Antonie van Leeuwenhoek (2012) 101:105–124
DOI 10.1007/s10482-011-9628-4

ORIGINAL PAPER

Molecular signatures (conserved indels) in protein sequences that are specific for the order Pasteurellales and distinguish two of its main clades

Hafiz Sohail Naushad · Radhey S. Gupta

Received: 21 July 2011 / Accepted: 29 July 2011 / Published online: 10 August 2011
© Springer Science+Business Media B.V. 2011

Abstract The members of the order Pasteurellales are currently distinguished primarily on the basis of their branching in the rRNA trees and no convincing biochemical or molecular markers are known that distinguish them from all other bacteria. The genome sequences for 20 *Pasteurellaceae* species/strains are now publicly available. We report here detailed analyses of protein sequences from these genomes to identify conserved signature indels (CSIs) that are specific for either all Pasteurellales or its major clades. We describe more than 23 CSIs in widely distributed genes/proteins that are uniquely shared by all sequenced *Pasteurellaceae* species/strains but are not found in any other bacteria. Twenty-one additional CSIs are also specific for the Pasteurellales except in some of these cases homologues were not detected in a few species or the CSI was also present in an isolated non-*Pasteurellaceae* species. The sequenced *Pasteurellaceae* species formed two distinct clades in a phylogenetic tree based upon concatenated sequences for 10 conserved proteins. The first of these clades consisting of *Aggregatibacter, Pasteurella, Actinobacillus succinogenes, Mannheimia succiniciproducens*, *Haemophilus influenzae* and *Haemophilus somnus* was also independently supported by 13 uniquely shared CSIs that are not present in other *Pasteurellaceae* species or other bacteria. Another clade consisting of the remaining *Pasteurellaceae* species (viz. *Actinobacillus pleuropneumoniae*, *Actinobacillus minor, Haemophilus ducryi, Mannheimia haemolytica* and *Haemophilus parasuis*) was also strongly and independently supported by nine CSIs that are uniquely present in these bacteria. The order Pasteurellales is presently made up of a single family, *Pasteurellaceae*, that encompasses all of its genera. In this context, our identification of two distinct clades within the Pasteurellales, which are supported by both phylogenetic analyses and by multiple highly specific molecular markers, strongly argues for and provides potential means for the division of various genera from this order into a minimum of two families. The genetic changes responsible for these CSIs were likely introduced in the common ancestors of either all Pasteurellales or of these two specific clades. These CSIs provide novel means for the identification and circumscription of these groups of Pasteurellales in molecular terms.

Keywords Conserved indels · Pasteurellales taxonomy and systematics · Pasteurellales clades · Phylogenetic analyses · *Pasteurellaceae* genomes · Comparative genomics · Molecular markers for Pasteurellales · Lateral gene transfers

Electronic supplementary material The online version of this article (doi:10.1007/s10482-011-9628-4) contains supplementary material, which is available to authorized users.

H. S. Naushad · R. S. Gupta (✉)
Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, ON L8N 3Z5, Canada
e-mail: gupta@mcmaster.ca

Table 1 Sequence characteristics of the Pasteurellales genomes

Organism	GenBank accession No.	Size (Mbp)	No. of proteins	% GC content	Reference
Actinobacillus pleuropneumoniae L20	CP000569	2.3	2012	41.3	Foote et al. (2008)
Actinobacillus pleuropneumoniae serovar 3 str. JL03	CP000687	2.2	2036	41.2	Xu et al. (2008)
Actinobacillus pleuropneumoniae serovar 7 str. AP76	CP001091	2.3	2131	41.2	STHH[b]
Actinobacillus succinogenes 130Z	CP000746	2.3	2079	44.9	DOE-JGI
Actinobacillus minor 202	ACFT00000000	2.1	2050	39.3	McGill University[c]
Aggregatibacter actinomycetemcomitans D11S-1	CP001733	2.2	2135	44.3	Chen et al. (2009)
Aggregatibacter aphrophilus NJ8700	CP001607	2.3	2219	42.2	Di Bonaventura et al. (2009)[c]
Haemophilus ducreyi 35000HP	AE017143	1.7	1717	38.2	Ohio State University[a]
Haemophilus influenzae 86-028NP	CP000057	1.9	1792	38.2	Harrison et al. (2005)
Haemophilus influenzae PittEE	CP000671	1.8	1613	38.0	Hogg et al. (2007)
Haemophilus influenzae PittGG	CP000672	1.9	1661	38.0	Hogg et al. (2007)
Haemophilus influenzae Rd KW20	L42023	1.8	1657	38.2	Fleischmann et al. (1995)
Haemophilus influenzae R2846	CP002276	1.8	1691	38.0	UW-BRI
Haemophilus influenzae R2866	CP002277	1.9	1817	38.1	UW-BRI
Haemophilus parasuis SH0165	CP001321	2.3	2021	40.0	Yue et al. (2009)
Haemophilus somnus 129PT	CP000436	2.0	1792	37.2	Barabote et al. (2009)
Haemophilus somnus 2336	CP000947	2.3	1980	37.4	Virginia Tech
Mannheimia haemolytica[c]	AASA01000000	2.6	2839	41.1	Gioia et al. (2006)
Mannheimia succiniciproducens MBEL55E	AE016827	2.3	2369	42.5	Hong et al. (2004)
Pasteurella multocida subsp. *multocida* str. Pm70	AE004439	2.3	2015	40.4	May et al. (2001)

UW-BRI University of Washington; Seattle Biomedical Research Institute, *DOE-JGI* Genome is sequenced by the Department of Education Joint Genome Institute

[a] Sequenced by Ohio State University

[b] Sequenced by Stiftung Tieraerztliche Hochschule Hannover (STHH)

[c] Draft genomes. The sequences for *Actinobacillus minor* 202 and NM305 are being sequenced by McGill University

Introduction

The members of the order Pasteurellales are Gram-negative, non-motile and aerobic to facultative anaerobic bacteria, which constitute one of the main orders within the Class Gammaproteobacteria (Pohl 1981; Mutters et al. 1989; Paster et al. 1993; Olsen et al. 2005; Christensen et al. 2007; Christensen and Bisgaard 2010). The order Pasteurellales presently contains a single family, *Pasteurellaceae*, that is made up of at least 15 genera and >70 species (see http://www.the-icsp.org/taxa/Pasteurellaceaelist.htm; Christensen and Bisgaard 2010). These bacteria are commonly present as commensals in the mucosal membranes of the respiratory, alimentary and reproductive tracts of various vertebrates (mainly birds and mammals) including humans (Bisgaard 1993; Olsen et al. 2005; Christensen and Bisgaard 2010). The presence of these bacteria in both healthy as well as diseased vertebrates indicates that they are opportunistic pathogens and several of them are important human and animal pathogens. For example, *Haemophilus influenzae, Haemophilus ducreyi* and *Aggregatibacter* (*Agg.*) *actinomycetemcomitans* are respectively involved in the causation of bacteremia, pneumonia and acute bacterial meningitis; the

sexually transmitted disease chancroid; and juvenile periodontitis in humans (Bisgaard 1993; Fleischmann et al. 1995; Spinola et al. 2002; Olsen et al. 2005; Christensen and Bisgaard 2010). Other species such as *Mannheimia* (*Man.*) *haemolytica, Pasteurella multocida* and *Actinobacillus* (*Act.*) *pleuropneumoniae* are causative agents of the shipping fever in cattle, fowl cholera and pleuropneumonia in pigs, respectively (Bisgaard 1993; Bosse et al. 2002; Gioia et al. 2006).

The Pasteurellales are presently distinguished from other bacteria primarily on the basis of their branching in 16S rRNA gene sequence trees, where they form a distinct cluster (Mutters et al. 1989; De Ley et al. 1990; Dewhirst et al. 1992; Dewhirst et al. 1993; Olsen et al. 2005; Christensen and Bisgaard, 2006; Christensen and Bisgaard, 2010). The species from this order/family also form a distinct clade in phylogenetic trees based on numerous other genes and protein sequences (Korczak et al. 2004; Christensen et al. 2004; Kuhnert and Korczak, 2006; Gao et al. 2009; Williams et al. 2010). Some morphological and nutritional characteristics such as lack of motility, requirement for sodium ions, V-factor and organic nitrogen sources for growth, are often used to distinguish these bacteria from other orders of Gammaproteobacteria (e.g. Vibrionales, Aeromonadales, Enterobacteriales and Alteromonadales) (Olsen 1993; Kainz et al. 2000; Olsen et al. 2005; Christensen and Bisgaard 2006; Hayashimoto et al. 2007). However, none of these characteristics are unique for the Pasteurellales and reliance only on them can lead to incorrect identification/placement of species in this group and its various genera (Christensen et al. 2004; Olsen et al. 2005; Christensen et al. 2007; Christensen and Bisgaard 2010). Presently, no convincing molecular or biochemical characteristic is known that is uniquely shared by various Pasteurellales and which can be used to clearly distinguish this group of bacteria from all others. Our current understanding of the phylogeny/taxonomy for these bacteria is also unsatisfactory (Olsen et al. 2005; Christensen and Bisgaard 2006). For example, several of the genera classified within Pasteurellales (viz. *Haemophilus, Actinobacillus* and *Mannheimia*) are not monophyletic and species from them branch in a number of different clusters with other members of this group (Olsen et al. 2005; Gioia et al. 2006; Redfield et al. 2006; Christensen and Bisgaard 2006; Christensen and Bisgaard 2010; Bonaventura et al. 2010). Although suggestions have been made to restrict these genera to a limited number of species (Olsen et al. 2005; Christensen and Bisgaard 2006), the taxonomy of members of the Pasteurellales/*Pasteurellaceae* is clearly unsatisfactory at present (Christensen et al. 2007; Christensen and Bisgaard, 2010; Bonaventura et al. 2010). Thus, it is important to identify other novel sequence based characteristics that could provide reliable means for the identification of species from this order and which could also prove useful in clarifying their taxonomy and evolutionary relationships.

Since the sequencing of first genome for *H. influenzae* in 1995 (Fleischmann et al. 1995), sequence data for more than 1500 bacteria covering all major bacterial phyla are now available (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html). Of these genomes, 20 genomes are from Pasteurellales species/strains representing five genera from this family (Table 1). These genome sequences provide an unprecedented and valuable resource for discovering novel molecular characteristics that are uniquely shared by either all Pasteurellales or specific groups/clades of these bacteria and could provide more reliable means for their identification (Shah et al. 2009). Using genomic sequences, our recent work has focused on identifying two different types of molecular markers that are specific for different groups of bacteria. One type of molecular markers consists of conserved signature inserts or deletions (i.e. Indels) (CSIs) in widely distributed proteins, that are specifically present in particular groups of bacteria (Gupta 2000; Gupta and Mok 2007; Gupta 2009; Gupta 2010). The whole proteins that are uniquely present in particular groups of bacteria provide another type of molecular markers that are useful for these studies (Gupta 2006; Gupta and Griffiths 2006; Gupta and Mathews 2010). Our recent work has identified large numbers of CSIs for a number of major taxa within bacteria (viz. Alphaproteobacteria, Epsilonproteobacteria, Chlamydiae, Actinobacteria, Cyanobacteria, Bacteroidetes-Chlorobi, Deinococcus-Thermus) and for many of their subgroups (Gupta and Griffiths 2006; Gupta and Mathews 2010). Recently, some molecular signatures for the Class Gammaproteobacteria as a whole were also identified (Gao et al. 2009).

In the present work, we have employed these comparative genomic approaches in conjunction with phylogenetic analysis for investigation of the

Table 2 Conserved Signature Indels that are specific for all Pasteurellales

Protein name	Gene name	Accession no.	Figure nos.	Indel size	Indel position[a]	Functional categories
Tetratricopeptide domain protein	–	YP_003006869	Fig. 2a	8 aa ins	44–91	Carbohydrate transport and metabolism
Murein transglycosylase C	mltC	YP_001343852	Supplementary Fig. 1	3 aa del	76–116	Cell wall/membrane biogenesis
Exoribonuclease II	rnb	NP_873703	Supplementary Fig. 2	10 aa ins	416–468	Transcription
Glycerol-3-phosphate acyltransferase	plsB	YP_003255015	Supplementary Fig. 3	2 aa ins	554–610	Lipid transport and metabolism
3-phosphoshikimate 1-carboxyvinyltransferase	aroA	YP_003256375	Supplementary Fig. 4	2 aa ins	360–402	Amino acid transport and metabolism
Hypothetical protein CGSHiEE_05875	–	YP_001290919	Supplementary Fig. 5	2 aa ins	32–71	General function prediction only
5-methylaminomethyl-2-thiouridine methyltransferase	mnmC	YP_003255458	Supplementary Fig. 6	2 aa del	122–152	Multifunctional
Adenylate cyclase[b]	cyaA	NP_873154	Supplementary Fig. 7	2 aa del	526–576	Nucleotide transport and metabolism
Murein transglycosylase A	mltA	NP_874023	Supplementary Fig. 8	1 aa del	241–286	Cell wall/membrane biogenesis
Lipoyltransferase	lipB	YP_001344010	Supplementary Fig. 9	1–5 aa ins	75–116	Coenzyme transport and metabolism
Transcription repair coupling factor	mfd	NP_873467	Supplementary Fig. 10A	1 aa ins	226–261	Replication, recombination and repair
Fumarate reductase flavoprotein subunit	frdA	NP_872657	Supplementary Fig. 10B	1 aa ins	287–331	Energy production and conversion
Hemolysin	corB	YP_003008000	Supplementary Fig. 11	1 aa ins	228–270	Inorganic ion transport and metabolism
Chaperonin HslO	hslO	ZP_05919977	Supplementary Fig. 12	1 aa ins	246–278	Posttranslational modification, protein turnover and chaperones
Exodeoxyribonuclease VII small subunit	xseB	ZP_01791820	Supplementary Fig. 13	1 aa ins	27–68	Replication, recombination and repair
Periplasmic serine peptidase DegS	degS	ZP_05850718	Supplementary Fig. 14	1 aa ins	190–216	Posttranslational modification, protein turnover and chaperones
Multidrug resistance protein MdtK	mdtK	YP_003007368	Supplementary Fig. 15	1 aa ins	200–249	Defense mechanisms
Glutamate-ammonia-ligase adenylyltransferase	glnE	YP_088470	Supplementary Fig. 16	1 aa ins	271–309	Multifunctional
Hypothetical protein PM0734	–	NP_245671	Supplementary Fig. 17	1 aa ins	184–212	Hypothetical
Hypothetical protein HD1793	–	NP_874155	Supplementary Fig. 18	1 aa ins	168–200	Hypothetical
Hypothetical protein HD1794	–	NP_874156	Supplementary Fig. 19	1 aa ins	75–109	Hypothetical
Peptidyl-prolyl cis–trans isomerase B	ppiB	ZP_06222848	Supplementary Fig. 20	6 aa ins	43–75	Posttranslational modification, protein turnover and chaperones

Table 2 continued

Protein name	Gene name	Accession no.	Figure nos.	Indel size	Indel position[a]	Functional categories
Peptidyl-prolyl cis–trans isomerase B	ppiB	YP_003007916	Supplementary Fig. 21	6 aa ins	100–137	Posttranslational modification, protein turnover and chaperones
Nicotinamide-nucleotide adenylyltransferase[c]	nadR	YP_003255205	Supplementary Fig. 22	1 aa ins	121–151	Coenzyme transport and metabolism
N-acetyl-D-glucosamine kinase (GlcNAc kinase)[c]	nagK	YP_003007117	Supplementary Fig. 23	1 aa ins	153–195	Multifunctional
Putative inner membrane protein[c]	–	ZP_02478497	Supplementary Fig. 24	1 aa ins	197–222	Cell wall/membrane
Galactokinase[c]	galK	YP_003007703	Supplementary Fig. 25	3 aa ins	240–276	Carbohydrate transport and metabolism
Deoxyguanosinetriphosphate triphosphohydrolase-like protein[c]	–	YP_001344904	Supplementary Fig. 26	17 aa ins	59–126	Nucleotide transport and metabolism
Inner membrane protein YicO[c]	yicO	YP_003007341	Supplementary Fig. 27	1 aa ins	199–237	General function prediction only
PTS system, fructose subfamily, IIC subunit[c]	fruA	YP_001343401	Supplementary Fig. 28	3 aa ins	241–281	Carbohydrate transport and metabolism
Anion transporter[c]	–	YP_001343337	Supplementary Fig. 29	7 aa ins	258–296	Inorganic ion transport and metabolism
Hypothetical protein PM0935[c]	–	NP_245872	Supplementary Fig. 30	4 aa ins	61–108	Hypothetical
23S rRNA (guanosine-2′-*O*-)-methyltransferase[d]	rlmB	ZP_05629947	Supplementary Fig. 31	1 aa ins	115–178	Posttranslational modification, protein turnover, chaperones
Glutamate ammonia ligase adenylyltransferase[d]	glnE	NP_874080	Supplementary Fig. 32	17 aa ins	381–436	Multifunctional
Murein transglycosylase C[d]	mltC	YP_001343852	Supplementary Fig. 33	1 aa ins	148–180	Cell wall/membrane biogenesis
ProS protein[d]	proS	AAU38670	Supplementary Fig. 34	1 aa ins	453–482	Translation
D-methionine-binding lipoprotein[d]	metQ	YP_003008527	Supplementary Fig. 35	1 aa ins	97–130	Inorganic ion transport and metabolism
DNA-dependent helicase II[e]	uvrD	YP_001293092	Fig. 2B	3–4 aa ins	61–104	Replication, recombination and repair
Hypothetical protein NT05HA_0747[e]	–	YP_003007227	Supplementary Fig. 36A	2 aa ins	36–68	Unknown
Lysyl-tRNA synthetase[e]	genX	NP_245139	Supplementary Fig. 36B	2 aa del	148–191	Translation
Protein cof[e]	–	YP_003008147	Supplementary Fig. 37	1 aa ins	45–80	General function prediction only
6-phosphogluconolactonase[e]	pgl	NP_873341	Supplementary Fig. 38	4 aa del	97–145	Carbohydrate transport and metabolism
Geranyltranstransferase[e]	ispA	ZP_04977790	Supplementary Fig. 39	2 aa del	112–150	Coenzyme transport and metabolism

Table 2 continued

Protein name	Gene name	Accession no.	Figure nos.	Indel size	Indel position[a]	Functional categories
DNA repair protein RecN[e]	recN	YP_002475883	Supplementary Fig. 40	3 aa ins	68–106	Replication, recombination and repair

[a] The indel position indicates the region of the protein where a given CSI is present

[b] A 1 aa deletion is present in *H. parasuis* rather than the 2 aa deletion found in all Pasteurellales

[c] Homologous sequences corresponding to this region were not identified in some Pasteurellales species

[d] The CSI is not present in 1–2 Pasteurellales species

[e] The CSI is also found in 1–2 non-Pasteurellales species

available Pasteurellales genomes. The primary objective of this work is to identify novel molecular markers consisting of conserved signature indels (CSIs) that are unique to either all Pasteurellales or its major subgroups/clades. Our work has identified >40 CSIs that are specific for all (or most) genome sequenced Pasteurellales species/strains. In addition, we also describe many CSIs that are specific for a number of distinct subclades of Pasteurellales, which are also supported by phylogenetic analyses. These molecular signatures provide valuable means for the identification of members of the Pasteurellales and a number of their subclades and for the division of Pasteurellales into two distinct groups.

Methods

Phylogenetic analysis

Phylogenetic analysis was performed on a concatenated sequence alignment for 10 highly conserved proteins (viz. 50S ribosomal protein L5, RNA polymerase subunit beta (RpoB), prolyl-tRNA synthetase, chaperone protein DnaK, threonyl-tRNA synthetase, valyl-tRNA synthetase, cell division protein FtsY, alanyl-tRNA synthetase, translation initiation factor IF-2, DNA gyrase subunit B) that are present in most extant bacteria (Harris et al. 2003) and which have been extensively used for phylogenetic studies (Korczak et al. 2004; Christensen et al. 2004; Gao et al. 2009; Gupta 2009). The sequences for these proteins for various Pasteurellales and several other Gammaproteobacteria, which served as outgroup, were retrieved and multiple sequence alignments for them were created using the CLUSTAL_X 1.83 program (Jeanmougin et al. 1998). After concatenation, the poorly aligned regions from the sequence alignment were removed using the Gblocks 0.91b program (Castresana 2000). The resulting alignment, which consisted of 6783 characters, was employed for phylogenetic analyses. A neighbour-joining (NJ) tree based upon 500 bootstrap replicates of this sequence alignment was constructed employing Kimura's distance calculation using the TREECON 1.3 program (Van de Peer and De Wachter 1994).

Identification of CSIs for members of the order Pasteurellales

To identify conserved indels in protein sequences that might be specific for the Pasteurellales, Blastp searches were performed on all proteins from the genome of *Aggregatibacter aphrophilus* NJ8700 (Di Bonaventura et al. 2009). For those proteins/ORFs for whom high scoring homologues were present in most Pasteurellales species/strains as well as certain outgroup species, sequences for 10–15 high scoring homologues were retrieved from diverse Pasteurellales and other bacteria and their multiple sequence alignments were constructed using the Clustal_X 1.83 program. These sequence alignments were visually inspected to identify any conserved inserts or deletions that were restricted to either all Pasteurellales or its major clades and which were flanked by at least 5–6 identical/conserved residues in the neighboring 30–40 amino acids on each side. The indels that were not flanked by conserved regions were not further studied as they do not provide useful molecular markers (Gupta 1998; Gupta 2000; Gupta 2009). The conserved indels, which in addition to the

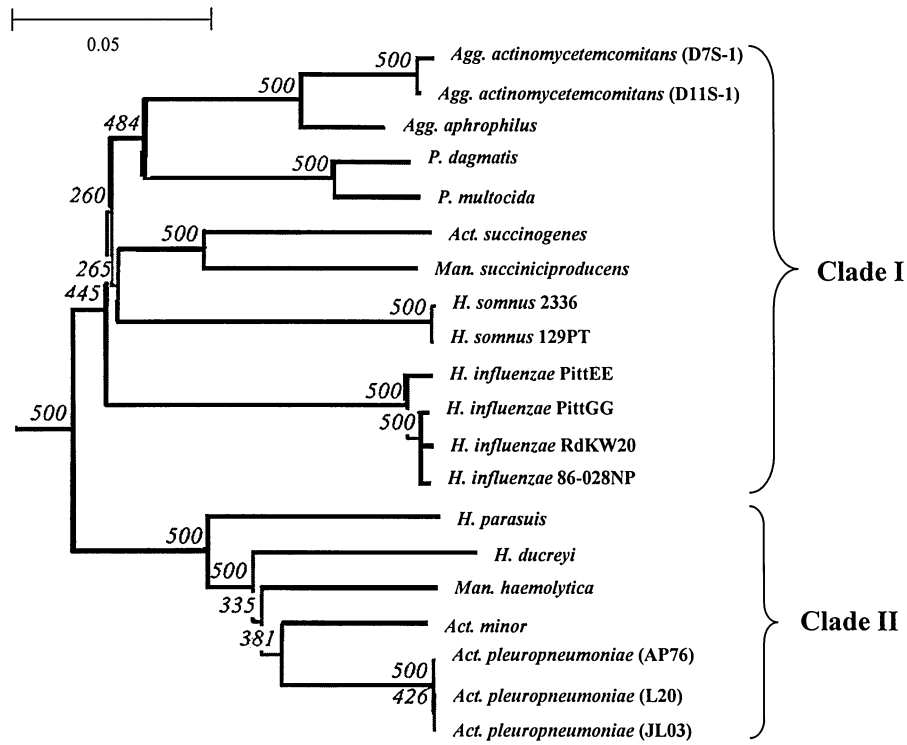


Fig. 1 A neighbor-joining distance tree for the sequenced Pasteurellales based upon concatenated sequences for 10 conserved proteins. The tree was rooted using sequences for other Gammaproteobacteria (viz. Vibrionales or Enterobacteriales) and the *numbers* on the nodes indicate the bootstrap values out of 500. The two main clades of Pasteurellales that are seen in the tree are marked

Pasteurellales were also present in a few other bacteria, were also retained. The indels for individual species or smaller clades were not analyzed in detail in the present work. The species distribution patterns of all such indels were further evaluated by detailed Blastp searches on short sequence segments containing the indels and their flanking conserved regions (Gupta 2009). The sequence information for various conserved indels from all Pasteurellales and some representative high scoring Gammaproteobacteria were compiled into signature files. Due to space consideration, sequence information for different strains of the same species is not shown, but the indicated CSIs were present in all of the sequenced strains. Further, unless otherwise noted, all of these CSIs are specific for the indicated groups.

Results

Phylogenetic analysis of Pasteurellales

The evolutionary relationships among Pasteurellales in the past was mainly examined on the basis of phylogenetic trees for the 16S rRNA gene and a number of individual protein sequences (Dewhirst et al. 1993; Korczak et al. 2004; Christensen et al. 2004; Olsen et al. 2005; Christensen and Bisgaard 2006). However, the availability of genome sequences now enables one to determine the branching order of these species based upon concatenated sequences for large numbers of proteins. The trees based upon large numbers of characters derived from multiple proteins provide more reliable indication of the phylogenetic relationships within a given group than those based on any single gene or protein (Rokas et al. 2003; Ciccarelli et al. 2006; Gao et al. 2009; Wu et al. 2009; Williams et al. 2010). Previously, Redfield et al. (2006) and Gioia et. al. (2006) have reported construction of phylogenetic trees for eight Pasteurellales species (viz. *H. influenzae, H. ducreyi*, *Haemophilus somnus*, *P. multocida*, *Act. pleuropneumoniae*, *Agg. actinomycetemcomitans*, *Mannheimia succiniciproducens* and *Man. haemolytica*, based upon concatenated sequences for 12 and 50 conserved proteins, respectively. More recently, Bonaventura et al. (Bonaventura et al. 2010) have carried out detailed phylogenetic analyses for 12 Pasteurellales genomes representing 10 species (the

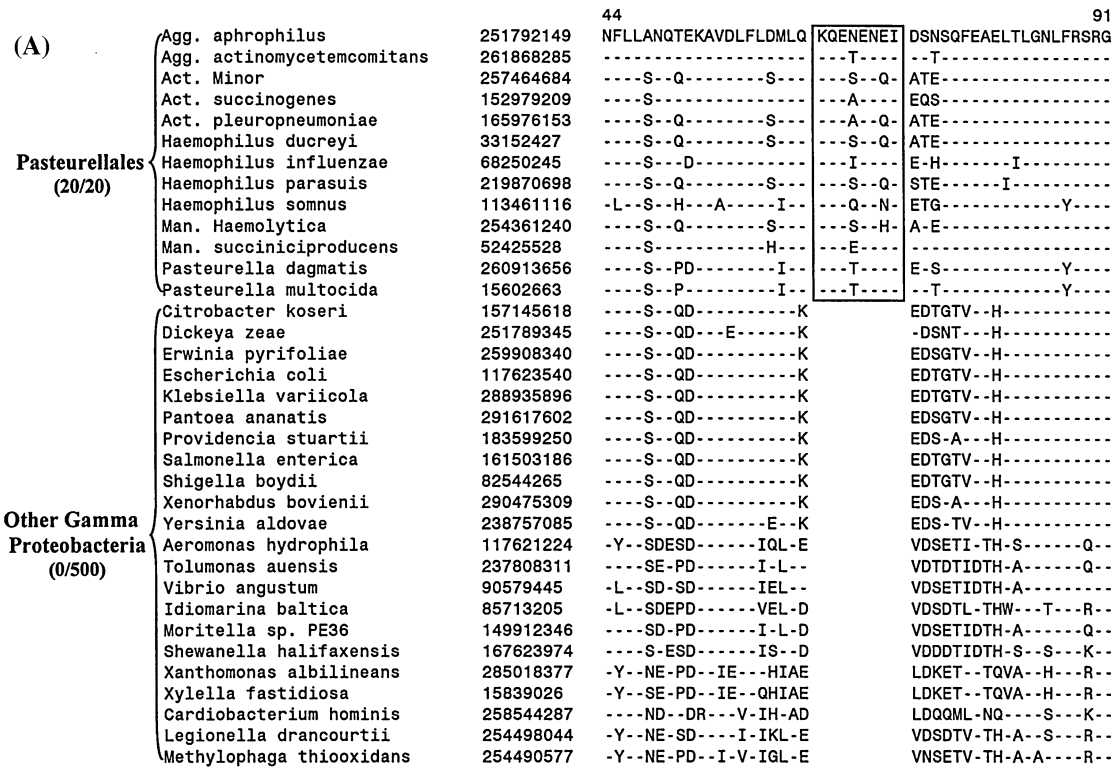
above eight species plus *Agg. aphrophilus* and *Actinobacillus succinogens*) based upon concatenated sequences for different orthologous proteins found in their genomes. Although, these trees provide useful resources for understanding the evolutionary relationships among the indicated *Pasteurellaceae* species/strains, in the past 2–3 years sequences for a number of new *Pasteurellaceae* species (viz. *Haemophilus parasuis, Actinobacillus minor* and *Pasteurella dagmatis*), as well as additional strains for several species, have become available in the NCBI database (Table 1). A few characteristics of these genomes, some of which are draft genomes, are listed in Table 1. In order to determine the evolutionary significances of various CSIs identified by our analyses, it was necessary to construct a phylogenetic tree that included sequence information for all of these Pasteurellales. In the present work, phylogenetic trees for 20 Pasteurellales species/strains representing 13 species were constructed based upon concatenated sequences for 10 conserved proteins.

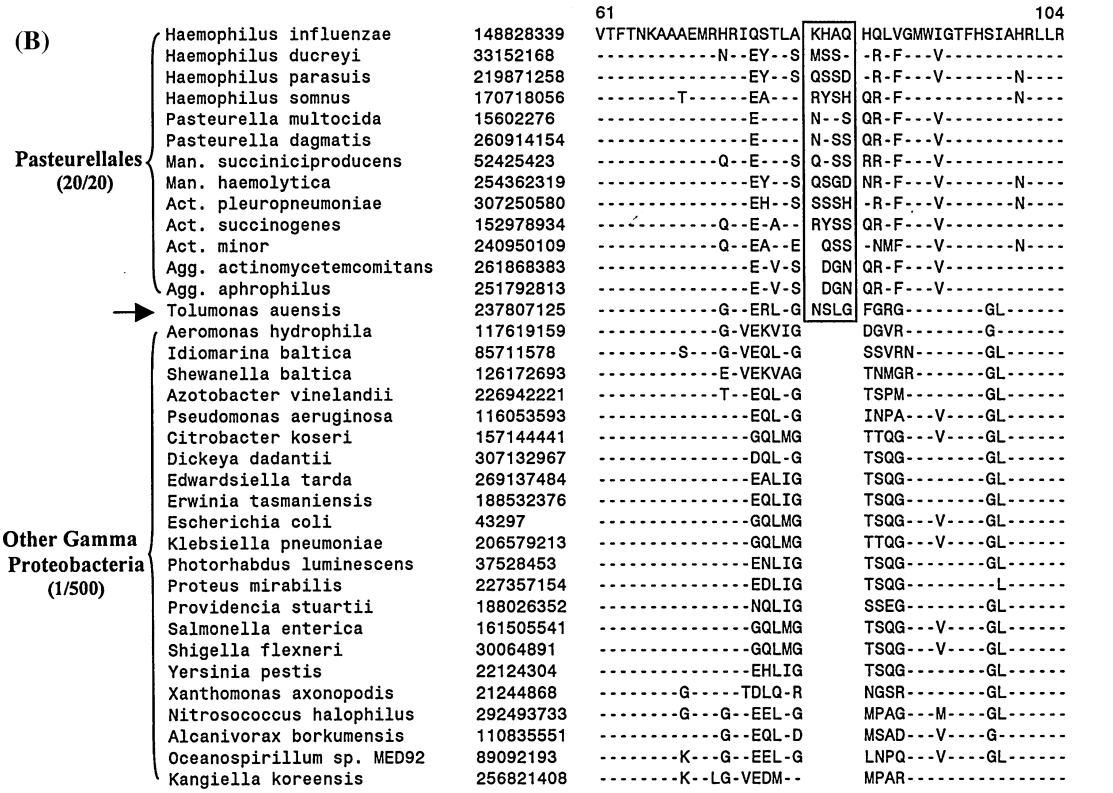
A NJ distance tree for the above Pasteurellales species that was rooted using other Gammaproteobacteria (viz. Vibrionales or Aeromonadales) is shown in Fig. 1. As expected, the Pasteurellales species formed a distinct and strongly supported clade in the tree. Further, as observed in earlier studies, species from a number of Pasteurellales genera viz. *Haemophilus, Actinobacillus* and *Mannheimia* branched in a number of different clusters, indicating that these genera are not monophyletic. In the NJ tree shown, the Pasteurellales species formed two main clades. The first of these clades (Clade I) consists of various *Aggregatibacter* and *Pasteurella* species and it also included *Act. succinogenes, Man. succiniciproducens* and various strains of *H. influenzae* and *H. somnus*. Within this clade, the grouping of *Aggregatibacter* with *Pasteurella* species and that of *Act. succinogenes* with *Man. succiniciproducens* was strongly supported. The second clade (Clade II) consisted of *H. ducryi, H. parasuis, Man. haemolytica* and various strains of *Act. pleuropneumoniae*. These two clades of Pasteurellales were also supported by earlier phylogenetic studies based upon different datasets of protein sequences (Gioia et al. 2006; Redfield et al. 2006; Bonaventura et al. 2010). These trees provide us a phylogenetic framework to understand/interpret the evolutionary significance of various identified CSIs.

Fig. 2 Partial sequence alignments of the proteins **a** a tetratricopeptide domain-containing protein showing a conserved CSI (*boxed*) that is uniquely present in all Pasteurellales species and **b** DNA-dependent helicase II, showing a conserved insert (*boxed*) that is largely specific for all Pasteurellales. However, in this case the CSI was also present in one non-Pasteurellales species (marked with *arrow*). The shared presence of the CSI in this species could be due to LGTs, however, other possibilities cannot be excluded. The *dashes* in the sequence alignments indicate identity with the amino acid on the *top* line. The numbers on the *top* lines indicate the regions of proteins where these CSIs are present in the species shown on the *top*. Sequence information for other bacteria is shown here for only a limited number of species. However, no other species within the first 500 blast hits contained the indicated indels. Information for many other CSIs that are specific for all Pasteurellales is provided in Table 2 ▶

Identification of conserved indels that are specific for the order Pasteurellales

Our analyses have identified 44 CSIs in broadly distributed proteins that are largely specific for most of the sequenced Pasteurellales species (Table 2). The CSIs in the first 23 proteins listed in this table are commonly shared by all sequenced Pasteurellales species/strains but they are not found in the homologues from any other bacteria (at least the top 500 blast hits). One example of these Pasteurellales-specific CSIs is shown in Fig. 2a. In this case, an 8 aa insert in a highly conserved region of a tetratricopeptide (TPR) domain-containing protein is uniquely present in all sequenced Pasteurellales. Although, sequence information is presented here for only a limited number of species, unless indicated otherwise, the CSI shown here as well as other molecular signatures shown are specific for the Pasteurellales group and not found elsewhere. Other CSIs that are uniquely present in all Pasteurellales are listed in Table 2 and the sequence alignments of these proteins showing the presence of the indicated CSIs are provided as Supplementary Figs. 1–21. Of these, the enzyme peptidyl-prolyl cis-tran isomerase B contains two 6 aa inserts in different positions that are specifically present in all sequenced Pasteurellales. However, there are two homologues of this protein in *P. multocida, P. dagmatis* and *Man. succiniciproducens* and these CSI are present in only one of the homologues (Supplementary Figs. 20, 21). Five other proteins listed in Table 2 (Supplementary Figs. 22–26), also contain CSIs that are specific for the *Pasteurellaceae* species. However, the homologues


(A)
44
91
Pasteurellales (20/20)
Other Gamma Proteobacteria (0/500)
Agg. aphrophilus 251792149 NFLLANQTEKAVDLFLDMLQ KQENENEI DSNSQFEAELTLGNLFRSRG
Agg. actinomycetemcomitans 261868285 ---T---- --T----------
Act. Minor 257464684 ----S--Q--------S--- ---S--Q- ATE----------
Act. succinogenes 152979209 ----S---------- ---A---- EQS----------
Act. pleuropneumoniae 165976153 ----S--Q--------S--- ---A--Q- ATE----------
Haemophilus ducreyi 33152427 ----S--Q--------S--- ---S--Q- ATE----------
Haemophilus influenzae 68250245 ----S---D---------- ---I---- E-H-------I---------
Haemophilus parasuis 219870698 ----S--Q--------S--- ---S--Q- STE------I----------
Haemophilus somnus 113461116 -L--S--H---A-----I-- ---Q--N- ETG-----------Y----
Man. Haemolytica 254361240 ----S--Q--------S--- ---S--H- A-E----------
Man. succiniciproducens 52425528 ----S---------H--- ---E---- ----------
Pasteurella dagmatis 260913656 ----S--PD--------I-- ---T---- E-S-----------Y----
Pasteurella multocida 15602663 ----S--P---------I-- ---T---- --T-----------Y----
Citrobacter koseri 157145618 ----S--QD----------K EDTGTV--H----------
Dickeya zeae 251789345 ----S--QD---E------K -DSNT---H----------
Erwinia pyrifoliae 259908340 ----S--QD----------K EDSGTV--H----------
Escherichia coli 117623540 ----S--QD----------K EDTGTV--H----------
Klebsiella variicola 288935896 ----S--QD----------K EDTGTV--H----------
Pantoea ananatis 291617602 ----S--QD----------K EDSGTV--H----------
Providencia stuartii 183599250 ----S--QD----------K EDS-A---H----------
Salmonella enterica 161503186 ----S--QD----------K EDTGTV--H----------
Shigella boydii 82544265 ----S--QD----------K EDTGTV--H----------
Xenorhabdus bovienii 290475309 ----S--QD----------K EDS-A---H----------
Yersinia aldovae 238757085 ----S--QD-------E--K EDS-TV--H----------
Aeromonas hydrophila 117621224 -Y--SDESD------IQL-E VDSETI-TH-S------Q--
Tolumonas auensis 237808311 ----SE-PD------I-L-- VDTDTIDTH-A------Q--
Vibrio angustum 90579445 -L--SD-SD------IEL-- VDSETIDTH-A---------
Idiomarina baltica 85713205 -L--SDEPD------VEL-D VDSDTL-THW---T---R--
Moritella sp. PE36 149912346 ----SD-PD------I-L-D VDSETIDTH-A------Q--
Shewanella halifaxensis 167623974 ----S-ESD------IS--D VDDDTIDTH-S--S---K--
Xanthomonas albilineans 285018377 -Y--NE-PD--IE---HIAE LDKET--TQVA--H---R--
Xylella fastidiosa 15839026 -Y--SE-PD--IE--QHIAE LDKET--TQVA--H---R--
Cardiobacterium hominis 258544287 ----ND--DR---V-IH-AD LDQQML-NQ----S---K--
Legionella drancourtii 254498044 -Y--NE-SD----I-IKL-E VDSDTV-TH-A--S---R--
Methylophaga thiooxidans 254490577 -Y--NE-PD--I-V-IGL-E VNSETV-TH-A-A----R--


(B)
61
104
Pasteurellales (20/20)
Other Gamma Proteobacteria (1/500)
Haemophilus influenzae 148828339 VTFTNKAAAEMRHRIQSTLA KHAQ HQLVGMWIGTFHSIAHRLLR
Haemophilus ducreyi 33152168 ----------N--EY--S MSS- -R-F---V----------
Haemophilus parasuis 219871258 ----------EY--S QSSD -R-F---V-------N----
Haemophilus somnus 170718056 --------T------EA--- RYSH QR-F----------N----
Pasteurella multocida 15602276 ----------E---- N--S QR-F---V----------
Pasteurella dagmatis 260914154 ----------E---- N-SS QR-F---V----------
Man. succiniciproducens 52425423 ----------Q--E---S Q-SS RR-F---V----------
Man. haemolytica 254362319 ----------EY--S QSGD NR-F---V-------N----
Act. pleuropneumoniae 307250580 ----------EH--S SSSH -R-F---V-------N----
Act. succinogenes 152978934 ----------Q--E-A-- RYSS QR-F---V----------
Act. minor 240950109 ----------Q--EA--E QSS -NMF---V-------N----
Agg. actinomycetemcomitans 261868383 ----------E-V-S DGN QR-F---V----------
Agg. aphrophilus 251792813 ----------E-V-S DGN QR-F---V----------
Tolumonas auensis 237807125 ----------G--ERL-G NSLG FGRG--------GL------
Aeromonas hydrophila 117619159 ----------G-VEKVIG DGVR--------G-------
Idiomarina baltica 85711578 --------S---G-VEQL-G SSVRN-------GL------
Shewanella baltica 126172693 ----------E-VEKVAG TNMGR-------GL------
Azotobacter vinelandii 226942221 ----------T--EQL-G TSPM--------GL------
Pseudomonas aeruginosa 116053593 ----------EQL-G INPA---V----GL------
Citrobacter koseri 157144441 ----------GQLMG TTQG---V----GL------
Dickeya dadantii 307132967 ----------DQL-G TSQG--------GL------
Edwardsiella tarda 269137484 ----------EALIG TSQG--------GL------
Erwinia tasmaniensis 188532376 ----------EQLIG TSQG--------GL------
Escherichia coli 43297 ----------GQLMG TSQG---V----GL------
Klebsiella pneumoniae 206579213 ----------GQLMG TTQG---V----GL------
Photorhabdus luminescens 37528453 ----------ENLIG TSQG--------GL------
Proteus mirabilis 227357154 ----------EDLIG TSQG---------L------
Providencia stuartii 188026352 ----------NQLIG SSEG--------GL------
Salmonella enterica 161505541 ----------GQLMG TSQG---V----GL------
Shigella flexneri 30064891 ----------GQLMG TSQG---V----GL------
Yersinia pestis 22124304 ----------EHLIG TSQG--------GL------
Xanthomonas axonopodis 21244868 --------G-----TDLQ-R NGSR--------GL------
Nitrosococcus halophilus 292493733 --------G---G--EEL-G MPAG---M----GL------
Alcanivorax borkumensis 110835551 ----------G--EQL-D MSAD---V----G-------
Oceanospirillum sp. MED92 89092193 --------K---G--EEL-G LNPQ---V----GL------
Kangiella koreensis 256821408 --------K--LG-VEDM-- MPAR----------

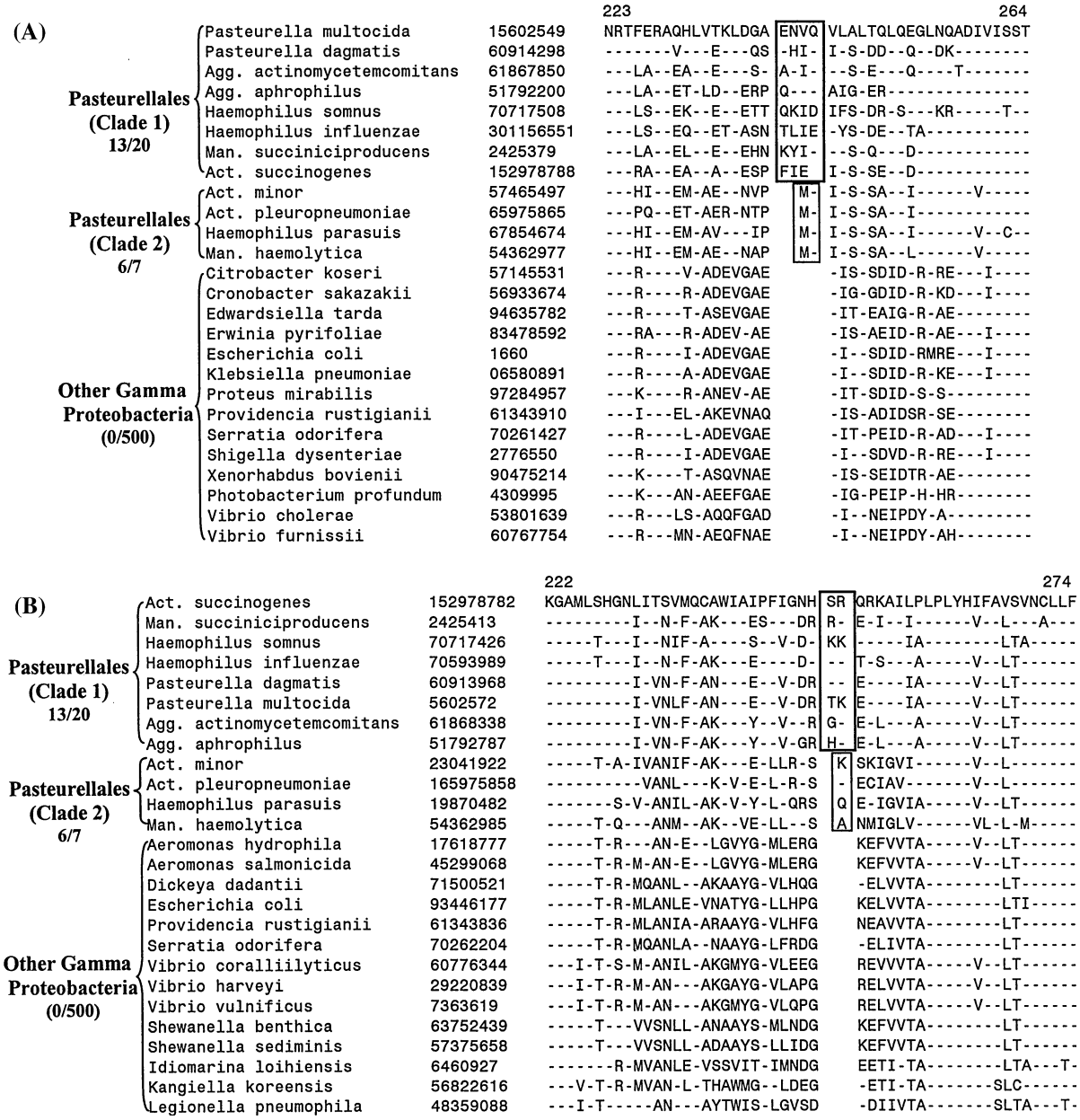


Fig. 3 Partial sequence alignments of **a** glutamyl-tRNA reductase and **b** long-chain-fatty-acid-CoA ligase, each containing two CSIs of different lengths (*boxed*) at the same positions that are specific for the two Pasteurellales clades. The *dashes* in the sequence alignments indicate identity with the amino acid on the *top* line. In the case of Glutamyl-tRNA reductase, a 4 aa insert is present in various Clade I species, while all of the Clade 2 species contain a 2 aa insert in this position. In the long-chain-fatty-acid-CoA ligase, 2 aa and 1 aa inserts are found in the Clades 1 and 2 species, respectively. The different lengths of CSIs in these proteins serve to distinguish the Clades 1 and 2 species from each other. Sequence information for only a limited number of species from other bacterial group is presented here

for these proteins were not detected in one of the Pasteurellales species (viz. *H. ducreyi* or *Agg. actinomycetemcomitans*). Similarly, for four other proteins that contained Pasteurellales specific CSIs, their homologues were not detected in a few species from this group (Supplementary Figs. 27–30).

In a number of additional proteins, while the CSIs of interest are specifically present in most

Pasteurellales, they are lacking in 1–2 species. For example the 1 aa insert in 23S rRNA (guanosine-2′-*o*)-methyltransferase and the 17 aa insert in glutamate ammonia ligase adenylyltransferase are specifically present in all Pasteurellales except *H. parasuis* (Supplementary Figs. 31–32). Likewise, the 1 aa inserts in murein transglycosylase C, ProS protein and D-methionine-binding lipoprotein are present in all Pasteurellales except *Act. minor* and the two *Pasteurella* species, respectively (Supplementary Figs. 33–35). The absence of CSIs in these Pasteurellales species could result from a variety of possibilities including deeper branching of these species in relation to other species or replacement of the gene containing CSI by a gene lacking the CSI by means of LGTs. However, at present these or other possibilities cannot be distinguished.

In addition to the above proteins that contained CSIs that were highly specific for either all or most Pasteurellales species, in a small number of cases the identified CSIs in addition to being shared by all or most Pasteurellales were also present in 1–2 isolated species from other Gammaproteobacteria. One example of such CSIs is a 3–4 aa insert in the DNA dependent helicase II (Fig. 2b), that is commonly shared by all sequenced Pasteurellales species as well as by *Tolumonas auensis*, belonging to the order Aeromonadales. However, this CSI is not present in other Aeromonadales. The other proteins containing Pasteurellales-specific CSIs with isolated exceptions include the presence of a 2 aa insert in the hypothetical protein NTO5HA_0747 that is also shared by *Psychrobacter* sp. PRwf-1 (Supplementary Fig. 36A); a 2 aa deletion in the Lysyl tRNA synthetase that is also shared by *Marinomonas* sp. MWYL1 (Supplementary Fig. 36B); a 1 aa insert in the protein Cof, a haloacid dehalogenease-like hydrolase, that is also present in *Pantoea* sp. At-9b (Supplementary Fig. 37); a 4 aa deletion in 6-phophogluconolactonase that is also found in *Cardiobacterium hominis* (Supplementary Fig. 38), a 2 aa deletion in the geranyltranstransferase also present in *Allochromatium vinosum, Marinobacter algicola and Marinobacter aquaeolei* (Supplementary Fig. 39); and lastly a 3 aa insert in the DNA repair protein RecN that in addition to all Pasteurellales is also present in *Cellvibrio japonicus and Psychromonas* sp. CNPT3 (Supplementary Fig. 40). The shared presence of these CSIs in isolated species from other groups could result from a variety of possibilities including lateral gene transfer from Pasteurellales to these species; independent occurrence of similar genetic changes in these species; or that some of these species might be more closely related to the Pasteurellales and that they have been incorrectly assigned to these other genera/orders. We are unable to distinguish between these possibilities based upon the available data.

Molecular signatures distinguishing two main clades of Pasteurellales

The order Pasteurellales currently consists of a single family *Pasteurellaceae* and the interrelationship among different species/genera within this family is poorly understood (Olsen et al. 2005; Christensen and Bisgaard 2006; Christensen and Bisgaard 2010). Thus, molecular markers that can provide reliable insights concerning the evolutionary relationships among these species should be of much interest. In phylogenetic trees, based upon two different large sets of protein sequences, the sequenced Pasteurellales species formed two distinct clades (Gioia et al. 2006; Redfield et al. 2006; Bonaventura et al. 2010), as confirmed in the present study (Fig. 1). Importantly, the existence of these two clades is independently strongly supported by the species distribution patterns of many CSIs that we have identified in the present work. A brief description of these CSIs is provided below.

The protein glutamyl-tRNA reductase, which catalyzes the NADPH-dependant reduction of glutamyl-tRNA to glutamyl-1-semialdehyde, contains two different lengths of CSIs in the same position that serve to distinguish various *Pasteurellaceae* species from all other bacteria and at the same time they also provide clear distinction between the Clades I and II species (Fig. 3a). In this case, a 4 aa insert in a conserved region is uniquely present in all of the Pasteurellales species that form Clade I (viz. *Agg. actinomycetemcomitans, P. multocida, P. dagmatis, Act. succinogenes, Man. succinoproducens, H. somnus* and *H. influenzae*), whereas in the various species that comprise Clade II, a 2 aa insert is present in the same position. Because these CSIs are related in sequence, the most likely explanation to account for them is that a 2 aa or 4 aa insert was initially introduced in a common ancestor of all Pasteurellales and it was followed by either a 2 aa insert in the Clade I species or a 2 aa deletion in the Clade II

species. Similarly to glutamyl-tRNA reductase, in the protein long chain fatty acid-CoA ligase, which plays an important role in the breakdown of fatty acids, different lengths of CSIs in a conserved region are uniquely present in the two Pasteurellales clades (Fig. 3b). In this case, a 2 aa insert is present in all of the Clade I species, whereas the Clade II species have a 1 aa insert in this position. The presence of different lengths of CSIs in this protein can also be explained as above. Interestingly, the homologues of both of these proteins were not detected in *H. ducreyi*.

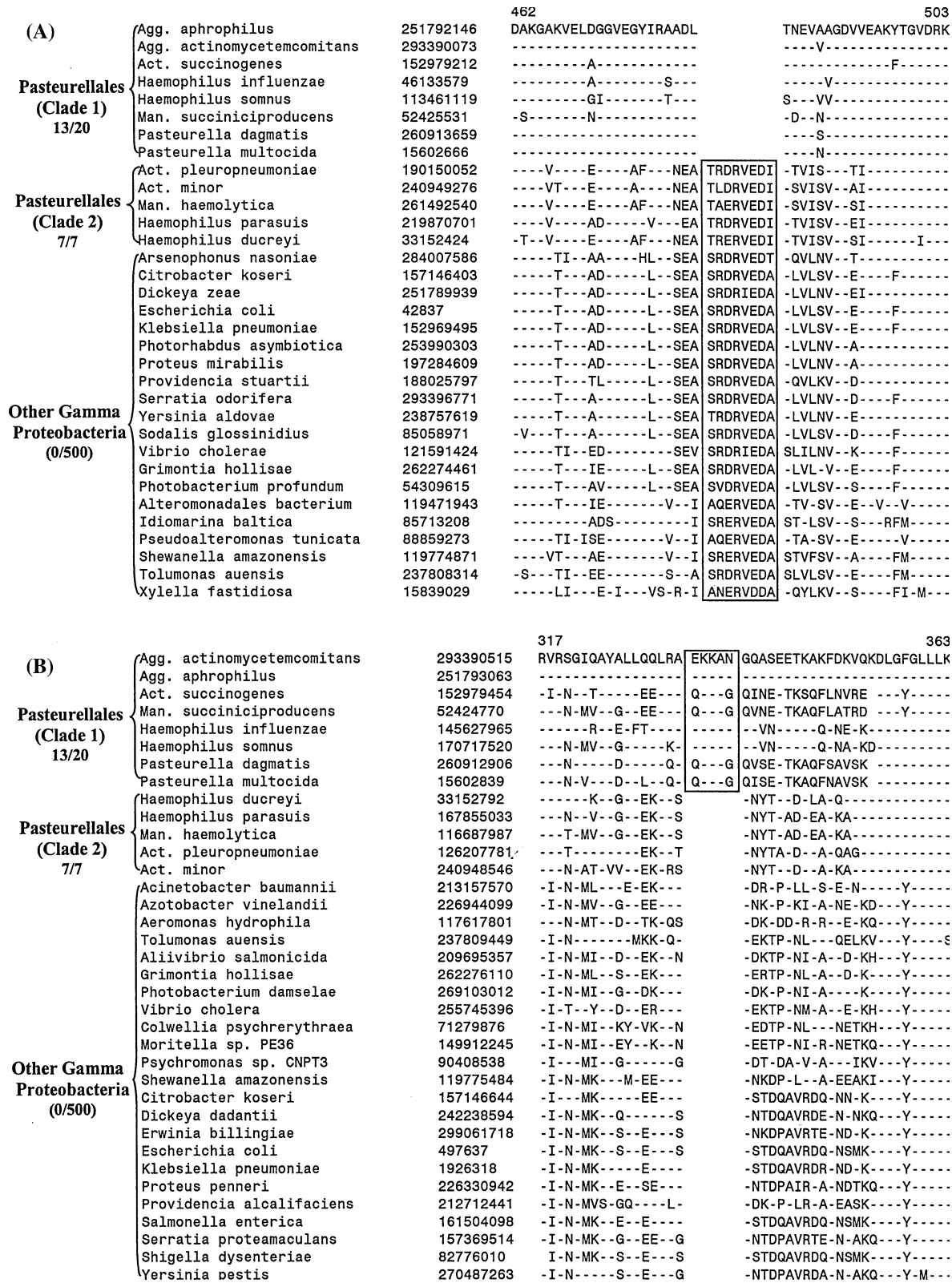
In addition to these CSIs that distinguish both Clades I and II species, we have also identified 11 CSIs in widely distributed proteins that are either uniquely or mainly found in the Clade I species (Table 3A). Two examples of such CSIs are presented in Fig. 4. In the universally distributed ribosomal protein S1, which plays a central role in protein synthesis, an eight amino acid deletion in a conserved region is uniquely present in all Clade I Pasteurellales species (Fig. 4a). The absence of this indel in all other Pasteurellales as well as other bacteria provides evidence that this indel represents a deletion in the Clade I species rather than an insert in other bacteria. Similarly, in the protein cytochrome-D-ubiquinol oxidase subunit 1, which is a component of the aerobic respiratory chain, a 5 aa insert in a conserved region is uniquely present in all Pasteurellales species belonging to Clade I, but not found in any other bacteria (Fig. 4b). Sequence alignments for other proteins which contain CSIs that are specific for Pasteurellales Clade I are presented in Supplementary Figs. 41–45. The CSIs in all of the above proteins are highly specific for Pasteurellales Clade I indicating that they were introduced in a common ancestor of this clade.

Four other proteins also contain CSIs that are largely specific for the Clade I. Within Clade I, *H. influenzae* shows deepest branching in the phylogenetic tree (Fig. 1). We have identified a 2 aa insert in the protein thiamine-monophosphate kinase that is commonly shared by all Clade I species except *H. influenzae* (Supplementary Fig. 46). The most likely explanation for this CSI is that the genetic change responsible for it occurred in a common ancestor of the remaining Clade I species after the branching of *H. influenzae*. For CSIs in three other proteins, the indels of interest are also present in an isolated species from Clade II in addition to the members of Clade I. For example, in the fumarate reductase iron-sulfur subunit, which is involved in the interconversion of fumarate and succinate, an 11 aa insert in a highly conserved region is uniquely present in various Clade I species and also *H. parasusis*, which shows deepest branching in the Clade II (Supplementary Fig. 47). Likewise, in the cell division protein FtsZ, a 3 aa insert is present in various Clade I species and also *Man. haemolytica* (Supplementary Fig. 48). The protein lysyl-tRNA synthetase also contains a 2 aa insert that is specific for the Clade I. However, in this case, only one of the *H. somnus* strain contains this CSI, whereas the other *H. somnus* strain has a more divergent homologue that lacks this indel (Supplementary Fig. 49). The species distribution patterns of these latter CSIs could result from a number of possibilities including LGT events or introduction of these genetic changes at various stages in the evolution of the Pasteurellales species that are not apparent from this tree.

Fig. 4 Excerpts from the sequence alignments for **a** ribosomal protein S1 and **b** cytochrome D ubiquinol oxidase subunit 1, showing two different CSIs in conserved regions of these proteins that are uniquely present in various Clade 1 Pasteurellales species. The other CSIs those are specific for the Clade I species are listed in Table 3A. The *dashes* in the sequence alignments indicate identity with the amino acid on the *top* line

The Pasteurellales species *Act. pleuropneumoniae*, *Act. minor, H. ducreyi, Man. haemolytica* and *H. parasuis* form Clade II in the phylogenetic tree (Fig. 1). As indicated above, the proteins glutamyl-tRNA reductase and long chain fatty acid-CoA ligase contain distinctive inserts that are specific for the Clade II species (Fig. 3). We have also identified a number of other CSIs that are specific for this clade (Table 3B). In the enzyme DNA adenine methylase, which is responsible for methylation of the newly synthesized strand of DNA, a 3 aa insert that is specific for the Clade II species is present in a highly conserved region (Fig. 5a). Other sequence alignments showing CSI specific to Pasteurellales Clade II (Table 3B) are shown in Supplementary Figs. 50–52. The genetic changes responsible for these CSIs were likely introduced in a common ancestor of the Clade II species and they strongly support the existence of this clade.

Within Clade II, the deepest branching in the phylogenetic tree is observed for *H. parasuis* (Fig. 1).


(A)
462
503
Pasteurellales (Clade 1) 13/20
Pasteurellales (Clade 2) 7/7
Other Gamma Proteobacteria (0/500)
Agg. aphrophilus 251792146 DAKGAKVELDGGVEGYIRAADL TNEVAAGDVVEAKYTGVDRK
Agg. actinomycetemcomitans 293390073 ----V----
Act. succinogenes 152979212 ---------A---------- ----------F------
Haemophilus influenzae 46133579 ---------A--------S--- -----V----------
Haemophilus somnus 113461119 ---------GI-------T--- S---VV----------
Man. succiniciproducens 52425531 -S-------N---------- -D--N----------
Pasteurella dagmatis 260913659 ----S----------
Pasteurella multocida 15602666 ----N----------
Act. pleuropneumoniae 190150052 ----V----E----AF---NEA TRDRVEDI -TVIS---TI----------
Act. minor 240949276 ----VT---E----A----NEA TLDRVEDI -SVISV--AI----------
Man. haemolytica 261492540 ----V----E----AF---NEA TAERVEDI -SVISV--SI----------
Haemophilus parasuis 219870701 ----V----AD-----V---EA TRDRVEDI -TVISV--EI----------
Haemophilus ducreyi 33152424 -T--V----E----AF---NEA TRERVEDI -TVISV--SI------I---
Arsenophonus nasoniae 284007586 -----TI--AA----HL--SEA SRDRVEDT -QVLNV--T----------
Citrobacter koseri 157146403 -----T---AD-----L--SEA SRDRVEDA -LVLSV--E----F------
Dickeya zeae 251789939 -----T---AD-----L--SEA SRDRIEDA -LVLNV--EI----------
Escherichia coli 42837 -----T---AD-----L--SEA SRDRVEDA -LVLSV--E----F------
Klebsiella pneumoniae 152969495 -----T---AD-----L--SEA SRDRVEDA -LVLSV--E----F------
Photorhabdus asymbiotica 253990303 -----T---AD-----L--SEA SRDRVEDA -LVLNV--A----------
Proteus mirabilis 197284609 -----T---AD-----L--SEA SRDRVEDA -LVLNV--A----------
Providencia stuartii 188025797 -----T---TL-----L--SEA SRDRVEDA -QVLKV--D----------
Serratia odorifera 293396771 -----T---A------L--SEA SRDRVEDA -LVLNV--D----F------
Yersinia aldovae 238757619 -----T---A------L--SEA TRDRVEDA -LVLNV--E----------
Sodalis glossinidius 85058971 -V---T---A------L--SEA SRDRVEDA -LVLSV--D----F------
Vibrio cholerae 121591424 -----TI--ED--------SEV SRDRIEDA SLILNV--K----F------
Grimontia hollisae 262274461 -----T---IE-----L--SEA SRDRVEDA -LVL-V--E----F------
Photobacterium profundum 54309615 -----T---AV-----L--SEA SVDRVEDA -LVLSV--S----F------
Alteromonadales bacterium 119471943 -----T---IE-------V--I AQERVEDA -TV-SV--E--V--V-----
Idiomarina baltica 85713208 ---------ADS---------I SRERVEDA ST-LSV--S---RFM-----
Pseudoalteromonas tunicata 88859273 -----TI-ISE-------V--I AQERVEDA -TA-SV--E-----V-----
Shewanella amazonensis 119774871 ----VT---AE-------V--I SRERVEDA STVFSV--A----FM-----
Tolumonas auensis 237808314 -S---TI--EE-------S--A SRDRVEDA SLVLSV--E----FM-----
Xylella fastidiosa 15839029 -----LI---E-I---VS-R-I ANERVDDA -QYLKV--S----FI-M---
(B)
317
363
Pasteurellales (Clade 1) 13/20
Pasteurellales (Clade 2) 7/7
Other Gamma Proteobacteria (0/500)
Agg. actinomycetemcomitans 293390515 RVRSGIQAYALLQQLRA EKKAN GQASEETKAKFDKVQKDLGFGLLLK
Agg. aphrophilus 251793063 ----------
Act. succinogenes 152979454 -I-N--T-----EE--- Q---G QINE-TKSQFLNVRE ---Y-----
Man. succiniciproducens 52424770 ---N-MV--G--EE--- Q---G QVNE-TKAQFLATRD ---Y-----
Haemophilus influenzae 145627965 ------R--E-FT---- ----- --VN-----Q-NE-K----------
Haemophilus somnus 170717520 ---N-MV--G-----K- ----- --VN-----Q-NA-KD---------
Pasteurella dagmatis 260912906 ---N-----D-----Q- Q---G QVSE-TKAQFSAVSK ----------
Pasteurella multocida 15602839 ---N-V---D--L--Q- Q---G QISE-TKAQFNAVSK ---------
Haemophilus ducreyi 33152792 ------K--G--EK--S -NYT--D-LA-Q----------
Haemophilus parasuis 167855033 ---N--V--G--EK--S -NYT-AD-EA-KA----------
Man. haemolytica 116687987 ---T-MV--G--EK--S -NYT-AD-EA-KA----------
Act. pleuropneumoniae 126207781 ---T--------EK--T -NYTA-D--A-QAG----------
Act. minor 240948546 ---N-AT-VV--EK-RS -NYT--D--A-KA----------
Acinetobacter baumannii 213157570 -I-N-ML---E-EK--- -DR-P-LL-S-E-N-----Y-----
Azotobacter vinelandii 226944099 -I-N-MV--G--EE--- -NK-P-KI-A-NE-KD---Y-----
Aeromonas hydrophila 117617801 ---N-MT--D--TK-QS -DK-DD-R-R--E-KQ---Y-----
Tolumonas auensis 237809449 -I-N-------MKK-Q- -EKTP-NL---QELKV---Y----S
Aliivibrio salmonicida 209695357 -I-N-MI--D--EK--N -DKTP-NI-A--D-KH---Y-----
Grimontia hollisae 262276110 -I-N-ML--S--EK--- -ERTP-NL-A--D-K----Y-----
Photobacterium damselae 269103012 -I-N-MI--G--DK--- -DK-P-NI-A----K----Y-----
Vibrio cholera 255745396 -I-T--Y--D--ER--- -EKTP-NM-A--E-KH---Y-----
Colwellia psychrerythraea 71279876 -I-N-MI--KY-VK--N -EDTP-NL---NETKH---Y-----
Moritella sp. PE36 149912245 -I-N-MI--EY--K--N -EETP-NI-R-NETKQ---Y-----
Psychromonas sp. CNPT3 90408538 -I---MI--G------G -DT-DA-V-A---IKV---Y-----
Shewanella amazonensis 119775484 -I-N-MK---M-EE--- -NKDP-L--A-EEAKI---Y-----
Citrobacter koseri 157146644 -I---MK-----EE--- -STDQAVRDQ-NN-K----Y-----
Dickeya dadantii 242238594 -I-N-MK--Q------S -NTDQAVRDE-N-NKQ---Y-----
Erwinia billingiae 299061718 -I-N-MK--S--E---S -NKDPAVRTE-ND-K----Y-----
Escherichia coli 497637 -I-N-MK--S--E---S -STDQAVRDQ-NSMK----Y-----
Klebsiella pneumoniae 1926318 -I-N-MK-----E---- -STDQAVRDR-ND-K----Y-----
Proteus penneri 226330942 -I-N-MK--E--SE--- -NTDPAIR-A-NDTKQ---Y-----
Providencia alcalifaciens 212712441 -I-N-MVS-GQ----L- -DK-P-LR-A-EASK----Y-----
Salmonella enterica 161504098 -I-N-MK--E--E---- -STDQAVRDQ-NSMK----Y-----
Serratia proteamaculans 157369514 -I-N-MK--G--EE--G -NTDPAVRTE-N-AKQ---Y-----
Shigella dysenteriae 82776010 I-N-MK--S--E---S -STDQAVRDQ-NSMK----Y-----
Yersinia pestis 270487263 -I-N-----S--E---G -NTDPAVRDA-N-AKQ---Y-M---

A clade consisting of the remaining Clade II species (all except *H. parasuis*) is strongly supported in the phylogenetic tree. We have identified three CSIs that are specific for this subclade of the Clade II. Information for one of these CSIs is presented in Fig. 5b, which shows a 5 aa insert in the enzyme tRNA-(uracil-5-)-methyltransferase. Similar to this CSI, a 2 aa insert in a highly conserved region of the ribosomal proteins S4 (Supplementary Fig. 53) and a 7 aa deletion in the enzyme adenylate cyclase is also specific for this subclade of the Clade II species (Supplementary Fig. 54). The genetic changes for these CSIs were likely introduced in a common ancestor of the remaining Clade II species after the branching of *H. parasuis*. In the enzyme DNA gyrase B, which contains a 2 aa insert specific for the Clade II species, in the same position where this insert is found, a 5 aa insert is also uniquely present in the two succinic acid producing bacteria *Act. succinogenes* and *Man. succiniciproducens* (Supplementary Fig. 51). The latter two bacteria form a strongly supported cluster in the phylogenetic tree and the shared presence of this insert support that they are specifically related (Fig. 1). The different lengths and species specificity of these inserts indicate that the genetic changes responsible for them occurred independently in the common ancestors of these two groups of Pasteurellales species.

Discussion

The members of the Order Pasteurellales are presently distinguished from other bacteria primarily on the basis of their distinct branching in phylogenetic trees (Olsen et al. 2005; Christensen and Bisgaard 2006; Christensen and Bisgaard 2010). Furthermore, although this order is comprised of at least 15 genera, due to a lack of reliable information about their interrelationships, all of them are placed into a single family (Olsen et al. 2005; Christensen and Bisgaard 2006; Christensen and Bisgaard 2010). We report here for the first time >60 molecular signatures that are distinctive characteristics of either all sequenced Pasteurellales species/strains or a number of well-defined subclades within this order. Of the signatures described here, 23 CSIs in widely distributed proteins are uniquely found in all of the sequenced Pasteurellales species/strains (Table 2) and they are not found in any other bacteria. Due to their specificity to the Pasteurellales, the rare genetic changes responsible for them were likely introduced only once in a common ancestor of these bacteria and then passed on to various descendent species (Gupta 1998; Rokas and Holland 2000; Gupta and Mathews 2010). The presence of these CSIs in all Pasteurellales and their absence in all other bacteria strongly indicates that the genes for these proteins have not been laterally transferred from Pasteurellales to other bacterial groups or vice versa (Gogarten et al. 2002; Christensen and Bisgaard 2010). Thus, these CSIs provide potentially useful molecular markers (synapomorphies) for the identification and circumscription of species from the order Pasteurellales in molecular terms.

Fig. 5 Partial sequence alignments for the proteins **a** DNA adenine methylase showing a 3 aa insert that is specific for Clade 2 Pasteurellales species and **b** tRNA (uracil-5-)-methyltransferase, showing a 5 aa insert, that is uniquely found in all Clade 2 species except *H. parasuis*, which is the deepest branching species in Clade 2 (Fig. 1). Other CSIs showing similar specificity are listed in Table 3B. The *dashes* in the sequence alignments indicate identity with the amino acid on the *top* line

In addition to these CSIs that are uniquely found in all sequenced Pasteurellales, 21 other CSIs were identified that are also largely specific for this order of bacteria. However, in some of these cases the homologues for these genes/proteins were not detected in 1 or 2 Pasteurellales species, whereas in some others an isolated species from other bacterial groups was also found to contain these CSIs. Because these CSIs are commonly present in all (or most) Pasteurellales, with only isolated exceptions showing no specific pattern, it is highly likely that the genetic changes responsible for them also occurred in a common ancestor of the Pasteurellales. This was likely followed by loss of the genes from a few species and their acquisition by isolated species from other groups by LGTs (Gogarten et al. 2002). However, the possibility that sequence information for some of these observed exceptions might be incorrect in the public databases cannot be entirely ruled out.

All of the genera within the order Pasteurellales are currently placed into a single family, *Pasteurellaceae* (Olsen et al. 2005; Christensen and Bisgaard 2006; Christensen and Bisgaard 2010). However, the present work has also identified many CSIs that are

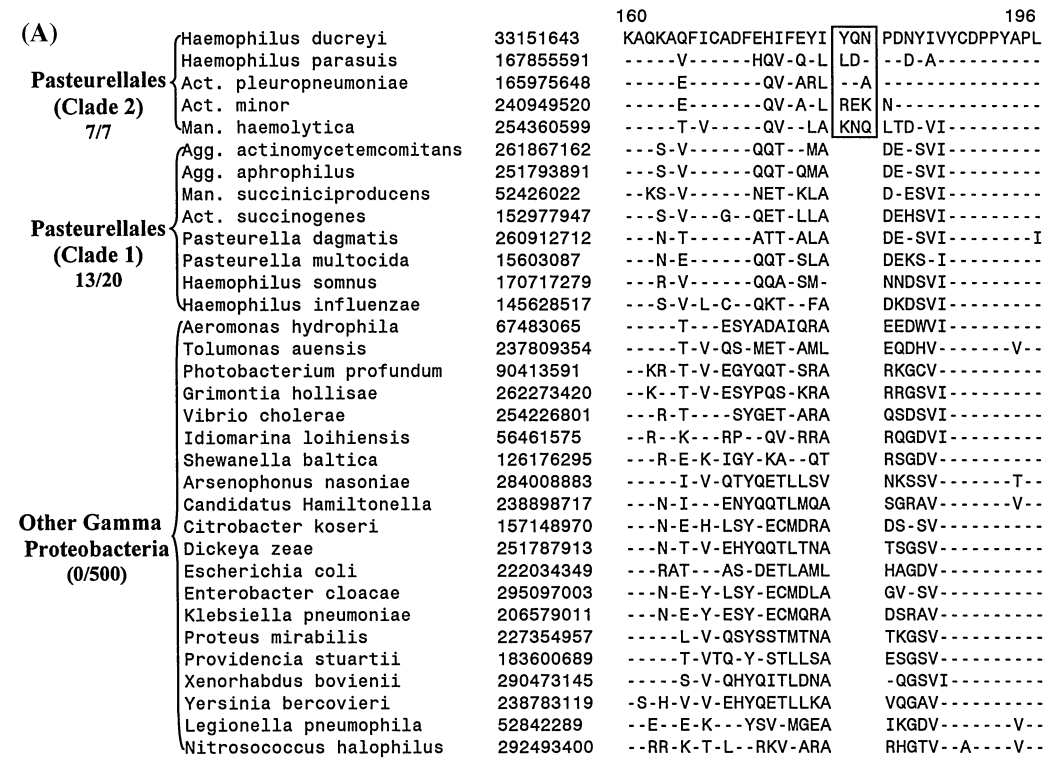

(A)
160
196
Pasteurellales (Clade 2) 7/7
Pasteurellales (Clade 1) 13/20
Other Gamma Proteobacteria (0/500)
Haemophilus ducreyi 33151643 KAQKAQFICADFEHIFEYI YQN PDNYIVYCDPPYAPL
Haemophilus parasuis 167855591 -----V------HQV-Q-L LD- --D-A----------
Act. pleuropneumoniae 165975648 -----E-------QV-ARL --A --------------
Act. minor 240949520 -----E-------QV-A-L REK N--------------
Man. haemolytica 254360599 -----T-V-----QV--LA KNQ LTD-VI---------
Agg. actinomycetemcomitans 261867162 ---S-V------QQT--MA DE-SVI---------
Agg. aphrophilus 251793891 ---S-V------QQT-QMA DE-SVI---------
Man. succiniciproducens 52426022 --KS-V------NET-KLA D-ESVI---------
Act. succinogenes 152977947 ---S-V---G--QET-LLA DEHSVI---------
Pasteurella dagmatis 260912712 ---N-T------ATT-ALA DE-SVI--------I
Pasteurella multocida 15603087 ---N-E------QQT-SLA DEKS-I---------
Haemophilus somnus 170717279 ---R-V------QQA-SM- NNDSVI---------
Haemophilus influenzae 145628517 ---S-V-L-C--QKT--FA DKDSVI---------
Aeromonas hydrophila 67483065 -----T---ESYADAIQRA EEDWVI---------
Tolumonas auensis 237809354 -----T-V-QS-MET-AML EQDHV-------V--
Photobacterium profundum 90413591 --KR-T-V-EGYQQT-SRA RKGCV----------
Grimontia hollisae 262273420 --K--T-V-ESYPQS-KRA RRGSVI---------
Vibrio cholerae 254226801 ---R-T----SYGET-ARA QSDSVI---------
Idiomarina loihiensis 56461575 --R--K---RP--QV-RRA RQGDVI---------
Shewanella baltica 126176295 ---R-E-K-IGY-KA--QT RSGDV----------
Arsenophonus nasoniae 284008883 -----I-V-QTYQETLLSV NKSSV-------T--
Candidatus Hamiltonella 238898717 ---N-I---ENYQQTLMQA SGRAV-------V--
Citrobacter koseri 157148970 ---N-E-H-LSY-ECMDRA DS-SV----------
Dickeya zeae 251787913 ---N-T-V-EHYQQTLTNA TSGSV----------
Escherichia coli 222034349 ---RAT---AS-DETLAML HAGDV----------
Enterobacter cloacae 295097003 ---N-E-Y-LSY-ECMDLA GV-SV----------
Klebsiella pneumoniae 206579011 ---N-E-Y-ESY-ECMQRA DSRAV----------
Proteus mirabilis 227354957 -----L-V-QSYSSTMTNA TKGSV----------
Providencia stuartii 183600689 -----T-VTQ-Y-STLLSA ESGSV----------
Xenorhabdus bovienii 290473145 -----S-V-QHYQITLDNA -QGSVI---------
Yersinia bercovieri 238783119 -S-H-V-V-EHYQETLLKA VQGAV----------
Legionella pneumophila 52842289 --E--E-K---YSV-MGEA IKGDV-------V--
Nitrosococcus halophilus 292493400 --RR-K-T-L--RKV-ARA RHGTV--A----V--
(B)
42
77
Pasteurellales (Clade 2) 6/7
Pasteurellales (Clade 1) 13/20
Other Gamma Proteobacteria (0/500)
Haemophilus ducreyi 33151895 RMRAEFRVWHQ KNEQG ENDLYHIMFDPTTKQRYRVD
Act. minor 223041752 ----------D ---K- --E--------E--------
Act. pleuropneumoniae 32034047 ----------D T--V- --E-------QE-----C-E
Act. succinogenes 152978112 -------L--D YS-NR GGN-------K---------
Haemophilus parasuis 219872016 -L---------D NG--------QA--------
Haemophilus influenzae 145628253 -------I--E QD-F------QA-LK-----
Haemophilus somnus 170718062 -------I--E QD-F------QK----F-I-
Man. haemolytica 261494651 ----------E -GE------N-E--A-----
Agg. Actinomycetemcomitans 293391475 -------I--E QD-F------QQS-------
Agg. aphrophilus 251793563 -------I--E GD-F------QQS-------
Pasteurella dagmatis 260914579 -------I--N QG-F------QQ-R------
Pasteurella multocida 15603668 -------I--D KG-F------QR-R------
Man. succiniciproducens 52426422 ----------D KG-------NQQ--------
Aeromonas hydrophila 117621352 -------I--D GD----C-YA-A--EII---
Grimontia hollisae 262273152 ----------E GD---Y---NQE---K----
Photobacterium damselae 269103682 ----------E GE---Y---NQQ-REK----
Vibrio cholerae 229527105 -------I--E GD-M-Y---NQE-REK----
Idiomarina baltica 85711838 ----------E Q----Y---N-E-REKI-M-
Psychromonas ingrahamii 119944400 -----L----D GDE--Y----KK---KF--E
Shewanella amazonensis 119773334 ---C------D GD---YC---NVA-EKV-T-
Arsenophonus nasoniae 284009190 -------L--E GQ--F-----Q-----I-I-
Citrobacter koseri 157147235 -------L--D GD-----I-EQQ--S-I---
Dickeya zeae 251787782 -------I--D GD--------QQ----I---
Edwardsiella ictaluri 238921674 -------I--D GD---------QS---I---
Enterobacter cloacae 296105348 -------I--D GD-----I--QQ--S-I--N
Erwinia tasmaniensis 188532302 -------I--E GD-----I--QQ-RE-I---
Escherichia coli 188496220 -------I--D GD-----I--QQ--S-I---
Klebsiella pneumoniae 206579514 -------L--D GD-----I--QQ-RS-I---
Proteus mirabilis 227354694 -------I--E QDA-------QE----I---
Providencia stuartii 188025467 ----------D GD--F-----KE--E-V-I-
Salmonella enterica 161505380 -------L--D GD--------QQ--S-I---
Shigella dysenteriae 82778852 -------I--D GD-----I--QQ--S-I---
Sodalis glossinidius 85060132 -------I--D -D---------Q--A-I-I-
Xenorhabdus bovienii 290477206 -------I--E QD--------QQ----I-I-
Yersinia aldovae 238760129 ----------D -D--------QQ----I--E

Table 3 Conserved signature indels that are specific for two different pasteurellales clades

Protein name	Gene name	Accession no.	Figure no.	Indel size	Indel position[a]	Functional categories
(A) Conserved indels that are specific for the Clade I Pasteurellales						
Glutamyl-tRNA reductase[b]	hemA	NP_245621	Fig. 3a	4 aa ins	223–264	Coenzyme transport and metabolism
Long-chain-fatty-acid–CoA ligase[b]	fadD	YP_001344411	Fig. 3b	2 aa ins	222–274	Lipid transport and metabolism
Ribosomal protein S1	rpsA	YP_003006866	Fig. 4a	8 aa del	462–503	Translation
Cytochrome D ubiquinol oxidase subunit 1	cydA	ZP_06634849	Fig. 4b	5 aa ins	317–363	Energy production and conversion
Glucose-6-phosphate isomerase	pgi	ZP_05920623	Supplementary Fig. 41	1 aa ins	351–384	Carbohydrate transport and metabolism
TldD protein	tldD	YP_087972	Supplementary Fig. 42	2 aa ins	75–127	General function prediction only
Acriflavin resistance protein	acr	YP_001343841	Supplementary Fig. 43	1 aa del	111–154	Defense mechanisms
Guanylate kinase	gmk	YP_003007858	Supplementary Fig. 44	1 aa ins	51–105	Nucleotide transport and metabolism
GTP-binding protein EngA	engA	YP_003255506	Supplementary Fig. 45	1 aa ins	93–128	General function prediction only
Thiamine-monophosphate kinase	thiL	YP_001344779	Supplementary Fig. 46	2 aa ins	190–230	Coenzyme transport and metabolism
Fumarate reductase iron-sulfur subunit[c]	frdB	YP_003007744	Supplementary Fig. 47	11 aa ins	117–165	Energy production and conversion
Cell division protein FtsZ[d]	ftsZ	YP_001345210	Supplementary Fig. 48	3 aa ins	239–276	Cell cycle control, mitosis and meiosis
Lysyl-tRNA synthetase[e]	lysS	YP_001290934	Supplementary Fig. 49	2 aa ins	316–366	Translation
(B) Conserved indels that are specific for the Clade II Pasteurellales						
Glutamyl-tRNA Reductase[b]	hemA	NP_245621	Fig. 3a	2 aa ins	223–264	Coenzyme transport and metabolism
Long-chain-fatty-acid–CoA ligase[b]	fadD	YP_001344411	Fig. 3b	1 aa ins	222–274	Lipid transport and metabolism
DNA adenine methylase	dam	NP_872996	Fig. 5a	3 aa ins	160–196	Replication, recombination and repair
tRNA (uracil-5-)-methyltransferase[f]	trmA	NP_873248	Fig. 5b	5 aa ins	42–77	Translation
S-ribosyl-homocysteinase	luxS	NP_872951	Supplementary Fig. 50	1 aa ins	85–137	Signal transduction mechanisms
DNA gyrase subunit B	gyrB	YP_001343447	Supplementary Fig. 51	2 aa ins	281–326	Replication, recombination and repair
Cysteine desulfurase	iscS	NP_873559	Supplementary Fig. 52	2 aa ins	274–319	Amino acid transport and metabolism
Ribosomal protein S4[f]	rpsD	ZP_00134833	Supplementary Fig. 53	2 aa ins	22–51	Translation
Adenylate cyclase[f]	cyaA	NP_873154	Supplementary Fig. 54	7 aa del	175–221	Nucleotide transport and metabolism

[a] The indel position indicates the regions of the proteins where CSIs are present

[b] Homologous sequences corresponding to this region could not be identified in *H. ducreyi*

[c] The CSI is also found in *H. parasuis* of Clade II

[d] The CSI is also found in *Man. haemolytica* of Clade II

[e] One *H. somnus* strain was found without indel

[f] The CSI is not found in *H. parasuis*

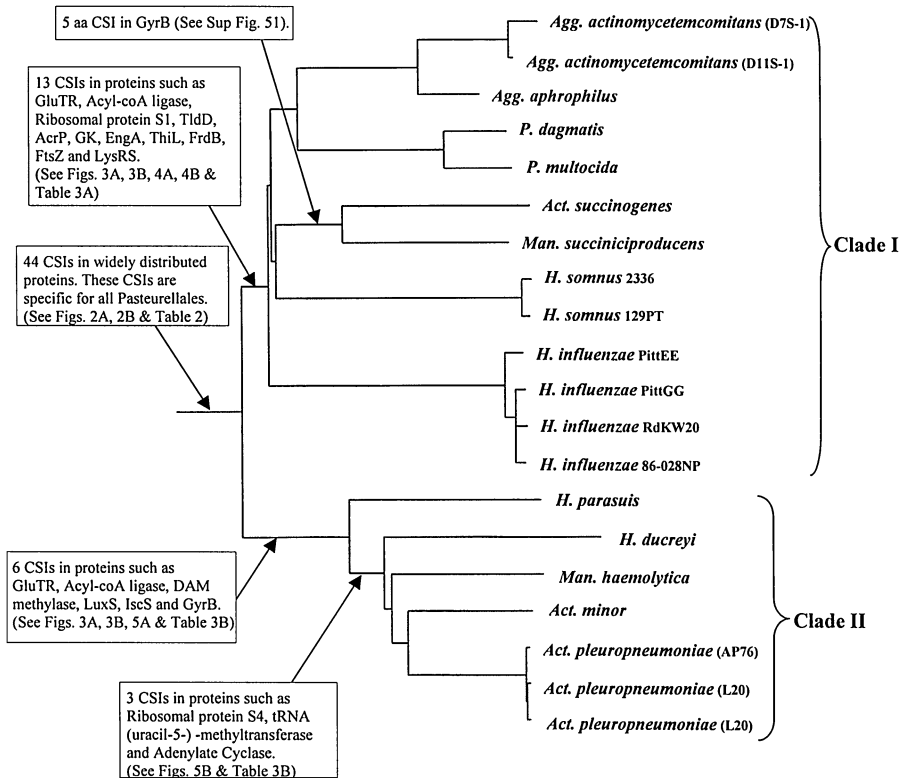


Fig. 6 A summary diagram showing the distribution patterns of various Pasteurellales-specific CSIs indicating the evolutionary relationships among Pasteurellales species. The different clades within this order that are supported by both phylogenetic studies and the identified molecular signatures are shown

specific for two distinct clades of Pasteurellales, which are also supported by our phylogenetic analyses (Fig. 1) and that of others (Gioia et al. 2006; Redfield et al. 2006; Bonaventura et al. 2010). The first of these clades, supported by 13 CSIs (Table 3A), includes *Aggregatibacter* and *Pasteurella* species and also *Act. succinogenes, Man. succiniciproducens* and various strains of *H. influenzae* and *H. somnus*. The remaining Pasteurellales species (viz. *Act. pleuropneumoniae*, *Act. minor, H. ducryi, Man. haemolytica* and *H. parasuis*) formed the second clade, which was supported by nine uniquely shared CSIs (Table 3B). Within Clade II, several CSIs also supported the deeper branching of *H. parasuis* in comparison to other species. The mutually exclusive presence of many of these CSIs in species from these two clades make a persuasive case that these clades are evolutionarily distinct and the genetic changes responsible for these CSIs were introduced in their common ancestors as indicated in Fig. 6. It should be noted that in contrast to numerous CSIs that supported the existence of these two clades, we have not come across significant numbers of CSIs that support any other alternative clades. Therefore, the identified CSIs, independently of phylogenetic analyses, provide strong evidence for the existence of these two Pasteurellales clades. We suggest that these two Pasteurellales clades, whose existence is supported by both phylogenetic analyses and by many discrete molecular signatures, should be recognized as distinct higher taxonomic groupings (i.e. families) within this order.

Sequence information for all of the identified CSIs is presently limited to only those Pasteurellales species/strains, whose genomes have been sequenced. Hence, to fully understand the evolutionary and taxonomic significance of these CSIs, it is of much importance to obtain sequence information for them from other Pasteurellales species, notably including the appropriate type strains. For the CSIs that are specific for all Pasteurellales, due to their exclusive presence in all sequenced species/strains from this order and no other (>1500) prokaryotic or eukaryotic organisms, it is highly likely that they will also be present in other Pasteurellales species/strains for whom no sequence information is presently available. Our earlier work on many CSIs for other prokaryotic groups indicates that the CSIs of this kind have a high

degree of predictive ability (Griffiths and Gupta 2002; Gupta 2005; Gao and Gupta 2005; Griffiths and Gupta 2006; Gupta 2009) and many of them will provide reliable molecular markers for the entire Pasteurellales order as sequence information for other species becomes available. However, for those CSIs that are specific for the two subclades of Pasteurellales, further studies to obtain sequence information from additional species/strains should be very informative. Based upon the presence or absence of the CSIs that are specific for the two subclades, it should be possible to assign/place other species into these subclades. This should help in determining more clearly the taxonomic boundaries of these two subclades. It is also possible that some species of Pasteurellales may be lacking both Clades I and II specific CSIs. This would suggest that such species might be parts of other higher taxonomic clades within the order Pasteurellales that have yet to be identified.

The *Pasteurellaceae* species are important human and animal pathogens and new species related to them are continually being discovered (Christensen and Bisgaard 2010). The identification of these medically important bacteria at present primarily relies upon culture-based nutritional and phenotypic characteristics (Olsen et al. 2005; Christensen and Bisgaard 2006; Christensen and Bisgaard 2010). However, such tests are unable to reliably distinguish members of Pasteurellales species from some other orders of Gammaproteobacteria (Olsen et al. 2005; Christensen and Bisgaard 2006; Christensen and Bisgaard 2010). In this context, the Pasteurellales-specific CSIs described here provide a novel means for the identification of these bacteria. Degenerate PCR primers based on conserved regions of these CSIs-containing genes, should provide novel and specific means for the detection of both previously known as well as novel Pasteurellales species (or isolates) in different environments.

In the present study, our focus has been mainly on identifying CSIs that are specific for either all Pasteurellales or its larger clades. Although our work has identified many CSIs of these kinds, further detailed studies on other Pasteurellales genomes could lead to identification of additional signatures of this kind. In the present work, we have not analyzed CSIs that were specific for individual species/genera or for the smaller clades of Pasteurellales. We have also not yet looked for the presence of signature proteins (CSPs) that are specific for either all Pasteurellales or its different subgroups. Such studies will form the focus of our future work. A number of Pasteurellales genera (viz. *Haemophilus, Actinobacillus* and *Mannheimia*) are not monophyletic and it is important to develop reliable means to reorganize them (Olsen et al. 2005; Christensen and Bisgaard 2006; Christensen and Bisgaard 2010). The identification of large numbers of CSIs and CSPs those that are specific for individual species or smaller clades, in addition to their diagnostic values, should prove very helpful in the reorganization and circumscription of various Pasteurellales genera.

Most of the CSIs identified in this work are present in conserved regions of various proteins that are involved in wide variety of essential cellular functions. Our recent work on a number of CSIs in the GroEL and DnaK proteins show that these CSIs are essential for the group of organisms where they are found (Singh and Gupta 2009). Any deletions or significant changes in them lead to failure of cell growth, indicating that they are playing essential roles in these organisms (Singh and Gupta 2009). Based upon these observations and the evolutionary conservation of these CSIs for the Order Pasteurellales, it is expected that these CSIs also play important (and possibly essential) functional roles in these bacteria. Hence, further studies on understanding the cellular functions of these CSIs could provide important insights into novel genetic, biochemical and physiological characteristics of members of Pasteurellales or their different clades.

Acknowledgments This work was supported by a research grant from the Natural Science and Engineering Research Council of Canada. HSN was partly supported by a scholarship from the Islamia University of Bahawalpur.

References

Barabote RD, Xie G, Leu DH et al (2009) Complete genome of the cellulolytic thermophile *Acidothermus cellulolyticus* 11B provides insights into its ecophysiological and evolutionary adaptations. Genome Res 19:1033–1043

Bisgaard M (1993) Ecology and significance of Pasteurellaceae in animals. Zentralbl Bakteriol 279:7–26

Bonaventura MP, Lee EK, DeSalle R, Planet PJ (2010) A whole-genome phylogeny of the family Pasteurellaceae. Mol Phylogenet Evol 54:950–956

Bosse JT, Janson H, Sheehan BJ et al (2002) *Actinobacillus pleuropneumoniae*: pathobiology and pathogenesis of infection. Microbes Infect 4:225–235

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–552

Chen C, Kittichotirat W, Si Y, Bumgarner R (2009) Genome sequence of *Aggregatibacter actinomycetemcomitans* serotype c strain D11S-1. J Bacteriol 191:7378–7379

Christensen H, Bisgaard M (2006) The genus *Pasteurella*. In: Dworkin M (ed) The prokaryotes: a handbook on the biology of bacteria. New York, Springer, pp 1062–1090

Christensen H, Bisgaard M (2010) Molecular classification and its impact on diagnostics and understanding the phylogeny and epidemiology of selected members of Pasteurellaceae of veterinary importance. Berl Munch Tierarztl Wochenschr 123:20–30

Christensen H, Kuhnert P, Olsen JE, Bisgaard M (2004) Comparative phylogenies of the housekeeping genes atpD, infB and rpoB and the 16S rRNA gene within the Pasteurellaceae. Int J Syst Evol Microbiol 54:1601–1609

Christensen H, Kuhnert P, Busse HJ, Frederiksen WC, Bisgaard M (2007) Proposed minimal standards for the description of genera, species and subspecies of the Pasteurellaceae. Int J Syst Evol Microbiol 57:166–178

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life. Science 311:1283–1287

De Ley J, Mannheim W, Mutters R et al (1990) Inter- and intrafamilial similarities of rRNA cistrons of the Pasteurellaceae. Int J Syst Bacteriol 40:126–137

Dewhirst FE, Paster BJ, Olsen I, Fraser GJ (1992) Phylogeny of 54 representative strains of species in the family Pasteurellaceae as determined by comparison of 16S rRNA sequences. J Bacteriol 174:2002–2013

Dewhirst FE, Paster BJ, Olsen I, Fraser GJ (1993) Phylogeny of the Pasteurellaceae as determined by comparison of 16S ribosomal ribonucleic acid sequences. Zentralbl Bakteriol 279:35–44

Di Bonaventura MP, DeSalle R, Pop M et al (2009) Complete genome sequence of *Aggregatibacter* (Haemophilus) *aphrophilus* NJ8700. J Bacteriol 191:4693–4694

Fleischmann RD, Adams MD, White O et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd*. Science 269:496–512

Foote SJ, Bosse JT, Bouevitch AB, Langford PR, Young NM, Nash JH (2008) The complete genome sequence of *Actinobacillus pleuropneumoniae* L20 (serotype 5b). J Bacteriol 190:1495–1496

Gao B, Gupta RS (2005) Conserved indels in protein sequences that are characteristic of the phylum Actinobacteria. Int J Syst Evol Microbiol 55:2401–2412

Gao B, Mohan R, Gupta RS (2009) Phylogenomics and protein signatures elucidating the evolutionary relationships among the Gammaproteobacteria. Int J Syst Evol Microbiol 59:234–247

Gioia J, Qin X, Jiang H et al (2006) The genome sequence of *Mannheimia haemolytica* A1: insights into virulence, natural competence, and Pasteurellaceae phylogeny. J Bacteriol 188:7257–7266

Gogarten JP, Doolittle WF, Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. Mol Biol Evol 19: 2226–2238

Griffiths E, Gupta RS (2002) Protein signatures distinctive of chlamydial species: horizontal transfer of cell wall biosynthesis genes *glmU* from Archaebacteria to Chlamydiae, and *murA* between Chlamydiae and Streptomyces. Microbiology 148:2541–2549

Griffiths E, Gupta RS (2006) Molecular signatures in protein sequences that are characteristics of the Phylum Aquificales. Int J Syst Evol Microbiol 56:99–107

Gupta RS (1998) Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. Microbiol Mol Biol Rev 62:1435–1491

Gupta RS (2000) The phylogeny of proteobacteria: relationships to other eubacterial phyla and eukaryotes. FEMS Microbiol Rev 24:367–402

Gupta RS (2005) Protein signatures distinctive of alpha proteobacteria and its subgroups and a model for alpha proteobacterial evolution. Crit Rev Microbiol 31:135

Gupta RS (2006) Molecular signatures (unique proteins and conserved indels) that are specific for the epsilon proteobacteria (Campylobacterales). BMC Genomics 7:167

Gupta RS (2009) Protein signatures (molecular synapomorphies) that are distinctive characteristics of the major cyanobacterial clades. Int J Syst Evol Microbiol 59:2510–2526

Gupta RS (2010) Molecular signatures for the main phyla of photosynthetic bacteria and their subgroups. Photosynth Res 104:357–372

Gupta RS, Griffiths E (2006) Chlamydiae-specific proteins and indels: novel tools for studies. Trends Microbiol 14: 527–535

Gupta RS, Mathews DW (2010) Signature proteins for the major clades of cyanobacteria. BMC Evol Biol 10:24

Gupta RS, Mok A (2007) Phylogenomics and signature proteins for the alpha proteobacteria and its main groups. BMC Microbiol 7:106

Harris JK, Kelley ST, Spiegelman GB, Pace NR (2003) The genetic core of the universal ancestor. Genome Res 13:407–412

Harrison A, Dyer DW, Gillaspy A et al (2005) Genomic sequence of an otitis media isolate of nontypeable *Haemophilus influenzae*: comparative study with *H. influenzae* serotype d, strain KW20. J Bacteriol 187:4627–4636

Hayashimoto N, Ueno M, Tkakura A, Itoh T (2007) Biochemical characterization and phylogenetic analysis based on 16S rRNA sequences for V-factor dependent members of Pasteurellaceae derived from laboratory rats. Curr Microbiol 54:419–423

Hogg JS, Hu FZ, Janto B et al (2007) Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. Genome Biol 8:R103

Hong SH, Kim JS, Lee SY et al (2004) The genome sequence of the capnophilic rumen bacterium *Mannheimia succiniciproducens*. Nat Biotechnol 22:1275–1281

Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ (1998) Multiple sequence alignment with Clustal x. Trends Biochem Sci 23:403–405

Kainz A, Lubitz W, Busse HJ (2000) Genomic fingerprints, ARDRA profiles and quinone systems for classification of Pasteurella sensu stricto. Syst Appl Microbiol 23:494–503

Korczak B, Christensen H, Emler S, Frey J, Kuhnert P (2004) Phylogeny of the family Pasteurellaceae based on rpoB sequences. Int J Syst Evol Microbiol 54:1393–1399

Kuhnert P, Korczak BM (2006) Prediction of whole-genome DNA–DNA similarity, determination of G+C content and phylogenetic analysis within the family Pasteurellaceae by multilocus sequence analysis (MLSA). Microbiology 152:2537–2548

May BJ, Zhang Q, Li LL, Paustian ML, Whittam TS, Kapur V (2001) Complete genomic sequence of Pasteurella multocida, Pm70. Proc Natl Acad Sci USA 98:3460–3465

Mutters R, Mannheim W, Bisgaard M (1989) Taxonomy of the group. In: Adam C, Rutter JM (eds) Pasteurella and Pasteurellosis. Academic Press, London, pp 3–34

Olsen I (1993) Recent approaches to the chemotaxonomy of the Actinobacillus-Haemophilus-Pasteurella group (family Pasteurellaceae). Oral Microbiol Immunol 8:327–336

Olsen I, Dewhirst FE, Paster BJ, Busse HJ (2005) Family I. Pasteurellaceae Phol 1981b, 382^{VP} (Effective Publication: Pohl 1979, 81). In: Brenner DJ, Krieg NR, Staley JT, Garrity GM (eds) Bergey's manual of systematic bacteriology: the proteobacteria, part B: the gammaproteobacteria, 2nd edn. Springer, New York, pp 851–856

Paster BJ, Russell JB, Yang CM, Chow JM, Woese CR, Tanner R (1993) Phylogeny of the ammonia-producing ruminal bacteria *Peptostreptococcus anaerobius, Clostridium sticklandii*, and *Clostridium aminophilum* sp. nov. Int J Syst Bacteriol 43:107–110

Pohl S (1981) DNA relatedness among members of *Haemophilus, Pasteurella* and *Actinobacillus*. In: Kilian M, Frederiksen W, Bilberstein EL (eds) Haemophilus, pasteurella and actinobacillus. Academic Press, London, pp 245–253

Redfield RJ, Findlay WA, Bosse J, Kroll JS, Cameron AD, Nash JH (2006) Evolution of competence and DNA uptake specificity in the Pasteurellaceae. BMC Evol Biol 6:82

Rokas A, Holland PW (2000) Rare genomic changes as a tool for phylogenetics. Trends Ecol Evol 15:454–459

Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804

Shah HN, Olsen I, Bernard K, Finegold SM, Gharbia SE, Gupta RS (2009) Approaches to the study of the systematics of anaerobic, Gram-negative, non-spore-forming rods: current status and perspectives. Anaerobe 15:179–194

Singh B, Gupta RS (2009) Conserved inserts in the Hsp60 (GroEL) and Hsp70 (DnaK) proteins are essential for cellular growth. Mol Genet Genomics 281:361–373

Spinola SM, Bauer ME, Munson RS Jr (2002) Immunopathogenesis of *Haemophilus ducreyi* infection (chancroid). Infect Immun 70:1667–1676

Takatsuka Y, Chen C, Nikaido H (2010) Mechanism of recognition of compounds of diverse structures by the multidrug efflux pump AcrB of *Escherichia coli*. Proc Natl Acad Sci USA 107:6559–6565

Van de Peer Y, De Wachter R (1994) TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. Comput Appl Biosci 10:569–570

Williams KP, Gillespie JJ, Sobral BW et al (2010) Phylogeny of gammaproteobacteria. J Bacteriol 192:2305–2314

Wu D, Hugenholtz P, Mavromatis K et al (2009) A phylogeny-driven genomic encyclopaedia of bacteria and archaea. Nature 462:1056–1060

Xu Z, Zhou Y, Li L et al (2008) Genome biology of *Actinobacillus pleuropneumoniae* JL03, an isolate of serotype 3 prevalent in China. PLoS One 3:e1450

Yue M, Yang F, Yang J et al (2009) Complete genome sequence of *Haemophilus parasuis* SH0165. J Bacteriol 191:1359–1360