Collagen-like sequences as adhesion factors encoded by the pathogenic bacterium *Burkholderia cenocepacia*: computational analysis and experimental studies

Ricardo Estevens[1,2]

1.Departamento de Bioengenharia, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisboa, Portugal.
2. iBB-Instituto de Bioengenharia e Biociências e i4HB- Instituto para a Saúde e a Bioeconomia, Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisboa, Portugal.

Burkholderia cenocepacia is a multi-drug resistant pathogen capable of causing chronic infections in cystic fibrosis or immunocompromised patients. This species synthesizes a panoply of adhesins that are critical for adhesion to the host cell. A transcriptomic study revealed that during the first contact of *B. cenocepacia* K56-2 to giant plasma membrane vesicles derived from human bronchial cells, a group of "collagen-like genes" are overrepresented among the induced genes. Since the role in adhesion of these genes in *B. cenocepacia* is unexplored, in this study we provide a bioinformatic analysis. We identified 75 collagen-like proteins (CLPs) containing the Bacterial Collagen Middle Region (Col_mid_reg) in *Burkholderia cepacia* complex (Bcc) members, analyzed its phylogenetic distribution, identified that CLPs are formed by extensive intrinsically disordered regions, and discussed how these regions may increase the efficiency of CLPs as adhesion factors. Next, we analyzed 5 CLPs paralogs in *B. cenocepacia* J2315. Additionally, the functional analysis of gene *bcal1524* encoding a CLP was carried out. A mutant of *B. cenocepacia* K56-2 was constructed and a phenotypic evaluation was performed regarding biofilm formation, adhesion to host cells (bronchial cell line), and to components of the extracellular matrix (ECM), motility, and virulence using the model *Galleria mellonella*. Our data suggest that the absence of this protein decreases the motility capacity and increases the adherence to host cells. However, no differences were observed in biofilm formation, virulence or adhesion to different components of the ECM. Collectively, our results suggest that the negligible effects observed may represent functional redundancy and compensation effects of the paralogous copies. More studies are necessary to understand the role of CLPs in this species.

Keywords: *B. cenocepacia*; cell adhesion; bacterial collagen-like proteins; low complexity regions; intrinsic disorder proteins

Introduction

Bacterial adhesion is very complex and can be defined as the process by which bacteria adhere to other cells or surfaces[1]. It is a mechanism dependent on several factors, such as surface, surrounding environment, and the cell itself[2]. Adhesion to host cells is usually mediated by specific ligands, which may have different natures: lipids, proteins, or sugars. Several categories of adhesins have been identified through the years, from flagella[3] to trimeric autotransporter adhesins[4]. In more recent years, several proteins with disordered regions, tandem repeats (TRs) and low complexity regions have been identified as possessing an important role in the adhesion process and pathogenicity in bacteria[5,6,7]. The bacterial collagen, which possesses the typical triple helical structure composed by the repetition of the G-X-Y triplet of amino acids identified in metazoan collagen, was also described as important for this biological processes[8,9,10]. Regarding its origin in bacteria, two options were hypothesized: horizontal transfer of collagenous sequences between eukaryotes and bacteria[11,12], or the emergence of collagen repeats in bacteria *de novo*, resulting from spontaneous mutations, duplication of repetitive sequences, or domain organization[13,14]. Some proteins were studied, especially in *Streptococcus pyogenes*[15,16], with the first mentioned collagen-like protein (CLP) in bacteria. Other proteins have been identified in other organisms, like *Bacillus amyloliquefaciens*[17] and *Burkholderia pseudomallei*[18]. However, two types of CLPs seem to exist: one already mentioned, with the typical GXY triplet repeats, capable of forming structured triple helixes, and a second category with proteins possessing the Bacterial Collagen Middle Region (Col_mid_reg) and disordered regions.

The *Burkholderia cepacia* complex (Bcc) is a group of closely related gram-negative bacteria[19]. From this group of bacteria, some members emerged as opportunistic pathogens in immunocompromised and cystic fibrosis patients[20]. Infections caused by these pathogens are related to necrotizing

pneumonia and septicaemia, resulting in death[20]. Among the different pathogenic species, *Burkholderia cenocepacia* is one of the most virulent, responsible for most infections[21]. This species is resistant to antibiotics and possesses a large repertoire of virulence factors, including toxins, enzymes, invasins and adhesins[22,23]. Through the years, several components were identified regarding the adhesion process. Trimeric autotransporter adhesins (TAAs) are one of the most well-studied adhesins. Mil Homens *et al.*[24] identified a putative adhesion cluster in the *B. cenocepacia* genome. By studying the TAAs expressed by those genes, the authors were able to observe their importance in biofilm formation, motility, and invasion of host cells[24,25,26]. Pimenta *et al.*[27] studied the adhesion of *B. cenocepacia* K56-2 bacteria to giant plasma membrane vesicles derived from human bronchial cells. After performing a transcriptome analysis of the adhered bacteria, the authors were able to identify several genes involved in the bacterial adaptation to the adhesion process. They also observed the overexpression of various genes encoding pili, TAAs and other adhesins, revealing their role in the initial stage of infections. From these genes, some were identified as belonging to the CLPs family, namely three among the most overexpressed genes: *bcal 1523, bcal1524* and *bcam0695*.

Pimenta *et al*.[27] hypothesized that the role of CLPs is somehow related to adhesion. Building on that, the present work aims to study this class of proteins, performing computational analysis of the CLPs present in the Bcc and, specifically, in *B. cenocepacia*. Besides, we aim to understand the importance of the BCAM1524 and BCAM0695 proteins in the adhesion process and virulence. For that, a deletion mutant for each gene of *B. cenocepacia* was constructed and a phenotypic analysis *vs* the wild-type strain was performed. Several tests were conducted to evaluate differences in growth, motility, biofilm formation, adhesion to bronchial pulmonary host cells (16HBE14o- cell line), and components of the extracellular matrix (collagen type I and IV and fibronectin). Additionally, two truncates were constructed for each protein to continue the evaluation of the proteins' domains and their function. We intend to understand the importance of these two CLPs in the adhesion process of the pathogenic *B. cenocepacia*, as well as to increase the knowledge regarding bacterial collagen and its function.

Materials and Methods

Computational Analysis

Proteins possessing the Col_mid_reg family (PF15984) were identified and characterized in each species of the Bcc using the Burkholderia Genome Online Database[28]. The phylogenetic trees were obtained with phylogeny.fr[29]. InterPro online tool was used to search for specific domains and motifs, the XSTREAM Web Interface[30] to identify tandem repeats, and the flDPnn and SMART webservers[31,32] for disordered region prediction. The three-dimension structural prediction of each domain of the *B. cenocepacia* J2315 proteins was performed with the I-TASSER server[33].

Bacterial strains and growth conditions

B. cenocepacia clinical isolate K56-2, *Escherichia coli* αDH5, and *E. coli* One Shot® Mach1™-T1^R (Invitrogen, Thermo Fischer Scientific, USA) were used. Bacteria were cultured in Luria-Broth (LB) medium (NZYtech, Portugal) at 37ºC with orbital agitation at 250 rpm. When appropriate, the medium was supplemented with 150 mg/L of ampicillin, 50 mg/L or 100 mg/L of kanamycin (Sigma Aldrich, USA) for *E. coli*, and 100 mg/L, or 150 mg/L of trimethoprim (Sigma Aldrich, USA) for *B. cenocepacia* K56-2 mutants.

Cell lines and cell culture

An immortalized human bronchial epithelial cell line 16HBE14o-[34] was maintained in fibronectin-collagen-coated flasks in a minimum essential medium with Earle's salt (MEM) (Gibco, USA) supplemented with 10% fetal bovine serum (Gibco, USA), 0.292 g/L L-glutamine (Sigma-Aldrich, USA), penicillin (100 U/ml) (Gibco, USA), and streptomycin (100 µg/mL) (Gibco, USA) in a humidified atmosphere at 37 °C with 5% CO_2.

Construction of *bcam0695* and *bcal1524* *B. cenocepacia* mutants

Construction of mutants was performed as described in[25], with some modifications. A fragment of each gene (*bcal1524* and bcam0695) of *B. cenocepacia* K56-2 was amplified by PCR using primers frBCAM0695_fwd' and frBCAM0695_rev (Stabvida, Portugal) (Table 1), containing *BamHI* and *XbaI* restriction sites (underlined), respectively. The obtained fragments were cloned into pDrive (Qiagen), creating the pDrive695 and pDrive1524 plasmids. Next, the 1.1-kb fragment from pUC-Tp plasmid

containing the Trimethoprim (Tp) cassette was amplified by PCR using the primers frTp_Fwd and frTp_Rev (Stabvida, Portugal) (Table 1) with *XhoI* restriction sites (underlined). The pDrive695 and pDrive1524 vectors, and the Tp inserts were ligated and transformed into chemically competent *E. coli* αDH5 cells plated in an LB agar medium supplemented with 100 µg/mL of Ampicillin (Sigma Aldrich, USA) and 100µg/mL trimethoprim (Sigma Aldrich, USA). The transformations were kept at 37ºC overnight. To confirm the transformation, a colony PCR was performed using the primers for each fragment (Table 1).

The pDrive695Tp and pDrive1524Tp plasmids were inserted in electrocompetent *B. cenocepacia* K56-2 cells by electroporation. Transformants were selected on LB agar supplemented with 150 mg/L of trimethoprim at 37ºC. To distinguish between single- and double-crossover mutants, the genomic DNA was extracted, and a PCR using the respective primers (either for *bcam0695* or *bcal1524*) (Table 1) was performed, allowing the identification of the mutants.

Only *bcal1524* mutant was created and it was confirmed by whole-genome sequencing using an Illumina MiSeq system at Instituto Gulbenkian de Ciência (Portugal).

Motility assays

Swimming agar plates contained 0.3% (w/v) bacto agar (Difco, USA), 10 g/L tryptone (Gibco, USA), 5 g/L NaCl (Sigma Aldrich, USA), 5 g/L yeast extract (Difco, USA). Swarming plates composition was similar, with higher bacto agar concentration – 0.5 % - and supplemented with 5 g/ L of glucose (Sigma Aldrich, USA). Bacterial cultures were grown overnight under defined conditions and a 5 µL drop of culture was inoculated in each plate. The plates were incubated at 37ºC for 24h and the halo diameter was determined.

Bacterial adhesion to human bronchial epithelial cells

The experiment was performed on 16HBE14o- cells following the procedure described by Mil-Homens *et al*.[25], with modifications. Human bronchial epithelial cells were seeded on a 24-well plate (Orange Scientific, Belgium) at $5x10^5$ cells/well in MEM medium supplemented and incubated for 24h at 37ºC, in a humidified atmosphere with 5% CO_2. Wells were washed with Hepes Buffer Saline (HBS) pH 7.4 (Sigma Aldrich, USA) and maintained in MEM medium. The wt and mutant strains were grown overnight in previously described conditions and were used to infect the epithelial bronchial cells at a multiplicity of infection of 50:1 (bacteria to human cell). After infection, plates were incubated at 37ºC in a humidified atmosphere with 5% CO_2 for 30 minutes. Next, the MEM medium was removed, the cells were washed three times with PBS and 200 µL of lysis buffer containing 0.25% (v/v) Triton X-100, (Sigma Aldrich, USA) and 10 mM EDTA (Sigma Aldrich, USA) was added and incubated for 20 minutes at RT. Cell lysates were scraped and mixed with the lysis buffer. For quantification, serial dilutions of the cells' lysate mixture in PBS were conducted and plated in LB agar plates.

Production and purification of recombinant truncate proteins

The Champion™ pET SUMO Protein Expression System kit (Invitrogen, Thermo Fisher, USA) was used to construct the following truncates: BCAM0695_21_758, BCAM0695_320_758, BCAL1524_33_558 and BCAL1524_116_558, derived from the *bcam0695* and *bcal1524* genes, respectively. For that, the primers for each truncate were used (Table 1), to amplify each fragment by PCR. The construction of the vectors was performed with the fresh PCR products and the insertion into chemically competent *E. coli* αDH5 following manufacturers' instructions. The correct construction was confirmed by sequencing (GATC Biotech AG, Germany). Posterior transformation of chemically competent *E. coli* BL21-DE3 cells was performed. The truncates from the *bcam1524* gene were overexpressed testing the following conditions: induction with IPTG (0.2, 0.6 and 1 mM) for 4 hours at 37ºC, 250 rpm. For the optimal expression conditions, cells were harvested from 1L of culture by centrifugation at 7025 g for 10 minutes at 4ºC. The supernatant was discarded, and the pellet was resuspended in buffer A (20 mM sodium phosphate, 500 mM NaCl, 10 mM imidazole, pH 7.4). The cells were disrupted by sonication using the Branson Sonifier 250 (Emerson, USA) (8 cycles of 15 pulses with 70% duty cycle and output control 8) and centrifuged at 17600 g for 5 minutes and then the supernatant was transferred to new tubes and centrifuged again for 1 hour. The supernatant was collected and loaded into a 5 mL HisTrap column (GE Healthcare, Germany), which was equilibrated with buffer A using an AKTA Start system (GE Healthcare, Germany), following the manufacturer's instructions. The amount of

imidazole was altered by a continuous gradient from 10 mM to 300 mM. After elution, fractions containing the desired truncates were concentrated using an Amicon 10kDa (Merck Millipore, Ireland) and dialysed to PBS. The concentrated samples were quantified using the Nanodrop One (Thermo Fisher, USA) and stored at 4ºC. To separate the SUMO protein from our fragments of interest, SUMO protease was added (0.1% v/v) and the mixture was incubated overnight at 4ºC. Then, the mixture was loaded into a 1 mL HisTrap column (GE Healthcare, Germany) to collect the flowthrough that contain our fragments, using buffer A. Then SUMO protein was eluted with a phosphate buffer containing 300 mM imidazole. The fractions with the fragments were concentrated using an Amicon 10kDa (Merck Millipore, Ireland), dialysed to PBS, quantified using the Nanodrop One (Thermo Fisher, USA) and stored at 4ºC.

Table 1. List of primers used in this study.

frBCAM0695_fwd	**5'-CGGGATCCACAACCTCGGCAATGCAGTGA-3'**
frBCAM0695_rev	**5'-TACGTCTAGAGTCGTTCTGGTACGTGATGCGTTA-3'**
frTp_fwd	**5'-ACCGCTCGAGCGCCCACAGTCCATTGAACAAA-3'**
frTp_rev	**5'-ACCGCTCGAGTATGCTTCCGGCTCGTATGTTG-3'**
frBCAL1524_fwd	**5'- CGGGATCCAACCGTTCGCCACATCTGCT- 3'**
frBCAL1524_rev	**5' - TACGTCTAGATGATCGCCGAGCTGATCGGTT- 3'**
fr695_21_758_fwd	**5'-GGGAATTCCATATGTACGGTTGCGGCTCGGTCGAT-3'**
fr695_21_758_rev	**5'- CGCGGATCCTGTCGTTCTGGTACGTGATGCGTT-3'**
fr695_21_758_fwd2	**5'-TACGGTTGCGGCTCGGTCGAT-3'**
fr695_21_758_rev2	**5'-TGTCGTTCTGGTACGTGATGCGTT-3'**
fr695_320_758_fwd	**5'-CAGAAATCCGGCAACCTCGTGA-3'**
fr1524_33_558_fwd	**5'- GGCTCCATCAGCCAGGGTCT-3'**
fr1524_116_558_fwd	**5'-GTGGGAAATGTCCTCGCACAA-3'**
fr1524_33_558_rev	**5'-CCTCGCTGTCTTTTCGATGAATGG-3'**

Statistical Analysis

All experiments were performed in a minimum of three independent replicates. Statistical analysis was carried out by using GraphPad Prism 8-0-1 software. Relative comparisons were done among corrected values with Welch's t-test for significance. A p-value of <0.05 was considered statistically significant in all analyses.

Results and Discussion

Computational Analysis

CLPs are being identified in increasing numbers, namely in pathogens[8,13,35]. In this work, we began by identifying the presence and distribution of CLPs among a representative strain for each of the 19 members of the Bcc with the Col_mid_reg (PF15984) used as a reference. This domain is described as a conserved domain represented in bacterial collagen, and is mainly present in bacterial species, as represented in Figure 1 A. Inside the Bacteria domain, the Burkholderiaceae family, which englobes the *Burkholderia* genus and the Bcc, is the family with the most species represented, with a total of 38. This indicates the existence of a high number of proteins with the Col_mid_reg domain yet to study. Considering the search for CLPs in Bcc, a total of 75 putative CLPs were identified, whose sequences were then used to obtain the phylogenetic tree represented in Figure 1 B. The tree is divided into four different clusters (I, II, III and IV), with cluster IV subdivided into three subclusters (IV_A, IV_B, and IV_C). Analysing the tree, it is understandable that the sequences are not grouped by species, but the division is rather based on the identity of the Col_mid_reg. Thus, proteins from the same cluster possess identities of around 80-90%, while proteins from different clusters can have identities of 40%.

Next, an analysis regarding the presence and frequency of Tandem Repeats in these 75 proteins was performed, due to the described importance of TR in adhesion and infection processes[36,37]. The TRs were identified, and the total percentage of the protein occupied by these repetitions was calculated (Figure 2). Firstly, it is notable the existence of TR in 70 out of 75 proteins. Additionally, in this graph is visible that *B. cenocepacia* J2315 and *B. multivorans* ATCC 17616 possess all the proteins with TRs. These two species are known for being the two most prevalent Bcc pathogenic species[38]. There seems to exist a tendency that associates the TRs with an increase in pathogenicity in these species, but more studies are still necessary.

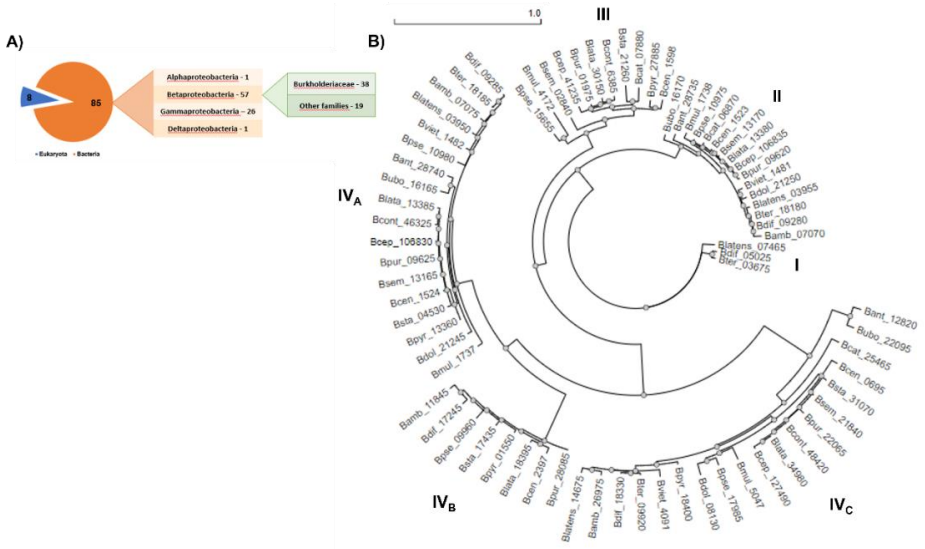


Figure 1. A) Distribution of the Col_mid_reg through different levels of taxonomy: domains, classes, and families, respectively. **B)** Phylogenetic tree of the 75 putative CLPs from the Bcc, divided into four clusters: I, II, III and IV. The division is based on the identity of the Col_mid_reg.

We next analyzed the sequences' distribution by species and by subcellular localization of the product. We can observe that the number of CLPs per species is between 3 and 5 and that the majority are described as membrane proteins. There are a considerable number of proteins with predicted localization in the cytoplasmic membrane or the outer membrane. Is it also observable that there is at least one lipoprotein per species, except in *B. vietnamensis* and *B. dolosa*. This even distribution of CLPs in all members of the Bcc may indicate the importance of these proteins in indispensable biological functions. Considering *B. cenocepacia* J2315, 5 CLPs were identified: BCAL1523, BCAL1524, BCAL2397, BCAM0695, and BCAM1598. These proteins are all characterized as putative lipoproteins, indicating their potential as having a role in the adhesion process. Additionally, these proteins are described as membrane proteins, as expected due to their lipoprotein nature, and the genes *bcal1523, bcal1524* and *bcam0695* appeared overexpressed in Pimenta et al experiment[27], as already referred, predicting their importance in the adhesion process. We then focused on these 5 CLPs and genes from *B. cenocepacia* J2315.

Firstly, we assessed the distribution of the genes on the chromosomes. *B. cenocepacia* J2315 possesses 3 chromosomes and 1 plasmid[39], although the 5 genes of interest are distributed only in chromosome 1 (*bcal1523, bcal1524,* and *bcal2397*) and chromosome 2 (*bcam0695* and *bcam1598*) (Figure 3 A). Next, we analysed the genes' neighbourhood (Figure 3 B). Most of the genes up and downstream of our genes of interest are not characterized yet, possessing any function attributed. However, two genes (*bcal1520* and *bcal1525*) were also overexpressed in Pimenta's experiment[27] which may indicate their importance and coordination with our target CLPs.

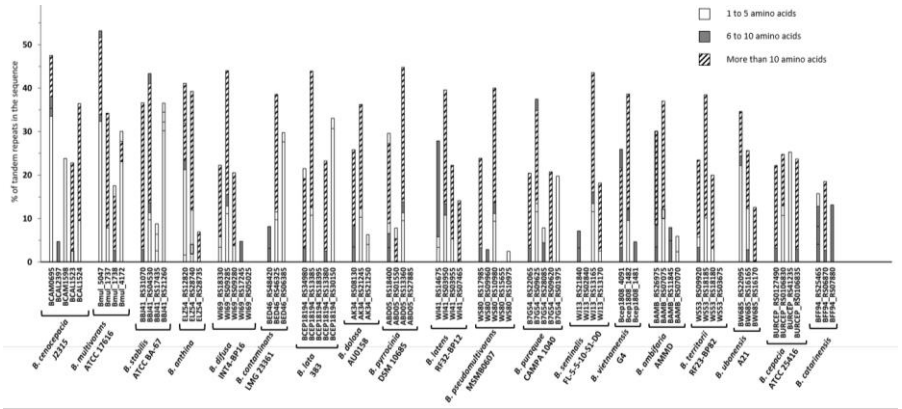


Figure 2. Distribution and frequency of tandem repeats in the 75 CLPs from the Bcc.

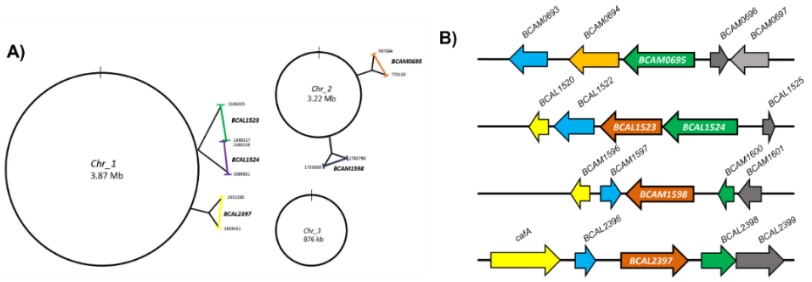


Figure 3. A) Distribution and localization of the 5 collagen-like encoding genes in the *B. cenocepacia* J2315 chromosomes. **B)** Neighbourhood of the 5 CLPs.

Noticeably, the *bcal1520* product is also described as a putative lipoprotein and *bcal1525* is already described as an Flp-type Pilus subunit, being a virulence trait described in other organisms[40,41]. Studying the up and downstream genes around the CLPs is important to assess the eventual coordination between them. Afterwards, the characterization of the five CLPs was performed. Using InterPro and the SMART web servers, we identified the size and localization of the Col_mid_reg and of low complexity regions along the protein, respectively (Figure 4). The Col_mid_reg is present in all 5 proteins, as expected, with a size of approximately 200 amino acids, and all 5 proteins possess several low complexity regions, especially in the areas not identified as the Col_mid_reg (Figure 4). In fact, the percentage of protein considered with low complexity is variable, with a maximum of 72% for the BCAM0695 protein and a minimum of 43% for the BCAL1523 (Table 2). Moreover, we searched for the presence of GXY repeats, typical of collagen-like domains. Manual visualization allowed the identification of GXY repeats in 4 out of 5 proteins, with BCAL2397 as the exception. The number of repeats varies between four in BCAL1523 and 83 in BCAM0695. A computational search of the GXY repeats was performed, but only domains with 20 copies of the GXY triplet are detected by InterPro. Thus, only BCAM1598 and BCAM0695 possess enough repetitions to be detected by Interpro. Analyzing the four G-X-Y-containing proteins, we can observe that the most frequent pattern is G-T-S (Glycine – Threonine – Serine), comprising 91% of the G-X-Y tripeptides present in the proteins. It is also noted that no other domains were detected. Focusing on the location of the GXY repeats and the existence of low complexity regions along the proteins, we can observe that coincidence. This may be due to the nature of the amino acids and of the repetitiveness of that area of the protein. It is known that the presence of a higher frequency of Glycine and Serine can lead to the formation of disordered regions in proteins, something that is also related to collagen domains[42]. Additionally, it is notable that the protein with the highest percentage of low complexity regions (BCAM0695) is also the protein with the highest number of GXY repeats, indicating the correlation between these two characteristics, even though the presence of low complexity regions is not completely dependent of the presence of GXY, as we may see in proteins BCAL1523 and BCAL2397, which possess few or no repetitions and present low complexity regions throughout the extension of the sequence.

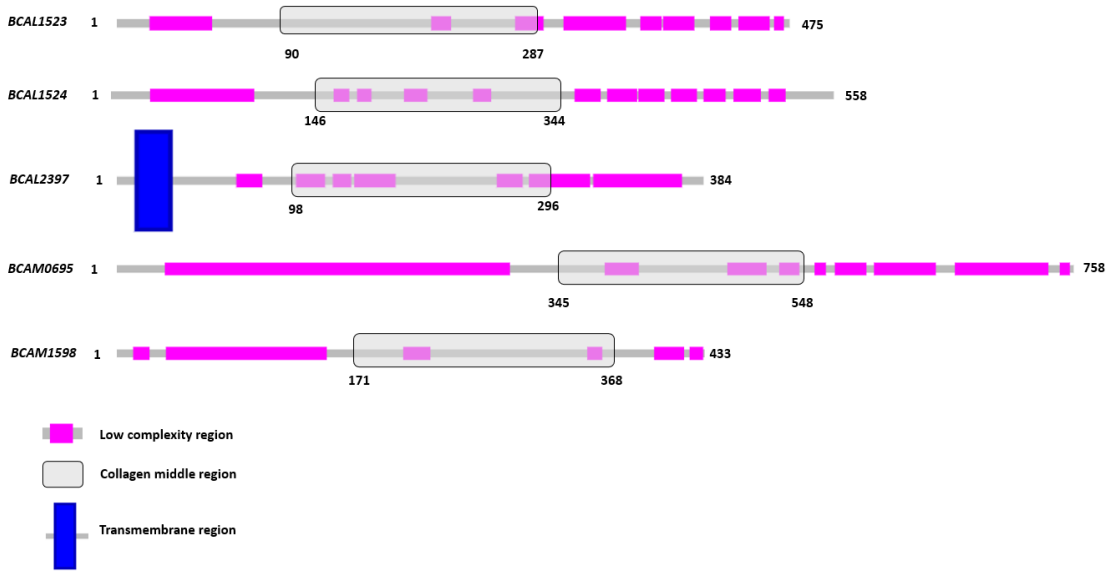
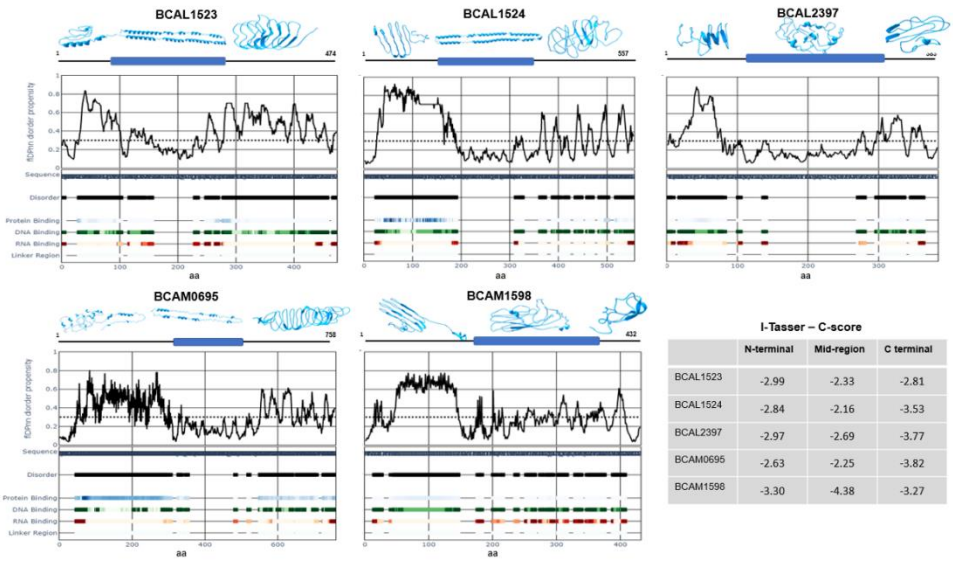


Figure 4. Representation of the low complexity regions and the Col_mid_reg predicted by SMART and InterPro webservers, respectively.

Table 2. Characterization of the 5 CLPs identified in *B. cenocepacia* J2315. The number of amino acids, the predicted localization of the protein, the significance of the existence of the Col_mid_reg, the percentage of low complexity region per protein, and the number and type of GXY repeats.

Protein	No. amino acids	Localization	Collagen Middle Region E-value	Low complexity region (%)	No. GXY repeats	GXY Type
BCAL1523	475	Cytoplasmic Membrane	$1.8e^{-63}$	43	4	GTS_3GSS_1
BCAL1524	558	Extracellular	$5.1e^{-68}$	49	18	$GSS_1GTS_{16}GVS_1$
BCAL2397	384	Extracellular	$9.4e^{-58}$	63	-	-
BCAM0695	758	Unknown/Cytoplasmic Membrane	$9.3e^{-56}$	72	83	$GSS_3GTS_{79}GTG_1$
BCAM1598	433	Extracellular/Cytoplasmic Membrane	$1.2e^{-56}$	45	36	$GTN_1GTG_1GTS_{31}GTP_1$ GTN_1GII_1

The existence of an elevated percentage of low-complexity regions pushed us to analyze the disorder of the proteins. Using the flDPnn webserver we searched for the disordered regions of the 5 CLPs and observed that the lowest disorder score coincided with the location of the collagen middle region (Figure 5). Moreover, the remaining sequence of the protein, both in the N-terminal and the C-terminal, may be considered disordered regions. The existence of a high frequency of tandem repeats (Figure 2) in these proteins explains the presence of low complexity and disordered regions. Furthermore, the disordered regions possess a high score for protein, DNA or RNA binding, indicating their putative importance in biological functions, as described in bacteria[5]. Using the I-TASSER web server we obtained a model for each protein, which allowed the visualization of the 3 regions of each protein (N-terminal, Collagen Middle Region, and C-terminal) (Figure 5). The collagen middle regions of BCAL1523, BCAL1524 and BCAM0695 show a more defined structure, with the N- terminal and C-terminal regions possessing no defined structure, being coincident with the information obtained regarding the disordered and low complexity regions. Considering BCAL2397 and BCAM1598, the three domains possess no defined structure. The disorder score for BCAL2397 is not high, while in BCAM1598 the disorder score may justify the representation obtained with I-TASSER. However, the quality of the predicted structures is not elevated, as expected due to the high amount of disordered and low complexity regions in these proteins (C-score table in Figure 5).



I-Tasser – C-score

	N-terminal	Mid-region	C terminal
BCAL1523	-2.99	-2.33	-2.81
BCAL1524	-2.84	-2.16	-3.53
BCAL2397	-2.97	-2.69	-3.77
BCAM0695	-2.63	-2.25	-3.82
BCAM1598	-3.30	-4.38	-3.27

Figure 5. Disorder scores obtained with the flDPnn webserver and I-TASSER structure predictions. C-score is a confidence score for the quality of the predicted structure, ranging from [-5, 2].

The presence of GXY repeats was then analyzed in the orthologues of the bcam0695 gene, since it is known that this type of tandem repeats is prone to variation, contraction, and extensions of this area of the proteins. Looking to the genomic sequence, we observed that the nucleotide tandem repeats can be classified as imperfect tandem repeats, being not totally conserved36, even though originating conserved amino acid repetitions, due to the degenerate nature of the genetic code. It is also described that varying the extension of tandem repeats is important for the pathogenicity of bacteria, leading to a more efficient adaptation36. In the different members of the Bcc we observed that only the Bmul_5047 gene from B. multivorans ATCC 17616 possesses the GXY repetitions conserved in the gene sequence. In fact, this gene possesses more nucleotide repetitions that originate GTS than *bcam0695* from *B. cenocepacia* J2315, demonstrating the existence of expansions in the type of region. However, when compared with the other members of the Bcc, no more consecutive GTS expressing repetitions are found in the gene sequence. *B. anthina* and *B. ubonensis* possess a sequence which encodes for a GTSX, with a fourth amino acid that impairs the existence of consecutive GTS repetitions. This is also common in regions enriched in tandem repeats, due to the enhanced probability of occurring mutations that alter the genomic sequence[36]. In the other members, no GTS encoding tandem repeats are annotated. This happens due to other typical phenomena in these tandem repeats: mutations that originate a frameshift. Searching upstream of the gene, we find other genes annotated possessing the tandem repeats belonging to the orthologue of *bcam0695* or the tandem repeats lost in intergenic regions. This happens due to sequence flexibility that tandem repeats confer to the genomic sequences, especially in pathogenic organisms that take advantage of these characteristics to adapt to a hostile environment more rapidly[36]. The study of these proteins and these disordered and flexible regions is important to understand the virulence and adaptative capacity of the Bcc members, namely *B. cenocepacia* as one of the most virulent and pathogenic.

Phenotypic characterization of the Δ1524::Tp *B. cenocepacia* mutant

The objective after constructing the mutant was to perform a phenotypic analysis of the mutant and understand the role of the BCAL1524 protein in different processes of the bacteria, namely processes related to pathogenicity and adhesion. Various assays were conducted to assess the importance of this protein for the bacteria, concerning bacterial growth, motility, biofilm formation, adhesion capacity to lung epithelial cells (16HBE14o- cell line) and extracellular matrix (ECM) components, namely fibronectin and collagen type I and IV, and virulence. The results showed no significant differences for all the experiments regarding growth, biofilm formation, adhesion to ECM components and virulence. However, the motility ability of the mutant strain was impaired when compared to the wild type (Figure 6). CLPs were described as part of flagella in *Bacillus amyloliquefaciens* FZB42 and their absence was described as

affecting motility, namely the swimming ability of the bacteria[17]. Additionally, several studies show that mutant bacteria with impaired motility present a decrease in adhesion, especially regarding flagellum impairment[43]. This result may indicate that the BCAL1524 is related to flagella functioning, but more assays are still necessary.

Additionally, adhesion assays using the cell line 16HBE14o- (bronchial epithelial cells) revealed that the mutant's ability to adhere is increased (Figure 6). As referred, the adhesion process is multifactorial, involving many proteins of different natures. It is plausible that the existence of several other proteins could mask the deletion of our gene of interest, with other proteins enhancing its activity, which would lead to the augmented adhesion capacity of the mutant and the wild-type. All in all, these results indicate the apparent importance of BCAL1524 for motility, but the influence of the absence of this protein for the other processes tested may be masked by the presence of other adhesins, including other CLPs of similar nature, which would also justify the increase in adhesion capacity of the mutant.

Expression and purification of BCAM0695 and BCAL1524 truncates

Along with the phenotypic characterization of the mutant, the expression and purification of several truncates from the proteins BCAM0695 and BCAL1524 was attempted: two for the BCAM0695, and two for the BCAL1524. For both proteins, the objective was to obtain a truncate of the protein except for the lipid box (BCAM0695_21_758 and BCAL1524_33_558) and a truncate without the GXY repetition domain (BCAM0695_320_758 and BCAL1524_116_558). Only the constructs from BCAL1524 were correct. The next step was to test the optimal expression conditions for each truncate. For that, overexpression assays were performed and the optimal conditions for each truncate were registered: for the BCAL1524_33_558, a time of incubation of 4 hours, with an IPTG concentration of 0.6 mM was the ideal. Concerning the BCAL1524_116_558 truncate, the optimal incubation time was also 4 hours, but with a needed concentration of 1 mM of IPTG.

The final aim was to obtain purified native truncates. For that, SUMO protease was incubated with the truncates for 1 hour at 30 ºC. Furthermore, a purification was performed to separate the cleaved SUMO protein from the native truncates. The efficient cleavage of the SUMO protein was achieved, which allowed the collection and storage of the native truncates, even though retaining some contamination.

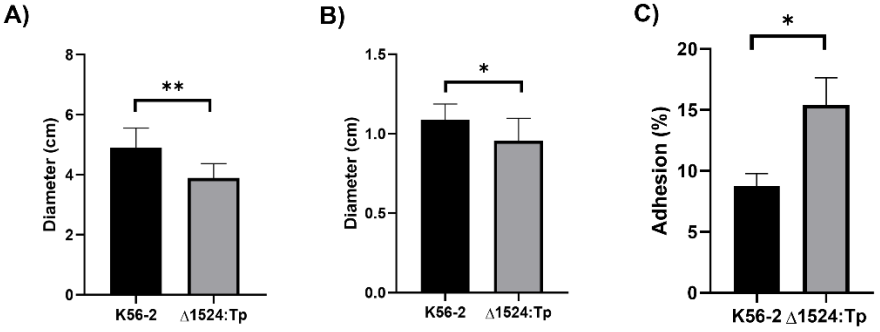


Figure 6. Results from the phenotypic characterization **A)** Swimming motility assay of wild-type *B. cenocepacia* K56-2 and *B. cenocepacia* Δ1524::Tp mutant. The graph translates the mean diameter with error bars corresponding to the standard deviation of the halo from three independent replicates. Swimming ability is impaired in the mutant (**, $P < 0.01$). **B)** Swarming motility assay of wild-type *B. cenocepacia* K56-2 and *B. cenocepacia* Δ1524::Tp mutant. The graph translates the mean diameter with error bars corresponding to the standard deviation of the halo from three independent replicates. Swarming ability is also altered in the mutant (*, $P < 0.05$). **C)** Adherence to 16HBE14o- epithelial cell line by the Δ1524::Tp mutant expressed as the percentage of adhered bacteria relative to the initial dose of bacteria applied to the monolayer of epithelial cells. Results are presented as the mean values with error bars from three independent experiments. A significative difference is observed between the wild-type and the mutant (*, $P < 0.05$).

Conclusions and Future Perspectives

Pimenta et al. performed a transcriptomic analysis during the first contact between *B. cenocepacia* K56-2 and human bronchial cells and identified that several collagen-like genes were present as overexpressed[27]. The aim of this work was then to study the existence and the characteristics of these CLPs in the Bcc. From the nineteen members of the Bcc, 75 proteins with the Col_mid_reg were identified. Of those, 93% possess TRs, a characteristic associated with pathogenicity[36,37]. Analysing the distribution per species, there seems to be a tendency for an increased percentage of tandem repeats and a higher level of pathogenicity. However, more studies are necessary to increase the understanding of these regions. Furthermore, we analyzed the cellular localization and nature of the proteins, concluding that the majority are membrane proteins, and being described as lipoproteins. Focusing on the proteins from *B. cenocepacia*, low complexity and disorder analysis showed a great prevalence of disordered regions in the 5 proteins, with these regions possessing high nucleic acid and proteins binding affinity, important for adhesion and infection processes[5,6]. These proteins, possessing low complexity regions and the Col_mid_reg, may suggest the existence of a different type of CLPs. Besides the CLPs possessing GXY domains that originate triple helixes and structured proteins, these unstructured proteins with a common central domain may possess interesting characteristics yet to be studied. These types of discoveries, adding to the existing evidence of overexpression of the genes *bcal1523*, *bcam0695* and *bcal1524* in adhesion assays[27] led us to create mutants in order to understand the eventual role and importance of these proteins during the adhesion process. A phenotypic analysis of the Δ1524::Tp mutant was performed, and the only impaired functions in the mutant when compared to the wild type *B. cenocepacia* K56-2 strain was motility (swimming and swarming were impaired) and the adhesion capacity to host cells, that was augmented in the mutant. This may indicate the importance of this gene for flagella's correct formation or function. The increase in adhesion capacity may be explained by the multitude of adhesins presented by the bacteria, especially the other 4 CLPs paralogs identified that may mask the absence of the BCAL1524. The possible alteration in the spatial composition of the membrane of the mutant is another hypothesis to explain this increase in the adhesion to human bronchial cells. The other processes of adhesion assessed were not altered. This can be explained by the multifactorial nature of the adhesion process, with the absence of just one being not enough to completely impair these processes. As mentioned before, the existence of four other CLPs with similar characteristics may result in a redundant activity that can compensate for the deletion of one specific protein. However, more studies are necessary to fully understand the importance of this protein. Further studies using protein truncates may be important to understand the structure, importance, and eventual potential of the different domains of the protein as a target to tackle this opportunistic pathogen. Additionally, more attempts to study other CLPs is necessary, since this will allow the study of the importance of GXY repeats (present in high number in the BCAM0695 protein) and the collagen-like domains in the virulence of *B. cenocepacia*, namely in adherence and invasion to the host cells or biofilm formation, as described in other organisms[8,10,35].

Acknowledgement: This document was written and made publicly available as an institutional academic requirement and as a part of the evaluation of the MSc thesis in Biotechnology of the author at Instituto Superior Técnico. The work described herein was performed at the Institute for Bioengineering and Biosciences of Instituto Superior Técnico (Lisbon, Portugal), during the period March-July 2022, under the supervision of Prof. Arsénio Fialho.

References

1. Stones, D. H. & Krachler, A. M. Against the tide : the role of bacterial adhesion in host colonization. *Biochem. Soc. Trans.* **44**, 1571–1580 (2016).
2. Berne, C., et al. Bacterial adhesion at the single-cell level. *Nat. Rev. Microbiol*. **16**, 616–627 (2018).
3. Haiko, J. & Westerlund-Wikström, B. The role of the bacterial flagellum in adhesion and virulence. *Biology (Basel)*. **2**, 1242–1267 (2013).
4. Girard, V. & Mourez, M. Adhesion mediated by autotransporters of Gram-negative bacteria: Structural and functional features. *Res. Microbiol*. **157**, 407–416 (2006).
5. Peng, Z. et al. Exceptionally abundant exceptions: Comprehensive characterization of intrinsic disorder in all domains of life. *Cell. Mol. Life Sci.* **72**, 137–151 (2015).
6. Xue, B. et al. Orderly order in protein intrinsic disorder distribution : disorder in 3500 proteomes from viruses and the three domains of life. *J. Biomol. Struct. Dyn*. **30**, 137-149 (2012).
7. Mier, P. & Andrade-Navarro, M. A. The conservation of low complexity regions in bacterial proteins depends on the pathogenicity of the strain and subcellular location of the protein. *Genes (Basel)*. **12**, (2021).
8. Yu, Z., et al. Bacterial collagen-like proteins that form triple-helical structures. *J Struct Biol* **186**, 451–461 (2014).

9. Paterson, G. K., et al.. PclA, a pneumococcal collagen-like protein with selected strain distribution, contributes to adherence and invasion of host cells. *FEMS Microbiol. Lett.* **285**, 170–176 (2008).
10. Chen, S. M. *et al.* Streptococcal collagen-like surface protein 1 promotes adhesion to the respiratory epithelial cell. *BMC Microbiol.* **10**, 320 (2010).
11. Rasmussen, M., et al. Genome-based Identification and Analysis of Collagen-related Structural Motifs in Bacterial and Viral Proteins. *J. Biol. Chem.* **278**, 32313–32316 (2003).
12. Koonin, E. V., et al. Horizontal Gene Transfer in Prokaryotes: Quantification and Classification. *Annu. Rev. Microbiol.* **55**, 709–742 (2001).
13. Lukomski, S., et al. Collagen-like proteins of pathogenic streptococci. *Mol. Microbiol.* **103**, 919–930 (2017).
14. McElroy, K, et al. Characterisation of a large family of polymorphic collagen-like proteins in the endospore-forming bacterium Pasteuria ramosa. *Res. Microbiol.* **162**, 701–714 (2011).
15. Lukomski, S. et al. Identification and characterization of the scl gene encoding a group A Streptococcus extracellular protein virulence factor with similarity to human collagen. *Infect. Immun.* **68**, 6542–6553 (2000).
16. Rasmussen, M., et al. SclA, a Novel Collagen-Like Surface Protein of Streptococcus pyogenes. *Infect. Immun.* **68**, 6370–6377 (2000).
17. Zhao, X., et al. The new flagella-associated collagen-like proteins ClpB and ClpC of Bacillus amyloliquefaciens FZB42 are involved in bacterial motility. *Microbiol. Res.* **184**, 25–31 (2016).
18. Bachert, B. et al. A Unique Set of the Burkholderia Collagen- Like Proteins Provides Insight into Pathogenesis , Genome Evolution and Niche Adaptation, and Infection Detection. *PLoS One* **10**, 1–36 (2015).
19. Coenye, T., et al. J. Taxonomy and identification of the Burkholderia cepacia complex. *J. Clin. Microbiol.* **39**, 3427–3436 (2001).
20. Isles, A. *et al.* Pseudomonas cepacia infection in cystic fibrosis: An emerging problem. *J. Pediatr.* **104**, 206–210 (1984).
21. Mahenthiralingam, E., et al. Burkholderia cepacia complex infection in patients with cystic fibrosis. *J. Med. Microbiol.* **51**, 533–538 (2002).
22. Drevinek, P. & Mahenthiralingam, E. Burkholderia cenocepacia in cystic fibrosis: Epidemiology and molecular mechanisms of virulence. *Clin. Microbiol. Infect.* **16**, 821–830 (2010).
23. McClean, S. & Callaghan, M. Burkholderia cepacia complex: Epithelial cell-pathogen confrontations and potential for therapeutic intervention. *J. Med. Microbiol.* **58**, 1–12 (2009).
24. Mil-Homens, D. & Fialho, A. M. Trimeric autotransporter adhesins in members of the Burkholderia cepacia complex: A multifunctional family of proteins implicated in virulence. *Front. Cell. Infect. Microbiol.* **1**, 1–10 (2011).
25. Mil-Homens, D. & Fialho, A. M. A BCAM0223 mutant of Burkholderia cenocepacia is deficient in hemagglutination, serum resistance, adhesion to epithelial cells and virulence. *PLoS One* **7**, (2012).
26. Pimenta, A. I., Mil-Homens, D. & Fialho, A. M. Burkholderia cenocepacia–host cell contact controls the transcription activity of the trimeric autotransporter adhesin BCAM2418 gene. *Microbiologyopen* **9**, 1–12 (2020).
27. Pimenta, A. I., et al M. *Burkholderia cenocepacia* transcriptome during the early contacts with giant plasma membrane vesicles derived from live bronchial epithelial cells. *Sci. Rep.* **11**, 1–16 (2021).
28. Winsor, G. L. *et al.* The Burkholderia Genome Database : facilitating flexible queries and comparative analyses. *Bioinformatics* **24**, 2803–2804 (2008).
29. Lemoine, F. *et al.* NGPhylogeny.fr: New generation phylogenetic services for non-specialists. *Nucleic Acids Res.* **47**, W260–W265 (2019).
30. Newman, A. M. & Cooper, J. B. XSTREAM: A practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC Bioinformatics* **8**, 1–19 (2007).
31. Hu, G. *et al.* flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. *Nat. Commun.* **12**, 1–8 (2021).
32. Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.* **46**, D493–D496 (2018).
33. Zheng, W. *et al.* Folding non-homologous proteins by coupling deep- learning contact maps with I-TASSER assembly simulations ll Folding non-homologous proteins by coupling deep-learning contact maps with I-TASSER assembly simulations. *Cell Reports Methods* **1**, 100014 (2021).
34. Cozens, A. L. *et al.* CFTR expression and chloride secretion in polarized immortal human bronchial epithelial cells. *Am. J. Respir. Cell Mol. Biol.* **10**, 38–47 (1994).
35. Xu, Y., et al. Streptococcal Scl1 and Scl2 Proteins Form Collagen-like Triple Helices. **277**, 27312–27318 (2002).
36. Zhou, K., Aertsen, A. & Michiels, C. W. The role of variable DNA tandem repeeats in bacterial adaptation. *FEMS Microbiol. Rev.* **10**, 119–141 (2014).
37. Saravanan, K. M. Sequence and structural analysis of fibronectin - binding protein reveals importance of multiple intrinsic disordered tandem repeats. *J. Mol. Recognit.* **32**, 1–9 (2019).
38. Mahenthiralingam, E., et al. The multifarious, multireplicon Burkholderia cepacia complex. *Nat. Rev. Microbiol.* **3**, 144–156 (2005).
39. Holden, M. T. G. *et al.* The genome of Burkholderia cenocepacia J2315, an epidemic pathogen of cystic fibrosis patients. *J. Bacteriol.* **91**, 261–277 (2009).
40. Nykyri, J. *et al.* Role and Regulation of the Flp/Tad Pilus in the Virulence of Pectobacterium atrosepticum SCRI1043 and Pectobacterium wasabiae SCC3193. *PLoS One* **8**, (2013).
41. Alteri, C. J. *et al.* The Flp type IV pilus operon of Mycobacterium tuberculosis is expressed upon interaction with macrophages and alveolar epithelial cells. *Front. Cell. Infect. Microbiol.* **12**, 1–15 (2022).
42. Habchi, J., et al. Introducing protein intrinsic disorder. *Chem. Rev.* **114**, 6561–6588 (2014).
43. Chaban, B., Hughes, H. V. & Beeby, M. The flagellum in bacterial pathogens: For motility and a whole lot more. *Semin. Cell Dev. Biol.* **46**, 91–103 (2015).