INSTITUT PASTEUR

Research in Microbiology 162 (2011) 701–714

www.elsevier.com/locate/resmic

Characterisation of a large family of polymorphic collagen-like proteins in the endospore-forming bacterium *Pasteuria ramosa*

Kerensa McElroy [a,*], Laurence Mouton [b], Louis Du Pasquier [c], Weihong Qi [d], Dieter Ebert [c]

[a] *School of Biotechnology and Biomolecular Science, University of New South Wales, Sydney, New South Wales 2052, Australia*
[b] *Université de Lyon, Université Lyon1, Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, 43 Bd du 11 Novembre 1918, 69622 Villeurbanne Cedex, France*
[c] *Zoologisches Insitut der Universität Basel, Evolutionsbiologie, Vesalgasse 1, 4051 Basel, Switzerland*
[d] *Functional Genomics Center Zurich, Winterthurerstrasse 190/Y32 H66, 8057 Zürich, Switzerland*

Received 18 February 2011; accepted 4 May 2011
Available online 21 June 2011

Abstract

Collagen-like proteins containing glycine-X-Y repeats have been identified in several pathogenic bacteria potentially involved in virulence. Recently, a collagen-like surface protein, Pcl1a, was identified in *Pasteuria ramosa*, a spore-forming parasite of *Daphnia*. Here we characterise 37 novel putative *P. ramosa* collagen-like protein genes (PCLs). PCR amplification and sequencing across 10 *P. ramosa* strains showed they were polymorphic, distinguishing genotypes matching known differences in *Daphnia/P. ramosa* interaction specificity. Thirty PCLs could be divided into four groups based on sequence similarity, conserved N- and C-terminal regions and G-X-Y repeat structure. Group 1, Group 2 and Group 3 PCLs formed triplets within the genome, with one member from each group represented in each triplet. Maximum-likelihood trees suggested that these groups arose through multiple instances of triplet duplication. For Group 1, 2, 3 and 4 PCLs, X was typically proline and Y typically threonine, consistent with other bacterial collagen-like proteins. The amino acid composition of Pcl2 closely resembled Pcl1a, with X typically being glutamic acid or aspartic acid and Y typically being lysine or glutamine. Pcl2 also showed sequence similarity to Pcl1a and contained a predicted signal peptide, cleavage site and transmembrane domain, suggesting that it is a surface protein.

Keywords: Daphnia; Collagen; Bacterial proteins; Virulence; *Pasteuri*a

1. Introduction

Collagens are a family of animal proteins with a triple-helical coiled structure, featuring a glycine-X-Y repeat motif where X and Y are any amino acid (Beck and Brodsky, 1998; Hulmes, 1992; Ramachandran, 1988). In addition to forming connective tissues, proteins containing G-X-Y repeat motifs (including mammalian lectins and human complement factor C1q) are implicated in host immune defences (Reid, 1993; Sellar et al., 1991).

Bacterial collagen-like proteins (CLPs) containing G-X-Y repeat motifs have also been identified, including in *Streptococcus* and *Bacillus* species (Lukomski et al., 2000, 2001; Sylvestre et al., 2002; Waller et al., 2005; Whatmore, 2001). Bacterial CLPs are typically localised to the cell surface (Rasmussen et al., 2000; Thompson and Stewart, 2008). SclC and BclA, CLPs found in *Streptococcus equi* and *Bacillus anthracis* respectively, are known to be antigenic (Karlstrom et al., 2006; Steichen et al., 2003). CLP from *Streptococcus* species are also involved in cell adhesion, invasion and intracellular signalling, interacting with human integrins, lipoproteins and complement proteins (Caswell et al., 2008; Han et al., 2006; Humtsoe et al., 2005). Collectively, these findings suggest bacterial CLP are involved in pathogenicity, immune response elicitation and host-parasite interactions, possibly evolving as mimics of host proteins containing G-X-Y motifs.

* Corresponding author.
E-mail addresses: kerensa@unsw.edu.au (K. McElroy), mouton@biomserv.univ-lyon1.fr (L. Mouton), dupasquier@dial.eunet.ch (L. Du Pasquier), weihong.qi@fgcz.ethz.ch (W. Qi), dieter.ebert@unibas.ch (D. Ebert).

0923-2508/$ - see front matter © 2011 Institut Pasteur. Published by Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.resmic.2011.06.009

A CLP, Pcl1a, was recently identified in the endospore-forming bacterium *Pasteuria ramosa* (Mouton et al., 2009). *P. ramosa* is a highly virulent, obligate parasite of freshwater crustaceans belonging to the *Daphnia* genus. Transmission is strictly horizontal, as infection causes complete sterilisation (Ebert et al., 2004). Infection is ultimately lethal, with several million spores released from the host's body cavity at death.

The high degree of genotype specificity in *Daphnia* clone/*P. ramosa* strain associations suggests these species may have coevolved (Carius et al., 2001; Decaestecker et al., 2007; Ebert, 2008). The mode of infection is not fully understood; however, spore surface proteins are likely to be involved in initial adhesion to the host. In the closely related species *Pasteuria penetrans* (a parasite of root-knot nematodes), it has been suggested that CLP on the endospore surface may interact with mucins on the nematode cuticle, in a velcro-like attachment process (Davies, 2009). The discovery of Pcl1a suggests that CLPs may also play a role in host adhesion and specificity in the *P. ramosa*/*Daphnia* model system. Pcl1a was isolated from the spore's surface and exhibits polymorphism between isolates with differential infectivity, supporting this hypothesis (Mouton et al., 2009). In the present study, we characterise the full extent of the *P. ramosa* collagen-like protein (PCL) gene family. To extend our understanding of the role of PCLs in *Daphnia/P. ramosa* interaction specificity, we investigate PCL polymorphisms for 10 *P. ramosa* strains extracted from *Daphnia* hosts with varying genotypes, geographical origins and *P. ramosa* susceptibility. A bioinformatic analysis of PCL relationships and structural features is also provided.

2. Materials and methods

2.1. P. ramosa *strains and spore production*

Five *P. ramosa* laboratory clones, two *P. ramosa* laboratory isolates and three *P. ramosa* field isolates were used, representing various geographical origins, host species and differential infectivity (Table 1). Clones were grown from single-spore infections achieved through either micromanipulation or limited dilution infections. Isolates were obtained from infected females, possibly containing more than one *P. ramosa* genotype. For spore production, *Daphnia magna* clones were kept under standardised conditions in artificial culture medium at 20 °C (Decaestecker et al., 2004). *Daphnia* were infected by exposing 20 juveniles of the appropriate clonal host population to 1×10^6 spores in a 400 mL jar. Daphnia were fed daily with 5×10^6 algal *Scenedesmus* spp. cells. Spores were harvested four weeks after infection.

2.2. Spore cleaning

For each *Pasteuria* strain, four infected *Daphnia* were crushed to extract spores. Spores were centrifuged at 10,000 g for 1 min and the supernatant discarded, before resuspension in 400 µl of lysis solution (Lysozyme 20 mg/µl, 50 mM Tris−Hcl, 10 mM EDTA, pH8.0). Spores were incubated at 37 °C with shaking for 60 min, followed by the addition of 40 µl 20% SDS and incubation for 60 min at 37 °C with shaking. Spores were washed twice as follows: pelleted by centrifugation for 2 min (maximum speed), supernatant discarded, then resuspension in 500 µl Tris−HCl (10 mM, pH7.0). Washed spores were pelleted again and resuspended in 300 µl proteinase K buffer (50 mM Tris−HCl, 5 mM $CaCl_2$, 4 M urea, 5 mM $CaCl_2$, pH8.0) to which was added 30 µl Proteinase K (20 mg/µl). Another two rounds of spore washing followed incubation for 1 h at 56 °C with shaking. Pelleted spores were resuspended in 450 µl H_2O and incubated at 95 °C for 15 min without shaking. 50 µl DNAseI buffer (100 mM Tris, 25 mM $MgCl_2$, 5 mM $CaCl_2$, pH7.5) and 2 µl DNAseI (10 mg/mL) was added to cooled samples, followed by incubation at 37 °C for 1 h without shaking. 2 µl of EDTA (50 mM) was then added, prior to 15 min heating (70 °C). Spores were washed twice more, before resuspension in 300 µl TE buffer (10 mM Tris−HCl, 1 mM EDTA, pH8.0).

Table 1
Summary of *P. ramosa* strains used for PCL polymorphism analysis.

Strain	Strain Type	Original host, site of origin	Current host clone (passaged)
C1	Clone	*Daphnia magna,* Moscow, RM 2, Russia	*Daphnia magna* HO2, Hungary
C14	Clone	*Daphnia magna,* VIW-2, Tvärminne, Finland	*Daphnia magna* AL1, Tvärminne Finland
C18	Clone	*Daphnia magna,* VIW-2, Tvärminne, Finland	*Daphnia magna* HO2, Hungary
$P4_{HO2}$	Lab isolate	*Daphnia magna,* OM2, Belgium	*Daphnia magna* HO2, Hungary
P12	Field isolate	*Daphnia magna,* VIW-4, Tvärminne, Finland	Not Applicable
C19	Clone	*Daphnia magna,* Garzerfeld, DG, North Germany	*Daphnia magna* Xinb3, Tvärminne Finland
C20	Clone	*Daphnia magna*, Kaimes UK	*Daphnia magna* HO2, Hungary
$P4_{FX}$	Lab isolate	*Daphnia magna,* OM2, Belgium	*Daphnia magna* FX, Tvärminne Finland
P8	Field isolate	*Daphnia longispina*, Mekkojärvi, Finland	Not Applicable
P10	Field isolate	*Daphnia longispina,* FO-21, Tvärminne, Finland	Not Applicable

2.3. DNA extraction

160 mg of 0.1 mm zirconian beads and 30 µl of proteinase K were added to spore suspensions, followed by beating at full speed for 20 s (FastPrep FP120, Bio101 Savant, Holbrook, NY, USA). After incubation at 56 °C for 30 min, 20 µl Rnase A (20 mg/mL) was added to supernatant followed by 5 min bench incubation. 350 mL phenol chloroform was then added and mixed by inversion, followed by centrifugation (full speed, 5 min). The aqueous layer was decanted and the phenol chloroform extraction repeated. This process was repeated once more using 350 µl of pure chloroform. 30 µl NaOAc (pH5.2 3 M) and 900 µl EtOH (100%) were added to the final preparation before overnight freezing (−20 °C). After 30 min centrifugation (full speed, 4 °C) supernatant was removed and discarded and the pellet washed twice in 70% EtOH. After

bench drying, pelleted DNA was resuspended in 25 µl TE buffer (10 mM Tris, 1 mM EDTA, pH8.0).

2.4. PCR

PCR amplifications were performed using the BD Advantage 2 PCR kit (BD Biosciences Clontech). Reaction concentrations (25 µl volume) were: 1× Advantage 2 PCR buffer, 200 µM of each dNTP, 400 µM of each primer and 1× Advantage 2 polymerase mix. Reactions were incubated at 95 °C (1 min), then cycled 35 times at 95 °C (30 s), 48–59 °C (30 s), and 68 °C (1 min/kb), followed by a final incubation at 68 °C (10 min). PCR products were purified (GenEluteTM PCR Clean-up Kit, Sigma) and sequenced (Macrogen, Korea; Fasteris SA, Inc., Plan-les-Ouates). Amplification primers and Genbank accessions are given in Table 2.

2.5. Bioinformatics

Bioinformatic analysis was performed on GeneMark gene predictions (Lukashin and Borodovsky, 1998) from the draft *P. ramosa* genome. The draft genome consists of 462812 454-FLX reads, assembled into 1641 contigs with length greater than 500 bp (average length 2219 bp, largest contig 46579 bp), covering 3642728 bp. HMMER [http://hmmer.janelia.org/] was used with a collagen Pfam alignment (PF01391) to search for potential PCLs. Predicted proteins from fully sequenced bacteria were downloaded from NCBI and searched for sequences annotated as CLPs using a custom made script.

Predicted amino acid sequences were aligned using ClustalX 2.0.12 (default parameters) (Larkin et al., 2007). This program creates a dendrogram of the most closely related groups of sequences at each branch level according to pairwise alignment, using this dendrogram as a guide for the final multiple alignment. This approach can reveal related clusters of proteins, even when there may be substantial protein length variation. Care must be taken, however, in the downstream interpretation of trees built from such alignments, and we make no assumption that the tree branch lengths accurately represent evolutionary distances. To validate the clustering of sequences and position of branches, bootstrapping (100 replicates) was performed on a neighbour-joining tree created from the ClustalX alignment using PHYLIP package programs (Felsenstein, 1989).

PCLs adhering to the genomic arrangement of a Group 1 PCL, followed by a Group 2 PCL, followed by a Group 3 PCL were used to investigate the hypothesis that these subfamilies arose through multiple rounds of genomic PCL triplet duplication. To facilitate analysis, PCLs clusters where only two of the three groups were represented were excluded. PHYLIP was used to generate maximum-likelihood trees for PCL DNA sequences within each group. The topology of these three trees was compared to a maximum-likelihood tree from a ClustalX alignment of the concatenated DNA sequences of the PCLs within each genomic triplet. Topological similarity between all four trees is interpreted as evidence for the gene duplication hypothesis. 100 bootstrapping replicates were performed for all trees described above.

Transmembrane regions were predicted using the TMHMM Server 2.0 (Krogh et al., 2001). Cleavage sites and signal peptides were predicted with SignalP and TatP (Bendtsen et al., 2004, 2005). G-X-Y repeat length structure and average X and Y composition were calculated manually. For polymorphism analysis, SNPs and percentage identity were identified using BLAST. Tandem repeat regions were predicted using the program XSTREAM (Newman and Cooper, 2007). Parameters were set to detect any tandem repeats with a basic repeat of at least 6 amino acids (Min period = 6), repeated at least 1.5 times (min length = 9) (this corresponds to a basic unit corresponding to two G-X-Y repeats). Up to two gaps were allowed in the tandem repeat and the minimum word match and minimum consensus match were both set to 0.7 (as recommended by the authors for moderately conserved repeats).

BINDIGO was used to detect SDs and calculate the optimal binding and free energy (ΔG_{SD}) between the anti-SD sequence AUCACCUCCUUU and the 30 bp immediately upstream of predicted start codons. As used by other authors, a $\Delta G_{SD} < -4.4$ kcal/mol defined a possible SD (Hodas and Aalberts, 2004; Ma et al., 2002).

3. Results

3.1. Identification of new PCLs

A Hidden Markov Model search of predicted genes from the *P. ramosa* draft genome revealed 41 potential PCLs in addition to Pcl1a. A manual search for G-X-Y regions revealed an extra two open reading frames (ORFs) containing G-X-Y repeats, bring the total number of potential PCLs to 43. Four of these genes were missing either their 5′ or 3′ end, due to the gene's location at the extreme of a contig. These partial genes were not analysed further (the position of one is given, however, in Fig. 1, to show its close proximity to other PCLs). The presence of the remaining 39 potential PCLs was confirmed by sequencing of PCR-amplified genes in the *P. ramosa* strain C1 (Table 2). For the two potential PCLs not identified in the HMM search, no in-frame start codons could be identified (ATG, GTG, and TTG were considered as possibilities – in the closely related species *Bacillus subtilis*, these codons account for >99% of start codon usage) (Rocha et al., 1999). These ORFs may be pseudogenes or they may start with rare alternative codons. One of these ORFs is presented in Fig. 1, as it lies in close proximity to other PCLs (marked with dashed outline). With this exception, both were excluded from further analysis, leaving 37 new putative PCLs (Table 2).

To place PCL genes in a broader context, we searched for CLPs in all fully sequenced bacterial genomes. 130 (10%) of 1298 genomes contained sequences annotated as CLPs. CLP copy numbers varied greatly, both between species and between strains of the same species. Most genomes ($n = 110$) contained only one to three CLPs. The soil bacterium *Conexibacter woesei* came closest to approaching the high number

Table 2
Summary of the 38 confirmed PCLs.

Daphnia strain		C1	C14	C18	$P4_{HO2}$	P12	C19	C20	$P4_{FX}$	P8	P10
Country of origin		RUS	SF1	SF1	B	SF1	D	UK	B	SF3	SF2
Pcl15 (HQ010378)	%id	100	100	100	100	100	98	98	96	86	87
GCCTTGGTTACCTTCAGGAC	a.a. no.	397	397	397	397	397	398	398	398	400	401
TGCCATGTGAGGTGAAAATC	Gxy no.	48	48	48	48	48	48	48	48	48	48
Pcl24 (HQ010387)	%id	100	100	100	100	100	84	84	84	84	89
AAATTGTCGAATGCCATGTG	a.a. no.	378	378	378	378	378	376	376	373	373	380
GACCTGTGGGACCTTGATCT	Gxy no.	49	49	49	49	49	49	49	49	49	49
Pcl29 (HQ010392)	%id	100	100	100	100	100	97	97	83	76	92
GCCACCAACCTCTTTTCGTA	a.a. no.	383	383	383	383	383	374	374	333	440	362
TCGATTGATCGCATTCAGTT	Gxy no.	81	81	81	81	81	78	78	65	100	74
Pcl3 (HQ010366)	%id	100	100	100	100	100	100	100	100		96
TGAAATTATGTAGCGATTTATTTTGTT	a.a. no.	417	417	417	417	417	417	417	417	-	417
AACAGCACTACCCCACAACC	Gxy no.	50	50	50	50	50	50	50	50		50
Pcl19 (HQ010382)	%id	100	100	100	100	100	99	99	100	96	
TCAATCTGGATGGTAGATTATGTGA	a.a. no.	254	254	254	254	254	254	254	254	254	-
GCCACAATGTCTATACATAGGGTTA	Gxy no.	15	15	15	15	15	15	15	15	15	
Pcl27 (HQ010390)	%id	100	100	100	100	100	100	100			77
CCCAAATCATTCATTGCTGT	a.a. no.	315	315	315	315	315	315	315	-	-	313
TCGTGCCCATAACAAACACT	Gxy no.	48	48	48	48	48	48	48			48
Pcl20 (HQ010383)	%id	100	100	100	100	100	100	100	100		
ATTGTGTTTCCGAAGTTTGC	a.a. no.	253	253	253	253	253	253	253	253	-	-
CAAGCAATGGATCATCTGAA	Gxy no.	15	15	15	15	15	15	15	15		
Pcl30 (HQ010393)	%id	100	100	100	100	100	99	99	99		
AATCAAAAAGGGTAAATAAGGTAAAAA	a.a. no.	251	251	251	251	251	251	251	251	-	-
TCCTAAATTCATAAAATGGCTTTGA	Gxy no.	18	18	18	18	18	18	18	18		
Pcl10 (HQ010373)	%id	100	100	100	100	100	99	99			
TGGAAATTACGGTGCCTACA	a.a. no.	332	332	332	332	332	332	332	-	-	-
GGTGACCCAAGCTTAATTCG	Gxy no.	46	46	46	46	46	46	46			
Pcl14 (HQ010377)	%id	100	100	100	100	100	100	100	-	-	-
AGTTTCCTGTATGGTGTGTTGC	a.a. no.	345	345	345	345	345	345	345	-	-	-
TTGATCTGGGTTTATATCCCTGT	Gxy no.	48	48	48	48	48	48	48	-	-	-
Pcl16 (HQ010379)	%id	100	100	100	100	100	99	99	-	-	-
TGCACAGAAGACGGTGAACT	a.a. no.	321	321	321	321	321	321	321	-	-	-
TCCAGCAAATAGGATAGGGAAT	Gxy no.	47	47	47	47	47	47	47	-	-	-
Pcl4 (HQ010367)	%id	100	100	100	100	100			100		
GGCCTAATTCATTCATTGCTG	a.a. no.	323	323	323	323	323	-	-	323	-	-
TCGAATAGCTGTATGAACCAACA	Gxy no.	50	50	50	50	50			50		
Pcl2 (HQ010365)	%id	100	100	100		100	97		97		
GTTGGGGGAGACATTATCCA	a.a. no.	371	371	371	-	371	380	-	380	-	-
TCTGGAATATAAAAAGATTGTAGTTGC	Gxy no.	43	43	43		43	46		46		
Pcl9 (HQ010372)	%id	100	100		100	100	100	100			
TGTAACTGCCGCAACTTTCT	a.a. no.	420	420	-	420	420	420	420	-	-	-
CAGATAAGGGCGGTATGGAA	Gxy no.	47	47		47	47	47	47			
Pcl11 (HQ010374)	%id	100			100	100	98	98			
AGTTTCCTGTATGGTGTGTTGC	a.a. no.	335	-	-	335	335	332	332	-	-	-
ATAACCAGCGGGACCTGTAG	Gxy no.	48			48	48	47	47			
Pcl1a (HQ010364)	%id	100	100				97				
TTATATAAATAAAGGGGATTGGATTTT	a.a. no.	637	637	*	*	*	625	*	*	*	*
TTGGAAATAATAGTGAAACCCAATC	Gxy no.	185	185				181				
Pcl28 (HQ010391)	%id	100	100				90				
ACCATATAGTCCTCTTAACGC	a.a. no.	412	412	*	*	*	415	*	*	*	*
GCATTGTGTTTTGTTCTGTT	Gxy no.	49	49				49				
Pcl5 (HQ010368)	%id	100					100		100		
AATAACAAAATGCCGTGTGAGG	a.a. no.	259	*	*	*	*	259	*	259	*	*
CTATAACGCTTGGTCCTGTCG	Gxy no.	33					33		33		
Pcl6 (HQ010369)	%id	100					98				
ACCACTTGGACCTTGATTACCC	a.a. no.	344	*	*	*	*	344	*	*	*	*
AAATAGTTTCCTGTATGGTGTAAAGC	Gxy no.	50					50				
Pcl7 (HQ010370)	%id	100					100				
GCATTTAATAGATGGCTCAATCC	a.a. no.	329	*	*	*	*	329	*	*	*	*
CAGGTGGCACAATCTTTGG	Gxy no.	48					48				

Columns are arranged according to genotypes defined by PCL polymorphism between *P. ramosa* strains from various *Daphnia* hosts. PCLs are given in rows according to the success rate of PCR amplification in tested strains (those amplified in all strains are given first). Genbank accession numbers follow the PCL name, with primers for amplification given below (reverse in italics). *Daphnia* sites of origin are labelled with a country-code (D-Germany, SF-Finland, RUS-Russia, B-Belgium and UK-United Kingdom) with the number giving the site within the country (for more details see Table 1). SF1 and SF2 are from two sampling locations very close to each other (about 3 km apart), but with differing *Daphnia* hosts. SF2 and SF3 are from *D. longispina*, all other samples were isolated from *D. magna*. Within each gene, strains identically shaded are also identical at the nucleotide level. '−' = no PCR product. '*' = PCR amplification not attempted. %id refers to the percentage of amino acid identities of an alignment of each gene in each strain to the corresponding gene in C1, rounded down to the nearest integer.
[a]Amplifies partial gene. Use with GCATTAAGATACCAGCCAG, *CAGATTAAGTTTTTGGGCAAGCTC*.
[b]Amplifies partial gene. Use with GGGAATACAGGCACAACAGG, *CGCTTGTGCCCAACTCTTT*.

Table 2 (*continued*).

Pcl8 (HQ010371)	%id	100					99				
ATATCGTAAAGAAATAAAAGAAGTTGC	a.a. no.	426	*	*	*	*	426	*	*	*	*
ATACCTCAGCATCAGCCTACG	Gxy no.	51					51				
Pcl13 (HQ010376)	%id	100					100				
AATCGGATAATAACCGTAAATGC	a.a. no.	323	*	*	*	*	323	*	*	*	*
AAAAGAAACAACAATTATGATAAGAGC	Gxy no.	47					47				
Pcl17 (HQ010380)	%id	100					99				
CCTAACACACAACCAACCATC[a]	a.a. no.	439	*	*	*	*	439	*	*	*	*
TTGGTTGCACCTCTTAATCC	Gxy no.	52					52				
Pcl18 (HQ010381)	%id	100					96				
GCAAATGATACAACCATATTCG	a.a. no.	302	*	*	*	*	312	*	*	*	*
GAAACAAGTTCATTATTGGAGACAAA	Gxy no.	18					18				
Pcl26 (HQ010389)	%id	100					94				
GTATATGGTTATGTAAATATCAGG[b]	a.a. no.	412	*	*	*	*	409	*	*	*	*
ATCGTTGGTAGCTGCATTTG	Gxy no.	45					46				
Pcl33 (HQ010396)	%id	100					95				
TATTCGGAAGGTGGTTGG	a.a. no.	140	*	*	*	*	140	*	*	*	*
GAACTTGTTAAGGTTTTTAATATAGG	Gxy no.	11					11				
Pcl34 (HQ010397)	%id	100					100				
AAGGATATGTGTGATAAATTATTTGG	a.a. no.	286	*	*	*	*	286	*	*	*	*
CTTTTTAGCCTATTTTATGTACG	Gxy no.	39					39				
Pcl32 (HQ010395)	%id	100					100				
GTTGAAAAAGATAAAACAATAAAGG	a.a. no.	256	*	*	*	*	256	*	*	*	*
GGCTCTGTTGCTACAAAACC	Gxy no.	19					19				
Pcl36 (HQ010399)	%id	100					92				
ATGTCTTCCCTAATATGTCAAG	a.a. no.	355	*	*	*	*	340	*	*	*	*
CGTTTATTCCTGTGTTACCTAC	Gxy no.	46					48				
Pcl37 (HQ010400)	%id	100					81				
TACTATGGATCAGGAACATTG	a.a. no.	326	*	*	*	*	326	*	*	*	*
AAACAGAATTACGCTCTTTATC	Gxy no.	47					47				
Pcl38 (HQ010401)	%id	100					98				
AAACAAACATCCAACTACAACC	a.a. no.	421	*	*	*	*	421	*	*	*	*
ATTAGGTTTTTGAACTCCTTTTAC	Gxy no.	46					46				
Pcl12 (HQ010375)	%id	100									
TATTGGGACCATCCTCCTGG	a.a. no.	421	*	*	*	*	-	*	*	*	*
TAAAGCAATTAACTCTTCTAATGTCG	Gxy no.	46									
Pcl21 (HQ010384)	%id	100									
ACAATATCTACTCGGTACTTTACC	a.a. no.	342	*	*	*	*	-	*	*	*	*
GGCGAAGTTTTCCTATCC	Gxy no.	50									
Pcl22 (HQ010385)	%id	100									
AAAAGATGGGATAGGAAAAC	a.a. no.	325	*	*	*	*	-	*	*	*	*
TTCAACAAATGACTCAAACC	Gxy no.	47									
Pcl23 (HQ010386)	%id	100									
TCACCGCAGGATATACTATTGG	a.a. no.	421	*	*	*	*	-	*	*	*	*
TATCTGCCAGTGATCTTTTACAAG	Gxy no.	48									
Pcl25 (HQ010388)	%id	100									
ATCCCATTTATCTCCCTTCAGTAAC	a.a. no.	316	*	*	*	*	-	*	*	*	*
TTCAATACATGACTTAAACCTACGG	Gxy no.	47									
Pcl31 (HQ010394)	%id	100									
CATCACTAGATAAAAATGCTATACAAGG	a.a. no.	231	*	*	*	*	-	*	*	*	*
TATGAAATGGTTGCGTAAATGG	Gxy no.	8									
Pcl35 (HQ010398)	%id	100									
CCAATGGAAGGTAATACGG	a.a. no.	320	*	*	*	*	-	*	*	*	*
ATGGCAATAGAAGCAAGC	Gxy no.	49									
Genotype ID:		1	1	1	1	1	2	2	3	4	5

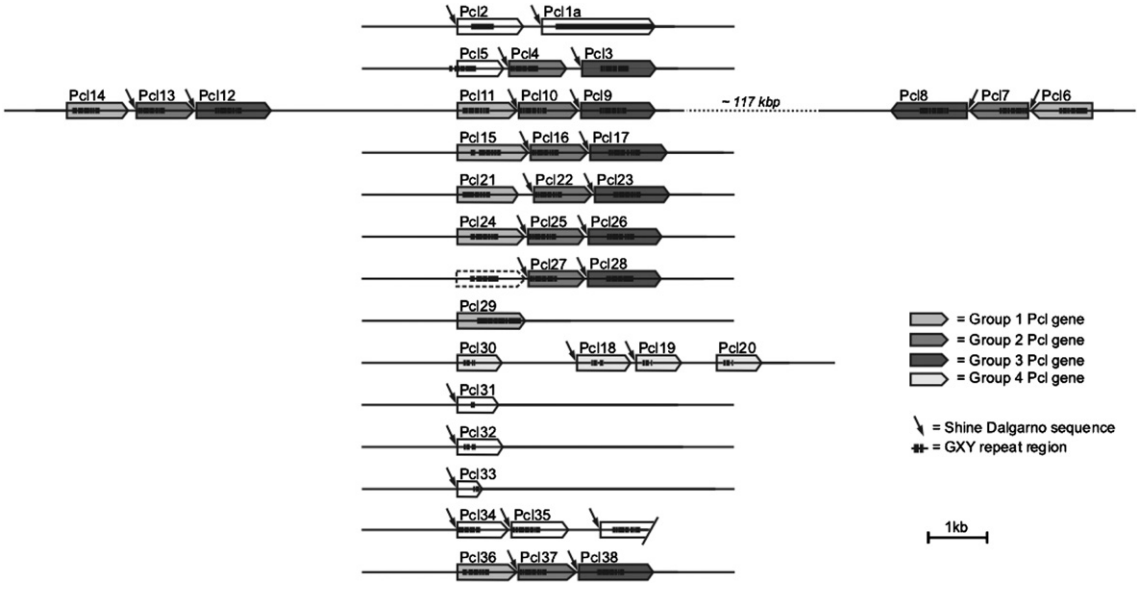

Fig. 1. Location of PCLs within the genome. One partial PCL is included, to show conservation of triplet clustering for Pcl34 and Pcl35. A potential PCL pseudogene has also been included, showing conservation of clustering for Pcl27 and Pcl28.

of CLP found in *P. ramosa*, with a total of 19 putative CLPs. Thus, *P. ramosa* is highly unusual, having by far the most CLPs known in any bacterium.

To reinforce that predicted PCLs represent translated genes, the 30 bp upstream of predicted start codons was searched for Shine-Dalgarno translation initiation sequences (SDs) (89.4% of *B. subtilis* genes feature SDs) (Ma et al., 2002). SDs were identified in close proximity to 28 of the 38 PCL genes (including Pcl1a) (Fig. 1). For Pcl5, an ATG start codon strongly supported by an SD was identified within the G-X-Y region. An alternative start codon, GTG, was also identified upstream of the G-X-Y region; however, this start codon was not supported by an SD, so the ATG start codon is accepted as the most likely start codon.

In comparison to Pcl1a, which is 637 amino acids long and has 185 G-X-Y repeats, the 37 novel PCLs have both fewer amino acids and G-X-Y repeats. Total putative protein length varied from 140 to 431 amino acids, whilst G-X-Y repeat number varied from eight to 81, although the majority of PCLs have around 50 G-X-Y repeats (Table 2).

3.2. PCL family structure

A neighbour-joining tree shows that 30 of the 38 PCLs can be grouped into subfamilies based on amino acid sequence similarity (Fig. 2). Bootstrap values at branches defining Group 1, Group 2, Group 3, and Group 4 PCLs strongly support the subfamily structure. Additionally, the tree provides support for a relationship between Pcl2 and Pcl1a, and Pcl31 and Pcl33. Pcl5, Pcl35, Pcl34, and Pcl29 were not assigned to any subfamily and their positions within the tree are ambiguous based on bootstrap values.

Strong sequence conservation at the 5′ and 3′ ends of each PCL also supports the subfamily structure. For Group 1, the N-terminal consensus sequence is MS*IT* (100%, 57%, *, 43%, 43%, *), and the C-Terminal consensus sequence is SVQQIG (57%, 57%, 86%, 100%, 100%, 86%). For Group 2, the N-terminal consensus sequence is MSILIG (100%, 78%, 89%, 100%, 89%, 100%), and the C-terminal consensus sequence is LIRKI* (44%, 89%, 89%, 100%, 100%, *). For Group 3, the N-terminal consensus sequence is MSQANI (100%, 100%, 100%, 100%, 100%, 100%), and the C-terminal consensus sequence is TVTKL* (56%, 56%, 56%, 78%, 56%,*). Group 4 does not display the same degree of conservation, although the last three amino acids of Pcl19, Pcl20, and Pcl30 are all QIA. Pcl2 and Pcl1a have extremely similar N-terminal sequences, differing only by one amino acid (M*KYKK).

Within the genome, PCLs are mostly clustered into groups of three, with very small intergenic distances (Fig. 1). Pcl27 and Pcl28 appear to form an exception; however, a potential PCL pseudogene 5′ of Pcl27 completes the cluster. Pcl29, Pcl31, Pcl32, and Pcl33 sit alone. Pcl2 clusters with Pcl1a, while Pcl30, Pcl18, Pcl19, and Pcl20 form a cluster of four.

Comparing Fig. 2 to Fig. 1, it can be seen that within each cluster of three PCLs, the first belongs to Group 1, the second to Group 2 and the third to Group 3. Group 4 PCLs, on the other hand, form their own distinctive cluster on the genome. The Group 1–Group 2–Group 3 arrangement leads to the hypothesis that PCL subfamilies arose through multiple duplications of clusters of three PCLs within the genome. This hypothesis is supported by a comparison of maximum-likelihood trees describing the relationships between genes within each subfamily, compared to between each genomic cluster of three PCLs (Fig. 3). The tree topologies are identical for genomic PCL clusters Pcl11-Pcl10-Pcl9, Pcl14-Pcl13-Pcl12, Pcl36-Pcl37-Pcl38, and Pcl21-Pcl22-Pcl23. For the remaining three genomic PCL clusters, the genomic cluster

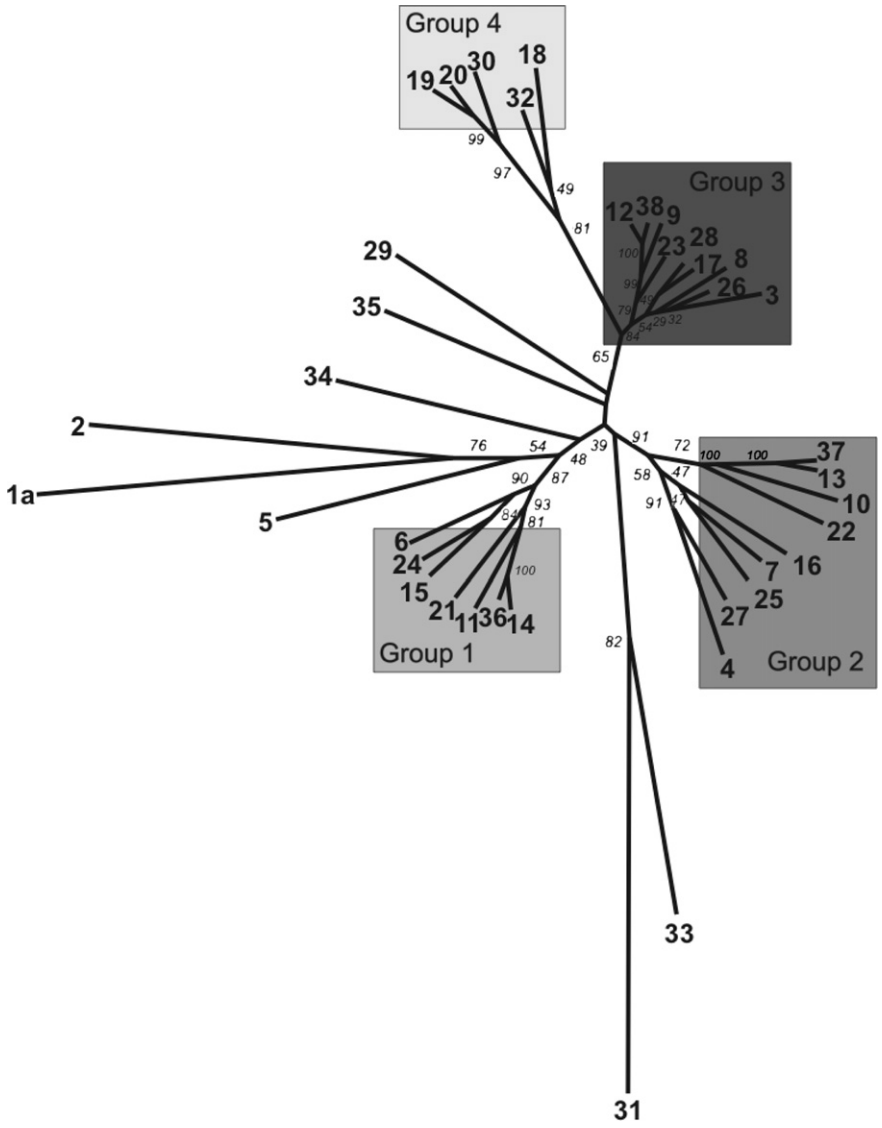

Fig. 2. Neighbour-joining tree showing groups of related PCLs. Numbers in bold correspond to PCL names. Small numbers in italics next to branches give bootstrap values.

tree shows Pcl24-Pcl25-Pcl26 and Pcl15-Pcl16-Pcl17 to be most closely related to each other, followed by Pcl6-Pcl7-Pcl8. Trees based on Group 1 and Group 2 PCLs support this arrangement; however the Group 3 tree has a conflicting topology with Pcl8 most closely related to Pcl17, followed by Pcl26. Bootstrap values suggest, however, that confidence in the relationships between these three genes is not high. The deviation in the Group 3 tree is therefore not taken as strong evidence against the duplication hypothesis.

3.3. Signal peptides and transmembrane domains

An in silico search revealed two potential transmembrane domains in only one PCL, Pcl2. The first spans positions seven to 26 and is predicted to run from inside to outside the cell, while the second runs from outside to inside the cell and is located just nine amino acids from the C-terminal, spanning positions 340 to 362. The first transmembrane domain corresponds to a predicted signal peptide (positions one to 28), immediately followed by a predicted cleavage site between positions 28 and 29.

Transmembrane domains were not identified in any of the other 36 new PCLs. All Group 3 PCLs, however, contained cleavage sites according to Signal P. While the amino acids prior to the cleavage sites were not identified as signal peptides, in most cases, the signal peptide prediction value was close to the threshold. Bacterial proteins may also be transported to the cell surface via the twin-arginine translocation (Tat) pathway. To check this possibility, a search for Tat pathway signal peptides was performed. Pcl5, Pcl6, Pcl16 and Pcl35 had predicted Tat signal peptides and cleavage sites above the threshold values. Pcl4, Pcl7, Pcl10, Pcl11, Pcl13, Pcl14, Pcl19, Pcl20, Pcl21, Pcl25, Pcl27, Pcl30, Pcl34, Pcl36, and Pcl37 also had appropriate stretches of amino acids above the signal peptide threshold value, but lacked cleavage sites. All PCLs lacked the conserved twin-arginine Tat motif, however.

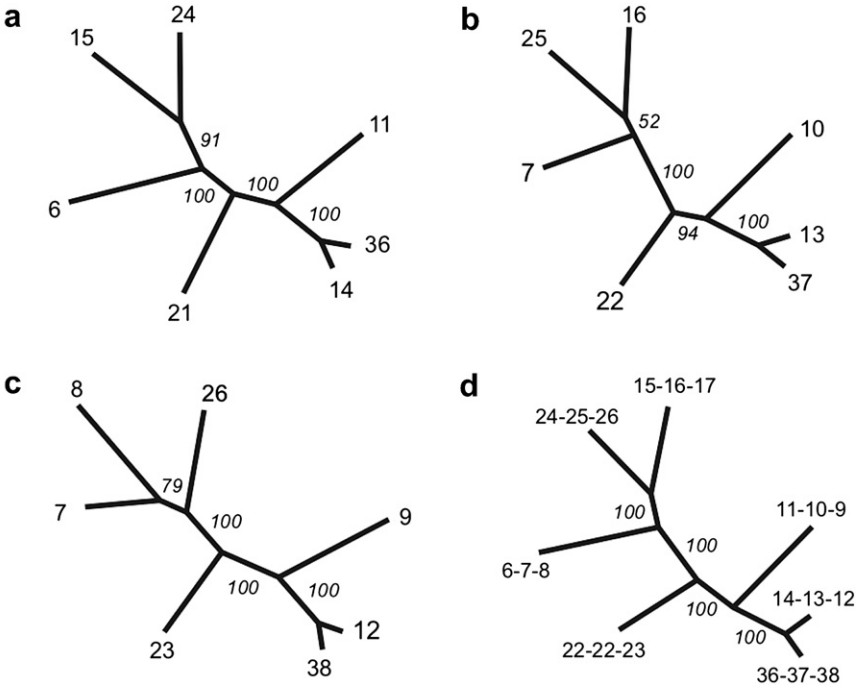

Fig. 3. Relationships between genomic clusters of three PCL genes. Maximum-likelihood trees based on the DNA sequences of (a) Group 1 PCLs, (b) Group 2 PCLs, (c) Group 3 PCLs, and (d) The concatenated sequences of genomic clusters of three PCLs (see Fig. 1). Topological similarity suggests that Group 1, Group 2 and Group 3 PCL subfamilies arose through multiple duplications of clusters of three PCLs. To facilitate comparison, only those genes falling neatly into a Group 1–Group 2–Group 3 cluster on the genome are included. Bootstrap values are given in small italics, while larger numbers at ends of branches refer to PCL names.

3.4. Putative protein repeat structure

With the exception of Pcl2 and Pcl31, the G-X-Y region of the new PCLs is not as uniform as Pcl1a, often interrupted by one or more amino acids. Interruptions form a distinctive pattern, separating G-X-Y regions into segments of conserved length. Fig. 4 graphically displays segment lengths for each gene. Segment conservation loosely supports the PCL subfamily structure. For most Group 2 and Group 3 PCLs, the G-X-Y region is separated into seven segments. Regions one to seven typically have 5, 4, 12, 9, 5, 3, and 12 G-X-Y repeats (there is some variation around these values; for instance, Pcl10 has segments lengths 5, 4, 12, 9, 4, 4, and 8). For some Group 3 genes, however, the pattern is changed by the absence of either the last (Pcl8) or two last (Pcl28, Pcl3) interruptions, giving a longer final G-X-Y segment. For Pcl17, also from Group 3, segment six has been duplicated, giving a pattern of 5, 4, 12, 9, 5, 4, 4, and 9 G-X-Y repeats. The Group 2 genes Pcl27 and Pcl4 are also missing some interruptions. Pcl4 follows the same pattern as Pcl28 and Pcl3, while Pcl27 is missing the interruption between segments 5 and 6.

Group 1 PCLs, on the other hand, resemble the typical G-X-Y segmentation described above, but are missing either the first interruption (Pcl14, Pcl36, Pcl11, Pcl15, and Pcl24) or first two interruptions (Pcl21, Pcl6), giving them a longer first segment. Group 4 PCLs do not resemble the other groups; they have both shorter and fewer G-X-Y segments. Interestingly, Pcl35 and Pcl5, despite not clustering with any subfamily in the neighbour-joining tree, have G-X-Y segmentation resembling both Group 2 and Group 3 PCLs. Pcl33, also not a member of a subfamily, has similar G-X-Y segmentation to Group 4 PCLs, while Pcl34 and Pcl29 stand out as being unique. Pcl31, Pcl2, and Pcl1a are the only PCLs with uninterrupted G-X-Y regions.

As a repeated G-X-Y motif domain is the defining feature of PCLs, it is reasonable to assume that PCLs may have distinctive tandem repeat (TR) structures. A search for moderately conserved TRs revealed 43 TRs within the 38 known PCLs (Table 3). Thirty of these were present in only one PCL. Copy numbers were mostly quite low, varying from two to four repeats. Pcl1a and Pcl2 form an exception, having TRs with copy number 26 and 11.78 respectively. Pcl29 also has a distinctive TR structure; it has four TRs in total, one of which has a copy number of five, another of which, although having a copy number of only two, has an unusually large period of 39aa, leading this TR to span a total of 78aa. Support for the PCL subfamilies based on TR structure is ambiguous; while PCLs within a subfamily may share a TR, they may also share TR with PCLs from other subfamilies. For example, Pcl6 shares TR26 with Pcl15 and Pcl36 (all Group 1 PCLs); however, Pcl6 also shares TR27 with Pcl7 (Group 2), Pcl17 (Group 3), and Pcl32 (Group 4).

Amino acid usage analysis for the X and Y positions of G-X-Y repeats for each PCL revealed proline to be the most common X amino acid, followed by alanine and isoleucine, with Y most commonly being threonine, closely followed by glutamine (Table 4). This mirrors the amino acid

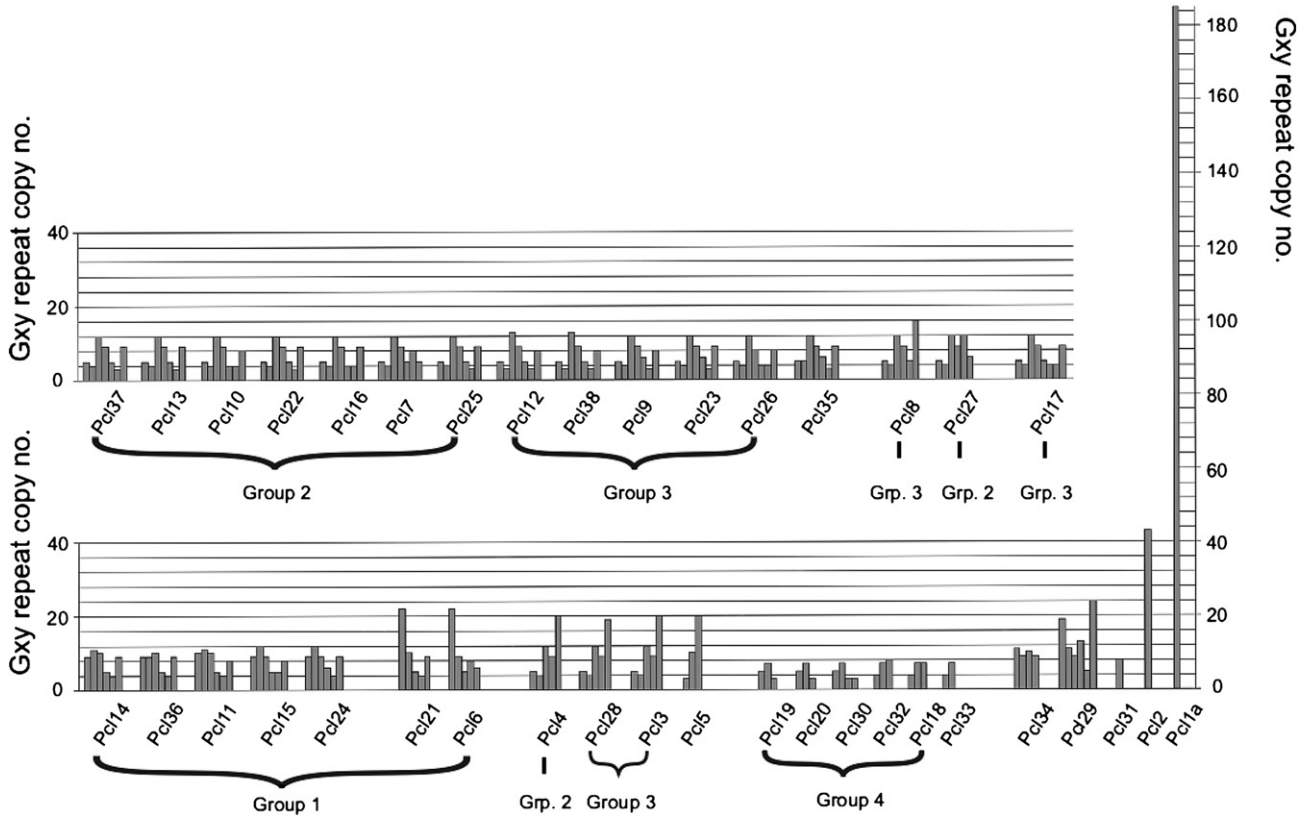

Fig. 4. G-X-Y repeat motif segment lengths. Consecutive G-X-Y segments within each gene are given adjacent to each other along the horizontal axis. The vertical axis gives the number of G-X-Y repeats for each segment.

composition of CLPs in other bacteria – in *B. anthracis*, BclA has GPT as its most common repeat motif, while bacterial CLPs in general have an overabundance of proline for X and threonine for Y, compared to collagens from multicellular organisms (Rasmussen et al., 2003; Sylvestre et al., 2002). Again, Pcl1a and Pcl2 stand out as notably different from the other PCLs – for Pcl1a, X is most commonly aspartic acid closely followed by glutamic acid and Y is most commonly either proline or lysine closely followed by glutamine, while for Pcl2, X is most commonly glutamic acid closely followed by aspartic acid and Y is most commonly lysine closely followed by glutamine.

3.5. PCL polymorphism

PCLs were amplified and sequenced for 10 *P. ramosa* strains, extracted from a range of *Daphnia* hosts of different species and geographic origin (Tables 1 and 2). All *P. ramosa* strains included in this study were confirmed as one species, based on a 625 bp section of the *P. ramosa* 16s rRNA gene. P10 showed the most divergence in its 16s rRNA sequence compared to C1 (the sequenced strain), with a sequence identity of more than 99%. In contrast, sequence identity between the 16s rRNA gene sequence of *P. ramosa* strain C1 and the corresponding 16s rRNA gene sequence of *P. penetrans* [GenBank: AF077672.1], the closest relative of *P. ramosa*, was only 90%.

Polymorphisms resulting in varying numbers of G-X-Y repeats, predicted protein length and amino acid substitution were all observed. Based on these differences, five genotypes are distinguishable (Table 2). Of these five genotypes, three are from *P. ramosa* samples obtained from infected *D. magna* (1–3), while two others (4 and 5) are from infected *Daphnia longispina*. Genotype 1 is the most widely spread, being found in samples from Russia, Finland and Belgium. The Belgium sample includes two genotypes (1 and 3) which were discovered and separated by infecting different *D. magna* clones with this isolate. Our hypothesis is that these differences did not arise by mutation during growth within the host, but rather, that they were present in the original isolate and had different affinities to the different host clones used for passaging.

4. Discussion

CLPs have been described in several bacterial species, including *B. anthracis*, *Streptococcus pyogenes*, *S. equi* (Karlstrom et al., 2004; Lukomski et al., 2000; Sylvestre et al., 2002) and in the model organism *P. ramosa* (Mouton et al., 2009). Our exploration of the *P. ramosa* draft genome revealed 37 further putative PCLs, bringing the total PCL number to 38. The actual number of PCLs is probably even larger, as the genome is incomplete and some partial PCL sequences were identified. The extent of the PCL family is

Table 3
Tandem repeats.

Repeat	Sequence	Gene	Pos.	Period	Copy Number	Consensus Error
1	GPSGEQGIPGPIGPIGPTGPSDGPPGPIGPTGGMGIGPT	Pcl29	210–289	39	2.00	0.11
2	PGEPGPQGRQGEAGLPGPPGKQGEPGLSGPKGEKGD	Pcl1a	117–242	36	3.50	0.12
3	GPIGPTGADFIGDTGPSFTGPTGDI	Pcl32	38–93	25	2.24	0.16
4	GPTGPTGPMGPAGNAVRLNAIIN	Pcl17	183–235	23	2.30	0.04
5	GSNGGSVSNGSNGSNGSNGSNN	Pcl33	76–119	22	2.05	0.10
6	TGITGITGPTGETGFTGI	Pcl7	50–98	18	2.72	0.14
7	TGPQGPPGLQGNTGI	Pcl21	95–124	15	2.00	0.10
8	GPPGPTGDIGIQGPI	Pcl29	111–142	15	2.13	0.03
9	GLTVTGAQGPIGPI	Pcl24	137–163	14	2.00	0.17
10	GATGPTGPTGLMGD	Pcl28	119–146	14	2.00	0.14
11	QGEAGMPGPKGD	Pcl1a	198–509	12	26.00	0.05
12	TGITGPTGATGI	Pcl8	219–257	12	3.25	0.13
13	TGPTGDLGLTGD	Pcl18	119–147	12	2.42	0.10
14	PGNTGPTGLIG	Pcl6	48–74	11	2.45	0.17
15	TGISITGATGA	Pcl9	228–249	11	2.00	0.12
	ITGPTGATGIS	Pcl12	220–241	11	2.00	0.09
	ITGPTGETGPS	Pcl23	220–241	11	2.00	0.14
16	TGPQGLASIGL	Pcl22	124–145	11	2.00	0.18
17	KGDKGEQGL	Pcl2	100–205	9	11.78	0.02
18	TGPTGIPGI	Pcl6	164–182	9	2.11	0.10
	SIGPTGPIG	Pcl35	134–158	9	2.78	0.16
	ITGPTGDTG	Pcl38	220–251	9	3.33	0.16
19	TGAIGATGS	Pcl14	167–184	9	2.00	0.11
20	GPQGPQGLI	Pcl21	76–103	9	2.89	0.14
21	TGLGITGA	Pcl17	127–142	8	2.00	0.12
22	KGPTGISI	Pcl19	65–83	8	2.38	0.11
	GISIKGPT	Pcl20	71–87	8	2.12	0.00
	GPTGASIT	Pcl30	65–84	8	2.50	0.05
23	TGPTGYGP	Pcl24	165–180	8	2.00	0.11
	TGPTGPST	Pcl36	131–146	8	2.12	0.06
24	TGPTGPA	Pcl23	112–125	7	2.00	0.14
	GPTGPSD	Pcl29	162–178	7	2.43	0.11
25	IGPKGP	Pcl4	5–16	6	2.00	0.08
	IGPKGI	Pcl5	22–33	6	2.00	0.08
	GPKGIT	Pcl14	99–111	6	2.17	0.08
26	GFQGPQ	Pcl6	79–91	6	2.17	0.08
	QGPQGV	Pcl15	161–180	6	3.33	0.05
	QGPQGA	Pcl36	72–83	6	2.00	0.08
27	TGPTGS	Pcl6	129–154	6	4.00	0.19
	ITGSTG	Pcl7	120–137	6	3.00	0.11
	GNTGPT	Pcl17	174–190	6	2.83	0.12
	GPTGAT	Pcl32	89–102	6	2.33	0.07
28	QGPSGI	Pcl7	41–52	6	2.00	0.08
29	ITGDTG	Pcl8	131–142	6	2.00	0.08
	GITGAT	Pcl24	214–230	6	2.83	0.12
	TGDTGA	Pcl31	74–87	6	2.33	0.07
30	GNTGET	Pcl8	171–183	6	2.17	0.08
	TGNTGA	Pcl28	236–249	6	2.33	0.00
31	GINGPV	Pcl9	176–188	6	2.17	0.00
32	GATGPQ	Pcl13	36–48	6	2.17	0.08
	GPTGPQ	Pcl15	80–99	6	3.33	0.10
33	GMTGAT	Pcl13	83–95	6	2.17	0.08
	TGATGT	Pcl26	231–244	6	2.33	0.07
34	GITGAP	Pcl13	148–159	6	2.00	0.08
35	QGIQGV	Pcl16	138–154	6	2.83	0.06
36	TGTTGL	Pcl17	259–278	6	3.33	0.00
37	IGNTGN	Pcl19	50–63	6	2.33	0.07
	IGNTGN	Pcl20	52–63	6	2.00	0.08
	IGPTGN	Pcl38	103–114	6	2.00	0.08
38	TGANGT	Pcl21	173–184	6	2.00	0.08
39	QGPTGA	Pcl22	41–58	6	3.00	0.11
40	QGNQGL	Pcl24	127–138	6	2.00	0.08
41	TGPMGN	Pcl26	135–146	6	2.00	0.08

(*continued*)

Table 3 (*continued*)

Repeat	Sequence	Gene	Pos.	Period	Copy Number	Consensus Error
42	PGCQGP	Pcl29	306–335	6	5.00	0.03
43	LSLNNE	Pcl34	262–273	6	2.00	0.00

Table of tandem repeats within the collagen-like genes of *P. ramosa*. Repeats are uniquely identified by number. Position gives the location of the tandem repeat in base pairs within each gene. The period refers to the length of individual repeat units (bp), while copy number gives the number of times each basic unit is repeated. Consensus error × 100 gives the percentage of disagreement between the 'consensus' repeat unit, and the full length tandem repeat sequence.

exceptional – our search for CLP in completed bacterial genomes showed that only 10% of sequenced bacteria have CLP and of these, most have three or fewer CLP.

The large number of PCLs, combined with their subfamily organisation and triplet arrangement within the genome, suggests that these genes are highly mobile. CLP may have been introduced into *P. ramosa* via horizontal gene transfer or may have evolved directly through mutational change; however, following this initial introduction, Group 1, Group 2 and Group 3 genes clearly evolved through multiple duplications of three-gene-long PCL clusters. This is supported by extremely close topologies between separate trees of Group 1, Group 2, and Group 3 PCLs, and a tree based on the concatenated sequences of each three-gene-long cluster (Fig. 3).

Table 4
Conserved terminal amino acids.

Family	Gene	N-term	C-term
1	Pcl11	MSTITT	CIQQIN
1	Pcl14	MSSLTC	AIQQIG
1	Pcl36	MSSLIC	AIQQIG
1	Pcl21	MDAIVR	SVQQIG
1	Pcl6	MSYIAG	SVQQIG
1	Pcl15	MYGVTQ	SVQQIG
1	Pcl24	MYGVKQ	SVEQIG
2	Pcl37	MSILIG	LIRKIK
2	Pcl13	MSILIG	LIRKIK
2	Pcl7	MSILVG	TIRKIK
2	Pcl27	MSILIG	VFRKIG
2	Pcl4	MSILIG	VIRKIG
2	Pcl16	MSILIG	TIKKIG
2	Pcl25	MSILIG	NIRKIS
2	Pcl22	MGILIG	LIRKIS
2	Pcl10	MNNLIG	LIRKIA
3	Pcl12	MSQANI	TVTRIQ
3	Pcl38	MSQANI	TVTRIQ
3	Pcl9	MSQANI	NITKIN
3	Pcl23	MSQANI	TVTKIA
3	Pcl17	MSQANI	TVVKLA
3	Pcl3	MSQANI	NILKLA
3	Pcl28	MSQANI	DVTKLS
3	Pcl8	MSQANI	NILKLS
3	Pcl26	MSQANI	TILKLS
4	Pcl19	MSKNFL	SFTQIA
4	Pcl20	MNHKPQ	TITQIA
4	Pcl30	MLYEYL	TLTQIA
4	Pcl18	MHYNQD	SIHRVN
4	Pcl32	MYIPEH	YQINTI
–	Pcl35	MTTIGP	TIYKIS
–	Pcl5	MQGPKG	LIQQLM
–	Pcl29	MSNPNI	VVHPAA
–	Pcl34	MSVKGA	NVHKIG
–	Pcl31	MFMKEN	IVNKIG
–	Pcl33	MSTLKD	VRGLSS
–	Pcl2	MYKYKK	RLLYKK
–	Pcl1a	MNKYKK	LLALKG

Conserved initial six N-terminal and final six C-terminal amino acids, arranged according to subfamily groupings.

Why so many PCLs have been retained in the genome remains elusive. Davies (2009) speculates that in the related *P. penetrans*/nematode system, CLPs are involved in attachment of spores to the host and that CLP polymorphisms are responsible for the strong specificity of host-*Pasteuria* interactions. In the *P. ramosa/Daphnia* system, infection specificity is so strong that a binary outcome is created – for a given *P. ramosa* clone/*Daphnia* clone, interaction is either compatible (resulting in infection) or not (Luijckx et al., 2011). Polymorphic PCL genes may play a role in this highly specific interaction.

Amplification and sequencing of PCL genes across a range of *P. ramosa* strains demonstrates that these genes are indeed highly polymorphic, discriminating five genotypes among 10 tested strains. Polymorphic CLPs have been proposed as candidate genes for *Bacillus* strain and species discrimination (Castanha et al., 2006; Sylvestre et al., 2003). The degree of PCL polymorphism observed suggests they may also be useful for *P. ramosa* genetic fingerprinting. It should be noted, however, that only Pcl15, Pcl24 and Pcl29 were successfully amplified for all strains. This could be due to polymorphism in the primer sites or to the absence of some PCLs in some strains, hinting at even higher levels of polymorphism than reported in Table 2.

As noted above, previous studies have demonstrated a high degree of *P. ramosa/Daphnia* genotype specificity. For those *P. ramosa* clones where the pattern of *D. magna/P. ramosa* interaction was known (namely C1, C14, C18, C19, C20), we found a perfect match between genotypes distinguished by PCL polymorphism and the pattern of infection specificity. C19 and C20 are indistinguishable in their *Daphnia* specificity, but differ markedly from C1, C14 and C18, which do not differ among themselves (Luijckx et al., 2011). This is reflected in their genotypes, with C19 and C20 belonging to genotype 2, and C1, C14 and C18 belonging to genotype 1. This genotype/phenotype relationship is particularly remarkable, as these five *P. ramosa* clones originate from four different countries (Tables 1 and 2). However, the correlation is based on only five clones, which is too few to reach a strong conclusion. Despite the small sample size, though, it is obvious that the host genotype has a stronger effect on the *P. ramosa* genotype than the geographic origin of the *P. ramosa* strain.

Genotype 3 makes this point most clearly. $P4_{HO2}$ (genotype 1) and $P4_{FX}$ (genotype 3) are strains originating from the same Belgian field isolate. $P4_{HO2}$ has been passaged through

D. magna clone HO2, originally from Hungary, whilst $P4_{FX}$ has been passaged through *D. magna* clone FX, originally from Tvärminne, Finland. The genotype of $P4_{HO2}$ is indistinguishable from the genotypes of *P. ramosa* clones C1 and C14, which were also extracted from *D. magna* HO2. Genotype $P4_{FX}$, however, is unique, differing markedly from all other genotypes. The degree of polymorphism between $P4_{HO2}$ and $P4_{FX}$ makes it unlikely that $P4_{FX}$ has evolved through new mutations during passaging. A more likely explanation is that $P4_{FX}$ was present in low population numbers in the original isolate. Passaging through FX has selected for this genotype in a kind of genotype-specific cloning. A previous cross-infection study, using *P. ramosa* and *D. magna* from different locations, concluded that most variation in host-parasite interactions is found locally (Ebert et al., 1998), which is consistent with balancing selection within the population (Ebert, 2008). The $P4_{HO2}$ and $P4_{FX}$ genotypes support this view.

Bioinformatic analysis of predicted PCL amino acid sequences identified only one putative protein, Pcl2, with properties comparable to Pcl1a. A neighbour-joining tree of all putative PCLs showed that Pcl2 clusters most closely with Pcl1a according to amino acid similarity. Pcl2 also lies in close proximity to Pcl1a on the genome, a characteristic shared only by Group 4 PCLs (which cluster both according to amino acid similarity and physically on the genome). Pcl1a and Pcl2 are also distinctive in that both have a G-X-Y amino acid composition that differs markedly from other PCLs, both have uninterrupted G-X-Y repeat regions characterized by high copy number TRs and both share a conserved N-terminal six amino acids that differ from the N-terminal consensus sequences found in other PCL subfamilies. Significantly, Pcl2 and Pcl1a are also the only PCLs with predicted transmembrane regions, and classical signal peptides with cleavage sites. Pcl2 is predicted to have two transmembrane regions; based on the position of the cleavage site, the N-terminal transmembrane region is most likely cleaved after signal peptide targeting to the cell surface, while the predicted C-terminal transmembrane domain may serve as a cell surface anchor. As Pcl1a's only transmembrane region is closely followed by a cleavage site (Mouton et al., 2009), and the G-X-Y repeat region of CLP can form triple-helices in association with other CLP (Xu et al., 2010), Pcl2 may also anchor Pcl1a, binding it to the spore surface after signal peptide cleavage. Both Pcl2 and Pcl1a are therefore likely to be targeted to the spore surface, potentially being involved in initial host adhesion and host–parasite interaction specificity.

Definitive transmembrane regions and signal peptides were not found in the remaining 36 PCLs. Group 3 PCLs had N-terminal regions similar to signal peptides (but below the prediction threshold), immediately followed by predicted cleavage sites. Pcl5, Pcl6, Pcl16 and Pcl17 also had predicted cleavage sites and signal peptides with properties similar to Tat signal peptides, while a number of the remaining PCLs had predicted Tat signal peptides lacking cleavage site. All these proteins, however, lacked the Tat conserved RR motif. These proteins and the Group 3 PCLs may have evolved from cell-surface targeted proteins, with their signal peptides being gradually degraded. Alternatively, as understanding of the Tat pathway and other non-classical cell-surface targeting mechanisms is still in its infancy, it is possible that these proteins are indeed targeted to the spore surface, via a modified Tat pathway or as yet undescribed pathways.

Between *P. ramosa* strains, the overall G-X-Y repeat region length is highly conserved, differing by less than two amino acids in otherwise polymorphic PCLs in all cases except for Pcl1a and Pcl29. In fact, conservation of G-X-Y repeat length is strictly maintained in all PCLs except Pcl1a, Pcl2, Pcl11, Pcl26, Pcl36, and Pcl29. Pcl29, however, varies wildly in G-X-Y repeat number from 65 to 100. Interestingly, despite its high degree of length polymorphism, Pcl29 (which does not belong to a subfamily) was one of the few genes able to be amplified across all *P. ramosa* strains tested, suggesting this gene is non-redundant. In contrast to Pcl29, PCLs belonging to a subfamily may display some redundancy. This is suggested by the presence of potential PCL pseudogenes and by the inability to amplify some PCLs in some strains. Taken together, these findings hint that PCLs may be subjected to varying evolutionary forces, leading to functional differentiation and specialisation. The striking relationship between Pcl1a and Pcl2, combined with their structural and sequence divergence from other PCLs, lends weight to this theory.

Despite possible redundancy, most PCLs are associated with strong SDs, indicating that most PCLs are probably expressed. All Group 2 and Group 3 PCLs have clear SDs. Group 1 PCLs, on the other hand, do not. This is not strong evidence against expression, however, as genes clustered in close proximity to upstream genes are known to be more likely to have SDs sequences than genes at the start of a cluster (Ma et al., 2002).

Structurally, G-X-Y repeat segment length emerges as a dominant feature, more so than TRs or G-X-Y amino acid usage. The pattern of interruptions may affect protein folding and triple-helix formation. An understanding of the effects of interruptions in the G-X-Y region may give clues to the function of the various PCL subfamilies.

While Pcl2 joins Pcl1a as a candidate gene mediating host attachment and specificity, the function of PCLs in general is yet to be determined. In *B. anthracis*, immunodominant CLPs are involved in spore structure and integrity (Thompson et al., 2007; Steichen et al., 2003). A recombinant CLP isolate from *S. pyogenes* forms aggregates, in a kind of bundled fibrilar structure (Yoshizumi et al., 2009). In the nematode parasite *P. penetrans*, which is closely related to *P. ramosa*, immunogold labelling of adhesin-associated epitopes targeted parasporal fibres during sporogenesis (Brito et al., 2003). At later stages of sporogenesis, the exosporium is also heavily labelled. Although purely speculative, it is conceivable that parasporal fibres consist of bundles of CLP, with surface-associated CLP expressed later in development. Spore development in *P. ramosa* may follow a similar pattern. Further studies, including immunomicroscopy, expression analysis and structural studies, will help shed light on the localisation and function of *P. ramosa*'s large and unique PCL family.

Acknowledgements

We would like to thank Urs Stiefel, Jürgen Hottinger and Pepijn Luijckx for their assistance. Work was financially supported by the Swiss National Science Foundation, the Freiwillige Akademische Geselschaft, Basel University, and an Australian Postgraduate Award.

References

Beck, K., Brodsky, B., 1998. Supercoiled protein motifs: the collagen triple-helix and the alpha-helical coiled coil. J. Struct. Biol. 122, 17–29.

Bendtsen, J.D., Nielsen, H., von Heijne, G., Brunak, S., 2004. Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. 340, 783–795.

Bendtsen, J.D., Nielsen, H., Widdick, D., Palmer, T., Brunak, S., 2005. Prediction of twin-arginine signal peptides. BMC Bioinforma. 6, 167.

Brito, J.A., Preston, J.F., Dickson, D.W., Giblin-Davis, R.M., Williams, D.S., Aldrich, H.C., Rice, J.D., 2003. Temporal formation and immunolocalization of an endospore surface epitope during Pasteuria penetrans sporogenesis. J. Nematol 35, 278–288.

Carius, H.J., Little, T.J., Ebert, D., 2001. Genetic variation in a host-parasite association: potential for coevolution and frequency-dependent selection. Evolution 55, 1136–1145.

Castanha, E.R., Swiger, R.R., Senior, B., Fox, A., Waller, L.N., Fox, K.F., 2006. Strain discrimination among *B. anthracis* and related organisms by characterization of bclA polymorphisms using PCR coupled with agarose gel or microchannel fluidics electrophoresis. J. Microbiol. Methods 64, 27–45.

Caswell, C.C., Han, R., Hovis, K.M., Ciborowski, P., Keene, D.R., Marconi, R. T., Lukomski, S., 2008. The Scl1 protein of M6-type group A Streptococcus binds the human complement regulatory protein, factor H, and inhibits the alternative pathway of complement. Mol. Microbiol. 67, 584–596.

Davies, K.G., 2009. Understanding the interaction between an obligate hyperparasitic bacterium, *Pasteuria penetrans* and its obligate plant-parasitic nematode host, Meloidogyne spp. Adv. Parasitol. 68, 211–245.

Decaestecker, E., Lefever, C., De Meester, L., Ebert, D., 2004. Haunted by the past: evidence for dormant stage banks of microparasites and epibionts of Daphnia. Limnol. Oceanogr 49, 1355–1364.

Decaestecker, E., Gaba, S., Raeymaekers, J.A.M., Stoks, R., Van Kerckhoven, L., Ebert, D., De Meester, L., 2007. Host-parasite 'Red Queen' dynamics archived in pond sediment. Nature 450, 870–873.

Ebert, D., 2008. Host-parasite coevolution: insights from the Daphnia-parasite model system. Curr. Opin. Microbiol. 11, 290–301.

Ebert, D., Zschokke-Rohringer, C.D., Carius, H.J., 1998. Within- and between-population variation for resistance of *Daphnia magna* to the bacterial endoparasite Pasteuria ramosa. Proc. R. Soc. B. Biol. Sci. 265, 2127–2134.

Ebert, D., Carius, H.J., Little, T., Decaestecker, E., 2004. The evolution of virulence when parasites cause host castration and gigantism. Am. Nat. 164 (5), 19–32.

Felsenstein, F., 1989. PHYLIP – phylogeny inference package (version 3.2). Cladistics 5, 164–166.

Han, R., Caswell, C.C., Lukomska, E., Keene, D.R., Pawlowski, M., Bujnicki, J.M., Kim, J.K., Lukomski, S., 2006. Binding of the low-density lipoprotein by streptococcal collagen-like protein Scl1 of *Streptococcus pyogenes*. Mol. Microbiol. 61, 351–367.

Hodas, N.O., Aalberts, D.P., 2004. Efficient computation of optimal oligo-RNA binding. Nucleic Acids Res. 32, 6636–6642.

Hulmes, D.J., 1992. The collagen superfamily–diverse structures and assemblies. Essays Biochem. 27, 49–67.

Humtsoe, J.O., Kim, J.K., Xu, Y., Keene, D.R., Hook, M., Lukomski, S., Wary, K. K., 2005. A streptococcal collagen-like protein interacts with the alpha2beta1 integrin and induces intracellular signaling. J. Biol. Chem. 280, 13848–13857.

Karlstrom, A., Jacobsson, K., Flock, M., Flock, J.-I., Guss, B., 2004. Identification of a novel collagen-like protein, SclC, in *Streptococcus equi* using signal sequence phage display. Vet. Microbiol. 104, 179–188.

Karlstrom, A., Jacobsson, K., Guss, B., 2006. SclC is a member of a novel family of collagen-like proteins in *Streptococcus equi* subspecies equi that are recognised by antibodies against SclC. Vet. Microbiol. 114, 72–81.

Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L., 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 305, 567–580.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. Clustal W and clustal X version 2.0. Bioinformatics 23, 2947–2948.

Luijckx, P., Ben-Ami, F., Mouton, L., Du Pasquier, L., Ebert, D., 2011. Cloning of the unculturable parasite *Pasteuria ramosa* and its Daphnia host reveals extreme genotype-genotype interactions. Ecol. Lett. 14, 125–131.

Lukashin, A.V., Borodovsky, M., 1998. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. 26, 1107–1115.

Lukomski, S., Nakashima, K., Abdi, I., Cipriano, V.J., Ireland, R.M., Reid, S.D., Adams, G.G., Musser, J.M., 2000. 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.

Lukomski, S., Nakashima, K., Abdi, I., Cipriano, V.J., Shelvin, B.J., Graviss, E. A., Musser, J.M., 2001. Identification and characterization of a second extracellular collagen-like protein made by group A Streptococcus: control of production at the level of translation. Infect. Immun. 69, 1729–1738.

Ma, J., Campbell, A., Karlin, S., 2002. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. J. Bacteriol. 184, 5733–5745.

Mouton, L., Traunecker, E., McElroy, K., Du Pasquier, L., Ebert, D., 2009. Identification of a polymorphic collagen-like protein in the crustacean bacteria *Pasteuria ramosa*. Res. Microbiol. 160, 792–799.

Newman, A.M., Cooper, J.B., 2007. XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. BMC Bioinforma. 8, 382.

Ramachandran, G.N., 1988. Stereochemistry of collagen. Int. J. Pept. Protein Res. 31, 1–16.

Rasmussen, M., Eden, A., Bjorck, L., 2000. SclA, a novel collagen-like surface protein of *Streptococcus pyogenes*. Infect. Immun. 68, 6370–6377.

Rasmussen, M., Jacobsson, M., Bjorck, L., 2003. Genome-based identification and analysis of collagen-related structural motifs in bacterial and viral proteins. J. Biol. Chem. 278, 32313–32316.

Reid, K.B., 1993. Structure/function relationships in the collectins (mammalian lectins containing collagen-like regions). Biochem. Soc. Trans. 21, 464–468.

Rocha, E.P., Danchin, A., Viari, A., 1999. Translation in Bacillus subtilis: roles and trends of initiation and termination, insights from a genome analysis. Nucleic Acids Res. 27, 3567–3576.

Sellar, G.C., Blake, D.J., Reid, K.B., 1991. Characterization and organization of the genes encoding the A-, B- and C-chains of human complement subcomponent C1q. The complete derived amino acid sequence of human C1q. Biochem. J. 274 (Pt 2), 481–490.

Steichen, C., Chen, P., Kearney, J.F., Turnbough, C.L., 2003. Identification of the immunodominant protein and other proteins of the *Bacillus anthracis* exosporium. J. Bacteriol. 185, 1903–1910.

Sylvestre, P., Couture-Tosi, E., Mock, M., 2002. A collagen-like surface glycoprotein is a structural component of the *Bacillus anthracis* exosporium. Mol. Microbiol. 45, 169–178.

Sylvestre, P., Couture-Tosi, E., Mock, M., 2003. Polymorphism in the collagen-like region of the *Bacillus anthracis* BclA protein leads to variation in exosporium filament length. J. Bacteriol. 185, 1555–1563.

Thompson, B.M., Stewart, G.C., 2008. Targeting of the BclA and BclB proteins to the *Bacillus anthracis* spore surface. Mol. Microbiol. 70, 421–434.

Thompson, B.M., Waller, L.N., Fox, K.F., Fox, A., Stewart, G.C., 2007. The BclB glycoprotein of *Bacillus anthracis* is involved in exosporium integrity. J. Bacteriol. 189, 6704–6713.

Waller, L.N., Stump, M.J., Fox, K.F., Harley, W.M., Fox, A., Stewart, G.C., Shahgholi, M., 2005. Identification of a second collagen-like glycoprotein produced by *Bacillus anthracis* and demonstration of associated spore-specific sugars. J. Bacteriol. 187, 4592–4597.

Whatmore, A.M., 2001. *Streptococcus pyogenes* sclB encodes a putative hypervariable surface protein with a collagen-like repetitive structure. Microbiology 147, 419–429.

Xu, C., Yu, Z., Inouye, M., Brodsky, B., Mirochnitchenko, O., 2010. Expanding the family of collagen proteins: recombinant bacterial collagens of varying composition form triple-helices of similar stability. Biomacromolecules 11, 348–356.

Yoshizumi, A., Yu, Z., Silva, T., Thiagarajan, G., Ramshaw, J.A.M., Inouye, M., Brodsky, B., 2009. Self-association of *Streptococcus pyogenes* collagen-like constructs into higher order structures. Protein Sci. 18, 1241–1251.