LBD, Ligand-Binding Domain

MapA, Medium adhesin protein A

MARTX, Multifunctional Autoprocessing Repeats-in-ToXin

NRPS, Non-Ribosomal Peptide Synthetases

PBD, Peptide-Binding Domain

RM, Retention Module

RTX, Repeat-in-ToXin

T1SS, Type I Secretion System

vWFA, von Willebrand Factor A-like domain

Introduction (rough draft)

Fibrillar adhesins (FAs) are long, single-chain bacterial surface proteins that fold as a string of domains [1-3]. Growing evidence suggests that FAs make initial contact with various surfaces through ligand-binding domains (LBDs) found at the distal end of their FAs [4-8]. These initial contacts include abiotic surfaces [9-11], as well as the surfaces of host cells [8, 12]. Once colonization is underway, FAs also contribute to the formation of biofilms, which exacerbate infections and shield the bacteria from both host defenses and antibiotics [11, 13-21].

Repeat-in-toxin (RTX) adhesins are a subset of FAs that are secreted through the type I secretion system (T1SS) [22-24], and a generic schematic of their domain architecture is shown in Fig. 1A. They have a T1SS signal at the C terminus preceded by a characteristic Ca^{2+}-binding nonapeptide repeat sequence [24] that folds into a β-roll/solenoid [25]. It has been suggested that the Ca^{2+}-dependent folding of this domain and those that emerge subsequently prevent reversal of secretion and may help extrude the adhesin into the Ca^{2+}-rich milieu outside of the bacterium [24, 26]. Only the very N-terminal retention module (RM) folds in the absence of Ca^{2+}, which enables it to serve as a periplasmic plug to terminate secretion and anchor the adhesin to the cell surface [27, 28].

Structure-function studies on RTX adhesins show that most of their domains are Bacterial Immunoglobulin-like (BIg) beta-sandwich modules that extend the distal ligand-binding region (LBR) outward from the bacterial surface to help contact binding partners [3, 29-31]. LBRs that have been examined in detail typically have two or three LBDs projected outward from the adhesin by 'split' domains [32]. For example, *Marinomonas primoryensis*, a bacterium isolated from a frozen lake in Antarctica [33], has an adhesin with three LBDs. The most distal of the three is the ice-

binding module that anchors *M. primoryensis* to the underside of ice [34], which is also where diatoms and other photosynthetic organisms receive the most sunlight. Adjacent to the ice-binding module are two other LBDs, one of which has affinity for fucosylated surface glycans [35]. The other binds to the C terminus of peptides, with nM affinity for the tripeptide sequence -Tyr-Thr-Asp-COOH [36]. These latter two LBDs bind to the diatom *Chaetoceros neogracile* and can be competed off by mM concentrations of fucose and sub-mM concentrations of tripeptide, respectively [35, 36].

In another example, Flagellar regulated hemagglutinin A (FrhA) is a key RTX adhesin for the infectivity of *Vibrio cholerae* classical O strain. This relatively short FA has two of the three LBDs found in the *M. primoryensis* ice-binding adhesin, these being the carbohydrate-binding module (CBM) that targets fucosylated surface glycans and the peptide-binding domain (PBD). Again, millimolar concentrations of fucose and sub-millimolar concentrations of tripeptide can block the *V. cholerae* LBDs from binding their targets [7, 37].

A third example is an RTX adhesin from *Aeromonas hydrophila*, which is a Gram-negative opportunistic pathogen found in fresh and brackish waters [38]. This adhesin has three putative LBDs, one of which is a CBM that binds to Lewis B and Y antigens [8]. These glycans are fucosylated and millimolar concentrations of free fucose were found to block CBM binding to a wide variety of targets such as yeasts, diatoms, erythrocytes, and human endothelial cells.

Small molecule ligand mimics have the potential to block the initial colonization of bacterial targets before growth can get established and lead to biofilm formation. For example, mannose-based antagonists that block FimH, an adhesin from *Escherichia coli*, can disrupt binding to mannose-containing glycoproteins in the uroepithelium, thereby reducing urinary tract infections

[39, 40]. This approach could be an alternative or adjuvant to antibiotics, which are rapidly losing their effectiveness [41]. However, it is becoming clear that bacterial species can have more than one RTX adhesin. One example of a species with two RTX adhesins is *Pseudomonas fluorescens*, a soil bacterium that is rarely pathogenic [42]. Here, the long adhesin protein (LapA) and the medium adhesin protein (MapA) have complementary roles in adherence and the formation of biofilms [20]. A second example is the pathogenic bacterium *Salmonella enterica*. The adhesin BapA is important for biofilm formation [43] whereas SiiE binds to glycans on epithelial cells [12]. These genes are located on pathogenicity islands that are sometimes laterally transferred or not found in all strains of a species [44], which suggests that the complement of adhesins might vary by strain within a species. Therefore, to develop strategies to block colonization by a bacterial pathogen species, it is important to know the number of adhesins they make, how they are distributed between strains, and the variety of LBDs they collectively possess.

Due to their large size and repetitive nature, RTX and other large adhesins are often poorly annotated in genome databases. Indeed, short-read assemblies cannot reliably be used for documenting long tandemly repeated proteins [45]. Here we have relied on the increasing availability of long-read genome assemblies to provide high-quality data with which to discover and characterize new adhesins, with a focus on RTX adhesins. Once large proteins are identified, InterProScan [46] is used to identify their domains using an integrated set of predictive models from multiple domain annotation databases. Although some key adhesin domains are frequently missed, BIg domains, as well as the N-terminal retention domain and the C-terminal T1SS that bracket full-length RTX adhesins are frequently found. This flags them for further analysis using AlphaFold3 where the structures of all the adhesin domains are predicted, which enormously helps with their identification. Moreover, the discovery of split domains [32] has facilitated the

recognition of LBDs, although their function and ligands often remain unknown. The application of this bioinformatic process can reliably find and sort the RTX adhesins from a bacterial species that are annotated on long-read genomic sequences. Here we have illustrated this methodology by applying it to seven bacterial pathogens of humans and other animals. This process can easily be adapted to catalogue other large proteins of interest, such as multifunctional autoprocessing repeats-in-toxin (MARTX) toxins [47] and non-ribosomal peptide synthetases (NRPS) [48].

Methods

Selection of quality genomes from which to extract RTX adhesins

A pipeline was written in Python that uses the NCBI Datasets v2 API [49] to find genomes for a user-specified species assembled using long-read technology, fetch all unique protein sequences within those genomes, cluster those proteins by shared identity, then produce a human-readable report. NCBI Datasets v2 was compiled as a Python library with OpenAPI. The first step of the pipeline was to fetch assemblies at the level of "complete genome" or "chromosome" (ignoring "contig" and "scaffold") for the given species from NCBI's Genome database. Assemblies that failed the completeness check or were noted as contaminated were excluded, as were those that failed the average nucleotide identity (ANI) test for automatic species determination [50]. In the case of redundant assemblies (GCF and GCA accession with the same ID), only the assembly from RefSeq (GCF) was retained. Next, only genomes obtained using long read sequencing technology were included by filtering the 'sequencing technology' portion of the record for any of the following keywords: ONT, Nanopore, Flongle, MinION, GridION, PromethION, PacBio, SRMT, RS II, or Sequel.

Each step of this pipeline is adjustable by the user through a human-readable config file, which is set up ahead of each run. The source code is publicly available (GitHub).

Sampling of filtered genome assemblies

In cases where more than 50 assemblies passed the filter criteria, the pipeline selected a sample of up to 50 representative assemblies. Sampling was stratified by the country of origin and

collection date of the DNA, using a fixed random seed for reproducibility. To maximize diversity, the sampling method avoided duplicates from the same stratum (combinations of country and date) unless it was necessary to reach the target sample size. In such cases, remaining slots were filled by re-sampling the remaining genomes.

Protein sequences and clustering

For each assembly in the sample, the pipeline retrieved its gene annotation package (a table of genes and their associated name, symbol, accession, and locus) from NCBI. A list of protein-coding sequences was generated by extracting the unique accession numbers from those packages. The corresponding amino acid sequences of all proteins were then fetched from NCBI as a single FASTA file.

To account for point mutations and other minor variations across samples, the pipeline clustered proteins by sequence identity using MMseqs2 [51]. Clustering was performed based on the 800 C-terminal residues with a sequence identity threshold of 80%. For each cluster, the most frequently observed sequence was selected as the representative sequence. In the event of a tie, the longest sequence was chosen. The protein and assembly accessions within each cluster were presented alongside the representative sequence for downstream analysis.

Counting of occurrences of cluster members within the genomes selected

The abundance of the protein clusters across genomes was computed by summing the number of occurrences of each cluster member within the sampled assemblies. To account for gene duplication, the number of assemblies represented by the cluster was also counted. The pipeline generated a human-readable table clearly summarizing this clustering and counting.

Identification of RTX adhesins within the output from the pipeline

Clusters containing proteins longer than 1,500 aa in length were selected for further human-guided analysis. Protein name annotations were used to exclude obvious non-adhesins. The representative sequence from candidate clusters was evaluated using the domain prediction tool InterProScan with default parameters. If domains commonly found in RTX adhesins were present, the sequence was selected for predictive modelling.

Generation of models and verification of protein clusters

AlphaFold3 was used to model the structure of the representative protein from any protein cluster suspected of being RTX adhesin using the AlphaFold3 Server [52]. If the sequence exceeded the 5,000-residue limit of the server, the protein was modelled in overlapping segments. The model was visually inspected for the presence of the retention module and β-roll/T1SS domains. The boundaries of all domains and their structural similarity to known domain types was determined from the model and used to generate schematic diagrams of the adhesins. The sequence of the LBR plus the β-roll/T1SS, beginning with the first split domain and ending at the C terminus, was used in BLASTp searches [53] against all proteins in the spreadsheet. Clusters were adjusted to include any additional sequences with > 75% identity to the LBR and to exclude those with <75% identity.

Identification of LBDs by type

Domains were identified as vWFA domains by InterProScan. Putative CBMs and PBDs were identified by strand topology as described below. β-rolls were identified by their structure and the presence of at least one nonapeptide repeat motif (Gly-Gly-x-Gly-x-Asp-x-Φ-x), where x is

any residue and Φ is a hydrophobic residue. Domains that did not resemble these three types were classified as unknown domains.

Phylogenetics

The 'Bacterial Genome Tree' service [54], which uses RAxML version 8.2.12, at the Bacterial and Viral Bioinformatics Resource Center (BV-BRC) was used to generate a phylogenetic tree for the species mentioned in the manuscript. Representative chromosome-level assembly genomes were selected, and default parameters were used except that the number of genes selected for comparison was set to 500.

Results

Part 1: Semi-automated bioinformatics pipeline eases RTX-adhesin detection

Rationale behind using the pipeline to filter genome sequences

The difficulties encountered in identifying adhesins in sequence databases led us to develop a pipeline to simplify the process (Fig. 2). Overall, over 50 adhesins were identified in the 29 Gram-negative bacterial species shown in Fig. 3, with the pipeline enabling rapid recognition and characterization of 29 RTX adhesins from seven species that are described in detail herein (Fig. 4). Briefly, the pipeline accepts a species as input, identifies genomes sequenced with long-read technology, retrieves all annotated protein sequences from those genomes, then clusters them based on sequenced identity (Fig. 2). The steps are summarized in the Methods section, and the rationale behind each stage in the process is outlined below.

The first issue that came to light was that RTX adhesins are so poorly conserved between species that they cannot be reliably detected by BLASTp searches. For example, the highest scoring hit when the first adhesin found in *Bordetella parapertussis* (*Bp*1) was compared to all adhesins in Fig. 4 was 29% identity to *Vp*4 over 2/3rds of its length. The same was sometimes observed even within a species, as the closest match to *Ab*3 is *Ab*1, with only 35% identity. Despite this, they all share some common domains and are all far longer than the average protein. Therefore, survey methods that focus on these characteristics rather than sequence similarity are more likely to yield results.

The second issue was that RTX adhesin ORFs were seldom assembled accurately due to their extreme length (approx. 4,500 to over 40,000 bp) and repetitiveness. For example, around a third of the adhesins shown in Fig. 4 contain over 20 identical repeats, ranging in size from ~250 to ~650 bp per unit, that encode the tandemly arrayed BIg-like domains of the extender region (Fig. 4). Misassembly was particularly evident within genomes assembled from short-read sequencing (e.g. Illumina), while long-reads from single-molecule real-time (SMRT) sequencing (Pacific Biosciences) or nanopore (Oxford Nanopore Technologies) usually spanned the entire gene and were properly assembled. Most submitted genomes are assembled from short reads, exemplified by *Acinetobacter baumannii* where only 833 of 40,037 genomes (2.1%) were high-quality assemblies from long reads (Table 1). The highly repetitive *A. baumannii* adhesin 3 (*Ab*3) (Fig. 4A), provides a clear example of the problem. A BLASTp search of the GenBank non-redundant (nr) protein database identified 84 obvious homologues with >98% identity to the non-repetitive C-terminal region of the protein. Of these, only five were greater than 2,000 aa in length and all of these originated from long-read assemblies. The longest (WP_269006110.1, 6,421 a.a) has 46 identical 108-aa extender domains. Of the remaining 79 homologues, just over half (43) were described as partial, and the rest are unlikely to be complete or accurate. Therefore, to maximize the chances of finding properly assembled full-length RTX adhesins, the first step of the pipeline is to filter available genome assemblies to only include those that used long-read sequencing technology and meet quality criteria.

If over 50 assemblies remain after quality filtering, they are subsampled by the pipeline for both practical and statistical reasons. Practically, protein sequences are fetched by assembly accession and processed locally. This means that for some organisms, like *Acinetobacter baumanii* which has 981 assemblies meeting our criteria, processing all of them would greatly inflate file size,

runtime, and server load to NCBI while yielding diminishing returns in analytical value. Statistically, large sequencing projects often submit many assemblies from a single strain or outbreak, causing their sequences to be overrepresented in the dataset. In *Vibrio parahaemolyticus*, 127 long-read assemblies passed our filter (Table 1) with twelve of these assemblies generated by a sequencing project investigating infection of a bivalve on the Pacific coast of Colombia [55]. These data are excellent, but to consider all these assemblies would bias the analysis toward samples collected in a specific region that infect a specific organism. Therefore, the subsample is generated using stratified random sampling of the filtered assemblies, drawing across country-and-date strata.

Clustering and filtering of proteins from the selected genomes

All annotated proteins encoded by the selected genomes are downloaded from the NCBI server by the pipeline. MMseqs2 then clustered all sequences at an 80% identity threshold based only on their 800 C-terminal residues. Clustering was limited to the C terminus because we noticed that even among closely related isoforms, the extender array varies in length. We had to consider a large enough window to compare more than the C-terminal secretory β-roll (120 to 450 aa) and after some experimentation, an 800-residue window was found to be sufficient. With all proteins clustered by their LBRs, we could then limit our search to sequences >1,500 residues, a minimal size for an RTX adhesin, thereby reducing the number of candidate sequences to screen (Table 1). The net effect was that rather than manually searching through hundreds of genome assemblies for adhesin-like sequences, a report for each species with fewer than fifty candidate sequences was quickly generated.

Identification of RTX adhesins from protein clusters and InterProScan

The prokaryotic genome annotation pipeline (PGAP) assigns algorithmically generated names for each encoded protein that can help narrow our focus [56]. These names are based on features detected in the sequence, so large well-characterized proteins such as 'non-ribosomal peptide synthetase' or 'DEAD/DEAH box helicase' can be further pruned from the dataset. For multidomain proteins such as RTX adhesins however, the name is typically generated based on just one of those domains. As such, adhesins are often assigned names like 'Ig-like domain containing protein' or 'retention module-containing protein', which is less reliable and necessitates further characterization.

The domain structure of any sequence suspected of being an RTX adhesin can be screened using using InterProScan [46], and the web-based version will accept up to 100 sequences at once. A variety of domain types are recognized within each of the four regions of a typical RTX (Fig. 1) with the types that are commonly observed listed in Table 2. Domains within each region are sometimes not recognized, but at least three different domain types were found in the adhesins described herein (Fig. 4). Detailed examples of InterProScan results are provided in Supplementary Results and shown in Supplementary Fig. 1. This screening provides a rapid means of excluding proteins that have a domain structure inconsistent with RTX adhesins and selecting those that do.

Modelling and verification of protein clusters using AlphaFold3

The representative sequence from any cluster suspected of being an adhesin was modelled using AlphaFold3 [52]. Any candidate that exceeds 5,000 aa in length (the limit imposed by the AlphaFold server at the time of writing) was modelled in segments. Generally, the domains that were not detected using InterProScan were consistently modelled in AlphaFold3. Proteins that

did not contain the canonical domains found in RTX adhesins, which are, at a minimum, the N-terminal retention module, a tandem array of Ig-like extenders, and a C-terminal β-roll with a T1SS signal, were excluded. The boundaries between individual domains were ascertained from these models and used to draw the domain maps in Fig. 4.

Within protein clusters, the number and sequence of extender domains was found to vary widely between homologues (described in more detail in Part 2), whereas the RM, LBR and β-roll/T1SS sections were usually highly conserved. However, as the LBRs of the different adhesins were often longer than the 800 aa length used to generate the protein clusters, proteins were sometimes misclassified. So, to verify that all cluster members are indeed homologous, the entire portion of the representative sequence C-terminal to the extender array (beginning with the first split domain) was used as a query in a BLASTp search against all members of the cluster, using the accession numbers provided in the spreadsheet from the pipeline. A second BLASTp was performed using the accession numbers of the representative sequences from all of the other clusters, to ensure that all homologues were included. Any isoforms that matched with >75% identity across this entire region, with the same domain architecture, were considered homologues.

Determination of the prevalence of specific adhesins

By tracking the number of times a member of a cluster is found across sampled genome assemblies, the pipeline also gives a sense of the prevalence of RTX adhesins, as well as their isoforms, in a species. For example, for *Ab*3 mentioned above, the dearth of hits is surprising, given that there are almost 1,000 long-read assemblies and over 40,000 assemblies total for *A. baumannii*. This suggests that *Ab*3 is an uncommon adhesin in this species and indeed, this

isoform was found only twice in the 50 long-read genomes examined (Supplementary Table 2). Conversely, *Ab*5 is found 14 times. It should be noted that these numbers could be underestimated, as even with long-read genomes, the chances of an assembly error leading to a frameshift increases in proportion to the length and repetitiveness of the adhesin [45]. However, if identical frameshift mutations are observed in multiple genomes, that would indicate that the adhesin has been lost in these strains.

Part 2: Adhesins from seven Gram negative bacterial species identified using the bioinformatics pipeline

General characteristics of the adhesins identified

A total of 29 RTX-adhesins, possessing both an N-terminal RM and C-terminal β-roll/T1SS, were identified in the seven species. The schematic diagrams of these adhesins shown in Fig. 4 enable comparisons between proteins. The length of the bars representing the extenders, split domain segments, and β-roll/T1SS segments are proportional to the length of these segments and serve to demonstrate that these domain types vary in molecular weight. Numerous domain types serve as 'split' domains, which project putative LBDs away from the main axis of the adhesin, the most common of which are the BIg split extenders shown in pink. Regular extenders lie between the RM and LBR and their numbers vary widely, from five in *As*1 to over 70 in *Vv*6.

The cartoon model of one of the shortest (1,681 aa) adhesins identified, *As*1, serves to illustrate the utility of the schematic diagrams (Fig. 5A). Here, both the vWFA (dark yellow) and CBM (purplish blue) can be seen to project out from the split domains. This adhesin also illustrates a frequent variation whereby proteins that have the same domain architecture and share >75%

identity throughout the LBR can have very different extender regions. Here, the short isoform (*As*1) has only five variable extenders, whereas the longest isoform (*As*1 long) has 37, and only the first two share any appreciable sequence identity. The opposite is sometime seen, as with *Ah*1 and *Ah*4 (Fig. 5B). The two adhesins share 97% sequence identity over most of their length and have the same domain architecture in the LBR, but the entire region from the last extender, through the entire LBR and partway into the β-roll was likely swapped in from another adhesin as the sequence identity drops to 28%. An example of an even smaller swap is *Lp*1 and *Lp*1a (Fig. 4B and Supplementary Table 1). These two highly similar proteins have CBMs that share only 28% identity (not shown).

Characteristics and variations of the various domain types

The RM domains of all adhesins described herein fall within a narrow length range of 127-196 aa (Fig. 1), with some variation in the number of strands within the RM and minor variation in the length of the linker. This domain was always followed by a region containing a string of extender domains, which can vary widely in number, with only five as in *As*1 to 73 as in *Vv*4. Additionally, the sequence similarity between extenders from different adhesins, and even between extenders within the same adhesin, can be so low that they are unrecognizable. Despite this, they all share an Ig-like fold consisting of a two-layer β-sandwich with an odd number of strands (most commonly seven or nine) such that the N and C termini lie at opposite ends of the domain, allowing the extenders to form a chain. The similarity of the fold is evident in two extenders at opposite extremes of the ~2-fold length variation. The 85-aa extender 7 from *Ab*5 and the 178-aa extender 5 from *Ah*1 (Fig. 4) share a pattern consisting of seven segments, coloured in rainbow hues and numbered from N to C terminus (Fig. 6A,B). Around the

circumference, the segment pattern is as follows; 1 (blue), 2 (cyan), 5 (yellow), 4 (light green), 3 (dark green), 6 (orange), and 7 (red). The main structural difference between the two are that the smaller extender domain has segments consisting of short continuous β-strands (Fig. 6A), whereas the segments in the larger extender (Fig. 6B) are sometimes split into two strands with an intervening loop or short α-helix. Additionally, the longer extender has two additional strands (dark grey). Split extender domains (Fig. 4, pink) within the LBR also share this fold (not shown).

The extenders that lie at either end of the extender region, along with those within the LBR, are usually dissimilar to other extenders in the same protein (Fig. 4, light grey). whereas those in the interior are often highly similar (Fig. 4, darker shades or colours). The most common pattern is for adjacent extenders to be similar, as in *Vp*3, where a set of 18 near-identical extenders is flanked on either side by a dissimilar extender. Another common pattern, found in *Ah5* for example, is for two dissimilar extenders to form a pair that is repeated in series. More complex patterns occur, sometimes consisting of adjacent blocks of near-identical repeats exemplified by *Ah*1, or consisting of a mixture of patterns as in *Ab*3 and *Vp*2. The other extreme is exemplified by *Vv*2, where each of the 37 extenders shares less than 80% identity with any of the others.

LBDs are generally recognized due to the presence of the split domains from which they project. The most common of these are the putative CBMs (Fig. 4, purplish blue triangles). These domains are highly variable, even within the same protein. For example, pairwise sequence identity between the six CBMs in *Ab*1 range from a low of 20% (CBM2 and CBM6) to a high of 56% (CBM1 and CBM3) (Supplementary Fig. 2). However, these domains can be recognized due to a common strand pattern (Fig. 6C-F). They have at least eight strands and around the

circumference, the segment pattern is as follows: 1 (violet), 2 (blue), 7 (orange), 4 (dark green), 5 (light green), 6 (yellow), 3(cyan), and 8 (red). They are only occasionally recognized as a PA14 domain by InterProScan.

The vWFA domains are easy to recognize as they consist of a six-stranded β-sheet with three α-helices on each face (dark yellow, Fig. 5a), and they are readily identified by domain search tools such as InterProScan. The rarer PBD domain is more difficult to recognize and to differentiate from CBM, until the stands pattern is examined. The PBD has two sets of strands in sequence in the pattern 6,7,8,9,10,5,4,3,2,1, in contrast to the less regular sequence for the CBM given above. Internal β-roll domains are also readily identified by InterProScan, in part due to the nonapeptide repeat (Gly-Gly-x-Gly-x-Asp-x-Φ-x).

Characteristics of the adhesins of each species

Acinetobacter baumannii is an aerobic coccobacillus non-flagellar bacterium that is causing increasing numbers of hospital-acquired infections [57, 58]. This includes pneumonia, as well as infection of the urinary tract, wounds, and blood. Its ability to rapidly acquire antibiotic resistance and to form biofilms complicates treatment and leads to a high mortality rate [59].

The adhesins of *A. baumannii* were characterized in a 2024 study where they were grouped into four types based upon their extender domains [60]. Here, we have taken an alternative approach by grouping them based upon their C-terminal LBRs. By this method, we have found adhesins with six distinct architectures (Fig. 4A). Both *Ab*1 and *Ab*2 were classified as the same type in the previous study as some of their extenders are identical. However, their LBRs share less than 50% identity, with *Ab*1 having six putative CBMs while *Ab*2 has only four. The highest similarity

between any two of these CBMs is 60% (CBM1 of *Ab*1 to CBM1 of *Ab*2). A third type (*Ab*3) has three CBMs, shares little sequence similarity with the others outside of the RM, and one of the CBMs acts as a split domain for a domain predicted to contain a bundle of α-helices (Fig. 6D). Given this, it is possible that these LBRs have different ligand specificities. Additionally, these adhesins have more putative CBMs than found in any of the other adhesins we have characterized (Fig. 4).

A fourth domain architecture is found in *Ab*4, where a single vWFA domain buds off a split β-roll domain. This is a common architecture in RTX-adhesins (Fig. 4), but it is typically found in combination with other LBDs. The fifth LBR architecture resembles the fourth, except instead of a vWFA domain, an unknown domain consisting of two long antiparallel strands tipped by a small bundle of three strands and one α-helix is found instead (not shown). The sixth LBR architecture is unique. It has a simple CBM lacking any additional strands that is projected well away from the protein by a split domain in combination with two long β-strands (Fig. 6E). This is followed by four apparent β-roll/T1SS domains in parallel (Fig. 4a).

Aeromonas hydrophila is a Gram-negative, motile, rod-shaped bacterium prevalent in fresh and brackish water [61]. It causes disease in humans, fish, and other aquatic animals. In humans, it is linked to wound infections, bacteremia, and primarily gastroenteritis [62-65]. *A. hydrophila* is a well-known pathogen in fish – including tilapia, perch, catfish, salmon, bass, and carp – causing diseases such as aeromonad septicemia and ulcerative infections [38, 66]. It is also implicated in infections in other aquatic animals, such as red leg syndrome in bullfrogs [67]. These infections contribute to significant economic losses in aquaculture [68].

We have found five distinct RTX adhesins in this species (Fig. 4A), one of which (*Ah*3) entirely lacks an LBR. Two others, *Ah*1 and *Ah*4, have structurally similar LBRs that each contain a vWFA-split domain combination followed by one extender domain and a putative CBM-split domain combination, yet they share only 36% sequence identity. The second RTX adhesin has been previously partially characterized as *Ah*Lap [8]. It has a CBM-split domain combination that is followed soon after by an RTX domain projected from a split extender domain and then a vWFA domain that emerges from the main RTX domain. The CBM of this adhesin has been shown by glycan-array analysis to bind Lewis B and Y antigens. The fifth *(Ah*5) is the shortest one and has two LBDs, a putative CBM and a vWFA domain that has an insert that resembles a CBM in its β-strand topology.

Aeromonas salmonicida is a Gram-negative, non-motile bacterium and one of the earliest identified fish pathogens, affecting wild and farmed fish in freshwater, brackish, and marine environments [69, 70]. It causes furunculosis – a septicemic disease of salmonids – characterized by skin ulcers and gill and fin lesions [71, 72]. Despite current control measures using vaccines and antibiotics, *A. salmonicida* remains one of the most destructive pathogens in salmonid aquaculture. Rising antibiotic resistance highlights the need for alternative treatment strategies [69, 73, 74]. There are four distinct RTX adhesins in this species that have different distributions of LBDs that share similarity to those found in *Aeromonas hydrophila* (Fig. 4A). *As*1 appears to be a condensed version of *Ah*1 and *Ah*4, with one fewer extender between the two LBDs. The LBR of *As*2 resembles that of *Ah*5, and the LBR of *As*3 is like that of *Ah*Lap. The LBR of the longest adhesin, *As*4, is also like that of *Ah*Lap but without the CBM domain.

Bordetella parapertussis (Bp), a rod-shaped, Gram-negative bacterium, is a homologue of *B. bronchiseptica* and a close relative of *B. pertussis*. While *B. parapertussis* and *B. pertussis* are

433 not direct homologues of each other, having evolved independently from *B. bronchiseptica*, both
434 serve as causative agents of whooping cough in humans, more notably in infants [75, 76]. Whooping
435 cough is a severe, acute respiratory illness known for being highly contagious due to its ease of
436 transmission through aerosol droplets and is associated with complications such as pneumonia [77].
437 Unlike *B. pertussis*, immunization against *B. parapertussis* has not yet been developed [78].
438 Consequently, recent reports show a significant increase in *B. parapertussis* incidence,
439 accounting for 95% of detected *Bordetella* spp. infections [79], highlighting the need for inhibitor
440 development. An RTX adhesin has been reported in *B. parapertussis* as being involved in biofilm
441 formation and attachment to respiratory epithelial cells [80]. Here we show that this RTX adhesin is
442 the only one in this species. It has a domain with a CBM-like fold and an LBR that resembles
443 those of *Ah*5 and *As*3 except with two extender domains between the CBM- and vWFA-split
444 domain pairs (Fig. 4B).

445 The bacterium *Legionella pneumophila (Lp)*, a waterborne, Gram-negative rod, is the causative
446 agent for Legionnaires' disease, which is a severe often fatal pneumonia [81-83]. *L. pneumophila* is
447 also known to cause Pontiac fever, a milder, self-limiting flu-like illness [84]. Infection occurs
448 through inhalation of aerosolized contaminated water sources and poses high risks to
449 immunocompromised individuals and those exposed in healthcare settings [83, 85, 86]. Humans are
450 considered dead-end hosts in the transmission cycle of *L. pneumophila* [83]. Here we show that
451 there are two main adhesin isoforms in this species (Fig. 4B), occurring with similar frequency
452 (56% vs 44%, Supplementary Table 1). They reside at a single gene locus, so each isolate has but
453 one copy. Both have a vWFA domain, but in *Lp*1 it is set off by an internal RTX split domain,
454 whereas in *Lp*2 it protrudes from a split extender domain. These two vWFA domains share less
455 than 30% identity. The only CBM is found in *Lp*1, and there are two different subforms (*Lp*1 and

*Lp*1a, Supplementary Table 1) where the CBMs share only 28% identity while the rest of the LBR is highly similar. The unknown domain (green, Fig. 4) is a 12-membered helical bundle that is not found in any other adhesin identified herein. Unlike the other domains within the LBR, this domain is >92% identical in the proteins from all 50 genomes.

V. parahaemolyticus is a Gram-negative, autochthonous bacterium widely distributed in temperate and tropical marine and coastal waters [87]. It is a major foodborne pathogen, typically causing diarrhea following the consumption of raw or undercooked seafood [88, 89]. We recognize five different RTX adhesins in this species (Fig. 4B), one of which (*Vp*1) lacks an LBR and resembles *Ah*3 in having only extender domains in its interior. *Vp*2 has three LBDs in its LBR, and its domain architecture resembles both *Ah*Lap and *As*3. *Vp*3 has only two LBDs. The first is a putative CBM-split domain combination followed by one extender domain and a putative CBM-split domain combination followed by one extender domain and then a vWFA-split domain combination. *Vp*4 is the shortest of the *Vp* RTX adhesins and has only one putative CBM-split domain combination while the unique feature of *Vp*5 is that it is the only adhesin herein that has two vWFA domains.

Like *Vp*, *Vibrio vulnificus (Vv)* is an aquatic, Gram-negative bacterium from the Vibrionaceae family and is a leading cause of seafood-related fatalities. Additionally, it is a major contributor to disease in aquaculture systems [90, 91]. *Vv* infections can quickly disseminate into the blood through wound or intestinal epithelium infections, causing the rapid onset of life-threatening symptoms in humans - like necrotizing fasciitis and septicemia [91, 92]. Yet, there are currently no studies that have elucidated the full complement of RTX-adhesin isoforms found in this pathogen. Using this bioinformatic pipeline, six unique adhesin isoforms were identified (Figure 4). *Vv*2 resembles *Ah*3 and *Vp*1 in lacking an LBR. *Vv*1, 3, 4 and 6 contain a similar order and

pattern of putative LBD-split domain combinations, including a CBM, an internal β-roll, and a vWFA domain, two of which have insertion sequences. *Vv*4 and *Vv4a* (Supplementary Table 1) are another example where most of the LBR is conserved (Fig. 4B, dashed lines) with an intervening portion with the same domain types (no underline) that is quite distinct (32% identity). The rare *Vv*5 isoform contains a completely different C-terminal LBR in which there is a putative peptide-binding domain [36] and a putative CBM separated by two extender domains. This is a similar pattern of LBDs to that seen in the FrhA RTX-adhesin of *Vibrio cholerae* and one RTX adhesin of *Aeromonas veronii* [37]. These six isoforms occupy two loci, with the highly conserved extender-only *Vv*2 (>90% identity) found intact in 29 of 30 genomes. The 30th genome has a large deletion that includes the *Vv*2 gene. The other five isoforms were found at the second locus, indicating that most isolates have two adhesins – a highly conserved adhesin with extenders only, and a variable adhesin with an LBR.

Discussion

By only considering genomes sequenced with long-read technology, applying strict inclusion criteria to those assemblies, and employing a sampling strategy that prioritizes diversity, we have found full-length sequences for the most common RTX adhesins in each bacterial species considered. Furthermore, we have identified and annotated the domains responsible for adhesion. Linking these adhesins to strains of those species may help explain the phenotype differences between strains and even permit the tailored design of strain-specific blocking strategies. Blocking strategies will most likely target the LBDs within the LBR of the adhesin because these domains are responsible for attachment to the bacteria's host. LBRs are also the most variable parts of the adhesin. Thus one might expect antibodies raised to the LBR will only target the one species or quite likely even the one strain.

The modularity of RTX adhesins makes their detection and classification a complicated task. A minimal RTX adhesin would comprise an N-terminal retention domain, a flexible linker that can span the T1 secretion channel, one or more tandemly repeated Ig-like domains, and a C-terminal β-roll/T1SS recognition domain. While detecting the members of this adhesin family, we have noted the degree to which each of these elements is prone to variation. Indeed, at the domain level, there is a “mix-and-match" appearance to the adhesins as though they have been assembled from a choice of retention domain variants, different types and quantities of Ig-like extenders, and a selection of LBDs budding off from different “split” extenders that are present in different arrangements and positions in the LBR (Fig. 4). Even the C-terminal T1SS recognition domain has variable length, exemplified by the very short (135-aa) version in *Ab1* and the much longer one (317-aa) version in *Vv*5. This mix and match quality might reflect DNA shuffling under selection pressure for the host bacteria to colonize different niches.

At a sequence level, we notice very low pairwise identity between like domains, not only across adhesins, but within adhesins. For example, in *Ab*1, the tandem array of Ig-like extenders is a mix of at least four different versions, two of which are shown (Fig. 5a). Poor conservation here supports the hypothesis that these play a simple structural role of projecting the binding domains away from the bacteria and toward their target. If these domains fold correctly, it is largely unimportant which residues they are comprised of. In the LBR of *Ab*1, six putative CBMs are present and the highest match in a pairwise identity matrix was 56%. Here the lack of conservation might support a different hypothesis – that these Ig-like binding domains mutate adaptively to bind different targets.

It was mentioned that only one RTX adhesin was found in *B. parapertussis*, which is unique among the species considered. At first we thought this was indicative of an error in the pipeline, but closer inspection revealed that *B. parapertussis* has astonishingly low diversity of proteins throughout the entire species. Fifty sampled *Bp* genomes contained an average of 4,221 protein-coding genes and across the entire sample, there were only 4,652 unique sequences. By comparison, *V. parahaemolyticus* with an average of 4,756 protein-coding genes per genome had 83,972 unique sequences across fifty genomes (Table 1). This makes it likely that an inhibitor against the solitary RTX adhesin would be effective against all cases of this human pathogen.

Despite the extreme conservation of *Bp* genes, the diversity of proteins in *Vp* is typical of the species considered. This diversity highlights a key challenge in sequence-based protein annotation: standard tools may fail to recognize domains that diverge sufficiently. Indeed, attempts to use InterProScan offer mixed results, occasionally missing retention domains and extenders, and almost entirely overlooking the putative CBMs. Custom-trained HMMs, structure-based methods applied to predicted models, or some combination thereof could be developed to better annotate these features. Improved detection will be an important first step toward functional studies, which are needed to understand to what degree the sequence variation in binding domains affects their affinity for ligands. More reliable identification of adhesins and annotation of their binding domains will therefore be essential in developing blocking strategies.

RTX-adhesins show evidence of DNA uptake and incorporation, DNA swapping, and gene duplication and divergence.

There is considerable plasticity in the complement and structure of RTX adhesins present in these bacterial species consistent with DNA shuffling. Bacteria are adept at incorporating DNA

that provides them with a selective advantage, which in this case would presumably be for the colonization and/or infection of various hosts. As the RTX adhesins were described in the Results section it was convenient to draw parallels between similar LBRs in different species and strains. For example, the combination of a peptide-binding domain and a CBM separated by two extender domains that is seen here in *Vv*6 is reproduced in *V. cholerae* and *A. veronii* [32]. The domain arrangement in the LBR of *Ah*Lap where WFA, RTX and putative CBM domains are present in that order is also seen in the LBR of *Vp*2 and closely resembles that present in four different *Vv* RTX adhesins (*Vv*1, *Vv*3, *Vv*4, and *Vv*5). A third example might be the tandemly arrayed putative CBM-split domain modules seen in some *Ab* RTX adhesins (*Ab*1, *Ab*2, and *Ab*3) that are similarly arranged in a *Citrobacter amalonaticus* RTX.

RTX adhesins without an LBR

Three of the seven bacterial species (*Ah*, *Vp*, *Vv*) featured in Fig. 4 contain single RTX adhesins (*Ah*3, *Vp*1, *Vv*2) that completely lack LBDs. A fourth species (*Ab*) has a small insertion budding from the C-terminal RTX domain of adhesin *Ab*5 but is otherwise devoid of any LBR elements. These LBR-lacking RTX adhesins are not rare but are well represented in the genomes of these species and are typically paired with an LBR-containing RTX adhesin in the same genome (Supplementary Table 1). For example, *Vv*2 is found in the same strains as *Vv*1. This situation is reminiscent of RTX adhesins in the classic strain of *V. cholerae* where FrhA is the RTX adhesin with a LBR containing both a peptide-binding domain and a CBM separated by two extender domains, and CraA is an RTX adhesin that completely lacks LBDs. The latter has been postulated to have a role in biofilm formation [11].

Generalized used of the bioinformatic pipeline

The bioinformatic pipeline used here can be readily adapted to find other large proteins like MARTX toxins [22, 93, 94] and NRPS [95, 96]. Indeed, these two protein types are frequently seen in the outputs alongside RTX adhesins in the >1,500 aa size category and feature a similar mix-and-match architecture. To this end, we simplified the installation and user configuration so that various protein families can be targeted. Basing the pipeline on long-read genomic sequencing makes it ideal for finding proteins with repetitive sequences.

Acknowledgements

This work was funded by a GlycoNet Collaborative Team Grant to PLD, and SG, and an NSERC Discovery Grant to PLD,

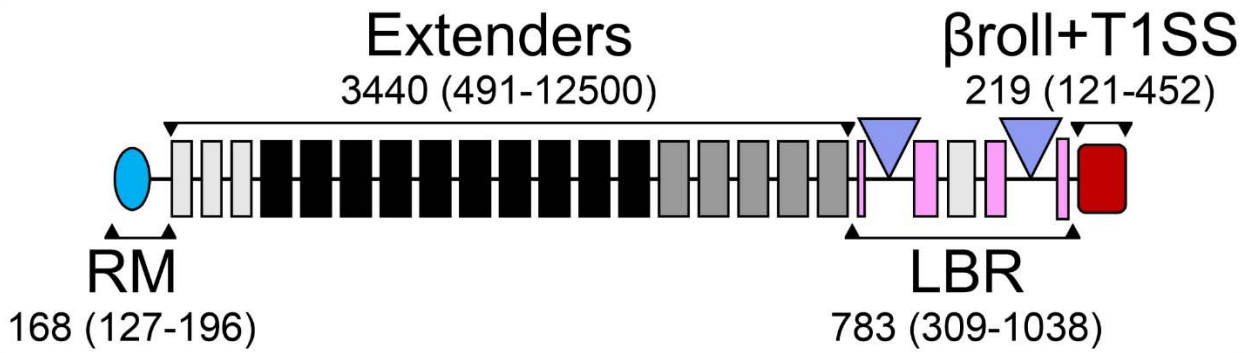


Figure 1. **Schematic showing the general characteristics of RTX adhesins.** The four sections common to most RTX adhesins are indicated with horizontal brackets and include the retention module (RM) at the N terminus, a string of extenders, the ligand-binding region (LBR) and the C-terminal β-roll and type I secretion signal (T1SS). The numbers indicate the average length of these segments in residues from a sample of 50 adhesins, with the numbers in brackets indicating the range.

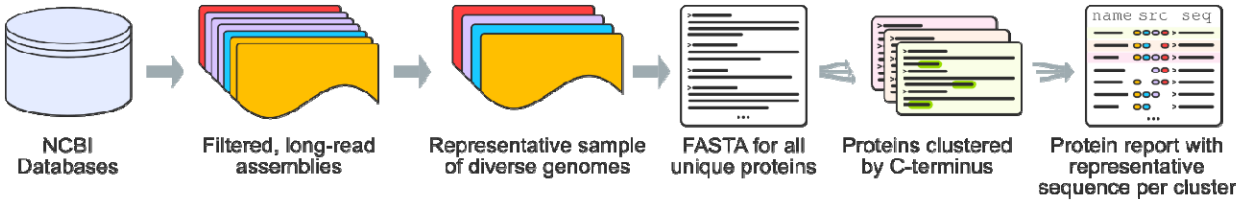


Figure 2. Flow of the bioinformatic pipeline used to identify unique RTX adhesin sequences for a given species. NCBI genomic databases are queried by species, and genome assemblies are then filtered for inclusion by sequencing technology, completeness, best average nucleotide identity (ANI) species match, and lack of notes for the assembly. A subset is taken from those filtered genomes to prevent over-sampling in cases where multiple assemblies are submitted from the same project, as well as to limit the size of the data for practical purposes. Clustering is performed on a specified stretch of residues with MMseqs2 to accommodate minor mutations across samples. Finally, a report is generated with per-cluster statistics, genome origin, and representative sequence indicating the most frequent isoform.

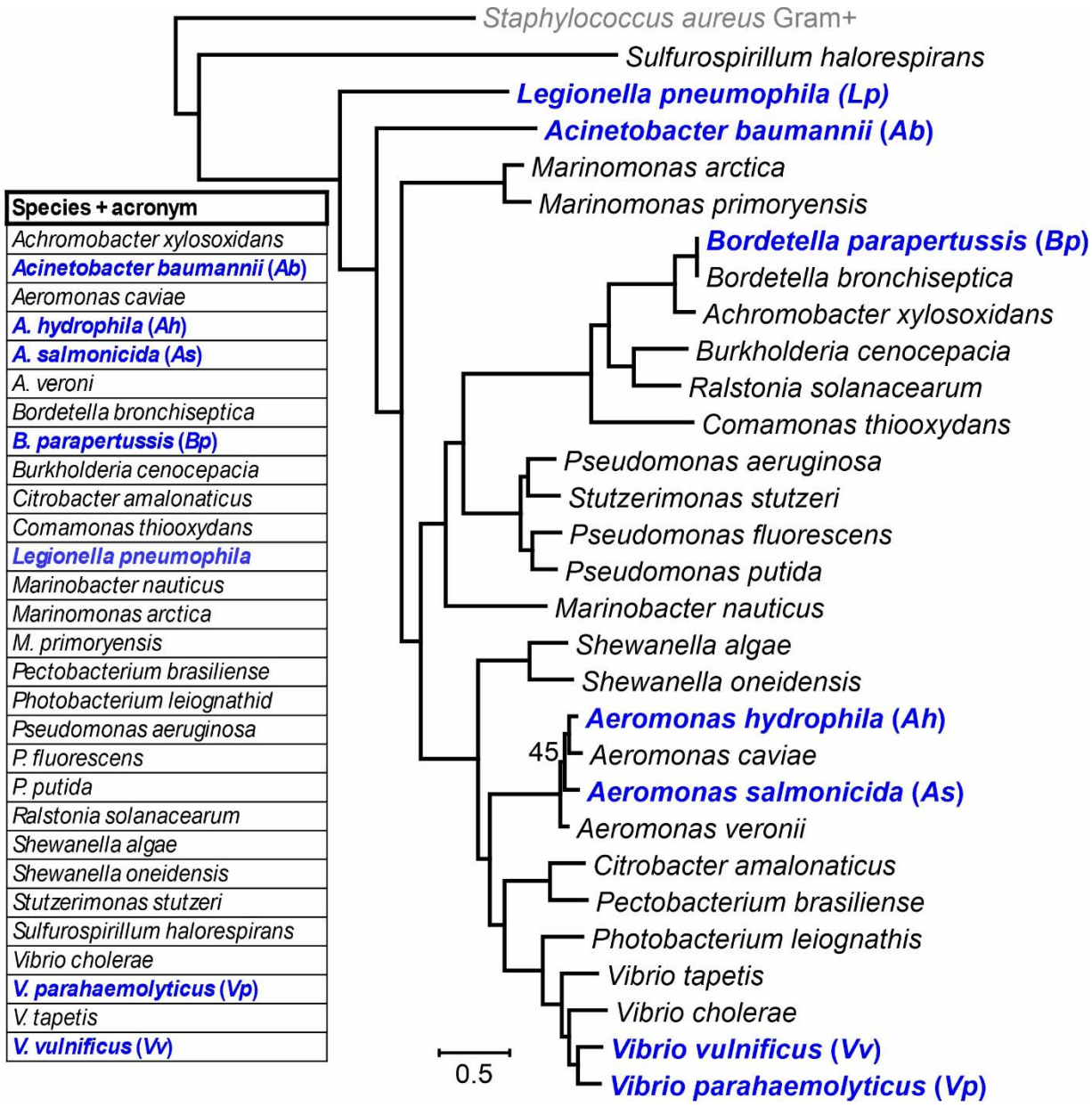
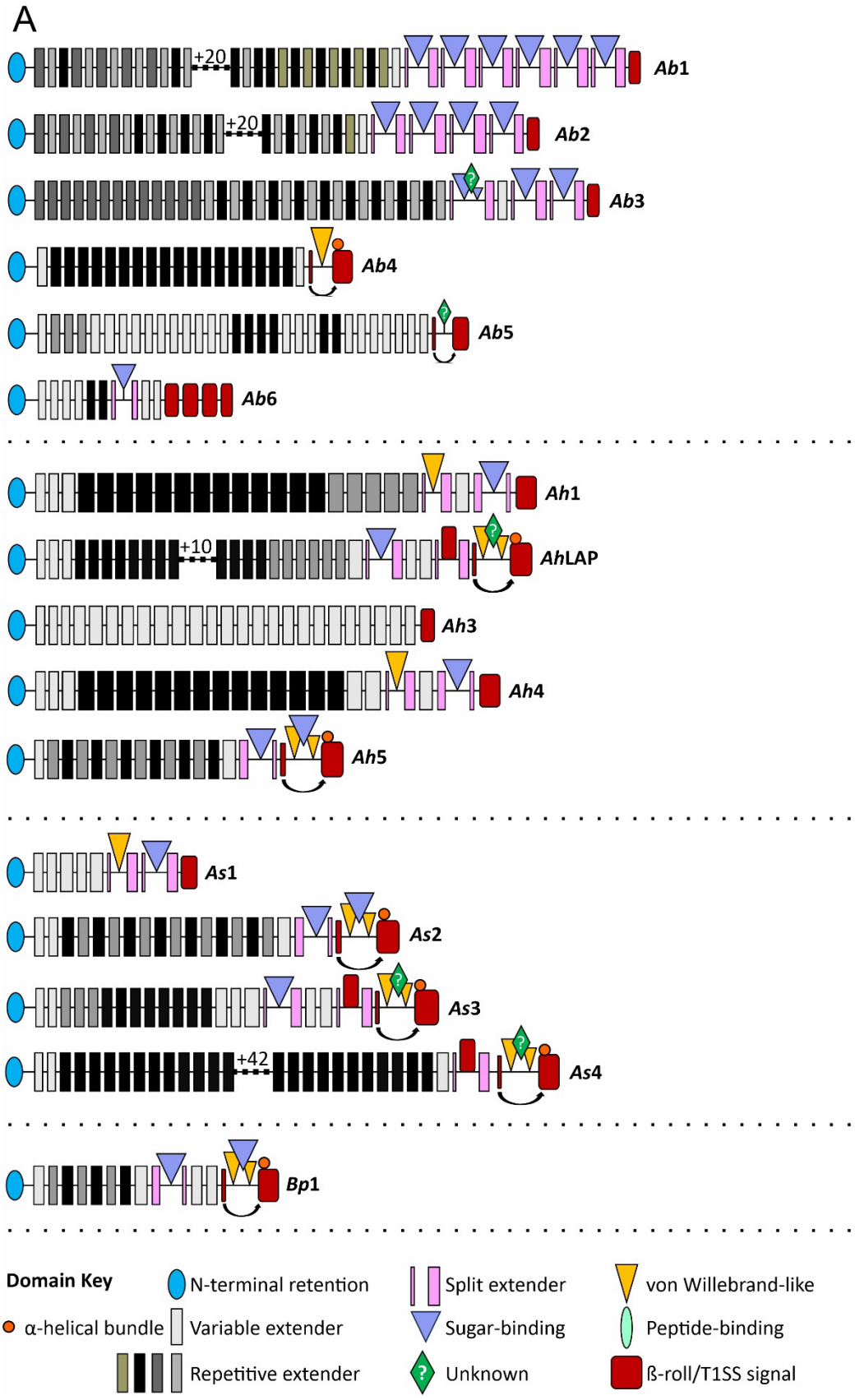


Figure 3. **Gram-negative species (Kingdom Pseudomonadati) found to contain adhesins using the pipeline.** Those whose adhesins are described in detail in this study indicated in bolded blue font. The species are indicated in alphabetic order in the table whereas the tree shows the phylogenetic relationships between the species, based upon 198 single copy genes (of 500 searched). Nodes did not change across all 100 bootstrap trials, excepting one within the *Aeromonas* grouping, so this is the only value shown. The tree was rooted using the Gram-

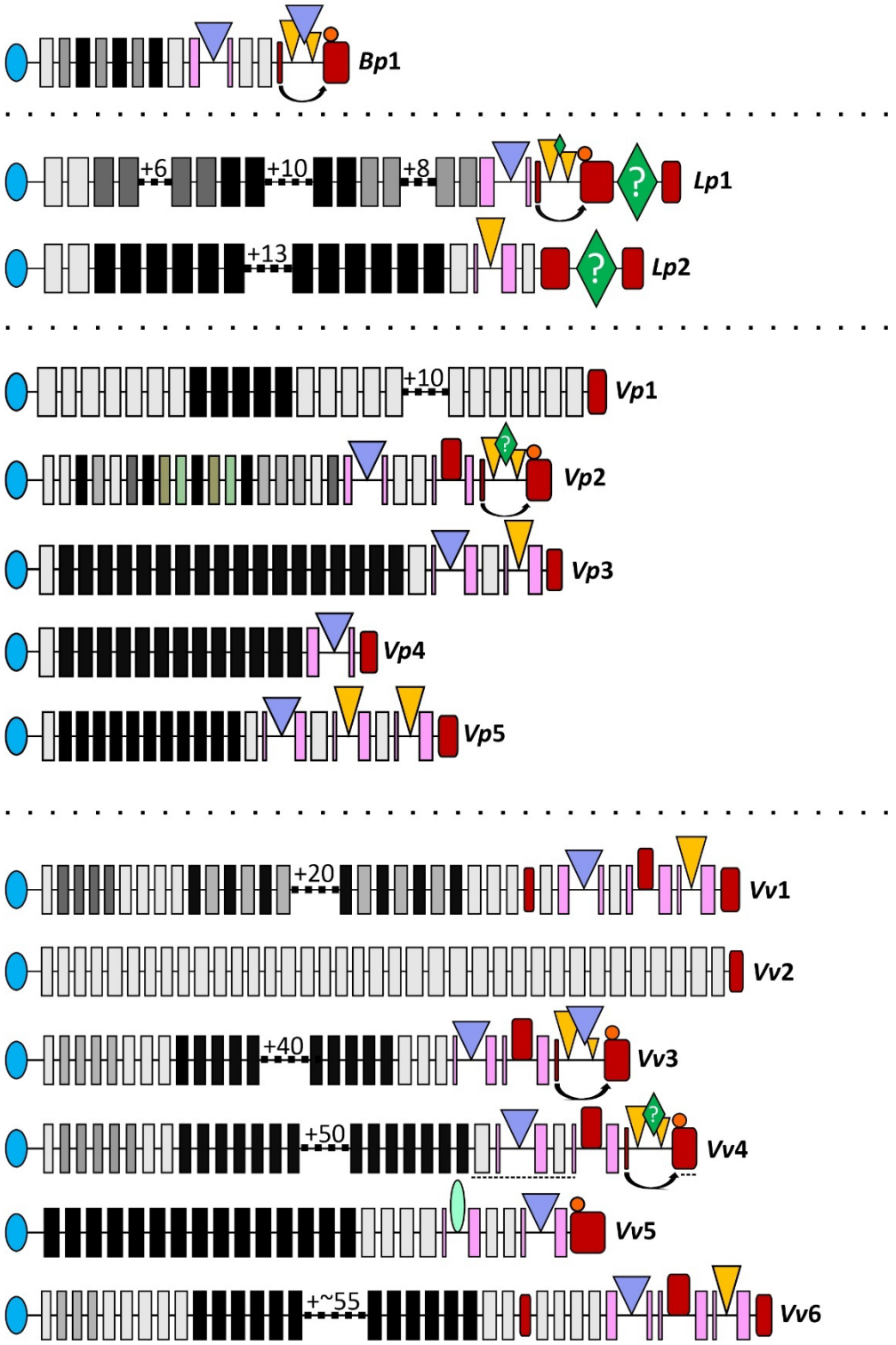
positive species *Staphylococcus aureus*. The scale bar indicates a phylogenetic distance of 0.5 nucleotide substitutions per site.


A
+20
Ab1
+20
Ab2
Ab3
Ab4
Ab5
Ab6
Ah1
+10
AhLAP
Ah3
Ah4
Ah5
As1
As2
As3
+42
As4
Bp1
Domain Key
N-terminal retention
Split extender
von Willebrand-like
α-helical bundle
Variable extender
Sugar-binding
Peptide-binding
Repetitive extender
Unknown
ß-roll/T1SS signal

608

609

B


Bp1
+6
+10
+8
?
Lp1
+13
?
Lp2
+10
Vp1
?
Vp2
Vp3
Vp4
Vp5
+20
Vv1
Vv2
+40
Vv3
+50
?
Vv4
Vv5
+~55
Vv6

610

Figure 4. **Domain maps for the adhesins found within the six bacterial species indicated in Figure 1.** Extenders are indicated with rectangles, where those in the lightest shade of grey are less than 80% identical to any other extender within the same protein. Darker shades of grey or black, as well as muted shades of green, are used to indicate groups of extenders that share at least 80% identity. A gap with a number indicates the number of extenders that have been omitted, which continue whatever pattern the extenders on either side show. The split extenders (pink) have a narrow bar at the left of the LBD if this projects after the first strand of the domain, or on the right if it projects just prior to the last two strands of the extender domain. The β-roll, vWFA and CBM domains can also act as split domains, indicated either by an arrow between the two segments (β-roll), or by two adjacent triangles with another projection (vWFA or CBM). The accession number of each representative sequence is given in Table 3. The dashed lines in *Vv*4 indicate portions of the LBR that are highly similar between *Vv*4 and *Vv*4a (Supplementary Table 1).

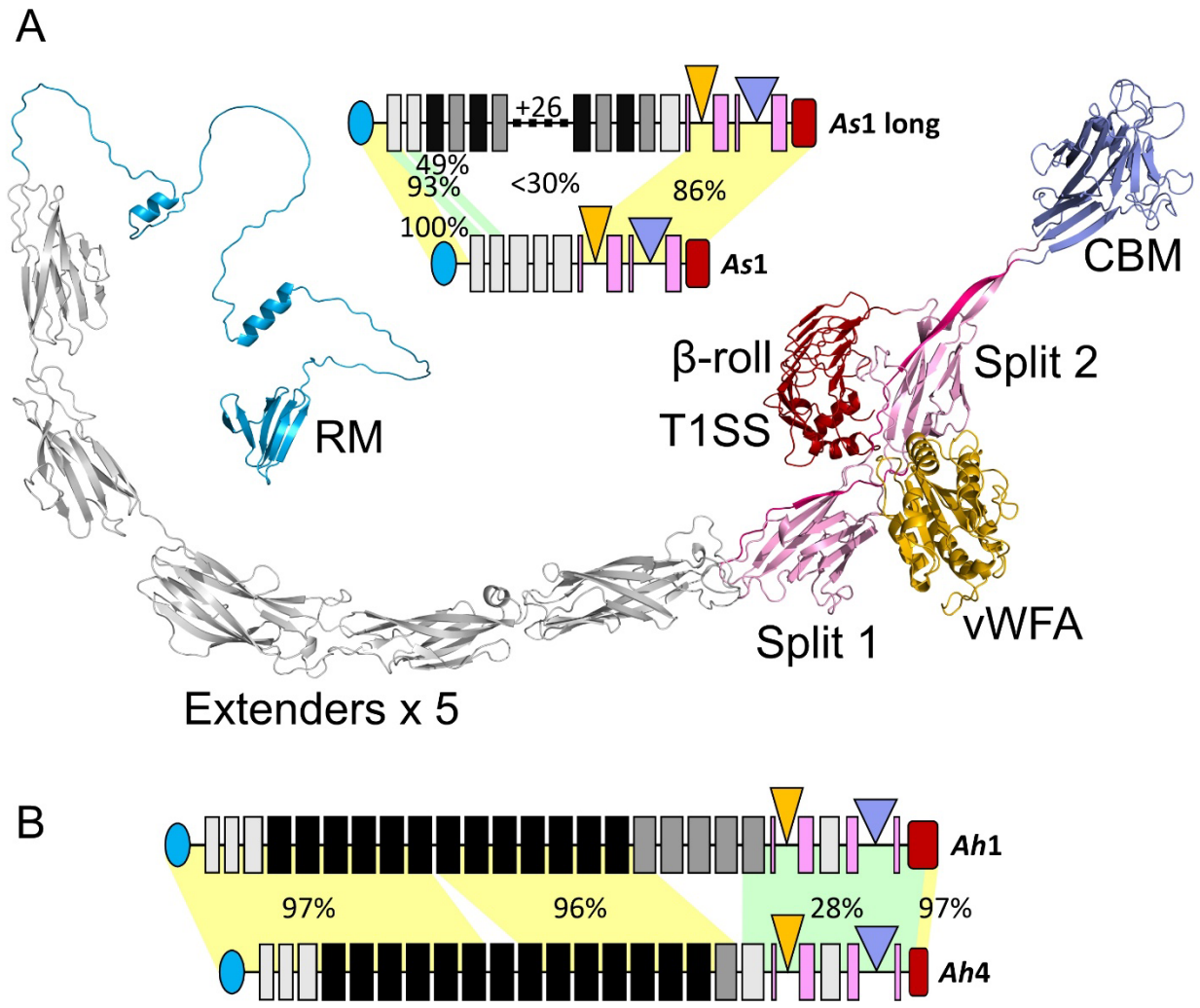


Figure 5. Comparison between structural and schematic depictions of adhesins and different patterns of sequence conservation. A) Cartoon depiction of the AlphaFold3 model of *As*1 with domains coloured as in Fig. 4, except that the first segment of the split domains are darker pink. The schematic of this short isoform is included, along with a schematic of the longest isoform of this type. The percent identity between sections of the two isoforms are indicated atop pale shading indicating the shared segments. **B)** Schematics of two *Ah* adhesins with percent identity indicated as in A.

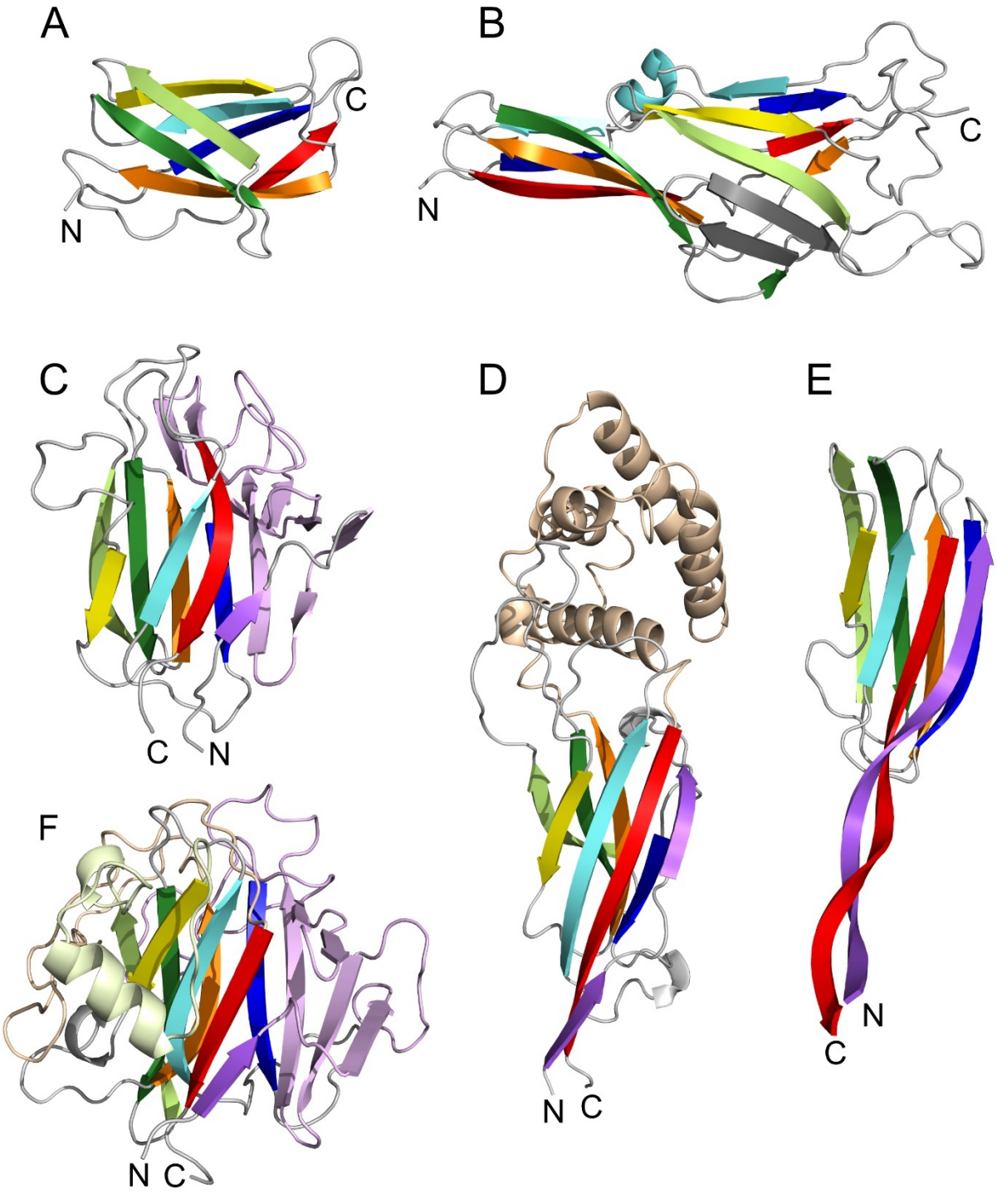


Figure 6. **Shared features and differences between the models of representative domains identified as extenders and CBMs.** A core set of seven strands are coloured in a rainbow pattern (blue, cyan, dark green, light green, yellow, orange, red) from N to C terminus for both

domain types. The CBMs have eight strands and the additional N-terminal strand is coloured violet. A) *Ab*5 extender 7 B) *Ah*1 extender 5 C) *Ah*Lap CBM D) *Ab*3 CBM E) *Ab*6 CBM F) *Ah*4 CBM. Extra strands in the extender are coloured dark grey and in the CMBs they are coloured in a pale shade of the strand they follow, such as pale orange for the unknown helical domain in *Ab*3.

Table 1. Summary of bioinformatics analysis of selected bacterial cpecies

Species acronym	*Ab*	*Ah*	*As*	*Bp*	*Lp*	*Vp*	*Vv*
Genome Categories							
Sequenced genomes	40,037	596	490	1,610	8,153	12,918	2,591
Complete or chromosome[1]	1,048	100	58	92	185	225	40
Long read + filters[2]	833	43	36	90	114	127	30
Number sampled	50	43	36	50	50	50	30
Protein Categories							
Total Unique Proteins	34,220	84,343	46,459	4,652	32,171	83,972	60,743
Clusters ≥ 800 a.a.	358	567	409	118	264	579	463
Clusters > 1500 a.a.	66	64	39	13	32	67	57
Unique RTX-adhesins[3]	6	5	4	1	2	5	6

[1] Excludes assemblies that are at the contig or scaffold assembly level.

[2] Excludes genomes from species that did not meet the Average Nucleotide Identity criteria (ANI) of 96□% identity over 90□% of the genome (PMID: 29792589) or that had warnings about completeness or accuracy. Includes genomes sequenced with either PacBio or Oxford Nanopore systems alone or in combination with short-read systems.

[3] Adhesins whose LBRs (from the start of the first split extender to the C-terminus) have the same architecture.

Table 2. Domains characteristic of RTX adhesins detected by InterProScan

Domain function	Automatic annotation name
Retention module	retention_LapA, BapA_N,
Extenders	Big_#, Ca_tandemer, Cadherin_#, CalX-beta, CalX-like, Ig-like_fold, LapA, tand_rpt_95, T1SS_rpt_143, VCBS_repeat, etc.
Ligand-binding	PA14 (rarely), VWF_A, vWFA, vWA-like, beta-Roll
β-roll	HemolysinCabind, beta-Roll, Serralysin-like_metalloprot_C
T1SS	T1SS_VCA0849, Peptidase_M10_C

References

1. Davies, P.L. & Guo, S. Structures of Fibrillar Adhesins That Mediate Bacteria–Host Interactions and Biofilm Formation, in *Biofilm Matrix*. (ed. C. Reichhardt) 153–185 (Springer Nature Switzerland, Cham; 2024).
2. Monzon, V., Lafita, A. & Bateman, A. Discovery of fibrillar adhesins across bacterial species. *BMC Genomics* **22**, 550 (2021).
3. Monzon, V. & Bateman, A. Large-Scale Discovery of Microbial Fibrillar Adhesins and Identification of Novel Members of Adhesive Domain Families. *J Bacteriol* **204**, e0010722 (2022).
4. Lofling, J., Vimberg, V., Battig, P. & Henriques-Normark, B. Cellular interactions by LPxTG-anchored pneumococcal adhesins and their streptococcal homologues. *Cell Microbiol* **13**, 186–197 (2011).
5. Melia, C.E. *et al.* Architecture of cell-cell junctions in situ reveals a mechanism for bacterial biofilm inhibition. *Proc Natl Acad Sci U S A* **118** (2021).
6. Reichhardt, C. The Pseudomonas aeruginosa Biofilm Matrix Protein CdrA Has Similarities to Other Fibrillar Adhesin Proteins. *J Bacteriol* **205**, e0001923 (2023).
7. Lloyd, C.J. *et al.* A peptide-binding domain shared with an Antarctic bacterium facilitates Vibrio cholerae human cell binding and intestinal colonization. *Proc Natl Acad Sci U S A* **120**, e2308238120 (2023).
8. Ye, Q. *et al.* Aeromonas hydrophila RTX adhesin has three ligand-binding domains that give the bacterium the potential to adhere to and aggregate a wide variety of cell types. *mBio* **16**, e0315824 (2025).
9. El-Kirat-Chatel, S., Boyd, C.D., O'Toole, G.A. & Dufrene, Y.F. Single-molecule analysis of Pseudomonas fluorescens footprints. *ACS Nano* **8**, 1690–1698 (2014).
10. Guo, S., Garnham, C.P., Whitney, J.C., Graham, L.A. & Davies, P.L. Re-evaluation of a bacterial antifreeze protein as an adhesin with ice-binding activity. *PloS one* **7**, e48805 (2012).
11. Kitts, G. *et al.* A Conserved Regulatory Circuit Controls Large Adhesins in Vibrio cholerae. *mBio* **10** (2019).
12. Wagner, C., Barlag, B., Gerlach, R.G., Deiwick, J. & Hensel, M. The Salmonella enterica giant adhesin SiiE binds to polarized epithelial cells in a lectin-like manner. *Cell Microbiol* **16**, 962–975 (2014).
13. Syed, K.A. *et al.* The Vibrio cholerae flagellar regulatory hierarchy controls expression of virulence factors. *J Bacteriol* **191**, 6555–6570 (2009).
14. Giglio, K.M., Fong, J.C., Yildiz, F.H. & Sondermann, H. Structural basis for biofilm formation via the Vibrio cholerae matrix protein RbmA. *J Bacteriol* **195**, 3277–3286 (2013).
15. Wu, C. *et al.* Oxygen promotes biofilm formation of Shewanella putrefaciens CN32 through a diguanylate cyclase and an adhesin. *Sci Rep* **3**, 1945 (2013).
16. Boyd, C.D. *et al.* Structural features of the Pseudomonas fluorescens biofilm adhesin LapA required for LapG-dependent cleavage, biofilm formation, and cell surface localization. *J Bacteriol* **196**, 2775–2788 (2014).
17. Yoshida, K., Toyofuku, M., Obana, N. & Nomura, N. Biofilm formation by Paracoccus denitrificans requires a type I secretion system-dependent adhesin BapA. *FEMS Microbiol Lett* **364** (2017).

18. Reichhardt, C., Wong, C., Passos da Silva, D., Wozniak, D.J. & Parsek, M.R. CdrA Interactions within the Pseudomonas aeruginosa Biofilm Matrix Safeguard It from Proteolysis and Promote Cellular Packing. *mBio* **9** (2018).
19. Kaus, K. *et al.* The 1.9 A crystal structure of the extracellular matrix protein Bap1 from Vibrio cholerae provides insights into bacterial biofilm adhesion. *J Biol Chem* **294**, 14499–14511 (2019).
20. Collins, A.J., Pastora, A.B., Smith, T.J. & O'Toole, G.A. MapA, a Second Large RTX Adhesin Conserved across the Pseudomonads, Contributes to Biofilm Formation by Pseudomonas fluorescens. *J Bacteriol* **202** (2020).
21. Huang, X. *et al.* Vibrio cholerae biofilms use modular adhesins with glycan-targeting and nonspecific surface binding domains for colonization. *Nat Commun* **14**, 2104 (2023).
22. Satchell, K.J. Structure and function of MARTX toxins and other large repetitive RTX proteins. *Annual review of microbiology* **65**, 71–90 (2011).
23. Guo, S., Vance, T.D.R., Stevens, C.A., Voets, I.K. & Davies, P.L. RTX Adhesins are Key Bacterial Surface Megaproteins in the Formation of Biofilms. *Trends Microbiol* **27**, 470 (2019).
24. Spitz, O. *et al.* Type I Secretion Systems-One Mechanism for All? *Microbiol Spectr* **7** (2019).
25. Baumann, U., Wu, S., Flaherty, K.M. & McKay, D.B. Three-dimensional structure of the alkaline protease of Pseudomonas aeruginosa: a two-domain protein with a calcium binding parallel beta roll motif. *Embo J.* **12**, 3357–3364 (1993).
26. Bumba, L. *et al.* Calcium-Driven Folding of RTX Domain beta-Rolls Ratchets Translocation of RTX Proteins through Type I Secretion Ducts. *Mol Cell* **62**, 47–62 (2016).
27. Smith, T.J., Font, M.E., Kelly, C.M., Sondermann, H. & O'Toole, G.A. An N-Terminal Retention Module Anchors the Giant Adhesin LapA of Pseudomonas fluorescens at the Cell Surface: a Novel Subfamily of Type I Secretion Systems. *J Bacteriol* **200** (2018).
28. Guo, S. *et al.* Conserved structural features anchor biofilm-associated RTX-adhesins to the outer membrane of bacteria. *FEBS J* **285**, 1812–1826 (2018).
29. Griessl, M.H. *et al.* Structural insight into the giant Ca(2)(+)-binding adhesin SiiE: implications for the adhesion of Salmonella enterica to polarized epithelial cells. *Structure* **21**, 741–752 (2013).
30. Guo, S. *et al.* Role of Ca(2)(+) in folding the tandem beta-sandwich extender domains of a bacterial ice-binding adhesin. *Febs J* **280**, 5919–5932 (2013).
31. Whelan, F. *et al.* Periscope Proteins are variable-length regulators of bacterial cell surface interactions. *Proc Natl Acad Sci U S A* **118** (2021).
32. Graham, L.A. *et al.* Adhesin domains responsible for binding bacteria to surfaces they colonize project outwards from companion split domains. *Proteins* (2024).
33. Gilbert, J.A., Hill, P.J., Dodd, C.E. & Laybourn-Parry, J. Demonstration of antifreeze protein activity in Antarctic lake bacteria. *Microbiology* **150**, 171–180 (2004).
34. Guo, S. *et al.* Structure of a 1.5-MDa adhesin that binds its Antarctic bacterium to diatoms and ice. *Science advances* **3**, e1701440 (2017).

35. Guo, S. *et al.* Structural Basis of Ligand Selectivity by a Bacterial Adhesin Lectin Involved in Multispecies Biofilm Formation. *mBio* **12** (2021).
36. Guo, S. *et al.* Molecular basis for inhibition of adhesin-mediated bacterial-host interactions through a peptide-binding domain. *Cell Rep* **37**, 110002 (2021).
37. Sherik, M. *et al.* Sugar-binding and split domain combinations in repeats-in-toxin adhesins from Vibrio cholerae and Aeromonas veronii mediate cell-surface recognition and hemolytic activities. *mBio*, e0229123 (2024).
38. Cai, W. *et al.* Biofilm and Sediment are Major Reservoirs of Virulent Aeromonas hydrophila (vAh) in Catfish Production Ponds. *J Aquat Anim Health* **31**, 112–120 (2019).
39. Mousavifar, L. *et al.* Insightful Improvement in the Design of Potent Uropathogenic E. coli FimH Antagonists. *Pharmaceutics* **15** (2023).
40. Montes-Robledo, A., Baldiris-Avila, R. & Galindo, J.F. D-Mannoside FimH Inhibitors as Non-Antibiotic Alternatives for Uropathogenic Escherichia coli. *Antibiotics (Basel)* **10** (2021).
41. Mancuso, G., Midiri, A., Gerace, E. & Biondo, C. Bacterial Antibiotic Resistance: The Most Critical Pathogens. *Pathogens* **10** (2021).
42. Scales, B.S., Dickson, R.P., LiPuma, J.J. & Huffnagle, G.B. Microbiology, genomics, and clinical significance of the Pseudomonas fluorescens species complex, an unappreciated colonizer of humans. *Clin Microbiol Rev* **27**, 927–948 (2014).
43. Latasa, C. *et al.* BapA, a large secreted protein required for biofilm formation and host colonization of Salmonella enterica serovar Enteritidis. *Mol Microbiol* **58**, 1322–1339 (2005).
44. Yu, M. *et al.* A systematic survey of type I secretion systems and their substrate proteins in Salmonella. *Virulence* **16**, 2533414 (2025).
45. Torresen, O.K. *et al.* Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Res* **47**, 10994–11006 (2019).
46. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
47. Satchell, K.J.F. Multifunctional-autoprocessing repeats-in-toxin (MARTX) Toxins of Vibrios. *Microbiol Spectr* **3** (2015).
48. Gulick, A.M. Nonribosomal peptide synthetase biosynthetic clusters of ESKAPE pathogens. *Nat Prod Rep* **34**, 981–1009 (2017).
49. O'Leary, N.A. *et al.* Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. *Sci Data* **11**, 732 (2024).
50. Ciufo, S. *et al.* Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *Int J Syst Evol Microbiol* **68**, 2386–2392 (2018).
51. Steinegger, M. & Soding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026–1028 (2017).
52. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

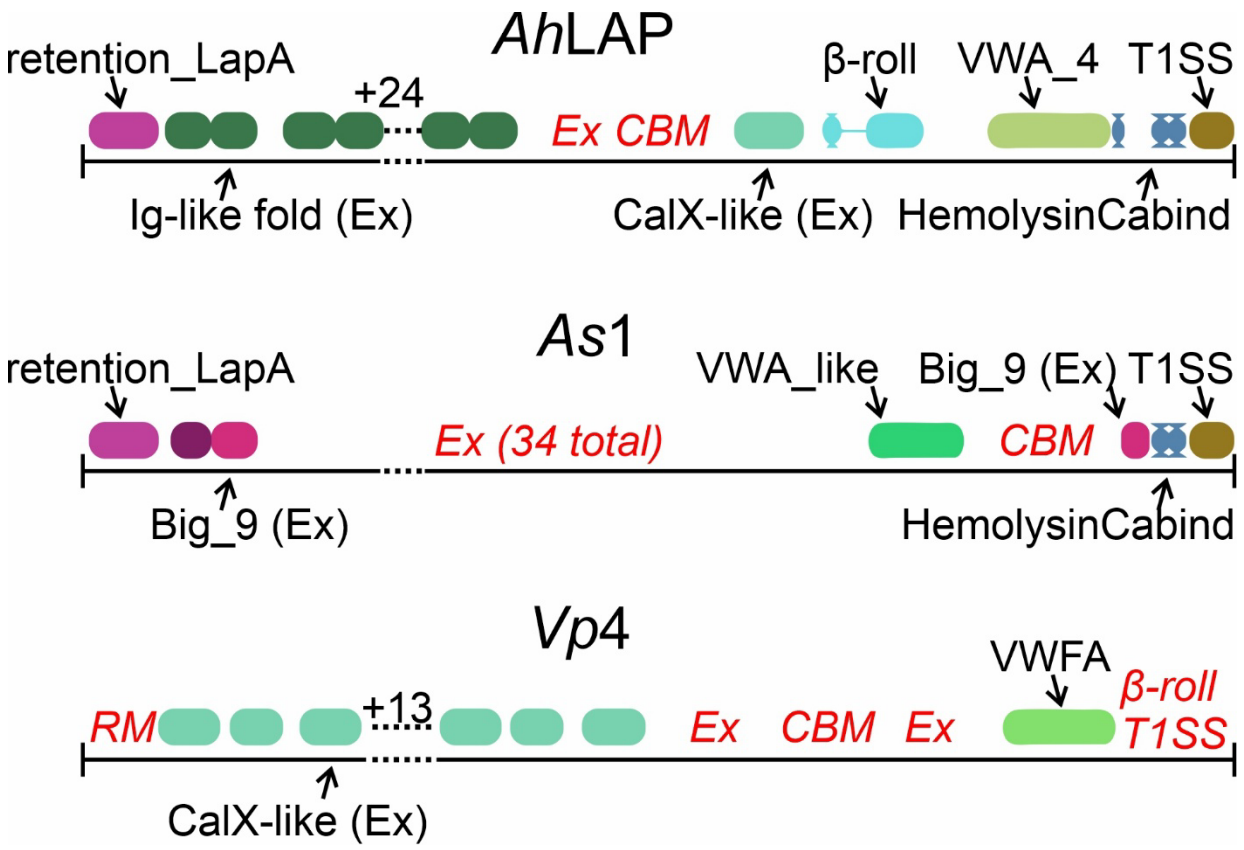
53. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).
54. Olson, R.D. *et al.* Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic Acids Res* **51**, D678–D689 (2023).
55. Restrepo-Benavides, M. *et al.* Unveiling potential virulence determinants in Vibrio isolates from Anadara tuberculosa through whole genome analyses. *Microbiol Spectr* **12**, e0292823 (2024).
56. Li, W. *et al.* RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res* **49**, D1020–D1028 (2021).
57. Peleg, A.Y., Seifert, H. & Paterson, D.L. Acinetobacter baumannii: emergence of a successful pathogen. *Clin Microbiol Rev* **21**, 538–582 (2008).
58. Morris, F.C., Dexter, C., Kostoulias, X., Uddin, M.I. & Peleg, A.Y. The Mechanisms of Disease Caused by Acinetobacter baumannii. *Front Microbiol* **10**, 1601 (2019).
59. Russo, A. & Serapide, F. The Multifaceted Landscape of Healthcare-Associated Infections Caused by Carbapenem-Resistant Acinetobacter baumannii. *Microorganisms* **13** (2025).
60. Upmanyu, K., Kumar, R., Rizwanul Haque, Q.M. & Singh, R. Exploring the evolutionary and pathogenic role of Acinetobacter baumannii biofilm-associated protein (Bap) through in silico structural modeling. *Arch Microbiol* **206**, 267 (2024).
61. Fernandez-Bravo, A. & Figueras, M.J. An Update on the Genus Aeromonas: Taxonomy, Epidemiology, and Pathogenicity. *Microorganisms* **8** (2020).
62. Tsai, Y.H. *et al.* Monomicrobial Necrotizing Fasciitis Caused by Aeromonas hydrophila and Klebsiella pneumoniae. *Med Princ Pract* **24**, 416–423 (2015).
63. Chao, C.M., Lai, C.C., Tang, H.J., Ko, W.C. & Hsueh, P.R. Biliary tract infections caused by Aeromonas species. *Eur J Clin Microbiol Infect Dis* **32**, 245–251 (2013).
64. Clark, N.M. & Chenoweth, C.E. Aeromonas infection of the hepatobiliary system: report of 15 cases and review of the literature. *Clin Infect Dis* **37**, 506–513 (2003).
65. Chan, F.K., Ching, J.Y., Ling, T.K., Chung, S.C. & Sung, J.J. Aeromonas infection in acute suppurative cholangitis: review of 30 cases. *J Infect* **40**, 69–73 (2000).
66. Baumgartner, W.A., Ford, L. & Hanson, L. Lesions caused by virulent Aeromonas hydrophila in farmed catfish (Ictalurus punctatus and I. punctatus x I. furcatus) in Mississippi. *J Vet Diagn Invest* **29**, 747–751 (2017).
67. Zepeda-Velazquez, A.P. *et al.* Bullfrogs (Lithobates catesbeianus) as a Potential Source of Foodborne Disease. *J Food Prot* **86**, 100067 (2023).
68. Griffin, M.J. *et al.* Rapid quantitative detection of Aeromonas hydrophila strains associated with disease outbreaks in catfish aquaculture. *J Vet Diagn Invest* **25**, 473–481 (2013).
69. Menanteau-Ledouble, S., Kumar, G., Saleh, M. & El-Matbouli, M. Aeromonas salmonicida: updates on an old acquaintance. *Dis Aquat Organ* **120**, 49–68 (2016).

70. Park, S.Y., Han, J.E., Kwon, H., Park, S.C. & Kim, J.H. Recent Insights into Aeromonas salmonicida and Its Bacteriophages in Aquaculture: A Comprehensive Review. *J Microbiol Biotechnol* **30**, 1443–1457 (2020).
71. Reith, M.E. *et al.* The genome of Aeromonas salmonicida subsp. salmonicida A449: insights into the evolution of a fish pathogen. *BMC Genomics* **9**, 427 (2008).
72. Pekala-Safinska, A. Contemporary Threats of Bacterial Infections in Freshwater Fish. *J Vet Res* **62**, 261–267 (2018).
73. Padra, J.T. *et al.* Aeromonas salmonicida binds differentially to mucins isolated from skin and intestinal regions of Atlantic salmon in an N-acetylneuraminic acid-dependent manner. *Infect Immun* **82**, 5235–5245 (2014).
74. Thornton, J.C., Garduno, R.A., Newman, S.G. & Kay, W.W. Surface-disorganized, attenuated mutants of Aeromonas salmonicida as furunculosis live vaccines. *Microb Pathog* **11**, 85–99 (1991).
75. Bergfors, E. *et al.* Parapertussis and pertussis: differences and similarities in incidence, clinical course, and antibody responses. *Int J Infect Dis* **3**, 140–146 (1999).
76. Bouchez, V. & Guiso, N. Bordetella pertussis, B. parapertussis, vaccines and cycles of whooping cough. *Pathog Dis* **73** (2015).
77. Watanabe, M. & Nagai, M. Whooping cough due to Bordetella parapertussis: an unresolved problem. *Expert Rev Anti Infect Ther* **2**, 447–454 (2004).
78. Remesh, A.T., Alagarasu, K., Jadhav, S., Prabhakar, M. & Viswanathan, R. Pertussis Vaccines Scarcely Provide Protection against Bordetella parapertussis Infection in Children-A Systematic Review and Meta-Analysis. *Vaccines (Basel)* **12** (2024).
79. Noble, B.A., Jiudice, S.S., Jones, J.D. & Timbrook, T.T. Reemergence of Bordetella parapertussis, United States, 2019-2023. *Emerg Infect Dis* **30**, 1058–1060 (2024).
80. Carrica, M.D.C. *et al.* BPP0974 is a Bordetella parapertussis adhesin expressed in the avirulent phase, implicated in biofilm formation and intracellular survival. *Microb Pathog* **193**, 106754 (2024).
81. Vogel, J.P. & Isberg, R.R. Cell biology of Legionella pneumophila. *Curr Opin Microbiol* **2**, 30–34 (1999).
82. LeChevallier, M.W., Prosser, T. & Stevens, M. Opportunistic Pathogens in Drinking Water Distribution Systems-A Review. *Microorganisms* **12** (2024).
83. Fields, B.S., Benson, R.F. & Besser, R.E. Legionella and Legionnaires' disease: 25 years of investigation. *Clin Microbiol Rev* **15**, 506–526 (2002).
84. Glick, T.H. *et al.* Pontiac fever. An epidemic of unknown etiology in a health department: I. Clinical and epidemiologic aspects. *Am J Epidemiol* **107**, 149–160 (1978).
85. Burlingame, G.A. & Bartrand, T.A. Laying the groundwork for a Legionella pneumophila risk management program for public drinking water systems. *J Water Health* **22**, 2385–2397 (2024).
86. Wickramasekaran, R.N., Sorvillo, F. & Kuo, T. Legionnaires' disease and associated comorbid conditions as causes of death in the U.S., 2000-2010. *Public Health Rep* **130**, 222–229 (2015).

87. Martinez-Urtaza, J. & Baker-Austin, C. Vibrio parahaemolyticus. *Trends Microbiol* **28**, 867–868 (2020).
88. Rince, A. *et al.* Occurrence of Bacterial Pathogens and Human Noroviruses in Shellfish-Harvesting Areas and Their Catchments in France. *Front Microbiol* **9**, 2443 (2018).
89. Scallan, E. *et al.* Foodborne illness acquired in the United States--major pathogens. *Emerg Infect Dis* **17**, 7–15 (2011).
90. Baker-Austin, C. *et al.* Stemming the rising tide of Vibrio disease. *Lancet Planet Health* **8**, e515–e520 (2024).
91. Kling, K. *et al.* Genetic Divergence of Vibrio vulnificus Clinical Isolates with Mild to Severe Outcomes. *mBio* **13**, e0150022 (2022).
92. Kumamoto, K.S. & Vukich, D.J. Clinical infections of Vibrio vulnificus: a case report and review of the literature. *J Emerg Med* **16**, 61–66 (1998).
93. Woida, P.J. & Satchell, K.J.F. Coordinated delivery and function of bacterial MARTX toxin effectors. *Mol Microbiol* **107**, 133–141 (2018).
94. Kim, B.S. The Modes of Action of MARTX Toxin Effector Domains. *Toxins (Basel)* **10** (2018).
95. Reimer, J.M., Haque, A.S., Tarry, M.J. & Schmeing, T.M. Piecing together nonribosomal peptide synthesis. *Curr Opin Struct Biol* **49**, 104–113 (2018).
96. Suring, W. *et al.* Nonribosomal Peptide Synthetases in Animals. *Genes (Basel)* **14** (2023).

Supplementary Results

The domain structure of candidate sequences can be ascertained by using InterProScan [46], accessed via the European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL–EBI) website. In some instances, domains corresponding to all four sections of an RTX adhesin (Fig. 1A) are detected (Supplementary Fig. 1, *Ah*Lap). Here, the first section was identified as a retention module (see Table 2 for a mapping of feature names to protein functions). Nearly all the extender domains (Ex) in section 2 were detected, as was the β-roll and von Willebrand factor A-like domain (vWFA) domain within section 3. Both the β-roll and the T1SS signal of section 4 that mediate secretion were detected. The domains that were not detected were the CBM (usually missed but occasionally detected as a PA14 domain) and one of the extenders. Split domains are seldom detected. In a second example, *As*1, domain detection was very similar except the bulk of the extenders were not detected. However, three were identified, two near the N terminus and one within the LBR. At least one extender was identified by InterProScan in all adhesins examined (not shown). In the third example (*Vp*4), domains were not detected in region 1 or 4. However, there are gaps in the domain map, indicating the possibility of undetected domains. Indeed, when modelled, *Vp*4 clearly possesses both a retention module and a β-roll/T1SS. Therefore, any proteins with domain maps that resemble the three shown here were selected for modelling.



Supplementary Figure 1. Representative InterProScan results for three RTX adhesins. The diagram was adapted from the actual output, including matching the colour, shape, and short descriptions of the domains found, but it has been simplified and modified. Domains that were not detected are indicated by red text in italics. Commonly detected domains are also listed in Table 2.

```
CLUSTAL O(1.2.4) multiple sequence alignment


CBM6      ----ADLPVDVASPVVSVNLIGSN-------SVVSDQFVVGAGTVANIDLSAVYTAQPLA   49
CBM5      -QGELGVAVGSATYLALVGITENL-NV-SLLGTPSVAFTIDAGHEADV--TFAYAPVLSL   55
CBM2      QLVGTE--TDSAFYGVLLNVGNIV-DL-QLLTVDTVDFTIAAGQEGVA--TFNFNSLIGA   54
CBM4      LQIGNDVAVGNTTYLALLTLTD---NFDFQAGGQSVNFTLTDATLNDV--TFNYSALISA   55
CBM1      ---GDDVSLGSTSYLVLTSVAGLDLQL----GGNSLGFTVAAGHEGDV--TFQYSGLIDA   51
CBM3      HLIEDDRALGSATYLALLSLAGINLQAPLPFVNSTVEFNVGAGETGTA--TFKYSSLINE   58
                   . :   .   :              :  * :  .       :  :

CBM6      SLLPTVSYVIQSWNGT--AWVNTIYSGSQTALGSVATIQAG-SVAFQDTVEHLAAGTYRV   106
CBM5      SLFNDYKVVLQQK-GADGAWHNIDGGS-STGLLNIGLLGNGG---IGVTVPDLGQGEYRA   110
CBM2      SALGDYNVVLQKYNDVTGQWEAVNGTG-DRSLLNLTLLGN----TPTAQIGGLTEGEYRA   109
CBM4      SLLADYVLVVQKFDTATNQWVAVNGTG-DADLLSLAAFGG-----NSVTLEGLAAGQYRA   109
CBM1      AVLSDYKLVVQKFNTTTNQWESIHGDA-NSSLISLHLLGIGTGNVPGAVLDGLDAGQYRA   110
CBM3      GALGDYQLVVQKFNTATNRWESITGSS-EASLLNLSVLGIGVNATPGVVVEGLDEGQYRA   117
          . :     *:*.   .   *      . .  * .:  :           :  *  * **.

CBM6      QYTLTGVSLGATSLDTNVSTTTVHL--- 131    CBM6
CBM5      FMVYTGL--GVGILGTMSVVKDDFDYTV 136    CBM5    25
CBM2      FLSFNGL-VGGAVAVTLNGSVDVYNP-- 134    CBM2    20   34
CBM4      YMTYAGSGVGVSLLGTLSVQKDVFD--- 134    CBM4    28   43   36
CBM1      FLAYDGL-LGLGVLGTLSATMDDYDL-- 135    CBM1    24   40   43   48
CBM3      FMTYNGL-YGKSILGTLSGTMDVYD--- 141    CBM3    25   40   44   47   56
               *   *     *       .
```

Supplementary Figure 2. Alignment and percent identity matrix of the six CBMs of *Ab*1. Asterisks indicate residues shared in all isoforms, with colons and dots representing conserved and semi-conserved residues, respectively.

Supplementary Table 1. Accession numbers and parameters of representative proteins from each adhesin group.

Adhesin	Representative accession	Length (a.a.)	Count[1]	% of Genomes
*Ab*1	WP_085762530.1	6180	15	30
*Ab*2	WP_115611016.1	5360	5	10
*Ab*3	WP_215738566.1	4650	2	4
*Ab*4	WP_228332306.1	2766	3	6
*Ab*5	WP_000250959.1	3270	34	68
*Ab*6	WP_004644391.1	1621	28	56
*Ah*1	WP_241876840.1[2]	4232	22	49
*Ah*2	WP_011707240.1	5047	38	88
*Ah*3	WP_135319613.1	3564	24	56
*Ah*4	WP_392389641.1[3]	3827	2	5
*Ah*5	WP_411679835.1	2881	6	14
*As*1	WP_225854360.1	1641	30	83
*As*1 long	WP_282683987.1[4]	5525	1	3
*As*2	WP_409011068.1	3224	5	14
*As*3	WP_088814302.1	3419	36	100
*As*4	WP_309266189.1	8957	1	3
*Bp*1	WP_010927777.1	2215	50	100
*Lp*1	WP_061784179.1	7931	28	56
*Lp*1a	WP_070322476.1[5]	7397	In above	
*Lp*2	WP_070065546.1	6420	22	44
Vp1	WP_021448081.1	5227	49	98
Vp2	WP_005477759.1[6]	3199	49	98
Vp3	WP_127913995.1	3550	4	8
Vp4	WP_021447549.1	2344	45	90
Vp5	WP_164853014.1	2947	1	2
*Vv*1	WP_223370862.1	6317	10	33
*Vv*2	WP_017422353.1	4654	29	97
*Vv*3	WP_061058482.1	7494	4	13
*Vv*4	WP_064490486.1	8523	6	20
*Vv*4a	WP_182289888.1[7]	5275	7	23
*Vv*5	WP_172571789.1	3769	1	3
*Vv*6	WP_185891060.1[8]	~10,703	1	3

[1] The number of isoforms that clustered within each group from a sampling of 50 or fewer genomes (see Table 1).

[2] This adhesin was misannotated in all genomes examined. The correct open reading frame is NZ_CP092356.1 from 1,823,727-1,838,066.

[3] Also misannotated, complete adhesin encoded by NZ_CP028568.2 from 3,178,854-3,165,730.

[4] Clusters with WP_225854360.1 but included as an example of extender variability.

[5] Very similar to *Lp*1, the exception is that the CBMs are only ~28% identical.

[6] Misannotated in many genomes, correct open reading frame is on NZ_CP187437.1 from 1723475-1735462.

[7] Similar to *Vv*4 with an identical domain architecture in the LBR region with 94% identity in two segments indicated by dashed lines in Fig. 4, but only 32% identity in the region between them.

[8] Likely a failed assembly due to at least 55 identical repeats, N-terminal portion with RM on WP_185891039.1.