Srivastava, A., Mohan, S., Mauchline, T. H. and Davies, K. G. 2019. Evidence for diversifying selection of genetic regions of encoding putative collagen-like host-adhesive fibers in Pasteuria penetrans. *FEMS Microbiology Ecology*. 95 (1), p. fiy217.

The publisher's version can be accessed at:

- [https://dx.doi.org/10.1093/femsec/fiy217](https://dx.doi.org/10.1093/femsec/fiy217)

The output can be accessed at:

[https://repository.rothamsted.ac.uk/item/95y02/evidence-for-diversifying-selection-of-genetic-regions-of-encoding-putative-collagen-like-host-adhesive-fibers-in-pasteuria-penetrans](https://repository.rothamsted.ac.uk/item/95y02/evidence-for-diversifying-selection-of-genetic-regions-of-encoding-putative-collagen-like-host-adhesive-fibers-in-pasteuria-penetrans).

© 30 October 2018, CC-BY terms apply
Evidence for diversifying selection of genetic regions of encoding putative collagen-like host-adhesive fibers in Pasteuria penetrans

Arohi Srivastava¹, Sharad Mohan², Tim H. Mauchline³ and Keith G. Davies¹,4,*

¹Department of Biological and Environmental Sciences, University of Hertfordshire, Hatfield, AL10 9AB, UK, ²Division of Nematology, Indian Agricultural Research Institute, Pusa Campus, New Delhi 110012, India, ³Department of AgroEcology, Rothamsted Research, Harpenden, AL5 2JQ, UK and ⁴Division of Biotechnology and Plant Health, Norwegian Institute of Bioeconomy Research, Postboks 115, Ås-1431, Norway

*Corresponding author: Keith G Davies, Department of Biological and Environmental Sciences, University of Hertfordshire, Hatfield, AL10 9AB, UK. E-mail: k.davies@herts.ac.uk

One sentence summary: Collagen-like genes that are expressed on the surface of spores are hypothesised to have been obtained by horizontal gene transfer and are involved in the spores’ adhesion to nematode cuticle.

Editor: Rolf Kümmerli

ABSTRACT

Pasteuria spp. belong to a group of genetically diverse endospore-forming bacteria (phylum: Firmicutes) that are known to parasitize plant-parasitic nematodes and water fleas (Daphnia spp.). Collagen-like fibres form the nap on the surface of endospores and the genes encoding these sequences have been hypothesised to be involved in the adhesion of the endospores of Pasteuria spp. to their hosts. We report a group of 17 unique collagen-like genes putatively encoded by Pasteuria penetrans (strain: Res148) that formed five different phylogenetic clusters and suggest that collagen-like proteins are an important source of genetic diversity in animal pathogenic Firmicutes including Pasteuria. Additionally, and unexpectedly, we identified a putative collagen-like sequence which had a very different sequence structure to the other collagen-like proteins but was similar to the protein sequences in Megaviruses that are involved in host-parasite interactions. We, therefore, suggest that these diverse endospore surface proteins in Pasteuria are involved in biological functions, such as cellular adhesion; however, they are not of monophyletic origin and were possibly obtained de novo by mutation or possibly through selection acting upon several historic horizontal gene transfer events.

Keywords: BclA; hair-like nap; endospore; adhesion; bacterial collagens; pathogenicity

INTRODUCTION

Pasteuria penetrans is a parasite of root-knot nematodes (Meloidogyne spp.). It is one of a number of species of Gram-positive bacteria that form endospores and has the potential to be used as an environmentally benign biological control agent of plant-parasitic nematodes (Davies 2009; Davies et al. 2018; Stirling 2014). A major constraint in its use as a bio-pesticide is its restricted host range; one population of P. penetrans will attach to and infect one population of root-knot nematode but...
not another (Stirling 1985; Davies, Kerry and Fylnm 1988; Espanol et al. 1997). The taxonomy of the Pasteuria group of bacteria is confused; this is because of its obligate nature and early characterisation had to rely on an isolate’s biology; essentially, its life-cycle, morphology and in particular, its host range. However, molecular techniques, including genomics and proteomics, are today playing an increasingly important and crucial role in characterising its diversity.

Currently, three species of Pasteuria have been characterised that parasitize plant-parasitic nematodes: namely P. penetrans (Sayre and Starr 1985) on hosts of Meloidogyne spp.; Pasteuria nishizawai (Sayre et al. 1991) on hosts of cyst nematodes Heterodera and Globodera spp.; Pasteuria thornei (Starr and Sayre 1988) on hosts of Pratylenchus, and one Pasteuria ramosa (Metcnikoff 1888) that is a parasite of Cladorcera and Daphnia, the water flea. Two other hyperparasites of plant-parasitic nematode species have been characterised and given Candidatus status: Pasteuria usgae isolated from Belonolaimus longicaudatus (Giblin-Davis et al. 2003) and P. aldrichii (Giblin-Davis et al. 2011) on Bursilla spp. (Stackebrandt 2014). Two other populations of Pasteuria have been characterised, P. hartismeralis, a parasite of Meloidogyne ardenensis and Pasteuria goettingiana, a parasite of Heterodera goettingiana but these names, according to Stackebrandt (2014) are invalid.

Early molecular methods of Pasteuria characterisation tended to focus on using 16S rRNA genes to differentiate species and populations (Sturhan 1988; Anderson et al. 1999; Duan et al. 2003). Interestingly, a closely related population of Pasteuria isolated from an Indian plant-parasitic cyst nematode Heteroderca cajani population (Sharma and Davies 1996) has recently been shown to be more promiscuous; endospores of this particular Pasteuria population attach to and infect not only the original cyst nematode host, H. cajani, but also nematodes from the genus Globodera (Mohan et al. 2012); more recent results using 16S rRNA showed that it was closely related to P. nishizawai with 98.6% base-pair similarity whilst endospores of P. nishizawai from Heteroderca glycines (soya bean cyst nematode) also adhered to a range of cyst nematodes, including Heteroderca espezedei (Lespedezia cyst nematode), Heteroderca schachtii (sugar beet cyst nematode), H. trifoli (clover cyst nematode) and Globodera rostochiensis (potato cyst nematode), but unlike the H. cajani population of Pasteuria from India, they neither infected nor complete their life cycle in any nematode other than H. glycines (Sayre et al. 1991; Atibalentja, Jakstys and Noel 2004; Noel, Atibalentja and Domier 2005). It would, therefore, appear that the relatedness of these Pasteuria from H. cajani and H. glycines, as constructed with 16S RNA, does not reflect host range. This is consistent with endospore attachment studies of root-knot nematodes where adhesion was not linked to nematode phylogeny (Davies et al. 2001).

Ribosomal RNA (16S rRNA) was used for the reconstruction of the ‘tree of life’ (Woese 1987) and it has been universally accepted as a tool for phylogenetic reconstruction and classification of prokaryotes. Apart from the known conservation of 16S rRNA at the level of nucleotide sequences and secondary structures, the classical concept presumed that these genes do not undergo horizontal gene transfer (HGT) events. However, several studies report the horizontal transfer of segments of this gene and advocate the need for alternative methods to avoid misidentification and interpretation of discordant phylogenies (Eardly, Wang and Van Berkum 1996; Yap, Zhang and Wang 1999; Schols, Schot and Jacobs 2003; Rajendhran and Gunasekaran 2011). Protein-coding housekeeping genes may have advantages over ribosomal RNA genes and a multilocus protein sequence approach, using 25 or more housekeeping genes translated from a genome survey sequence of strain RES147, produced a robust bacterial phylogeny that suggested the Pasteuria genus was ancestral to Bacillus (Charles et al. 2005); additionally, single nucleotide polymorphisms in protein-encoding genes provide increased phylogenetic discrimination than using 16S rRNA sequences (Mauchline et al. 2011).

Although these protein-coding genes may be more useful in determining the phylogenetic relatedness of Pasteuria between closely related Firmicutes than 16S RNA, they are unlikely to relate in any meaningful manner to host-range and pathogenicity.

Another protein-encoding gene with potentially the ability to characterise endospore-forming Firmicutes that have been found in animal parasitic bacteria is bclA. This gene, originally characterised in Bacillus anthracis, is expressed during the final stages of endospore formation and produces a collagen-like protein (CLP) on the surface of the endospore’s exosporium which forms a hair-like nap (Sylvestre, Mock and Couture-Tosi 2002; Steichen et al. 2003; Sylvestre, Couture-Tosi and Mock 2003; Todd et al. 2003). Genome survey sequences of P. penetrans have revealed that it too contain collagen-like genes (Davies and Opperman 2006) and it has been hypothesised that the expression of these collagen-like genes produce a similar hair-like nap on the surface of the Pasteuria endospore that is involved in a Velcro-like attachment process (Davies 2009). More recently, a large family of polymorphic collagen-like genes have been described in P. ramosa, the endospore-forming parasite of Daphnia spp. (Mouton et al. 2009; McElroy et al. 2011) and they are possibly involved in the interaction between P. ramosa and its Daphnia host (Luijckx et al. 2011; Luijckx et al. 2013).

As attachment of P. penetrans endospores to nematodes has been shown to be unrelated to nematode phylogeny (Davies et al. 2001) we would hypothesise that the phylogenetic relatedness using protein-encoding house-keeping genes of P. penetrans would be very different from the phylogenetic relatedness described by a protein involved in a key aspect of pathogenic-like endospore adhesion to its potential host. The present investigation used in-silico comparative bioinformatics to look at the phylogenetic relatedness between P. penetrans and other selected closely related Firmicutes based on their 16S rRNA phylogeny and to compare this to the phylogenetic trees using the low-complexity collagen-like repeat regions.

**MATERIALS AND METHODS**

Search for protein sequences coding for putative collagens in Pasteuria

Collagen-like genes were predicted in unpublished contigs for Pasteuria Res148 isolate, a related but host-selected sub-population, of Res147 (Mauchline et al. 2011).

Gene predictions were done using the RAST annotation web server (Aziz et al. 2008) and the annotations were searched for any predicted collagen-like sequences based on the comparison of contig annotations with the sequenced genome of a closely related Firmicute, Bacillus thuringiensis A1 Hakam. To look for any collagen-like sequences not predicted by RAST, the contigs were uploaded on to Artemis genome browser and annotation tool (Carver et al. 2012) and were manually searched for open reading frames containing ‘G-X-Y’ triplet amino acid repeats using the ‘navigator’ feature of Artemis (Rutherford et al. 2000) (Rutherford et al. 2000). The sequences were further tested for the presence...
Comparison of CL sequences in Pasteuria and other collagens

To search for similar sequences in other organisms, the putative Pasteuria collagens were used as query sequences for BLASTp searches targeting non-redundant (nr) protein database. The low complexity filter was turned on for these searches to avoid any random hits to low complex G-X-Y repeats of collagens. BLASTp hits with significantly low E-values ranging from 0 to $2^{-128}$ were pooled together. The G-X-Y repeat regions of selected putative CLPs (File I, Supporting Information) were analysed for the diversity in their percentage amino acid composition. The G-X-Y repeat regions were extracted from the sequences and a customised script was written in R (R Development Core Team 2010) to do the following analyses. The Manhattan distance (Kaufman and Roussseu 1990) between the percentage amino acid compositions for all possible pairs of sequences was computed. The resulting distance matrix was subjected to an agglomerative hierarchical clustering method using hclust function (Murtagh and Legendre 2011). The heatmap.2 function of gplots package in R (http://cran.r-project.org/web/packages/gplots/index.html) was used to generate a heatmap representation of the Manhattan distances between the percentage amino acid composition of the G-X-Y repeat regions of different collagen-like sequences. The custom R-script used for generating the heatmap can be found in the supplementary material (File II, Supporting Information for the script).

Conventional molecular phylogenetic tree based on 16S rRNA

The publicly available 16S rRNA sequences of the cladoceran parasitic P. ramosa and nematode parasitic Pasteuria spp. were compared with selected 16S rRNA sequences of Clostridium spp., Paenibacillus spp. and Pelosinus spp. and three major animal pathogenic Bacillus spp. (B. anthracis, B. thuringiensis and B. cereus) and the non-pathogenic Bacillus subtilis. Cyanobacterial species including Microcystis elabens, Arthrospira platensis, Cyanobium gracile and/or Gloeocapsa spp. were used as outgroups. The list of accession numbers of sequences used in the analyses can be found in the File III, Table I (Supporting Information) linked to this article (see).

As a standard comparator a phylogenetic tree was constructed using the classical 16S rRNA gene sequences in MEGA7 (Kumar, Stecher and Tamura 2016) using the Maximum Likelihood method based on Tamura-Nei model and JTT matrix-based model for nucleic acid and protein sequences respectively (Jones, Taylor and Thornton 1992; Tamura and Nei 1993). The initial tree was made using BioNJ algorithm (Gascuel 1997). The bootstrap consensus tree was inferred from 500 replicates (Felsenstein 1985).

RESULTS

Molecular phylogeny based on the 16S rRNA gene

The phylogenetic tree with the highest log likelihood (-11 556.6217) is shown in Fig. 1a. The analysis involved 33 nucleotide sequences. There was a total of 1821 positions in the final dataset. The 16S rRNA gene sequences from all the five-analysed species of the genus Pasteuria, including the cladoceran parasite P. ramosa, clustered together with high bootstrap support (100%). Pasteuria spp. were observed to be more closely related to Clostridium spp. than Bacillus species. However, their close-relatedness with Clostridium spp. was supported with a low bootstrap value of only 58%. All the Bacillus spp. and Paenibacillus spp. were grouped together supported by a high bootstrap value. The pathogenic Bacillus i.e. B. cereus, B. thuringiensis and B. anthracis were separated from the non-pathogen B. subtilis with 100% support. In 95% of the tree replicates, Pelosinus spp. was separately grouped as an outgroup with the cyanobacteria.

Putative collagen-like proteins in Pasteuria

Using the sequence-based comparison tool of RAST, 17 putative collagen coding genes were identified in different contigs obtained from the sequencing of Pasteuria Res148. Using a manual search, further 16 open reading frames containing G-X-Y repeat regions were found. These sequences (33 in total) were named as Ppcl for ‘Pasteuria penetrans collagen-like sequences’ and were numbered as Ppcl1 to Ppcl33. Out of these, 23 sequences were unique and intact (i.e. started with a start codon and ended with a stop codon). To confirm that these sequences are related to the collagen superfamily, these sequences were searched for collagen motifs in MOTIF search. Only 17 of these sequences hit the Pfam: collagen family and were considered as the putative CLPs in P. penetrans Res148 (Tables 1 and 2). These 17 Ppcl sequences were selected for further analyses. The G-X-Y repeat regions in some of these putative sequences were interrupted with one or more amino acids, while G-X-Y repeats in eight of the sequences were uninterrupted. See File I (Supporting Information) for the nucleic acid sequences of the 17-selected putative CLPs in P. penetrans Res148.

CL sequences similar to Pasteuria Ppcl sequences

In the BLASTp searches using putative CL sequences from P. penetrans, significant matches were found to Ppcl23, Ppcl25, Ppcl26 and Ppcl33. Ppcl23 hit CL sequences from Bacillus spp. (98 hits) Clostridium arbusi (1 hit) and Sediminibacillus albus (1 hit); all hits were in the E-value range of 0.008 to 3e-08. Ppcl25 had 108 BLASTp hits (E-value $\leq 1e^{-56}$) from Bacillus spp. (96 hits), Paenibacillus spp. (7 hits), Clostridium spp. (4 hits), Pithovirus sibericum (1 hit). All the hits to Ppcl23 and Ppcl25 were in the CTDs of the sequences (query cover: $<43\%$ for Ppcl23; $<49\%$ for Ppcl25). Ppcl26 hit CL proteins from a range of bacterial genera including Clostridium spp., Ruminococcus torques, Fusicatenibacter spp., Eisenbergiella spp., Desulfothermus acetoxidans, Lachnoclostridium spp., Eubacterium dolichum, Blautia producta, Bifidobacterium spp., Lactonitratilactor longivorans, Epulopiscium spp., Bacteroides gillainarum, Parabacteroides spp., Methanobrevibacter spp., Lactobacillus spp., Flavonifractor plautii, Leuconostoc spp., Veillonella dispar, Prevotella spp. tc2–28. The topmost hits were for proteins from Clostridium spp. that had $\geq90\%$ query cover and very low E-values $\leq 1e^{-62}$. Ppcl33 showed sequence similarity with Megaviruses and Ruminococcus spp. (E-value: 0 to $2e^{-126}$; as high as 99% identity for 77% query cover).

Characterisation of the low-complex G-X-Y repeat region

Of all the significant BLASTp hits to the Ppcl sequences, 52 CL sequences, which showed better query coverage and also had significantly low E-values (in the range of 0 to $2e^{-126}$), were
Figure 1. Molecular phylogeny of Pasteuria (A) The bootstrap consensus tree based on 16S rRNA gene sequences reconstructed using the Maximum Likelihood method. Numbers shown next to the branches are the percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates). Only bootstrap values 70% are shown. (B) Heatmap reconstructed using the cluster analysis of G-X-Y repeat regions of Ppcl sequences of P. penetrans, Pcl sequences of P. ramosa and other selected CLPs based on the percentage amino acid composition of the low complex G-X-Y repeat regions.

Table 1. List of 17 putative CLPs in P. penetrans Res148 (File I, Supporting Information for the complete sequences).

| Putative collagen | Length | N-terminal | C-terminal | Number of G-X-Y repeats | Interruptions within the G-X-Y repeat regions and their location within the region |
|-------------------|--------|------------|------------|-------------------------|---------------------------------------------------------------------------------|
| Ppcl1             | 1215 bp/405 aa | MSNLELLHRLCC | RQVVVIELPSGN | 83 | CVCPP (7..11)CVCPP (39..43) |
| Ppcl8             | 753 bp/251 aa | MPNHSGLRGSPL | GFGVLYERGRL | 30 | SPV (4..6) |
| Ppcl9             | 669 bp/223 aa | MISVVVTMTPSPL | SRSPAEMDYLP | 14 | |
| Ppcl16            | 606 bp/202 aa | MYHNDYQGKMSD | PCPPPYPHYREY | 28 | None |
| Ppcl17            | 1242 bp/414 aa | MKRSTKYFPFLAM | GQAANLIRRVF | 20 | ST (25..26) |
| Ppcl18            | 1170 bp/390 aa | MKKITDDLFLG | TTISNYVRQIA | 23 | None |
| Ppcl19            | 1143 bp/381 aa | MIMKAIINLYLI | TAAPLLIRIAS | 27 | NLQIT (73..76) |
| Ppcl20            | 780 bp/260 aa | MGRNARIGNL | RATARASVIRQIF | 28 | None |
| Ppcl21            | 1936 bp/646 aa |MLEFHELPSYI | SSGASPTIRVIA | 129 | IT (19..20) |
| Ppcl23            | 1158 bp/386 aa | MLAILESSAPLCA | SISAVLIRKIA | 26 | None |
| Ppcl24            | 1203 bp/401 aa | MNYTRLQSADY | GATSFLIRRLN | 34 | None |
| Ppcl25            | 1179 bp/393 aa | MKKIIYLUSL | SINASLIRQIS | 15 | None |
| Ppcl26            | 837 bp/279 aa | M A S L N K V R Q L | TATQANLFFKLV | 15 | None |
| Ppcl28            | 843 bp/281 aa | MILNLPPGCGF | VTTIKYSDSICS | 26 | None |
| Ppcl29            | 837 bp/279 aa | MILNLPPGCGF | VVFQSYNTRICQ | 29 | TTT (79..81) |
| Ppcl30            | 774 bp/258 aa | MLIGGNLNVGNT | GATSFLTIRLNL | 30 | I (3)IT (14..15) |
| Ppcl33            | 1770 bp/590 aa | MSRSQNNIINIVY | SQKTWILIEQY | 49 | AEK (124..126) |
Table 2. CL motifs in putative collagen-like sequences in P. penetrans Res148. CL motifs were predicted by GenomeNet MOTIF using the Pfam database. The location of each CL motif within the sequence is shown along with the E-value for each motif.

| Sequence | Location of CL motif | E-values | CL motifs along putative CL sequences in Pasteuria RES 147 |
|----------|----------------------|----------|----------------------------------------------------------|
| Ppcl1    | 21.57 62.119 110.166 181.224 220.278 | 0.00021 1.3e-09 9.3e-07 0.0041 5.2e-09 | ![Image](https://www.oup.com/us/academic) |
| Ppcl8    | 128.176 156.210 | 2e-06 1.7e-07 | ![Image](https://www.oup.com/us/academic) |
| Ppcl9    | 87.121 | 0.88 | ![Image](https://www.oup.com/us/academic) |
| Ppcl16   | 67.121 97.154 | 1.1e-09 6.3e-10 | ![Image](https://www.oup.com/us/academic) |
| Ppcl17   | 198.253 | 1.3e-05 | ![Image](https://www.oup.com/us/academic) |
| Ppcl18   | 166.210 180.234 | 2.2e-06 3.1e-08 | ![Image](https://www.oup.com/us/academic) |
| Ppcl19   | 154.207 176.228 | 2.2e-08 1.5e-07 | ![Image](https://www.oup.com/us/academic) |
| Ppcl20   | 28.75 53.111 | 3.4e-07 1.6e-09 | ![Image](https://www.oup.com/us/academic) |
| Ppcl21   | 387.444 432.490 | 3.4e-07 1.6e-09 | ![Image](https://www.oup.com/us/academic) |
| Ppcl23   | 159.209 191.246 | 1e-05 1.3e-08 | ![Image](https://www.oup.com/us/academic) |
| Ppcl24   | 145.197 190.246 | 2.8e-08 6.5e-08 | ![Image](https://www.oup.com/us/academic) |
| Ppcl25   | 169.217 190.245 | 8e-07 4.4e-09 | ![Image](https://www.oup.com/us/academic) |
| Ppcl26   | 124.189 232.256 | 1.2e-08 0.00019 | ![Image](https://www.oup.com/us/academic) |
| Ppcl28   | 22.70 46.97 | 2.1e-07 7.4e-07 | ![Image](https://www.oup.com/us/academic) |
| Ppcl29   | 22.73 51.97 | 5.5e-08 1.3e-07 | ![Image](https://www.oup.com/us/academic) |
| Ppcl30   | 25.80 59.112 | 1e-08 8.8e-08 | ![Image](https://www.oup.com/us/academic) |
| Ppcl33   | 34.89 79.130 130.181 | 7.6e-09 7.5e-08 4.1e-07 | ![Image](https://www.oup.com/us/academic) |

Selected for further analysis, along with 37 previously characterised CL sequences from P. ramosa, the BclA and ExsJ protein sequences from B. anthracis and B. cereus. See File III, Table II (Supporting Information) for the list of accession numbers of the CLP sequences selected for the comparative studies with Pasteuria Ppcl sequences. A total of 108 sequences including 17 Ppcl sequences were analysed for the percentage amino acid composition of the G-X-Y repeat regions. It was evident that after Glycine (G), Threonine (T) is the most commonly occurring amino acid, followed by Proline (P). However, other amino acids like Alanine (A) and Valine (V) also occur. The heatmap generated by the hierarchical clustering of the G-X-Y repeat regions is shown in Fig. 1b. The tree splits the sequences categorically into nine clusters.

The Ppcl sequences from P. penetrans were spread across five clusters (Cluster 3, 4, 5, 6, 7, 9). Ppcl28 and Ppcl29 clustered...
together (Cluster 3) with other collagen-like sequences from P. ramosa (Pcl12, Pcl17, Pcl18, Pcl18, Pcl32, Pcl30), Closstridium spp. (Clos12), P. thuringiensis (Bt4) and the Bc1A protein of B. anthracis. Ppcl26 was placed in Cluster 4 along with six collagen-like sequences from Clostridium (Clos10, Clos8, Clos7, Clos6, Clos4) and one sequence each from Fusobacterium (Fusica), Ruminococcus (Rumin1), Eisenbergiella (Eisen), Eubacterium (Eubac), Lachnospiraceae (Lachno) and Desulfotomaculum (Desulfo). Ppcl23, Ppcl20, Ppcl18, Ppcl17, Ppcl24, Ppcl30, Ppcl21, Ppcl19 and Ppcl25 along with a sequence from Proteoclamydia (Proto) formed Cluster 5 which was the cluster with the largest number of P. pene-
trans sequences. Ppcl1 clustered (Cluster 6) with four P. ramosa sequences Pcl4, Pcl3, Pcl35, Pcl5, two Clostridium sequences (Clos1, Clos5) and one sequence each from Bacillus spp. (B.LL01), Paeenbacillus (Paein19) and Pithovirus (Pitho). Ppcl33 most uniquely clustered, in Cluster 7, with all the seven Megavirus sequences that were used in the study along with two sequences from P. ramosa (Pcl1, Pcl2), one sequence from Clostridium spp. (Clos11) and one Ruminococcus sequence (Rumin02). Members of Cluster 7 were unique because they all had Lysine (K) and Aspartic acid (D) as the most occurring amino acid after Glycine. Ppcl8, Ppcl9 and Ppcl16 were separately clustered as Cluster 9. Most of the P. ramosa sequences (25 out of 37) clustered together in Cluster 1 with one Clostridium sequence (Clos3). Cluster 2 consisted of eleven Bacillus sequences (Bc2, BJH7, Bt3, B.wied, Bt2, Bc5, Bc3, Bc5, B.wiehen1, Bc5, Bc1) with one sequence each of Clostridium spp. (Clos2) and P. ramosa (Pcl19). Cluster 8 consisted of eight, three and one sequences of Bacillus spp., Paenbacil-
lus spp. and Sediminibacillus spp. (B.amylo1, B.velze, B.safensis1, B.amylo2, B.pumilus1, B.pumilus2, Bc.Exsf, B.acidicola, Paein4, Paein2, Paein1, Sedimin). See File III, Table I (Supporting Infor-
mation) for the codes used for the CLP sequences for different organisms.

**DISCUSSION**

The diversity of CLPs in P. penetrans Res148

The adhesive role of CLPs in the Pasteuria-nematode interac-
tion has been hypothesised for a long time (Davies and Danks 1993; Mohan, Fould and Davies 2001; Davies and Opperman 2006; Davies 2009). This hypothesis was built on the knowledge that the endospores of many Gram-positive bacteria are covered with a hair-like nap, the fibers of which are constructed of glyco-
proteins containing G-X-Y repeat sequences, the number of G-
X-Y repeats contributing to the length of the fibers (Sylvestre et al. 2003; Davies and Opperman 2006). While a diverse set of CLPs in the cladoseran parasite P. ramosa have been identified and characterised, the CLPs of P. penetrans and other nematode parasitic Pasteuria species remain unexplored. Here, we iden-
tify putative genes coding for CLPs in the genome of a highly host-selected isolate of P. penetrans (designated Res148). From a set of unpublished contigs, we predict 17 unique collagen-like sequences of which four putative CLPs (Ppcl23, Ppcl25, Ppcl26 and Ppcl33) were shown to have statistically significant similarities with 52 CLP sequences from 22 different bacterial species and two groups of viruses. A cluster analysis of these sequences along with selected previously characterised CL sequences from P. ramosa and the Bc1A and Exsf1 protein sequences from B. anthracis and B. cereus suggested that the CLPs in P. penetrans are extensively diverse. Since low complexity regions in proteins are more prone to non-erroneous replication slippage (DePristo, Zilversmit and Hartl 2006; Rado-Trilla and Alba 2012; Zilversmit et al. 2010) and are thereby susceptible to rapid evolution, these results suggest that the Ppcl sequences and their predicted homologs are evolutionary linked and they possibly evolved to serve similar biological functions. When compared with the phylogenetic diversity analyses using universal 16S rRNA gene, the cluster analysis of the G-X-Y repeat regions of CLPs suggest the latter to be an important source of variation and diversity on which evolution can act amongst Pasteuria spp. and other closely related pathogenic Firmicutes.

**Incongruencies in the phylogenetic resolution Pasteuria spp.**

Preliminary studies had shown slight incongruencies in the phy-
genetic trees inferred from different genes gyrB, groEL, spoOA and there was an observed phylogenetic biasness between the trees based on the nucleic acid and amino acid sequences. This is likely because proteins are under different selective constraints due to their functional roles and are likely to be conserved over geological timescales (Huynen and Bork 1998; Romero and Arnold 2009). Biasness in the phylogenetic res-
olution of Pasteuria spp. using CLPs suggests a real biological phenomenon more than just an artefact. One such known phenomenon could be due to incomplete lineage sorting, where a specific gene phylogeny is not congruent with the species phy-
genetic due to the evolutionary time-based selection pressures on different parts of a given genome (Degnan and Rosenberg 2006; Maddison and Knowles 2006). It is quite possible that this selection pressure did not allow the convergence of ancestries of individual genes of Pasteuria spp. to their overall observed phy-
genetic. Another phenomenon known to contribute to phyloge-
netic discordance is HGT which is known to facilitate bacte-
rial diversity and plays a major role in bacterial evolution (Dutta and Pan 2002; Philippe and Douday 2003; Maddison and Knowles 2006). HGT events are also known to be an important source of ecological variances between closely related taxa (Cohan and Koeppel 2008; Wiedenbeck and Cohan 2011).

Since the nematode parasitic bacteria form a shared habitat with other soil-inhabiting bacteria, and HGT events between differ-
ent species of soil bacteria is well-documented (Andrews et al. 2018), we suggest that these soil bacteria share a common gene pool, and therefore, they may have obtained these collagen-like sequences through HGT.

Interestingly, our results show an unanticipated sequence similarity of Ppcl33 with CLPs from Megaviridae. The members of the Megaviridae family are giant viruses (0.7 μm) that are known to infect protozoans in aquatic ecosystems. They have relatively large genomes (≥1.2 Mbp) which are predicted to encode more than a thousand protein coding genes including metabolic genes not found in any other viruses (Arslan et al. 2011; Colson et al. 2012; Legendre et al. 2012); they also possess hair-like structures on their capsids that are thought to be involved in adhesion and infection of their hosts. It has been postulated that Megaviruses acquired a large set of genes from bacteria during the course of their evolution through HGT (Filleul, Sigueri and Chandler 2007). The fact that P. ramosa is a bac-
terial endosymbiont of water fleas implies that both P. ramosa and Megaviruses possibly share the same ecological niche i.e. aquatic ecosystem and might support the view that HGT had occurred. However, our analysis did not identify any putative Megavirus CLPs in P. ramosa as might be expected. The CLP Ppcl33 was dissimilar to any other of the bacterial CLPs which
suggests a totally different phylogenetic origin. In total, our analysis revealed five clusters of CLPs which would, therefore, suggest that the CLPs in Pasteuria are certainly not of a monophyletic origin; they therefore could have arisen de novo through mutation and subsequent selection, or more speculatively through selection acting upon G-X-Y repeats that had been acquired through several historic HGT events.

SUPPLEMENTARY DATA

Supplementary data are available at FEMSEC online.

FUNDING

The authors would like to thank Indian Council for Agricultural Research for their support to AS as an ICAR International Fellowship and British Council for partial support of this work through their UKIERI grant DST-2013–14/059.

Conflicts of interest. None declared.

ACKNOWLEDGEMENTS

This research forms a part of the PhD thesis of AS carried out at the University of Hertfordshire.

REFERENCES

Anderson J, Preston J, Dickson D et al. Phylogenetic analysis of Pasteuria penetrans by 16S rRNA gene cloning and sequencing. J Natematol 1999;31:319.

Andrews M, De Meyer S, James E et al. Horizontal transfer of symbiosis genes within and between rhizobial genera: occurrence and importance. Genes 2018;9:321.

Arslan D, Legendre M, Seltzer V et al. Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. Proc Natl Acad Sci 2011;108:17486–91.

Atibalenatja N, Jakstys B, Noel G. Life cycle, ultrastructure, and host specificity of the North American isolate of Pasteuria penetrans. Int J Syst Evol Microbiol 2003;53:105–12.

Dutta C, Pan A. Horizontal gene transfer and bacterial diversity. J Biosci 2002;27:23–33.

Eardly B, Wang FS, Van-Berkum P. Corresponding 16S rRNA gene segments in Rhizobiaceae and Aeromonas yield discordant phylogenies. Current Issues in Symbiotic Nitrogen Fixation, In: Elkan, RG. (eds). Springer, Dordrecht. 1996;69–74.

Espanol M, Verdejo-Lucas S, Davies K et al. Compatibility between Pasteuria penetrans isolates and Meloidogyne populations from Spain. Biocon Sci Technol 1997;7:219–30.

Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 1985;39:783–91.

Fléée J, Siguiér P, Chandler M. I am what I eat and I eat what I am: acquisition of bacterial genes by giant viruses. Trends Genet 2007;23:10–15.

Gascuel O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 1997:14:685–95.

Giblin-Davis RM, Nong G, Preston JF et al. ‘Candidatus Pasteuria aldrichii’, an obligate endoparasite of the bacterivorous nematode Bursilla. Int J Syst Evol Microbiol 2011;61:2073–80.

Giblin-Davis RM, Williams DS, Bekal S et al. ‘Candidatus Pasteuria usgae’ sp. nov., an obligate endoparasite of the phytoparasitic nematode Belonolaimus longicaudatus. Int J Syst Evol Microbiol 2003;53:197–200.

Huynen MA, Bork P. Measuring genome evolution. Proc Natl Acad Sci 1998;95:5849–56.

Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Comp Appl Biosci CABIOS 1992;8:275–82.

Kanehisa M, Goto S, Kawashima S et al. The KEGG databases at GenomeNet. Nucleic Acids Res 2002;30:42–46.

Kaufman L, Rousseeuw PJ. Partitioning around medoids (proclustam). In: Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley, New York. 1990, 68–125.

Kumar S, Stecher G, Tamura K. MEGA7: molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol Biol Evol 2016;33:1870–4.

Legendre M, Arslan D, Abergel C et al. Genomics of Megavirus and the elusive fourth domain of Life. Commun Integr Biol 2012;5:102–6.

Luijckx P, Ben-Ami F, Mouton L et al. Cloning of the unculturable parasite Pasteuria ramosa and its Daphnia host.
reveals extreme genotype–genotype interactions. Ecol Letters 2011;14:125–31.
Luijkx P, Fienberg H, Duneau D et al. A matching-allele model explains host resistance to parasites. Curr Biol 2013;23:1085–8.
Maddison WP, Knowles LL. Inferring phylogeny despite incomplete lineage sorting. Syst Biol 2006;55:21–30.
Mauchline TH, Knox R, Mohan S et al. Identification of new single nucleotide polymorphism-based markers for inter- and intraspecies discrimination of obligate bacterial parasites (Pasteuria spp.) of invertebrates. Appl Environ Microbiol 2011;77:6388–94.
McElroy K, Mouton L, Du Pasquier L et al. Characterisation of a large family of polymorphic collagen-like proteins in the endospore-forming bacterium Pasteuria ramosa. Res Microbiol 2011;162:701–14.
Metchnikoff E. Pasteuria ramosa un représentant des bactéries à division longitudinale. Annales de l’Institut Pasteur (Paris) 1888;2:165–70.
Mohan S, Fould S, Davies KG. The interaction between the gelatin-binding domain of fibronectin and the attachment of Pasteuria penetrans endospores to nematode cuticle. Parasitology 2001;123:271–6.
Mohan S, Mauchline TH, Rowe J et al. Pasteuria endospores from Heterodera cajani (Nematoda: Heteroderidae) exhibit inverted attachment and altered germination in cross-infection studies with Globodera pallida (Nematoda: Heteroderidae). FEMS Microbiol Ecol 2012;79:675–84.
Mouton L, Traunecker E, McElroy K et al. Identification of a polymorphic collagen-like protein in the crustacean bacteria Pasteuria ramosa. Res Microbiol 2009;160:792–9.
Murtagh F, Legendre P. Ward’s hierarchical clustering method: clustering criterion and agglomerative algorithm. J Class. 2014;31:274–295.
Noel GR, Atibalenjia N, Domier LL. Emended description of Pasteuria nishizawae. Int J Syst Evol Microbiol 2005;55:1681–5.
Philippe H, Douady CJ. Horizontal gene transfer and phylogenetics. Curr Opin Microbiol 2003;6:498–505.
R Developement Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing, 2010, Retrieved from http://www.R-project.org
Radó-Trilla N, Albá M. Dissecting the role of low-complexity regions in the evolution of vertebrate proteins. BMC Evol Biol 2012;12:155.
Rajendran J, Gunasekaran P. Microbial phylogeny and diversity: small subunit ribosomal RNA sequence analysis and beyond. Microbiol Res 2011;166:99–110.
Romero PA, Arnold FH. Exploring protein fitness landscapes by directed evolution. Nature Rev Mol Cell Biol 2009;10:866–76.
Rutherford K, Parkhill J, Crook J et al. Artemis: sequence visualization and annotation. Bioinformatics 2000;16:944–5.
Sayre R, Starr M. Pasteuria penetrans (ex Thome, 1940) nom. rev., comb. n., sp. n., a mycellial and endospore-forming bacterium parasitic in plant-parasitic nematodes. Proc Helminthol soc Washington 1985;52:149–65.
Sayre R, Wergin W, Schmidt J et al. Pasteuria nishizawae sp. nov., a mycellial and endospore-forming bacterium parasitic on cyst nematodes of genera Heterodera and Globodera. Res Microbiol 1991;142:551–64.
Schouls LM, Schot CS, Jacobs JA. Horizontal transfer of segments of the 16S rRNA genes between species of the Streptococcus anginosus group. J Bacteriol 2003;185:7241–6.
Sharma S, Davies K. Characterisation of Pasteuria isolated from Heterodera cajani using morphology, pathology and serology of endospores. Syst Appl Microbiol 1996;19:106–12.
Stackebrandt E. The Family Pasteuriaceae. The Prokaryotes. Springer. 2014, 281–4.
Starr M, Sayre R. Pasteuria thornei sp. nov. and Pasteuria penetrans sensu stricto emend., mycellial and endospore-forming bacteria parasitic, respectively, on plant-parasitic nematodes of the genera Pratylenchus and Meloidogyne. Paper presented at the Annales de l’Institut Pasteur/Microbiologie 1988;139:11–31.
Steichen C, Chen P, Kearney JF et al. Identification of the immunodominant protein and other proteins of the Bacillus anthracis exosporium. J Bacteriol 2003;185:1903–10.
Stirling G. Host specificity of Pasteuria penetrans within the genus Meloidogyne. Nematol 1985;31:203–9.
Stirling G. Biological Control of Plant-parasitic Nematodes: Soil Ecosystem Management in Sustainable Agriculture. 2nd Edn. Wallingford: CAB International, 2014, 304–41.
Sturhan D. New host and geographical records of nematode-parasitic bacteria of the Pasteuria penetrans group. Nematol 1988;34:350–6.
Sylvestre P, Couture-Tosi E, Mock M. Polymorphism in the Collagen-Like Region of the Bacillus anthracis BclA Protein Leads to Variation in Exosporium Filament Length. J Bacteriol 2003;185:1555–63.
Sylvestre P, Mock M, Couture-Tosi E. A collagen-like surface glycoprotein is a structural component of the Bacillus anthracis exosporium. Mol Microbiol 2002;45:169–78.
Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 1993;10:512–26.
Todd SJ, Moir AJG, Johnson MJ et al. Genes of Bacillus cereus and Bacillus anthracis Encoding Proteins of the Exosporium. J Bacteriol 2003;185:3373–8.
Wiedenbeck J, Cohan FM. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. FEMS Microbiol Rev 2011;35:957–76.
Woese CR. Bacterial evolution. Microbiol Rev 1987;51:221.
Yap WH, Zhang Z, Wang Y. Distinct types of rRNA operons exist in the genome of the actinomycete Thermomonospora chromogena and evidence for horizontal transfer of an entire rRNA operon. J Bacteriol 1999;181:5201–9.
Zilversmit MM, Volkman SK, DePristo MA et al. Low-complexity regions in Plasmodium falciparum: missing links in the evolution of an extreme genome. Mol Biol Evol 2010;27:2198–209.