DISCOVERY AND APPLICATIONS OF NOVEL PROTEIN BASED MOLECULAR MARKERS FOR BACTERIAL CLASSIFICATION AND IDENTIFICATION

By

Sohail Naushad, M.Sc.

A Thesis

Submitted to the School of Graduate Studies

In Partial Fulfillment of the Requirements

For the Degree

Doctor of Philosophy

McMaster University

DOCTOR OF PHILOSOPHY (2015) McMaster University

(Biochemistry) Hamilton, Ontario

TITLE: Discovery and applications of novel protein based molecular markers for bacterial classification and identification

AUTHOR: Sohail Naushad, M.Sc. (McMaster University)

SUPERVISOR: Professor Radhey S. Gupta

NUMBER OF PAGES: xxii, 178

DEDICATION

This thesis is adoringly dedicated to my brilliant and supportive wife Misbah Sohail for her relentless support and infinite encouragement. She has provided many valuable suggestions to improve this work. Your patience, support and understanding have lightened up my spirit to finish this thesis. Thank you for your unconditional support with my studies.

ABSTRACT

The class *Gammaproteobacteria* and its different main orders are currently classified solely on the basis of their branching in phylogenetic trees. In most cases, no molecular, biochemical or physiological characteristics are known for their demarcation. The availability of genomic sequence data has enabled the discovery of two types of molecular characteristics in the forms of Conserved Signature Indels (CSIs) and Conserved Signature Proteins (CSPs) that provide novel means for identification and demarcation of prokaryotes. In the following work, numerous CSIs and CSPs have been identified for different orders within the class *Gammaproteobacteria,* with particular focus on *Pasteurellales, Xanthomonadales* and "Enterobacteriales". The order *Pasteurellales* contains a single family, *Pasteurellaceae*, harbouring many important human and animal pathogens. We have discovered a large number of novel CSIs that are specific for either all *Pasteurellales* or several distinct clades within this order of bacteria. Based upon these CSIs, we have been able to demarcate the "*sensu stricto*" members of the genera *Actinobacillus, Haemophilus,* and *Pasteurella* that are presently polyphyletic. Similarly, we have identified numerous CSIs for the phytopathogens-containing order *Xanthomonadales* and have used them in conjunction with phylogenetic analyses for the taxonomic reorganization of the members of this order. The *Xanthomonadales* species that branched monophyletically and shared CSIs were grouped into one of two families within the order *Xanthomonadales* while the other species were transferred to a new order. This work also reports many CSIs and CSPs for the phytopathogenic genera *Dickeya*, *Pectobacterium* and *Brenneria* from the order "Enterobacteriales" and this work also discusses the usefulness of

CSIs and CSPs for understanding prokaryotic systematics and taxonomy. Additionally, based upon the species distribution of CSIs, we have also examined the impact of LGT on prokaryotic phylogeny/systematics. The extensive work on CSIs that we have reviewed supports the notion that the genetic changes responsible for them have been inherited predominantly in a vertical manner following Darwinian mode of evolution. The molecular markers discovered in this work, because of their taxa specificities, provide valuable means for genetic and biochemical studies that should lead to discovery of novel biochemical and physiological characteristics of the studied groups of bacteria and they also provide new tools for their diagnostics and as potential drug targets for these bacteria.

ACKNOWLEDGEMENTS

First and foremost, I want to show my heartfelt gratitude to my research supervisor, Dr. Radhey S. Gupta. You were so wonderful to me. You made me believe that I had so much strength and courage to persevere even when I felt lost. You showed me light in a tunnel where everything seemed dark. You were very tolerant and determined to see me through. You were a wonderful motivator even when the coping seemed tough for me. I aspire to emulate you. Since I have joined your research group, I have greatly benefited from your mentorship and the freedom and support that you have provided me to research independently. More importantly, you have set wonderful example of being a good researcher and a great educator. The discipline I have learnt from you will stay with me as an asset for the rest of my life.

I would also like to acknowledge and thank the members of my supervisory committee, Dr. Herb Schellhorn and Dr. Paul Higgs. They have provided me with numerous helpful suggestions and feedbacks, throughout various meetings, we have had. Dr. Schellhorn has kindly provided me access and granted me freedom to work in his lab independently. He is always smiling and welcoming. The discussions with Dr. Higgs have helped me understand many fundamental issues in the field of microbiology.

My appreciation is further extended to past and current members of the Gupta Lab. I should acknowledge the support I received from Beile Gao, Amy Cui, Mathangi and Vaibhav Bhandari when I started my research in the Gupta lab. Among the current members, Bijendra Khadka and Mobolaji Adeolu are always willing to help and support. I always say it to Mobolaji, “man you are a gem”, whenever something is not working in lab,

I say “Mobolaji I need your help” and he is always there to fix it. He is very kind to my children, whenever they see him they get excited and shout “Uncle”!! The word “no” is not in his dictionary. He is always “oh yeah we can do it”.

I would also like to thank, one and only, Bilal Ahsan for his friendship and moral support. You have always helped me to successfully face the challenges in my personal and academic life. You have provided me with new knowledge and viewpoints, enriching my experience as a scientist.

I would like to extend my special thanks to David Aiken and Christine Aiken our family friends in Canada. Dave you have always cherished my success, and I hope you will be “super happy” in “heaven” to see that I am completing my doctoral studies. Christine I don’t have words to say thank you for all your unconditional support. We count on you; you are a blessing and a gift to this world. People say you are an artist, I say you are a magician who can turn stones into statues and make them immortal by your stories.

I would like to express a very special thanks to my father Naushad Ahmad and mother Naseem Akhtar, and my siblings who provided me unconditional love and support to pursue my studies. Life would not have been same without your support.

Finally, to my daughters Alyeen, Aaiza and Urvah, thanks for allowing me to follow my ambition. Without you, I most certainly would not be where I am today. Many a times you wonder where I spend my whole day. I hope one day you will get a chance to read this thesis and you shall get the answer to your question. Words cannot express how much I love you and how grateful I am for your support.

PREFACE

The following work is a sandwich thesis. Chapters 2, 3, 4, 5, 6 and 7 are the unaltered manuscripts, published in the years 2010 to 2014. The preface section in each chapter describes the details of the published article, as well as my contribution to the multiple-authored work. Chapter 1 provides an introduction to the field of systematics and the subjects of the various manuscripts to provide context for the significance of these manuscripts. Chapter 8 reflects on the presented data and shows the usefulness of the work. References for Chapters 1 and 8 are provided at the end of this thesis. All chapters have been reproduced with the consent of all co-authors. Irrevocable, non-exclusive license has been granted to McMaster University and to the National Library of Canada from all publishers. Copies of permission and licenses have been submitted to the School of Graduate Studies.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 2

CHAPTER 3

CHAPTER 7

LIST OF TABLES

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 7

LIST OF ABBREVIATIONS

aa .....Amino acid
AAT ..... Aminotransferase class I and II
AckA .....Acetate kinase
ADK .....Adenylate kinase
ADSS .....Adenylosuccinate synthase
AGPAT ..... 1-acyl-sn-glycerol-3-phosphate acyltransferase
ANI .....Average Nucleotide Identity
Anammox..... Anaerobic ammonium oxidation
ArgRS ..... Arginyl-tRNA synthetase
AspA .....Aspartate ammonia-lyase
AlaRS .....Alanyl-tRNA synthetase
BLAST .....Basic Local Alignment Search Tool
Blastp ..... Standard protein BLAST
CarA .....Carbamoyl phosphate synthase small subunit
CDP .....Cytidylyltransferase
CFBG .....Cytophaga-Flavobacterium-Bacteriodes group
COG .....Clusters of orthologous groups
CorA .....Mg2+ transporter protein
CoxI ..... Cytochrome-c oxidase subunit 1
CSI .....Conserved signature indel
CSP ..... Conserved signature protein
CysE ..... Serine O-acetyltransferase
CTPsyn ..... CTP synthetase
DSMZ ..... German Collection of Microorganisms and Cell Cultures
DAK ..... Dihydroxyacetone kinase phosphatase
DDH..... DNA-DNA hybridization
DGC.....Diguanylate cyclase
DnaA ..... Chromosomal replication initiation protein DnaA
DnaK ..... Chaperone DnaK
DXR .....1-Deoxy-Dxylulose-5-phosphate reductoisomerase
EF-G .....Elongation factor G
EF-Tu .....Elongation factor Tu
EngB .....Small GTP binding protein EngB
FGAM .....Synthase I Phosphoribosylformylglycinamidine synthase I
FGAM ..... Synthase II Phosphoribosylformylglycinamidine synthase II
FliM .....Flagellar motor switch protein FliM
E value .....Expect value
FabG .....Ketoacyl reductase
Ga .....Billion years
GI ..... GenBank Identifier
GidA .....Glucose inhibited division protein A
GidB .....Methyltransferase GidB

GK ..... Glycerol kinase
GlmM ..... Phosphoglucosamine mutase
GltB ..... Glutamate synthase
GlyA ..... Serine hydroxymethyltransferase
GlyS ..... Glycyl-tRNA synthetase, ß subunit
Gft ..... Glucosamine--fructose-6-phosphate aminotransferase
GluRS ..... Glutamyl-tRNA synthetase
GroEL ..... Chaperonin GroEL (also called Hsp60)
Gyrase A (or GyrA) ..... DNA gyrase subunit A
Gyrase B (or GyrB)..... DNA gyrase subunit B
GuaA ..... Bifunctional GMP synthase/glutamine amidotransferase
HGT ..... Horizontal gene transfer
HolB ..... DNA polymerase III subunit delta
IF-2 ..... Initiation factor IF-2
IclR ..... Transcriptional regulator IclR
IleRS ..... Isoleucine-tRNA synthetase
Indel ..... Insert or deletion
IPTG ..... Isopropyl ß-D-1-thiogalactopyranoside
JGI-IMG ..... Joint Genome Institute- Integrated Microbial Genomes
LGT ..... Lateral gene transfer
MEGA ..... Molecular Evolutionary Genetics Analysis
ML ..... Maximum likelihood
MLSA ..... Multilocus Sequence Analysis
MLST ..... Multilocus sequence typing
MraW ..... S-adenosylmethyltransferase MraW
MreB ..... Cell shape determining protein MreB
MurA ..... UDP-N-acetylglucosamine 1-carboxyvinyltransferase
MurB ..... UDP-N-actylenolpyruvoylglucosamine reductase
NCBI ..... National Center for Biotechnology Information
NDH..... NADH dehydrogenase
NCBI ..... National Center for Biotechnology Information
NJ ..... Neighbor joining
ODC ..... Ornithine decarboxylase
ORF ..... Open reading frame
PCR ..... Polymerase chain reaction
PDB ..... Protein Data Bank
PGM/PMM ..... Phosphoglucomutase/phosphomannomutase a/b/subunit
PMM ..... Phosphomannomutase
PNP ..... Purine nucleoside phosphorylase I
POCP ..... Percentage of conserved proteins
PolA ..... DNA polymerase I
PolC ..... DNA polymerase III
PolI ..... DNA polymerase I
PPDK ..... Pyruvate phosphate dikinase

ppGpp ..... Guanosine tetraphosphate
PrsA ..... Ribose phosphate pyrophosphokinase
Ptb ..... Phosphate butyryltransferase
PurD ..... Phosphoribosylamine-glycine ligase
PyrE ..... Orotate phosphoribosyltransferase
QueA ..... S-adenosylmethionine/tRNA ribosyltransferase-isomerase
QueF ..... 7-Cyano-7-deazaguanine reductase
RecR ..... Recombination protein RecR
RGC ..... Rare genomic change
RecA ..... Recombinase A
RecJ ..... Single stranded, DNA specific exonuclease
RmlA ..... Glucose-1-phosphate thymidyltransferase
RNR ..... Ribonucleoside diphosphate reductase
RplD ..... 50S Ribosomal protein L4
RplL ..... 50S Ribosomal protein L7/L12
RplM ..... 50S Ribosomal protein L13
RpsA ..... 30S ribosomal protein S1
RpsH ..... 30S Ribosomal protein S8
RpsI ..... 30S Ribosomal protein S9
RpoA ..... RNA polymerase a-subunit
RpoB ..... RNA polymerase ß-subunit
RpoC ..... RNA polymerase ß'-subunit
RppK ..... Ribose phosphate pyrophosphokinase
SecA ..... Protein translocase subunit SecA
SMC ..... Chromosome segregation protein SMC
SahH ..... S-adenosyl-L-homocysteine hydrolase
SeMet ..... Selenomethionine
SHMT ..... Serine hydroxymethyltransferase
SRP ..... Signal recognition particle
TrmD ..... tRNA (Guanine-1)-methyltransferase
ThrS ..... Threonyl-tRNA synthetase
TrmE ..... tRNA modification GTPase
TrpRS ..... Tryptophanyl-tRNA synthetase
TrxB ..... Thioredoxin reductase
TypA ..... GTP-binding protein TypA
UPRTase ..... Uracil phosphoribosyltransferase
UvrD ..... DNA helicase II
ValRS ..... Valyl-tRNA synthetase
Wzy ..... O-antigen polymerase

GLOSSARY

Analog: A feature that appears similar in two taxa which have originated from two different ancestors.

Ancestor: Any organism, population, or species from which some other organism, population, or species is descended.

Apomorphy: Specialized (derived) characters of an organism.

Bootstrapping: A statistical procedure to assess the reliability of a result (usually a phylogenetic tree) that involves sampling data into given number with replacement from the original data set.

Clade: A group of species including all the species descending from an internal node of a tree and no others. Originated from the Greek word "klados", meaning branch or twig

Cladogram: A diagram, resulting from a cladistic analysis, which depicts a hypothetical branching sequence of lineages leading to the taxa under consideration. The points of branching within a cladogram are called nodes. All taxa occur at the endpoints of the cladogram.

Convergence: Similarities which have arisen independently in two or more organisms that are not closely related. Contrast with homology.

Diversity: Term used to describe numbers of taxa, or variation in morphology.

Evolution: Darwin's definition: descent with modification. The term has been variously used and abused since Darwin to include everything from the origin of man to the origin of life.

Evolutionary tree: A diagram which depicts the hypothetical phylogeny of the taxa under consideration. The points at which lineages split represent ancestor taxa to the descendant taxa appearing at the terminal points of the cladogram.

Homologs: Sequences that are evolutionarily related by descent from a common ancestor (cf. orthologs and paralogs)

Homology: Two structures are considered homologous when they are inherited from a common ancestor who possessed the structure. This may be difficult to determine when the structure has been modified through descent.

Last universal common ancestor: The most recent organism from which all organisms now living on earth descend. Thus it is the most recent common ancestor of all current life on Earth.

Lineage: Any continuous line of descent; any series of organisms connected by reproduction by parent of offspring.

Long branch attraction: A phenomenon in phylogenetic analyses (most commonly those employing maximum parsimony) when rapidly evolving lineages are inferred to be closely related, regardless of their true evolutionary relationships.

Monophyletic: Adjective describing a group of species on a phylogenetic tree that share a common ancestor that is not shared by species outside the group. A clade is a monophyletic group.

Orthologs: Sequences from different species that are evolutionarily related by descent from a common ancestral sequence and that diverged from one another as a result of speciation.

Outgroup: A species (or group of species) that is known to be the earliest-diverging species in a phylogenetic analysis. Outgroup is added in order to determine the position of the root.

Paralogs: Sequences within the same organism that have arisen by duplication of one original sequence.

Phylogeny: An evolutionary tree showing the relationship between sequences or species.

Phylum: A taxonomic rank below Kingdom and above Class. The minimal requirement is that all organisms in a phylum should be related closely enough for them to be clearly more closely related to one another than to any other group.

Polyphyletic: Adjective describing a group of species on a phylogenetic tree for which there is no common ancestor that is not also shared by species outside the group. A polyphyletic group is evolutionarily ill-defined.

Rank -- In traditional taxonomy, taxa are ranked according to their level of inclusiveness. Thus a genus contains one or more species, a family includes one or more genera, and so on.

Synapomorphy: A character which is derived, and because it is shared by the taxa under consideration, is used to infer common ancestry (shared derived state).

Systematics: A field of biology that deals with the diversity of kinds. Systematics is usually divided into the two areas of phylogenetic and taxonomy.

Taxonomy: The science of naming and classifying organisms.

CHAPTER 1.

Introduction

Preface

"The affinities of all the beings of the same class have sometimes been represented by a great tree... As buds give rise by growth to fresh buds, and these, if vigorous, branch out and overtop on all sides many a feebler branch, so by generation I believe it has been with the great Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever branching and beautiful ramifications."

~Charles Darwin (the Origin of Species, Chapter IV, 1859)

Understanding prokaryotic evolution – The forefathers

A sound understanding of the prokaryotes, the lone dwellers of this planet for the first 2-2.5 billion years of life, has been the most captivating issue in life sciences (Gupta, 1998a; Fox et al., 1980; Woese et al., 1978; Sagan, 1967; Zotin et al., 1975; Uzzell and Spolsky, 1974; Gray, 2012; Novoselov et al., 2013). Throughout their evolution, prokaryotic species, have played a pivotal role in shaping this planet and its environments (Nishioka et al., 1970; Migita and Doi, 1970; Hall, 1971; Flavell, 1972; Novoselov et al., 2013; Nisbet and Sleep, 2001). Thus, to understand the most vital facets of the origin and history of life on earth and its spread to all life-permitting environments, a detailed and comprehensive understanding of prokaryotic evolutionary history is indispensable.

The knowledge gained by exploring prokaryotic evolution has provided many novel insights about various fundamental concepts such as the origination of cell, the origination of metabolic pathways and advent of information transfer processes (Hall, 1971; Flavell, 1972; Migita and Doi, 1970; Sagan, 1967; Kasting and Siefert, 2002; Novoselov et al., 2013; Natochin et al., 2012; Zhang et al., 2009; Nisbet and Sleep, 2001;

Harel et al., 2014). The prokaryotes are present in different environments, including those that are at the extreme of temperature, pressure, acidity, alkalinity, salinity etc. (Hauptmann et al., 2014; Colman et al., 2014; Sorokin et al., 2014). They have a diverse array of survival strategies and life histories; the studies on which, have broadened our understanding of many fundamental principles of life, including the evolution of oxygenic photosynthesis from anoxygenic photosynthesis, carbon and nitrogen fixation and their recycling, the beginning of symbiotic relationships leading to emergence of multicellular plants and animals, and the existence of beneficial, opportunistic, and pathogenic organisms (Raskin et al., 2006; Xiong, 2006; Gupta, 2000; Flavell, 1972; Knoll, 1999; Schopf, 1978; Kasting and Siefert, 2002; Nisbet and Sleep, 2001; Gray, 2012; Novoselov et al., 2013).

Microbial classification has long been a daunting challenge for scientists and taxonomists. The first notable attempt to classify microorganisms came at the hands of Carl Linnaeus in 1774. In his work "*Systema Naturae"*, he placed microbes, which he named "*Infusoria*" into one species that was judiciously baptized as "*Chaos infusoria*" (Linnaeus C, 1774; Pace et al., 2012; Oren, 2010). The classification of microbes saw little improvement after the Linnaean classification scheme was proposed, particularly because there was no consensus, in the early days, as to whether these microbes should be recognized as animals or small plants (Oren and Garrity, 2014; Pace et al., 2012; Oren, 2010). This plant-animal dualism was resolved by Ernst Haeckel in his famous work, *"Die Systematische Phylogenie"* in 1866 (English "Systematic Phylogeny"), in which he clearly defined the terms ontogeny, phylogeny and phylum and placed bacteria and blue

green algae into a separate division, he called Monera (Haeckel E, 1866; Pace et al., 2012; Oren, 2010). However, since it became evident that only few bacteria show close relationship to blue green algae, this division was quickly rejected (Sapp, 2005; Oren and Garrity, 2014; Stanier and Van Niel, 1962). Many efforts were put forward to create an accurate classification scheme for prokaryotes. However, the prokaryotes, due to limited means of observations, were poorly differentiated and placed into small number of groups, termed genera, based upon their cell shapes (Stanier and Van Niel, 1941).

The prokaryotes, for most part of the mid-20th century, were classified based upon morphological or physiological characteristics (Oren and Garrity, 2014; Ramasamy et al., 2014; Chun and Rainey, 2014; Stanier and Van Niel, 1941; Stanier and Van Niel, 1962; Whittaker, 1969). The use of morphological or physiological characteristics was later augmented with the addition of chemotypic and genotypic characteristics (Cowan, 1965; Oren and Garrity, 2014; Pace et al., 2012; Tindall et al., 2006; Whittaker, 1969). The morphological characters were limited to the observation of growth of microbes on culture plates, observing colony morphology or the monitoring of cell morphology, cell size, cell motility, flagellation type and Gram staining (Sapp, 2009; Stanier and Van Niel, 1941). The physiological characteristics used for classification included the growth temperature range, pH range, salinity tolerance, and acidity and alkalinity tolerance (Sapp, 2009; Oren and Garrity, 2014; Schleifer, 2009; Tindall et al., 2006). Much effort was expended on improving the understanding of bacterial phylogeny and classification schemes with numerous debates on whether morphological or physiological criteria were to take precedence in depicting prokaryotic relationships (Oren and Garrity, 2014;

Schleifer, 2009; Harris et al., 2003; Pace et al., 2012; Sapp, 2009; Oren, 2010). The criteria for prokaryotic classification were further expanded with the addition of cytological data into classification, which led to the distinction between prokaryotes and eukaryotes (Stanier and Van Niel, 1962). Prokaryotic classification, due to the diverse variety present in them, their simple morphology, their small sizes, and sharing of characteristics through convergent evolution, was difficult to establish. These difficulties in classifying bacteria based on simply physical criteria were widely discussed and acknowledged by the 1970s, leading to an era often described as "The Dark Age" (Whittaker, 1969; Stanier and Van Niel, 1962; Woese, 1987; Oren and Garrity, 2014; Oren, 2010). These discussions raised the importance of finding new, reliable methods of prokaryotic differentiation and classification.

The use of molecular data in classification – The genetic era

The first major innovation of the 20th century in prokaryotic classification were DNA-DNA hybridization (DDH) techniques, which allowed for more reliable differentiation of prokaryotes using genetic material that was not affected by phenotypic convergence (Mccarthy and Bolton, 1963; Gevers et al., 2005; Oren and Garrity, 2014; Sapp, 2009). DNA-DNA hybridization takes into consideration the entirety of the genetic material in a pair of organisms to estimate their relatedness and quickly became established as the "gold standard" for the differentiation of prokaryotic species. (Stackebrandt and Goebel, 1994; Oren and Garrity, 2014; Tindall et al., 2006; Chun and Rainey, 2014; Oren, 2010; Sapp, 2009; Tindall et al., 2010). A DNA-DNA hybridization value of 70% became the established cutoff threshold for species demarcation and was

widely used to rectify previously misclassified species designations (Stackebrandt and Goebel, 1994). However, the usage of DNA-DNA hybridization has its own limitations, as it is influenced by many factors including physiochemical parameters, genome size, plasmids, DNA purity (Stackebrandt and Ebers, 2006). Additionally, DNA-DNA hybridization is time-consuming and expensive (Pace et al., 2012; Gevers et al., 2005) and often requires special facilities, which are present in limited number of laboratories, making it difficult to establish a comparative database incrementally (Ramasamy et al., 2014; Oren and Garrity, 2014). Additionally, the cut off value is not applicable to all prokaryotic taxa, especially in cases where closely related species have DDH value >70%, as in case of *Rickettsia* species (Fournier and Raoult, 2009; Ramasamy et al., 2014). The hybridization analysis is only useful in differentiating among species and strains, relationships among distantly related groups (viz. genus and above) cannot be accurately ascertained through this methodology (Oren and Garrity, 2014). Furthermore, the analyses can only be performed on cultureable microbes, which are estimated to account for only about 1% of total prokaryotic diversity (Yarza et al., 2014; Amann et al., 1995; Pace, 1997). Additionally, the method is subject to experimental variability as different experiments, by different labs or experimenters, can produce different hybridization data (Goris et al., 2007).

A major revolution and advancement in the field of evolutionary sciences came with the advent of gene sequencing techniques, in particular sequencing of the 16S rRNA gene, and its use as a tool in identification of species relationships (Woese and Fox, 1977; Fox et al., 1980; Woese, 1987; Tindall et al., 2006; Tindall et al., 2010; Oren, 2010). In

the past 30 years, with the introduction of 16S rRNA gene sequences as phylogenetic marker, much has been learned about the diversity of prokaryotic organisms, which has revolutionized our understanding of the evolutionary history and systematics of the prokaryotes (Oren and Garrity, 2014; Ramasamy et al., 2014; Tindall et al., 2010; Tindall et al., 2006; Woese et al., 1985a; Woese, 1987; Olsen et al., 1994). 16S rRNA gene sequences are universally present and highly conserved among species of bacteria and archaea (Woese, 1987). The 16S rRNA gene contains variable regions which enable comparison among closely related species, and conserved regions which allow comparisons among more distantly related taxa. Some regions of 16S rRNA gene are completely conserved which enables the use of universal PCR primers for species detection (Greisen et al., 1994; Marchesi et al., 1998). Additionally, the 16S rRNA genes as being a part of a large functional complex (i.e. ribosome) are less likely to have undergone lateral gene transfer (Ramasamy et al., 2014; Pace et al., 2012; Yarza et al., 2014; Oren, 2010; Sapp, 2009; Woese, 1987). These characteristics make the 16S rRNA gene an ideal candidate for analysis in order to discover novel aspects of prokaryotic relationships, particularly through the use of phylogenetic trees or direct sequence comparisons. Empirically, 97% 16S rRNA gene homology corresponds to the 70% DNA-DNA hybridization threshold (Stackebrandt and Goebel, 1994), which has been widely used for classification of species (Tindall et al., 2010; Kim et al., 2014; Ramasamy et al., 2014; Yarza et al., 2014; Oren and Garrity, 2014; Sapp, 2009; Tindall et al., 2006). The use of 16S rRNA sequence analysis was instrumental in the introduction of three-domain classification system for cellular life forms including the division of the

prokaryotic species into Bacteria and Archaea (Woese, 1987; Woese and Fox, 1977). Subsequently, 16S rRNA gene based phylogenies have become extremely prevalent in prokaryotic systematics; the current definition of a prokaryotic species is based almost solely upon 16S rRNA sequence similarity (Tindall et al., 2006; Oren and Garrity, 2014; Ramasamy et al., 2014; Yarza et al., 2014; Tindall et al., 2010). The newly discovered bacterial isolates in most cases are defined as members of the same species if they share 97% or greater 16S rRNA gene sequence similarity (Tindall et al., 2010; Tindall et al., 2006; Oren and Garrity, 2014; Zhi et al., 2012). One should, however, keep in mind that the 16S rRNA gene similarity value of 97% was never intended as a cutoff for species demarcation (Oren and Garrity, 2014). This value was only shown as an equivalent of 70% DDH value that was obtained by comparing the 16S rRNA gene sequence data available at the time with DDH (Stackebrandt and Goebel, 1994). However, this cutoff value is widely used and has become an accepted standard for taxonomy and systematic of prokaryotes. The cutoff value has been subsequently reassessed and a cutoff of 98.7-99% was proposed in 2006, as a new threshold for species demarcation by comparing 380 organisms (Stackebrandt and Ebers, 2006). More recent analyses have suggested similarity value of 98.2-99% as cutoff, comparing 571 different strains of bacteria (Meier-Kolthoff et al., 2013).

However, whatever value we consider “accurate” for demarcation purposes, these values are only useful for the differentiation of closely related species. For higher taxa, no concept of “cutoff similarity value” is agreed upon (Wayne et al., 1987; Stackebrandt and Ebers, 2006; Oren, 2010; Zhi et al., 2012; Ramasamy et al., 2014; Yarza et al., 2014;

Stackebrandt et al., 2002). Many efforts were put forward to establish criteria that can be used for demarcation of higher taxa. Recently, an analysis of dataset containing 8,602 bacteria and archaeal species has been published which proposes 16S rRNA similarity criteria for the demarcation of higher taxonomic ranks within prokaryotes. A value of 94.5% or lower sequence similarity is suggested as a strong evidence to differentiate genera, a value of 86.5% or lower for demarcation of families, 82.0% or lower for distinction of orders, the values of 78.5% or lower and 75.0% or lower, have been proposed as boundaries to distinguish, classes and phyla, respectively (Yarza et al., 2014).

Although the use of 16S rRNA gene sequence analyses has allowed for a remarkable improvement in our understanding of prokaryotic taxonomy, there are a number of notable issues and limitations involving analyses of the 16S rRNA gene. For example, the high degree of sequence conservation in 16S rRNA gene limits the phylogenetic resolution of analyses based on the gene, leading to the misidentification of closely related species (Stackebrandt et al., 2002; Alperi et al., 2010; Oren and Garrity, 2014). The GC% contents of the 16S rRNA genes are strongly correlated with the optimal growth temperatures of prokaryotes, leading to a convergent GC% bias in organisms with similar optimal growth temperatures (Brenner et al., 2005; Stackebrandt et al., 2007; Stackebrandt et al., 2002). Another issue that arises when analyzing the 16S rRNA gene is that RNA gene in some prokaryotic species is present in multiple, sometimes highly divergent, copies (Oren and Garrity, 2014) that can produce different phylogenies (Janda and Abbott, 2007). Although rare, the 16S rRNA gene is also subjected to lateral gene transfer (Kitahara and Miyazaki, 2013). Most importantly, being

a single gene within genomes that contain hundreds or thousands of other genes, it is suggested that the 16S rRNA gene based phylogenies may not accurately reflect the true evolution of the whole genome of an organism (Ciccarelli et al., 2006; Oren and Garrity, 2014).

Due to the limitations of 16S rRNA gene based phylogenetic analysis, organisms can be misclassified as members of the incorrect taxonomic group based upon 16S rRNA gene analysis, while analysis of other genes and other characteristics may suggests contrary results (Janda and Abbott, 2007; Fox et al., 1992; Oren and Garrity, 2014). However, the current hierarchical classification of bacteria and archaea into different phyla and smaller taxa within these phyla is established based on the information, primarily deduced from their branching in 16S rRNA gene trees (Oren and Garrity, 2014; Tindall et al., 2006; Tindall et al., 2010; Chun and Rainey, 2014; Kim et al., 2014; Woese, 1998; Woese et al., 1990; Fox et al., 1980). Apart from their branching in phylogenetic trees no other criteria currently exists that can define these groups in more definitive terms.

The use of Genomic data in prokaryotic classification – The genomic era

The genomic era of prokaryotic research started with the availability of first complete genome sequence of *Haemophilus influenzae* in 1995 (Fleischmann et al., 1995). However, the use of complete genome sequences in prokaryotic taxonomy was very limited, because of high-cost and time consuming sequencing facilities. Ground breaking advancement in genomic field came with the establishment of next generation sequencing (NGS) technologies in 2005, with the development of Roche 454 sequencing

system (Margulies et al., 2005), followed by the Illumina DNA sequencing platforms, HiSeq and MiSeq (van Dijk et al., 2014). These two technologies were followed by a third NGS platform released in 2007, that worked on the principle of Sequencing by Oligo Ligation Detection (SOLiD), and a fourth NGS platform, the Ion Torrent, a semiconductor based sequencing technology (van Dijk et al., 2014). These technologies have provided means to sequence microbial organisms at a very low cost. A long awaited innovation in taxonomy and evolutionary sciences, and perhaps for all of biological sciences, has been the availability of these speedy and cost-effective genomic sequencing technologies, commonly referred to as NGS technologies. All of these NGS systems can generate massive amount of genomic data in a relatively short period of time, and, as the genome holds the complete genetic information of the organism, decoding the genome was expected to allow for insight into prokaryotic life and their relationships (Boussau and Daubin, 2010; Chun and Rainey, 2014; Ramasamy et al., 2014; Gupta, 2000). With the decrease in genome sequencing cost, a massive number of prokaryotic genomes have become available in public databases the last decade. As of December 2014, over 30,000 prokaryotic genome sequences are publically available in NCBI genome database and this number is increasing exponentially.

The availability of this huge amount of genomic data has provided us with wealth of information and has been useful in many aspects of biological sciences (Staudt, 2003). Genomic sequence data has also had a profound effect on the field of prokaryotic systematics, leading to the development of new methods of determining species relationships. Some of these methods include Average Nucleotide Identity (ANI), the

measure of mean nucleotide sequence similarity of shared genes between two species (Goris et al., 2007), Average Amino Acid Identity (AAI), the measure of mean amino acid similarity index between species (Konstantinidis and Tiedje, 2005), Genome BLAST Distance Phylogeny (GBDP), a method utilizing the all-against-all pairwise comparison by BLAST program to produce high-scoring segment pairs to infer phylogenetic relationships (Henz et al., 2005), Tetra-nucleotide regression analysis, a method based on the tetra-nucleotide usage patterns in different genomes (Karlin et al., 1994), and Maximum unique exact match index (MUMi), which involves the identification of regions of exact match between two genomes utilizing various algorithms (Deloger et al., 2009). All of these methods utilize whole genome sequence data for comparisons between two species and thus are termed "overall genome relatedness indices (OGRI) (Chun and Rainey, 2014).

Other sequence based methods that use part of the genome or sets of different genes, include multilocus sequencing typing (MLST), which is a technique to measure allelic profile or sequence types (ST) of 4-10 housekeeping genes to characterize different species based on the difference of their sequence types (Maiden et al., 1998) and multilocus sequence analysis (MLSA) is the use of multiple housekeeping genes to construct phylogenies. Other sequence based methods of determining species relationships include the comparison of gene content to identify differences in GC% and codon usage, comparisons of gene order differences, and the identification of rare genomic rearrangements (Coenye et al., 2005; Konstantinidis and Tiedje, 2005; Snel et al., 1999). All of these methods possess their own advantages and limitations, but they

have each contributed greatly in our understanding of prokaryotic evolution and have advanced the field of prokaryotic systematics (Sapp, 2009; Oren and Garrity, 2014; Tindall et al., 2006; Pace et al., 2012; Chun and Rainey, 2014). The large and increasing availability of prokaryotic genomes has led to the proposal to replace DNA-DNA hybridization (DDH) with in-silico genome comparison techniques, such as ANI, as the gold standard for taxonomic purposes (Konstantinidis and Tiedje, 2005). It has been estimated that an ANI of 95% between two genomes is equivalent to the 70% DDH, which is the standard for species cutoff (Tindall et al., 2010; Goris et al., 2007).

Many novel genera and species have been described using ANI analysis, including species of *Burkholderia*, *Streptococcus, Dehalococcoides maccartyi*, *Geobacter*, *vibrio*, *Sphaerochaeta globose* and *Sphaerochaeta acribbeanicus* (Goris et al., 2007; Chun and Rainey, 2014; Richter and Rossello-Mora, 2009; Camelo-Castillo et al., 2014; Hoffmann et al., 2012; Lee et al., 2012; Loffler et al., 2013; Chun and Rainey, 2014). Recently, a detailed analysis incorporating 6787 genomes from 22 different prokaryotic phyla, have been conducted to reassess the ANI cutoff value for intra - and interspecies relationships. Over one million comparisons were carried out to establish that an ANI of 95-96% should serve as a cutoff threshold for prokaryotic species demarcation, which also corresponds to 98.65% 16S rRNA sequence similarity, the current 16S rRNA based species cutoff (Kim et al., 2014). However, ANI analysis is limited to pairwise comparisons between two organisms and does not allow for the development of incremental database. Additionally, the results of ANI analysis do not always coincide with currently established phylogeny and thus should not be used as a sole tool for classification purposes (Ramasamy et al.,

2014). ANI values also cannot be used for the demarcation of higher taxa. A recent study analyzed the ANIs of genomes from 12 different prokaryotic families and orders to establish a cutoff for genus demarcation (Qin et al., 2014). However, based on observations of their results they concluded that ANI values are not consistent enough for genus level demarcation.

Unlike ANI, which compares whole genomes, phylogenetic inferences can also be obtained from genes that are conserved among different organisms. The two most widely used methods for such analyses are MLST and MLSA (Ramasamy et al., 2014; Chun and Rainey, 2014; Kim et al., 2014; Oren and Garrity, 2014). MLSA has been successfully used to elucidate phylogenetic structure of many important prokaryotic taxa (Chun et al., 2009; Haley et al., 2010; Chun and Rainey, 2014; Prado et al., 2014; Brady et al., 2014; Zhou et al., 2014; Gomila et al., 2014). MLSA uses conserved genes for phylogenetic analysis; often the genes *atpD, recA, glnII, dnaK, rpoB, gyrB*, *truA* and *thrA*. Currently no criteria exists to determine how many and which of the genes are good for phylogenetic studies. Many studies use a subset of the above genes in addition to other genes for MLSA based phylogenetic studies. However, it has been argued that phylogenies derived from single genes/proteins or even from concatenation of multiple genes, represents only a small fraction of whole genome, sometimes as little as 1% which limits their evolutionary significance (Dagan and Martin, 2006; Doolittle and Bapteste, 2007; Zhaxybayeva et al., 2006). However, the high level of correlation between phylogenetic trees based on a limited number of genes and phylogenetic trees based on

whole genomes or all conserved genes shared by a group of organisms have proven these concerns unfounded (Naushad et al., 2014b; Williams et al., 2010).

Although, similarity studies and phylogenetic tree construction methods are useful for inferring relationships among prokaryotic groups, they fail to provide distinct characteristics for defining a related group of organisms. All of the methodologies discussed above work on the principle of relative similarity and are based upon degrees of relatedness rather than providing unique characters that may distinguish groups of related organisms. Recently, a quantitative method to define the taxonomic unit "Genus" has been introduced. This method is based on finding the percentage of conserved proteins (POCP) between two strains to estimate their evolutionary and phenotypic distance. Two strains are considered the members of same genus if they have a POCP of more than 50% (Qin et al., 2014). However, this method is restricted to define only "genus" and cannot be used for species demarcation or identification of any higher taxonomic levels. Thus, there is a need to search for novel genomic features unique to phylogenetically related prokaryotic lineages.

Prokaryotic evolution and impact of Later Gene Transfer (LGT)

The tree-like evolutionary process, also known as the Darwinian mode of evolution, in which traits are transferred from ancestors to offspring, is the entrenched model for prokaryotic and eukaryotic evolution (Kurland, 2005; Gupta, 2000; Naushad et al., 2014a; Bhandari et al., 2012; Beiko et al., 2005; Puigbo et al., 2009). Hence, the term "tree of life" is used to elucidate the bifurcating connection linking all existing species to a last common universal ancestor (Darwin, 1859; Gogarten and Townsend, 2005).

Linnaean taxonomy reflects the recurrent bifurcation of ancestral lineages and represents the division of organisms in a ranked system so as to reflect their evolutionary history.

Recently, the Darwinian tree-like representation of relationships between species, have been questioned as lateral gene transfer (LGT), also known as horizontal gene transfer (HGT), has been implicated to affect this process (Bapteste et al., 2009; Boucher et al., 2003; Nelson et al., 1999). Lateral gene transfer (LGT) is the acquisition of foreign genetic material into the genome of a species through means other than vertical inheritance. The most common mechanisms of LGT are transformation, transduction or conjugation (Davison, 1999). It is believed, strongly among some investigators, that LGT events are so "rampant" that genes cannot be used as reliable phylogenetic markers (Boucher et al., 2003; Handy and Doolittle, 1999). The first experimental evidence for LGT as a mechanism for genetic transfer was demonstrated in 1951, in an experiment showing the lateral transfer of a virulence gene between different bacterial strains (Freeman, 1951). The primary role of LGT in prokaryotic communities was thought to be its involvement in the spread of antibiotic resistance in pathogenic bacteria (Boto, 2010; Akiba et al., 1960). However, the impact of LGT on bacterial evolution was not well explored until the availability of genome sequences (Boucher et al., 2003). The comparative analyses performed on genomic datasets have revealed that the prokaryotic relationships, inferred from phylogenies based on different genes and proteins are not congruent (Gogarten et al., 2002; Bapteste et al., 2009; Andam and Gogarten, 2011; Swithers et al., 2009; Bapteste et al., 2009; Dagan and Martin, 2006; Puigbo et al., 2009). Incidences of LGT are believed to be the main reason behind the phylogenetic

incongruence between different genes and proteins (Sjostrand et al., 2014). Although the contribution of LGT to genome evolution is well established, the frequency of such genetic events and the rate of successful incorporation of foreign genetic material into prokaryotic genomes has been the subject of much debate among evolutionary microbiologists (Naushad et al., 2014a; Bhandari et al., 2012; Naushad and Gupta, 2013; Daubin et al., 2003; Gogarten et al., 2002; Kurland et al., 2003; Doolittle and Bapteste, 2007).

The genes involved in large networks or performing essential functions were thought to be minimally affected by LGT (Jain et al., 1999; Rivera and Lake, 1992). However, it has been suggested that each gene has gone through one or more instances of LGT and that no gene is completely exempt from this process (Boucher et al., 2003; Bapteste et al., 2009; Brochier et al., 2000; Yap et al., 1999; Zhaxybayeva et al., 2006). Evidence for extensive lateral gene transfer in some prokaryotic organisms has led to the suggestion that prokaryotic genomes should not be thought of as coherent wholes, but as mosaics of genes with different evolutionary histories (Nelson-Sathi et al., 2015; Thiergart et al., 2014; Boto, 2010; Nelson et al., 1999; Koonin et al., 2001). Cases of LGT have been identified at the largest evolutionary distances, including instances of lateral transfers of genes from bacteria to archaea, bacteria to eukaryotes, archaea to eukaryotes and vice versa (Jaramillo et al., 2015; Suwastika et al., 2014; Thiergart et al., 2014; Nelson-Sathi et al., 2015; Thiergart et al., 2012; Boto, 2010). This indication of prevalent LGT among prokaryotes has led to the acceptance that perhaps LGT diminishes and conceivably eliminates, the ability to ascertain a Darwinian tree-like evolutionary history

for prokaryotic species (Bapteste and Boucher, 2008; Bapteste et al., 2009; Doolittle, 2000; Eisen, 2000). Thus, only a vague tangled web-like structure is believed to be present, representing phylogenetic histories (Swithers et al., 2009; Williams et al., 2011; Thiergart et al., 2012). One of the major issues in microbiology is the non-availability of discrete and reliable methods for the detection of LGT. Most methods of LGT detection are based on a number of explicit or implicit assumptions, thus different methods can produce different results using same dataset (Puigbo et al., 2009; Koonin et al., 2011; Bhandari et al., 2012).

The methods that are most routinely used for the detection of LGT are classified into three broad categories. These include sequence composition methods, similarity based or distance based methods, and phylogenetic tree construction methods (Sjostrand et al., 2014). Sequence composition based methods involve scanning of the genome sequences for regions of atypical base composition, such as GC content, codon usage pattern and different base composition in relation to others genes (Boto, 2010; Marri et al., 2006). Similarity based methods survey genes in the genome using BLAST to find their closest relatives (Nelson-Sathi et al., 2015; Nelson et al., 1999; Zhaxybayeva et al., 2006). Phylogenetic tree construction based methods search for evidence of discordance among single gene trees (Koonin et al., 2011; Boucher et al., 2003; Zhaxybayeva et al., 2006; Sjostrand et al., 2014; Akerborg et al., 2009). Phylogenetic tree construction based methods are the most widely used means of identifying instances of LGT (Sjostrand et al., 2014; Akerborg et al., 2009; Patterson et al., 2013). With the advancement of sequencing technologies, automated methods of identifying instances of LGT have also been

designed. The most popular of these are Bayesian Markov Chain Monte Carlo (MCMC) approaches, available in the forms of MrBayes and BEAST (Ronquist and Huelsenbeck, 2003; Drummond and Rambaut, 2007). Other methods such as PrIME-GSR, PrIME-DLTRS are also used to detect LGT. Both of these methods are based on constructing gene trees for different species. The first method is based on the GSR model, which incorporates Gene duplication, Sequence evolution, and a Relaxed molecular clock for substitution rates (Akerborg et al., 2009). The second method is based on DLTRS model (*duplication-loss-transfer model with independent and identically distributed rates across gene tree edges*) (Sjostrand et al., 2014). Another method, based on Detection of Coevolution with Lateral Transfers (DeCoLT), has also been widely used to identify instances of LGT from genome sequence data (Patterson et al., 2013).

Despite all these efforts, no consensus is present among different investigators regarding the prevalence and effect of LGT on evolutionary relationships. Many studies have been carried out suggesting a low incidence of lateral genetic transfers (Kurland et al., 2003; Kunin et al., 2005; Naushad and Gupta, 2013; Naushad et al., 2014a; Bhandari et al., 2012). It has been noted that several barriers to free genetic transfer among prokaryotic species exist (Jain et al., 1999; Kurland, 2005; Thomas and Nielsen, 2005). In an effort to quantify LGTs, Beiko et al. performed a comprehensive and detailed phylogenetic analysis on >220,000 proteins from 144 prokaryotic genomes. The inferred relationships suggest a pattern of vertical transfer of genetic material from ancestor to offspring, which supports the Darwinian mode of evolution. However, aberrant patterns were also observed in some closely related taxa and among distantly related organisms

living in convergent environments (Beiko et al., 2005). Additional studies attempting to quantify the effects of LGT have involved the reconstruction of phylogenetic trees for 6901 prokaryotic genes (Puigbo et al., 2009) and the reconstruction of single gene phylogenies for 315 prokaryotic and 85 eukaryotic genomes (Thiergart et al., 2014). Significant topological differences were observed among different trees in both studies representing possible incidences of LGT. However, a consistent phylogenetic signal was present in most of the trees, indicating a central trend of vertical inheritance, supporting the Darwinian mode of evolution. Many studies are beginning to suggest that prokaryotic evolutionary history follows both tree-like and network-like patterns of evolution (Koonin et al., 2011; Puigbo et al., 2009; Nelson-Sathi et al., 2015; Thiergart et al., 2014; Boto, 2010). The impact of LGT on prokaryotic phylogeny and its prevalence is further discussed in Chapters 2 and 7 of this thesis.

Conserved Molecular Signatures as phylogenic tools – Tree-independent phylogeny

The exponentially increasing availability of genome sequence data provides a means to perform different types of studies to find unique molecular features that serve as shared derived characters among prokaryotic taxa. These shared derived characters should be homologous, apomorphic characters, introduced only once during the course of evolution. Our lab has pioneered the usage and discovery of two such kind of molecular signatures for identifying prokaryotic phylogeny (Gupta, 1998b; Naushad et al., 2014a). The first type of these molecular marker are Conserved Signature Indels (CSIs) in widely distributed proteins, that are specific for the different prokaryotic taxa and are helpful in identifying different groups in molecular terms. The CSIs that serve as useful

phylogenetic marker for evolutionary studies are generally of defined size and are flanked on both sides by conserved regions to ensure reliability of signatures and to maintain that the CSI is not due to alignment errors or artifacts (Gupta, 1998b; Gupta and Griffiths, 2002).

The CSIs originate as a result of rare genomic events that occur once in a common ancestor and are then passed on to all descendants vertically. Hence, when CSIs of defined size are uniquely found in phylogenetically well-defined group(s) of species, they function as molecular synapomorphies that distinguish the group from other prokaryotic organisms (Gupta, 1998b). Due to the rarity of mutations affecting conserved regions within functionally important proteins, the shared presence of CSIs is most parsimoniously explained by the common inheritance of the rare genetic changes from an ancestor to its progeny (Gupta, 1998b). Also, since genetic changes leading to CSIs could be introduced at various stages during evolution, it allows for the identification of CSIs at different phylogenetic depths corresponding to various taxonomic rankings. Dr. R. S. Gupta and colleagues, over the course of the last two decades, have utilized these CSIs for the identification of different groups of prokaryotes ranging from the genus level (viz. *Clostridium*) to beyond the phyla level (e.g. Aquificae, Actinobacteria, Thermatogae and Synergistetes) including prokaryotic groups at the superphylum level (Gao et al., 2006; Gupta and Bhandari, 2011; Bhandari and Gupta, 2012; Gupta et al., 2012). Depending upon the presence or absence of a given CSI in the outgroup species a rooted phylogenic relationship can be deduced that is independent of phylogenetic trees.

The second type of taxonomic marker which provides powerful means to define different prokaryotic groups and their phylogenetic relationships are "whole proteins" or "Conserved Signature Proteins" (CSPs) that are uniquely present in particular prokaryotic taxa but not found anywhere else. These CSPs or lineage-specific proteins, arise throughout the evolutionary process (Naushad et al., 2014a; Bhandari et al., 2012; Kainth and Gupta, 2005). A large number of lineage-specific proteins, also called "ORFans", are introduced during speciation or strain divergence (Daubin and Ochman, 2004a; Daubin and Ochman, 2004b). Several studies have indicated the unique sharing of these CSPs from groups of organisms at different phylogenetic depths. Thus these CSPs serve as molecular markers for the identification of these groups from other prokaryotic taxa. Because these CSPs are restricted to particular taxa, it is likely that they are involved in some specialized functions that are limited to particular groups of organisms. Extensive comparative genomic analyses have been performed in Gupta lab to identify these lineage-specific proteins for many prokaryotic groups, such as *Alphaproteobacteria*, *Betaproteobacteria*, *Gammaproteobacteria*, *Epsilonproteobacteria*, *Chlamydia*, *Cyanobacteria*, *Deinococcus–Thermus*, *Bacteroidetes*, etc (Naushad et al., 2014a; Kainth and Gupta, 2005; Bhandari et al., 2012; Gao et al., 2009; Gupta and Griffiths, 2006; Griffiths and Gupta, 2006).

Molecular markers, in the form of CSIs or CSPs, when are found to be shared by distinct organisms, can most parsimoniously be explained by the Darwinian mode of evolution. "As Darwin wrote..."

"...when several characters, let them be ever so trifling, occur together throughout a large group of beings having different habits, we may feel almost sure, on the theory of descent, that these characters have been inherited from a common ancestor..." (Darwin, 1859)

Therefore, the CSIs and CSPs serve as useful characters for prokaryotic phylogenic and taxonomic studies. The utility of these molecular markers for discriminating different taxonomic groups is discussed in Chapters 2-7.

The diversity and phylogenetic overview of Gamma-Proteobacteria

The phylum Proteobacteria comprises the largest group within Bacteria. The members of this group, initially defined as "purple bacteria and relatives" were divided into 4 main groups or divisions (referred to by the Greek letters: Alpha, Beta, Gamma and Delta) (Woese, 1987; Woese et al., 1985b). The Proteobacteria are named after the Greek god Proteus, the god of the sea who is capable of assuming many different shapes (Stackebrandt E. et al., 1988), and not the genus *Proteus* which is a member of the Proteobacteria, but not the nomenclatural type of the phylum. Based on 16S rRNA gene trees, the phylum was later divided into 5 classes, *Alphaproteobacteria*, *Betaproteobacteria*, *Gammaproteobacteria*, *Deltaproteobacteria* and *Epsilonproteobacteria* (Brenner et al., 2005). Subsequent studies based upon 16S rRNA, RecA and GyrB sequences recognized a sixth class "*Zetaproteobacteria*" within phylum Proteobacteria (Emerson et al., 2007). Among these groups *Gammaproteobacteria* is the largest class and has >350 genera (NCBI 2015), accounting for about 46% of all known species of Proteobacteria and 17% of all known bacterial species (Cole et al., 2009). This

class encompasses members that show wide host range and great variety in terms of the phenotype and metabolic capabilities (Woese et al., 1985b). Metabolically they derive their energy through oxidation of sulfur, hydrogen or iron (Gupta, 2000).

The class includes several medically important groups of bacteria such as *Enterobacteriaceae* (including the most studied model organism *E. coli*), *Vibrionaceae* and *Pseudomonadaceae*. In addition, this groups also includes, human, animal and plant pathogens, e.g. *Salmonella* (enteritis and typhoid fever), *Yersinia* (plague), *Vibrio* (cholera), *Pseudomonas aeruginosa* (lung infections in hospitalized or cystic fibrosis patients), and major plant pathogens, e.g. *Xanthomonas* and *Xylella* (Williams et al., 2010; Helgerson et al., 2006). A large number of species belonging to this group reside endosymbiotically in humans, animals and plants (Williams et al., 2010; Brenner et al., 2005). Although this group is the most extensively studied group of bacteria because of its medical, ecological and agricultural importance, the taxonomy and systematics of this group remains problematic. Based on branching in 16S rRNA trees, the class *Gammaproteobacteria* is divided into 14 main orders or groups: *Aeromonadales*, *Alteromonadales*, *Cardiobacteriales*, *Chromatiales*, "Enterobacteriales", *Legionellales*, *Methylococcales*, *Oceanospirillales*, *Orbales*, *Pasteurellales*, *Pseudomonadales* (Type Order), "Salinisphaerales", *Thiotrichales*, "Vibrionales" and *Xanthomonadales* (Brenner et al., 2005; Parte, 2014).

The class *Gammaproteobacteria*, as well as all of its orders are presently identified solely based on their branching in the 16S rRNA or other gene trees, no other reliable biochemical or molecular characteristics are known that are specific for this class

and for its different orders, that can distinguish them from other bacteria. Additionally, the interrelationships among different orders within the class *Gammaproteobacteria* also remain to be determined with certainty. Because of the enormous diversity and the presence of important pathogens in the class *Gammaproteobacteria*, a large number of these bacteria have been sequenced and their complete genomes are deposited to public databases. As of January 2015 >10,000 complete/draft genomes belonging to class *Gammaproteobacteria* are present in the NCBI database (Tatusova et al., 2014).

Research Objective

The objective of my research work is to perform comparative genomic studies on available gammaproteobacterial genomes to discover novel molecular markers, in the form of CSIs and CSPs, to help identify groups (orders) within the class *Gammaproteobacteria* in molecular terms. Major focus of my work has been on the demarcation and taxonomic refinement of *Pasteurellales*, *Xanthomonadales* and "Enterobacteriales". With the help of identified molecular markers, major taxonomic revisions, particularly for *Xanthomonadales* and *Pasteurellales*, have been proposed. Additionally, the reliability of identified molecular markers has also been tested by sequencing new microbes. The impact and prevalence of LGT on prokaryotic genomes has been analyzed by utilizing CSIs and CSPs. It is estimated that these protein based molecular markers, provide strong evidence in favor of the Darwinian mode of prokaryotic evolution.

CHAPTER 2

Conserved Signature Indels and Signature Proteins as Novel Tools for Understanding Microbial Phylogeny and Systematics: Identification of Molecular Signatures that are Specific for the Phytopathogenic Genera *Dickeya*, *Pectobacterium* and *Brenneria*

This Chapter describes the usefulness of molecular markers (CSIs and CSPs) for the identification and classification of prokaryotic taxa. The phylogenetic trees were constructed along with the identification of CSIs and CSPs to understand phylogenetic relationships among *Dickeya, Pectobacterium* and *Brenneria,* the phytopathogens within Enterobacteriales. My contributions towards the completion of this chapter included the identification of CSIs and CSPs and construction of phylogenetic trees highlighted in the methods section. I was also involved in writing of the manuscript, including the figures and tables provided.

*Due to limited space, supplementary figures (1-23) and tables are not included in the chapter but can be accessed along with the rest of the manuscript at:

Naushad,H.S., Lee,B., and Gupta,R.S. (2014). Conserved signature indels and signature proteins as novel tools for understanding microbial phylogeny and systematics: identification of molecular signatures that are specific for the phytopathogenic genera Dickeya, Pectobacterium and Brenneria. Int J Syst. Evol. Microbiol 64, 366-383

International Journal of Systematic and Evolutionary Microbiology (2014), **64**, 366–383 **DOI** 10.1099/ijs.0.054213-0

Conserved signature indels and signature proteins as novel tools for understanding microbial phylogeny and systematics: identification of molecular signatures that are specific for the phytopathogenic genera *Dickeya*, *Pectobacterium* and *Brenneria*

Hafiz Sohail Naushad, Brian Lee and Radhey S. Gupta

Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, Ontario, Canada L8N 3Z5

Correspondence
Radhey S. Gupta
gupta@mcmaster.ca

Genome sequences are enabling applications of different approaches to more clearly understand microbial phylogeny and systematics. Two of these approaches involve identification of conserved signature indels (CSIs) and conserved signature proteins (CSPs) that are specific for different lineages. These molecular markers provide novel and more definitive means for demarcation of prokaryotic taxa and for identification of species from these groups. Genome sequences are also enabling determination of phylogenetic relationships among species based upon sequences for multiple proteins. In this work, we have used all of these approaches for studying the phytopathogenic bacteria belonging to the genera *Dickeya*, *Pectobacterium* and *Brenneria*. Members of these genera, which cause numerous diseases in important food crops and ornamental plants, are presently distinguished mainly on the basis of their branching in phylogenetic trees. No biochemical or molecular characteristic is known that is uniquely shared by species from these genera. Hence, detailed studies using the above approaches were carried out on proteins from the genomes of these bacteria to identify molecular markers that are specific for them. In phylogenetic trees based upon concatenated sequences for 23 conserved proteins, members of the genera *Dickeya*, *Pectobacterium* and *Brenneria* formed a strongly supported clade within the other *Enterobacteriales*. Comparative analysis of protein sequences from the *Dickeya*, *Pectobacterium* and *Brenneria* genomes has identified 10 CSIs and five CSPs that are either uniquely or largely found in all genome-sequenced species from these genera, but not present in any other bacteria in the database. In addition, our analyses have identified 10 CSIs and 17 CSPs that are specifically present in either all or most sequenced *Dickeya* species/strains, and six CSIs and 19 CSPs that are uniquely found in the sequenced *Pectobacterium* genomes. Finally, our analysis also identified three CSIs and one CSP that are specifically shared by members of the genera *Pectobacterium* and *Brenneria*, but absent in species of the genus *Dickeya*, indicating that the former two genera shared a common ancestor exclusive of *Dickeya*. The identified CSIs and CSPs provide novel tools for identification of members of the genera *Dickeya* and *Pectobacterium* and for delimiting these taxa in molecular terms. Descriptions of the genera *Dickeya* and *Pectobacterium* have been revised to provide information for these molecular markers. Biochemical studies on these CSIs and CSPs, which are specific for these genera, may lead to discovery of novel properties that are unique to these bacteria and which could be targeted to develop antibacterial agents that are specific for these plant-pathogenic bacteria.

INTRODUCTION

Phylogenetic and similarity studies based upon 16S rRNA genes have greatly advanced our understanding of the evolutionary relationships among prokaryotic organisms (Olsen & Woese, 1993; Yarza *et al.*, 2010). However, one

Abbreviations: CSIs, conserved signature indels; CSPs, conserved signature proteins; ML, maximum-likelihood; NJ, neighbour-joining.

One supplementary table and 23 supplementary figures are available with the online version of this paper.

central issue in microbial systematics that remains ill-defined concerns the methods used for identification and demarcation of prokaryotic taxa (Ludwig & Klenk, 2005; Stackebrandt, 2006; Oren, 2010). Except for a limited few, most prokaryotic taxa at different phylogenetic levels are currently identified solely on the basis of their branching in the 16S rRNA (gene) trees. For most taxa, no unique biochemical, molecular or other characteristics are known that are specific for them and could be used to distinguish them from all others. Because the branching pattern of the species in phylogenetic trees is influenced by large numbers of variables, demarcation of prokaryotic taxa based upon 'clustering in phylogenetic trees' is imprecise and constitutes only a 'statistical definition' (Ludwig & Klenk, 2005; Naum *et al.*, 2011). Additionally, defining prokaryotic taxa based upon their branching in phylogenetic trees provides no indication as to what properties might be commonly shared by different members of a given clade, and it suggests no experimental approaches to discover such characteristics. According to Woese (1998), a good classification scheme should have the following characteristics: 'A biological classification is in effect an overarching evolutionary theory that guides our thinking and experimentation, ...' Hence, to develop a more complete understanding of microbial phylogeny and systematics, it is necessary to discover other reliable markers or characters for different prokaryotic taxa that can supplement our current understanding of them (Gupta & Griffiths, 2002; Gupta, 2010; Gao & Gupta, 2012a).

Genome sequences provide a valuable resource for discovery of molecular markers that can be used for reliable classification of prokaryotic taxa and for understanding evolutionary relationships among them (Lerat *et al.*, 2005; Dutilh *et al.*, 2008; Gupta, 2010; Bhandari *et al.*, 2012; Gao & Gupta, 2012a). Conserved signature indels (CSIs) and conserved signature proteins (CSPs) represent two different types of molecular markers that are of great value in these regards. These markers have been extensively utilized in our work (Gupta, 2010; Bhandari *et al.*, 2012; Gao & Gupta, 2012a) and some of their important characteristics, which make them particularly useful for classification purposes, will be described later in this work. We report here the application of these markers and phylogenomic approaches for identification and classification of the plant-pathogenic bacteria belonging to the genera *Dickeya*, *Pectobacterium* and *Brenneria*.

Dickeya, *Pectobacterium* and *Brenneria* are important genera of plant-pathogenic bacteria belonging to the family *Enterobacteriaceae* (Hauben *et al.*, 2005; Samson *et al.*, 2005; Ma *et al.*, 2007). Of these bacteria, members of the genera *Pectobacterium* and *Dickeya* are considered broad-host-range, soft-rotting plant pathogens and affect many food crops including potato, tomato, onions, sugar beet, maize, pineapple, banana and sunflower, and many ornamental plants (Hauben *et al.*, 2005; Samson *et al.*, 2005; Charkowski, 2006; Ma *et al.*, 2007; Yishay *et al.*, 2008; Czajkowski *et al.*, 2011; Toth *et al.*, 2011; Costechareyre *et al.*, 2012). The taxonomy of the phytopathogenic bacteria belonging to the family *Enterobacteriaceae* has undergone much revision over the years (Hauben *et al.*, 1998, 2005; Samson *et al.*, 2005; Naum *et al.*, 2011; Denman *et al.*, 2012; Bull *et al.*, 2012). Initially, all Gram-negative, rod-shaped, non-spore-forming and peritrichous-flagellated plant pathogens were part of the genus *Erwinia* (Winslow *et al.*, 1917). However, this idea was not supported by later work as diverse pytopathogens were found in *Enterobacteriaceae*, such as the genus *Pantoea*, and *Enterobacter dissolvens* and *Brenneria salicis*, with varying biovars, morphovars and serovars (Dye, 1968; Brenner *et al.*, 1986; Gavini *et al.*, 1989; Hauben *et al.*, 1998). The detailed work conducted by Hauben *et al.* (1998) led to the reclassification of many species from the genus *Erwinia* (Winslow *et al.*, 1917) to the genus *Pectobacterium* (Waldee, 1945; Skerman *et al.*, 1980). This work also suggested division of *Pectobacterium carotovorum* into five subspecies, three of which were later elevated to species level (Gardan *et al.*, 2003). Subsequently, *Pectobacterium chrysanthemi* (formerly *Erwinia chrysanthemi*) and *Brenneria paradisiaca*, based upon their distinct branching in the 16S rRNA tree, were transferred to a new genus *Dickeya* (Samson *et al.*, 2005), and several additional members of this genus were also identified (Samson *et al.*, 2005; Parkinson *et al.*, 2009; Van Vaerenbergh *et al.*, 2012). The most recent development in this regard involves the transfer of *Brenneria quercina* to a new genus *Lonsdalea* and its division into three subspecies as well as the reclassification of *Dickeya dieffenbachiae* as *Dickeya dadantii* subsp. *dieffenbachiae* (Brady *et al.*, 2012).

Earlier phylogenetic studies based on individual gene/protein sequences indicate that the members of the genera *Pectobacterium*, *Dickeya* and *Brenneria* form a distinct clade within the family *Enterobacteriaceae* (Hauben *et al.*, 1998, 2005; Spröer *et al.*, 1999; Samson *et al.*, 2005; Ma *et al.*, 2007; Naum *et al.*, 2008, 2011; Parkinson *et al.*, 2009). However, the branching order of these genera has been found to be variable in these studies and it is not reliably resolved (Hauben *et al.*, 1998, 2005; Spröer *et al.*, 1999; Brown *et al.*, 2000; Samson *et al.*, 2005; Ma *et al.*, 2007; Young & Park, 2007; Naum *et al.*, 2008, 2011).

The members of the above three phytopathogenic genera are currently distinguished from each other as well as other groups of bacteria primarily on the basis of their branching in 16S rRNA trees (Hauben *et al.*, 2005; Samson *et al.*, 2005; Naum *et al.*, 2011). No molecular or biochemical property is known that is uniquely shared by species from these genera. Because these genera harbour many important plant pathogens, identification of molecular markers that are specific for members of these genera should prove very helpful in their reliable identification and classification. Currently, genome sequences are available for 13 species/strains from these three genera (Table 1). In addition, draft genome sequences for many *Dickeya* species/strains (>25) are also available in a number of databases (NCBI and EzGenome/EzBiocloud). In the present study,

Table 1. Some characteristics of the genomes of *Dickeya*, *Pectobacterium* and *Brenneria*

Organism	Reference sequence	Size (mbp)	No. of proteins	DNA G+C content (mol%)	Reference
Dickeya chrysanthemi Ech1591*	NC_012912.14367	4.8	4163	55	DOE-JGI†
Dickeya dadantii subsp. *dadantii* 3937	NC_014500.14687	4.9	4549	56	Glasner *et al.* (2011)
Dickeya dadantii Ech703*	NC_012880.14136	4.7	3970	55	DOE-JGI†
Dickeya zeae Ech586*	NC_013592.14318	4.8	4144	54	DOE-JGI†
Pectobacterium atrosepticum SCRI1043	NC_004547.2	5.1	4472	51	Bell *et al.* (2004)
Pectobacterium wasabiae WPP163	NC_013421.1	5.1	4437	51	DOE-JGI†
Pectobacterium wasabiae CFBP 3304	Draft genome	5.1	4636	51	Nykyri *et al.* (2012)
Pectobacterium carotovorum subsp. carotovorum PC1	NC_012917.1	4.9	4246	52	DOE-JGI†
Pectobacterium carotovorum subsp. carotovorum PCC21	NC_018525.1	4.8	4263	52	Park *et al.* (2012)
'*Pectobacterium carotovorum* subsp. *brasiliensis*' PBR1692	NZ_ABVX00000000	4.9	4836	52	(Glasner *et al.*, 2008)
Pectobacterium carotovorum subsp. carotovorum WPP14	NZ_ABVY00000000	4.8	4540	52	Glasner *et al.* (2008)
Pectobacterium sp. SCC3193	NC_017845.1	5.2	4705	50	Koskinen *et al.* (2012)
Brenneria sp. EniD312	NZ_CM001230.1	4.9	4388	56	DOE-JGI†

*The names of the following *Dickeya* species/strains, namely *D. dadantii* Ech586, *D. dadantii* Ech703 and *D. zeae* Ech1591 (as noted in the NCBI database) have recently been changed to *D. zeae* Ech586, *D. paradisiaca* Ech703 and *D. chrysanthemi* Ech1591, respectively (Marrero *et al.*, 2013). The names used in this manuscript reflect the new classification.
†DOE-JGI, US Department of Energy (DOE) Joint Genome Institute (JGI).

we have carried out detailed analysis of protein sequences from these genomes to identify molecular markers (i.e. CSIs and CSPs) that are either specific for members of the genera *Dickeya* and *Pectobacterium* or commonly shared by these genera and *Brenneria*, providing information regarding evolutionary relationships among them. We also describe multiple CSIs and CSPs that are uniquely shared by the members of these three genera defining a distinct clade of phytopathogenic bacteria.

METHODS

Reconstruction of phylogenetic trees. Phylogenetic trees were reconstructed based upon concatenated sequence alignment for 23 housekeeping and ribosomal proteins (namely GroEL, Gyrase A, Gyrase B, IleRS, MetRS, DnaK, EF-G, EF-P, ProRS, RpoA, RpoB, RpoC, SecY, GlyA, LeuRS, SerRS, ValRS, rRNA dimethyladenosine transferase, 30S ribosomal proteins S2, S8 and S9, and 50S ribosomal proteins L4 and L5). These proteins have been extensively used in phylogenetic studies (Ciccarelli *et al.*, 2006; Gupta, 2009; Wu *et al.*, 2009; Gao *et al.*, 2009a; Naushad & Gupta, 2013). Sequences for these proteins for various *Dickeya*, *Pectobacterium* and *Brenneria* species and different *Enterobacteriales* were retrieved from the NCBI (http://www.ncbi.nlm.nih.gov/) and EzGenome/EzBiocloud (EzGenome/EzBiocloud.net) databases. Multiple sequence alignments for individual proteins were created using CLUSTAL X 2.0 (Larkin *et al.*, 2007). After concatenation of these alignments into a single file, poorly aligned regions were removed using Gblocks 0.91 b (Castresana, 2000), leaving a total of 12 986 positions in the final dataset, which was used for reconstructing phylogenetic trees. The maximum-likelihood (ML) and neighbour-joining (NJ) trees based on 100 bootstrap replicates of this alignment were reconstructed using MEGA 5.1 (Tamura *et al.*, 2007) employing the Whelan and Goldman (Whelan & Goldman, 2001) and Jones–Taylor–Thornton (Jones *et al.*, 1992) substitution models, respectively.

Identification of CSIs. To identify CSIs that might be specific for *Dickeya*, *Pectobacterium* and *Brenneria*, BLASTP searches were carried out on each protein/ORF from the genome of *Dickeya zeae* Ech586 (indicated as *Dickeya dadantii* Ech586 in the NCBI database) (Table 1). For proteins bearing high scoring homologues (E values $<1e^{-20}$) in at least three to four species from the above genera and a number of other bacterial groups, such sequences were retrieved and multiple sequence alignments were reconstructed using CLUSTAL X 2.0 (Larkin *et al.*, 2007). These alignments were visually inspected for the presence of signature indels (insertions or deletions) that were restricted to *Dickeya*, *Pectobacterium* and *Brenneria*, and which were flanked on both sides by at least five to six conserved residues in the neighbouring 30–40 amino acids. Those indels that were not flanked by conserved regions were excluded as they do not provide useful molecular markers (Gupta, 1998, 2010; Gao & Gupta, 2012a). The species distribution of the indels thus identified was further examined by performing BLASTP searches on the NCBI non-redundant database on short sequence segments containing the indels and their flanking conserved regions. We report here the results of only those CSIs which are specific for species of the genera *Dickeya*, *Pectobacterium* and *Brenneria*, and where similar CSIs were absent in other bacteria in the top 250 BLAST hits. For a number of *Dickeya* species, for which

only draft genomes are available, the sequence information for different CSIs was obtained by downloading the genome sequences from the EzGenome/EzBiocloud database and then performing local TBLASTN searches against the indel queries. We present here sequence information for all sequenced species from the above three genera but only a limited number of representatives from other groups. For species of the genus *Dickeya* for which information was available from multiple strains, only the sequences for the type strains of the species are shown and unless otherwise indicated similar indels were present in all other strains. Note that the names of a number of *Dickeya* species/strains (namely *D. dadantii* Ech586, *D. dadantii* Ech703 and *D. zeae* Ech1591) for which sequence information was obtained from the NCBI database have recently been revised (Marrero *et al.*, 2013). The revised names of these species/strains, respectively, are *D. zeae* Ech586, *Dickeya paradisiaca* Ech703 and *Dickeya chrysanthemi* Ech1591. In the present work, we have used the revised nomenclatures of these species/strains rather than those indicated in the NCBI database. In addition, the following *Dickeya* species/strains (namely *Dickeya* sp. GBBC 2040, *Dickeya* sp. IPO 2222, *Dickeya* sp. MK10 and *Dickeya* sp. MK16) in the NCBI database are referred to as different strains of '*Dickeya solani*' (Pritchard *et al.*, 2013b). In the present work, we have used the original names of these species/strains, as '*Dickeya solani*' is, at the time of writing, not a validly published name. The differences in the names of *Dickeya* species/strains used in the present work and the corresponding names in the NCBI database are listed in Table S1, available in the online Supplementary Material.

Identification of signature proteins that are specific for *Dickeya*, *Pectobacterium* and *Brenneria*. These studies were carried out as described in our earlier work (Gao & Gupta, 2007; Gupta & Mok, 2007; Gupta & Mathews, 2010). Searches using the program BLASTP were carried out on individual proteins in the genomes of *D. zeae* Ech586 and *Pectobacterium wasabiae* WPP163. These searches were performed against the NCBI non-redundant database using default parameters and without the low-complexity filter (Altschul *et al.*, 1997). Proteins of interest were those where either all significant hits were from the members of these genera, or which involved a large increase in E values from the last hit belonging to these genera and the first hit from any other bacteria, and the E values for the latter were $>1e^{-3}$ (Gao & Gupta, 2007; Gupta & Mok, 2007; Gupta & Mathews, 2010). However, higher E values can be significant for smaller proteins, and hence the lengths of the query proteins and those of the observed hits were taken into consideration when analysing the results of these studies. For most of the CSPs identified in this work, the lengths of the observed hits were very similar to those of the query proteins. The proteins which were exclusively found only in a single species or strain are not reported here. For all identified signature proteins, their accession numbers, protein lengths and any information regarding cellular functions are presented.

RESULTS

Phylogenetic analysis based upon protein sequences

The evolutionary relationships among members of the genera *Dickeya*, *Pectobacterium* and *Brenneria* has thus far been studied mainly on the basis of either individual or a limited number of gene/protein sequences and have been found to be variable (Samson *et al.*, 2005; Ma *et al.*, 2007; Young & Park, 2007; Parkinson *et al.*, 2009; Naum *et al.*, 2011; Denman *et al.*, 2012). Sequences of several complete or draft genomes for species/strains belonging to the genera *Dickeya*, *Pectobacterium* and *Brenneria* are now available in the NCBI and EzGenome/EzBiocloud databases (Table 1, Table S1) (Glasner *et al.*, 2008, 2011; Koskinen *et al.*, 2012; Nykyri *et al.*, 2012; Pritchard *et al.*, 2013a, b). These genome sequences provide means to examine the relationship among these genera by reconstructing phylogenetic trees based upon concatenated sequences for multiple proteins. Such trees have proven more reliable in resolving the interrelationships among different taxa (Ciccarelli *et al.*, 2006; Wu *et al.*, 2009; Gao *et al.*, 2009a; Williams *et al.*, 2010; Gao & Gupta, 2012a). We have reconstructed phylogenetic trees for *Dickeya*, *Pectobacterium* and *Brenneria* species and other genome-sequenced species from the order *Enterobacteriales* based on concatenated sequences for 23 conserved proteins (see Methods). A bootstrapped ML tree based on these sequences is shown in Fig. 1. An NJ tree based on the same dataset is provided in Fig. S1. In both the ML and the NJ trees, the species from the order *Enterobacteriales* formed a number of distinct clades. One of these well-resolved clades consisted of species belonging to the genera *Dickeya*, *Pectobacterium* and *Brenneria*. This clade was strongly supported by both the NJ and the ML algorithms. Within this clade, species/strains from the genera *Dickeya* and *Pectobacterium* also formed distinct clades. For *Dickeya*, sequence information is now available for a large number of species/strain (see Table 1, Table S1) and the different strains of *D. dadantii*, *Dickeya dianthicola*, *D. chrysanthemi* and *D. zeae* as well as '*Dickeya* solani' (namely *Dickeya* sp. GBBC 2040, *Dickeya* sp. IPO 2222, *Dickeya* sp. MK10 and *Dickeya* sp. MK16) were found to form monophyletic clusters. *D. paradisiaca* formed the deepest branch among *Dickeya* species and it was separated from other species by a long branch. Marrero *et al.* (2013), on the basis of phylogenetic analysis of partial sequences of the *dnaA*, *dnaJ*, *dnaX*, *gybB* and *recN* genes, have proposed that *D. dadantii* Ech586, *D. dadantii* Ech703 and *D. zeae* Ech1591 should be reclassified as *D. zeae* Ech586, *D. paradisiaca* Ech703 and *D. chrysanthemi* Ech1591, respectively. The branching pattern observed in our tree supports their proposed reclassification. In addition to the above relationships, the tree based upon concatenated protein sequences also strongly supported a grouping of *Brenneria* sp. EniD312 with the *Pectobacterium* clade. Species from the other phytopathogenic genera, namely *Erwinia*, *Pantoea* and *Enterobacter*, were part of two other clades, which showed no relationship to the *Dickeya–Pectobacterium–Brenneria* clade. This tree provides a phylogenetic framework for the results obtained using other comparative genomic approaches that are described below.

Identification of CSIs that are specific for the genera *Dickeya*, *Pectobacterium* and *Brenneria*

As noted earlier, conserved inserts and deletions (i.e. indels or CSIs) in genes/proteins provide an important category of molecular markers, whose discovery is facilitated by genome sequences, which are proving very useful for systematic and taxonomic studies. The indels that are

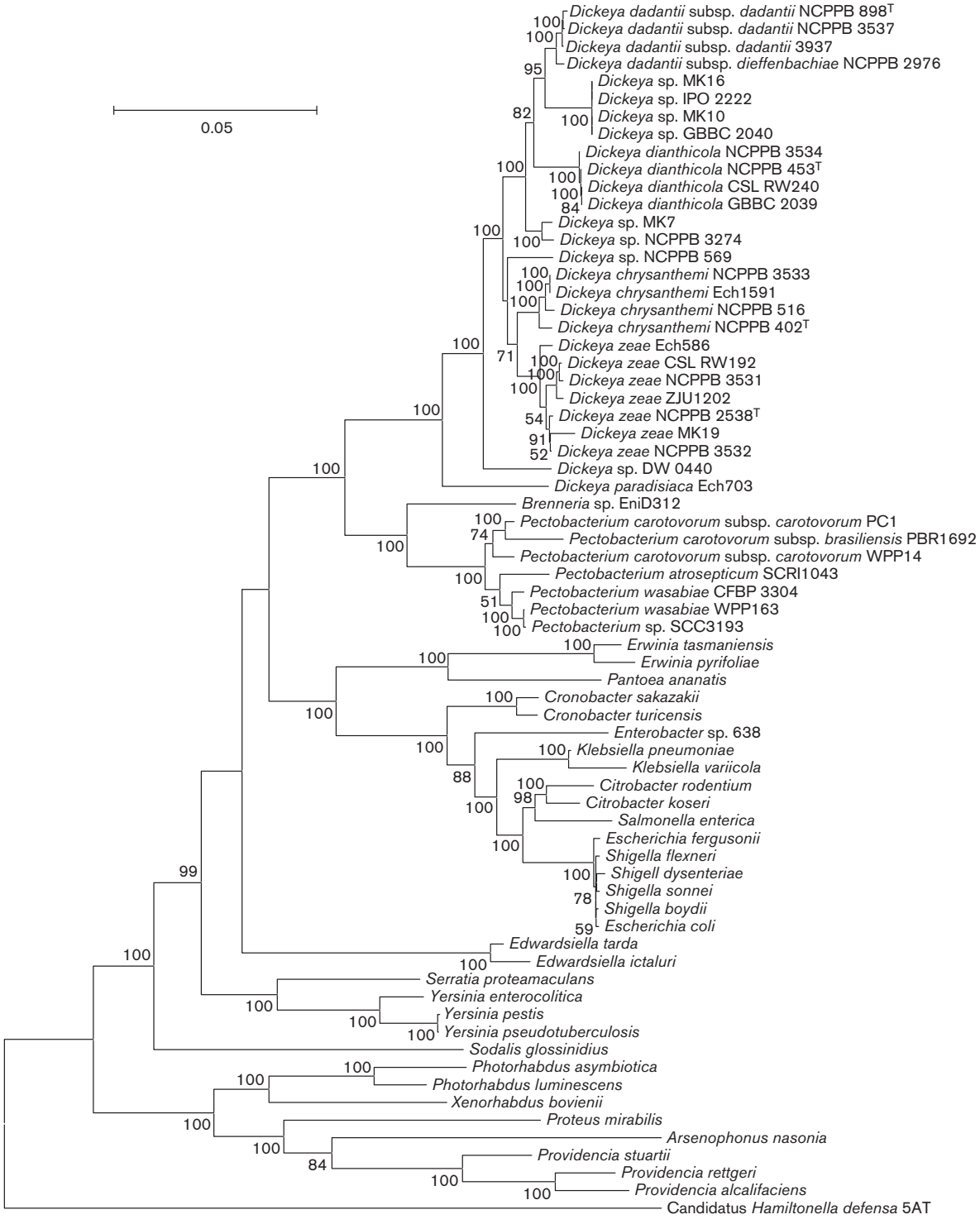


Fig. 1. ML distance tree for the genome-sequenced *Enterobacteriales* species based upon concatenated sequences for 23 conserved proteins. Bootstrap scores for different nodes are shown at branch points. The tree was rooted using sequences from *Xanthomonadales* species (i.e. *Xanthomonas* and *Xylella*). Bar, 0.05 changes per position. An NJ tree for this dataset is shown in Fig. S1.

useful phylogenetic markers are of defined size and are flanked on both sides by conserved regions to ensure that they constitute reliable characteristics (Gupta, 1998, 2000; Gupta & Griffiths, 2002; Ajawatanawong & Baldauf, 2013). Because of the highly specific nature of genetic changes that give rise to conserved indels, such changes are less likely to arise independently in different taxa by convergent or parallel evolution (Gupta, 1998; Rokas & Holland, 2000). Hence, when a CSI of defined size is uniquely found in a phylogenetically defined group(s) of species, the simplest explanation is that the genetic change responsible for it occurred once in a common ancestor of this group that then passed on to various descendants. Furthermore, depending upon the presence or absence of a given CSI in the outgroup species, it can be inferred whether a given indel is an insert or a deletion and based upon this information a rooted relationship among the species can be derived, independently of phylogenetic trees (Gupta, 1998; Griffiths & Gupta, 2004). Because genetic changes leading to CSIs can occur at various stages in evolution, the CSIs that are specific for taxonomic clades at different phylogenetic depths (phylum, order, family, genus or species) can be identified (Gao & Gupta, 2012b; Bhandari & Gupta, 2013). Lastly, the shared presence of CSIs in unrelated taxa also provides a means for identifying lateral gene transfer events (Griffiths & Gupta, 2006).

Detailed investigations on identification of CSIs carried out in this work have identified large numbers of CSIs that are specific for species of the genera *Dickeya*, *Pectobacterium* and *Brenneria* at multiple phylogenetic levels. Ten of these CSIs are uniquely found in all of the sequenced species/strains of *Dickeya* and two examples of them are shown in Fig. 2. In these examples, a 2 aa insert in the enzyme adenosine deaminase (Fig. 2a) and a 2 aa deletion in the multidrug resistance protein MdtA (Fig. 2b) are specifically found in all sequenced *Dickeya* species/strains, but they are not present in the homologues from any other bacteria, including *Pectobacterium* and *Brenneria*. Both these CSIs are located within conserved regions of the proteins, indicating that they constitute reliable molecular markers. Information for eight other CSIs in different proteins, which are also specific for the genus *Dickeya*, is summarized in Table 2 and their sequences are presented in Figs S2–S9. Note that our initial analysis, which identified these CSIs, was based only on a limited number of *Dickeya* species, for which complete genomes were available in the NCBI database (Table 1). The presence of these CSIs in the draft genomes of other *Dickeya* species (listed in Table S1) was examined during revision of the manuscript. The presence of these CSIs in all of the other sequenced *Dickeya* species strongly indicates that they provide useful molecular markers, with predictive ability, for identification of members of the genus *Dickeya*.

Similar to *Dickeya*, our work has identified six CSIs that are uniquely shared by all of the sequenced species/strains from the genus *Pectobacterium*. Two examples of these CSIs each involving independent 5 aa inserts in the glycine cleavage system T protein and the urea amidolyase related protein are shown in Fig. 3. Both these CSIs, which are present in conserved regions, are specific for the genus *Pectobacterium* and they are not found in the homologues from any other bacteria, including those from the genera *Dickeya* and *Brenneria*. Two other CSIs in the proteins glycerol-3-phosphate dehydrogenase subunit A and sigma E regulatory protein MucB/RseB (Figs S10 and S11) are also specifically present in all of the sequenced species/strains of *Pectobacterium*. For two additional CSIs, in the proteins RecJ and phosphoribosyl-formyl-glycinamidine synthase (Figs S12 and S13), which are also specific for *Pectobacterium*, homologues were not detected in *Pectobacterium atrosepticum* SCRI1043, which is the only species of the genus *Pectobacterium* that is restricted to a single host (potato) (Ma *et al.*, 2007). Some characteristics of the CSIs that are specific for the genus *Pectobacterium* are summarized in Table 3.

In the phylogenetic tree based upon concatenated protein sequences, *Brenneria* sp. EniD312 was found to form an outgroup of the *Pectobacterium* clade and this branching was statistically strongly supported (100 % bootstrap score). Three CSIs identified in this work independently support that the genera *Brenneria* (i.e. *Brenneria* sp. EniD312) and *Pectobacterium* shared a common ancestor exclusive of the genus *Dickeya*. One CSI depicting this relationship is presented in Fig. 4. In the periplasmic serine protease DegS, a 7 aa insert in a conserved region is commonly shared by species of the genera *Brenneria* and *Pectobacterium*, but it is absent in species of the genus *Dickeya* and other bacteria. Two other CSIs that are also specific for species from the genera *Brenneria* and *Pectobacterium* are found in the proteins 6, 7-dimethyl-8-ribityllumazin synthase and GCN5-like *N*-acetyltransferase (Figs S14 and S15). The main characteristics of these CSIs are also summarized in Table 3.

Our analyses have also identified 10 CSIs that in most cases are specifically shared by all of the sequenced species from these three genera. Two examples of these CSIs, which are found in the proteins phosphoglycerate mutase and seryl-tRNA synthetase, are shown in Fig. 5. The sequence information for other CSIs that are commonly shared by species from these genera is provided in Figs S16–S23 and some of their characteristics are summarized in Table 4. These CSIs provide strong evidence that species from these three genera shared a common ancestor exclusive of other bacteria.

Identification of CSPs that are specific for members of the genera *Dickeya*, *Pectobacterium* and *Brenneria*

CSPs, which are uniquely found in particular groups of organisms, provide another important category of molecular markers that are useful for systematic and evolutionary studies (Lerat *et al.*, 2005; Gao & Gupta, 2007; Gupta & Mok, 2007; Dutilh *et al.*, 2008; Gupta & Mathews, 2010;

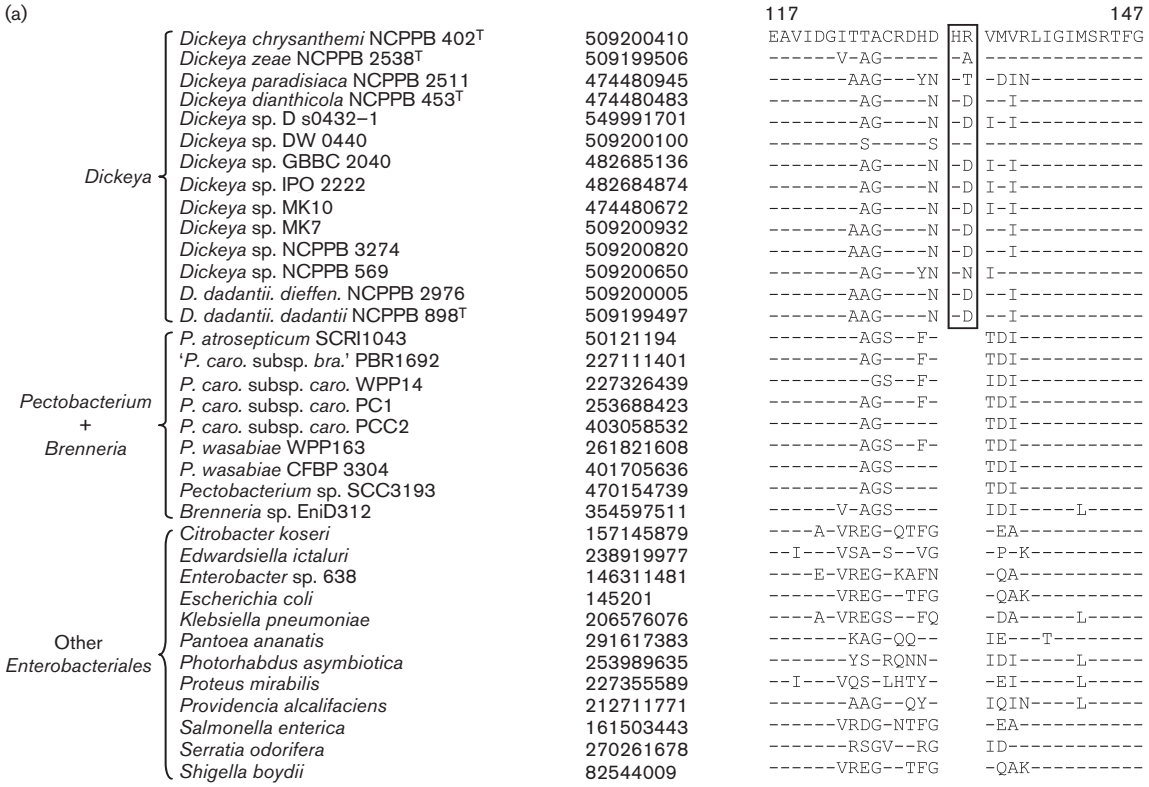

(a)
117
147
Dickeya chrysanthemi NCPPB 402T 509200410 EAVIDGITTACRDHD HR VMVRLIGIMSRTFG
Dickeya zeae NCPPB 2538T 509199506 ------V-AG----- -A ----------
Dickeya paradisiaca NCPPB 2511 474480945 -------AAG---YN -T -DIN------
Dickeya dianthicola NCPPB 453T 474480483 --------AG----N -D --I------
Dickeya sp. D s0432–1 549991701 --------AG----N -D I-I------
Dickeya sp. DW 0440 509200100 --------S-----S -- ----------
Dickeya sp. GBBC 2040 482685136 --------AG----N -D I-I------
Dickeya sp. IPO 2222 482684874 --------AG----N -D I-I------
Dickeya sp. MK10 474480672 --------AG----N -D I-I------
Dickeya sp. MK7 509200932 -------AAG----N -D --I------
Dickeya sp. NCPPB 3274 509200820 -------AAG----N -D --I------
Dickeya sp. NCPPB 569 509200650 --------AG---YN -N I------
D. dadantii. dieffen. NCPPB 2976 509200005 -------AAG----N -D --I------
D. dadantii. dadantii NCPPB 898T 509199497 -------AAG----N -D --I------
Dickeya
P. atrosepticum SCRI1043 50121194 --------AGS--F- TDI------
'P. caro. subsp. bra.' PBR1692 227111401 --------AG---F- TDI------
P. caro. subsp. caro. WPP14 227326439 ---------GS--F- IDI------
P. caro. subsp. caro. PC1 253688423 --------AG---F- TDI------
P. caro. subsp. caro. PCC2 403058532 --------AG----- TDI------
P. wasabiae WPP163 261821608 --------AGS--F- TDI------
P. wasabiae CFBP 3304 401705636 --------AGS---- TDI------
Pectobacterium sp. SCC3193 470154739 --------AGS---- TDI------
Brenneria sp. EniD312 354597511 ------V-AGS---- IDI-----L-----
Pectobacterium + Brenneria
Citrobacter koseri 157145879 ----A-VREG-QTFG -EA------
Edwardsiella ictaluri 238919977 --I---VSA-S--VG -P-K------
Enterobacter sp. 638 146311481 ----E-VREG-KAFN -QA------
Escherichia coli 145201 ------VREG--TFG -QAK------
Klebsiella pneumoniae 206576076 ----A-VREGS--FQ -DA-----L-----
Pantoea ananatis 291617383 -------KAG-QQ-- IE---T--------
Photorhabdus asymbiotica 253989635 -------YS-RQNN- IDI-----L-----
Proteus mirabilis 227355589 --I---VQS-LHTY- -EI-----L-----
Providencia alcalifaciens 212711771 -------AAG--QY- IQIN----L-----
Salmonella enterica 161503443 ------VRDG-NTFG -EA------
Serratia odorifera 270261678 -------RSGV--RG ID------
Shigella boydii 82544009 ------VREG--TFG -QAK------
Other Enterobacteriales
(b)
206
235
Dickeya chrysanthemi NCPPB 402T 509200451 GRVGLRQIDIGNYVTS TENLLVLTQTHPID
Dickeya zeae NCPPB 2538T 509199482 ---------- -D------
Dickeya dianthicola NCPPB 453T 474480454 ---------- -D------
Dickeya paradisiaca NCPPB 2511 474480936 -------V-V---I-- -TT------
Dickeya sp. D s0432–1 549989176 ---------- -D------
Dickeya sp. DW 0440 509200011 ---------- -D------
Dickeya sp. GBBC 2040 482685061 ---------- -D------
Dickeya sp. IPO 2222 482684846 ---------- -D------
Dickeya sp. MK10 474480620 ---------- -D------
Dickeya sp. MK7 509200910 ---------- -D------
Dickeya sp. NCPPB 3274 509200758 ---------- -D------
Dickeya sp. NCPPB 569 509200607 ---------- -D------
D. dadantii. dieffen. NCPPB 2976 509199927 ---------- -D------
D. dadantii. dadantii. NCPPB 898T 509199464 ---------- -D------
Dickeya
P. atrosepticum SCRI1043 50122106 --I--K-V-V---I-- GD -NGIV-I---Y---
'P. caro. subsp. bra.' PBR1692 227115074 --I--K-V-V---I-- GD -NGIV-I---Y---
P. caro. subsp. caro. WPP14 227329098 --I--K-V-V---I-- GD -NGIV-I---Y---
P. caro. subsp. caro. PC1 253689335 --I--K-V-V---I-- GD -NGIV-I---Y---
P. caro. subsp. caro. PCC21 403059444 --I--K-V-V---I-- GD -N-IV-----Y---
P. wasabiae WPP163 261820619 --I--K-V-V---I-- GD -NGIV-I---Y---
P. wasabiae CFBP 3304 401705164 --I--K-V-V---I-- GD -NGIV-I---Y---
Pectobacterium sp. SCC3193 470153689 --I--K-V-V---I-- GD -NGIV-I---Y---
Brenneria sp. EniD312 354596685 -----K-V-V---I-- GD -NG-V-I---Q---
Pectobacterium + Brenneria
Citrobacter koseri 157144977 -----K-V-V--QIS- GD -TGIV-I-------
Cronobacter turicensis 260598562 -----K-V----QIS- GD -TGIV---------
Edwardsiella ictaluri 238919101 --A----V-E---IS- AD -NG-V---------
Enterobacter cancerogenus 261340517 -----K-V----QIS- GD -TGIV-I-------
Erwinia amylovora 291199737 -----K-V-V----S- GD ANGIV-IA------
Escherichia coli 284922068 -----K-V-V--QIS- GD -TGIV-I-------
Escherichia fergusonii 218549491 -----K-V-V--QIS- GD -TGIV-I-------
Klebsiella pneumoniae 206580406 -----K-V----QIS- GD -TGIV---------
Pantoea ananatis 291153098 -----K-V-----I-- GD -NG-V-I-----V-
Salmonella enterica 161502717 -----K-V----QIS- GD -AGIV-I-------
Other Enterobacteriales

Fig. 2. Excerpts from the sequence alignments of adenosine deaminase (a) and multidrug resistance protein MdtA (b) showing two CSIs (boxed) that are commonly shared by all detected species/strains of the genus *Dickeya*. The dashes (–) in these as well as all other alignments indicate identity with the amino acid on the top line. The numbers on the top line represent the region of protein containing CSIs. The second column indicates GenBank identification numbers for the sequences. For the species, whose annotated genomes were not available (see Table S1), the numbers shown are for the contigs where these sequences are present. Due to space considerations, sequence information is shown for only a limited number of other bacteria. However, unless otherwise noted, all of the reported CSIs are specific for the indicated groups and similar CSIs were not observed in the top 250 BLASTP hits with the query sequences. Information for other *Dickeya*-specific CSIs is provided in Table 2 and Figs S2–S9. Abbreviations used in the species names are: *P.*, *Pectobacterium*; *P. caro.* subsp. *bra.*, '*Pectobacterium carotovorum* subsp. *brasiliensis*'; *P. caro.* subsp. *caro.*, *Pectobacterium carotovorum* subsp. *carotovorum*; *D. dadantii* subsp. *dieffen.* NCPPB 2976, *Dickeya dadantii* subsp. *dieffenbachiae* NCPPB 2976; *D. dadantii* subsp. *dadantii* NCPPB 898, *Dickeya dadantii* subsp. *dadantii* NCPPB 898.

Gao & Gupta, 2012b). Because of their unique shared presence by species from specific clades, the genes for these proteins probably first originated in the common ancestors of these groups and were then retained by all their descendants. Hence, these genes/proteins represent another distinct type of synapomorphic characters that are useful for identifying different groups of organisms in molecular terms and for understanding evolutionary relationships (Dutilh *et al.*, 2008; Gupta & Mathews, 2010; Gupta & Gao, 2010; Gao & Gupta, 2012b). The work on identification of CSPs, which was carried out as described in Methods, has identified five CSPs that are uniquely found in different sequenced members of the genera *Dickeya*, *Pectobacterium* and *Brenneria*, providing molecular markers for this group of phytopathogenic bacteria (Table 5). In addition, these studies have identified six CSPs whose homologues are uniquely found in all sequenced *Dickeya* species (Table 6). Eleven other CSPs are also specific for members of the genus *Dickeya*, except that their homologues were not detected in the deeper branching *D. paradisiaca* (Table 6). It is possible that the genes for these CSPs first originated in a common ancestor of the other *Dickeya* species after the divergence of *D. paradisiaca*. Lastly, 19 CSPs identified in this work are specific for members of the genus *Pectobacterium*, providing molecular markers for this genus (Table 7). Some characteristics of these proteins are listed in Tables 5–7. Most of these *Dickeya*- and *Pectobacterium*-specific proteins are annotated as hypothetical and their cellular functions are not known.

DISCUSSION

Members of the genera *Dickeya*, *Pectobacterium* and *Brenneria* are important plant pathogens that are currently distinguished primarily on the basis of their branching in phylogenetic trees (Hauben *et al.*, 2005; Samson *et al.*, 2005; Naum *et al.*, 2008, 2011). We have used comparative genomic approaches to examine their evolutionary relationships and to identify molecular markers that are specific for these bacteria. Information from available genomes was initially used to reconstruct phylogenetic trees for different species/strains of *Dickeya*, *Pectobacterium* and *Brenneria* as well as other *Enterobacteriales* based upon concatenated sequences of 23 conserved proteins. To date, this is the largest dataset used to examine the evolutionary relationship among these bacteria. In the resulting trees, species from these three genera formed a strongly

Table 2. CSIs specific for the genus *Dickeya*

Protein name	Gene name	GI number	Figure number	Indel size	Indel position
Adenosine deaminase	*add*	251789743	Fig. 2(a)	2 aa insert	117–147
Multidrug resistance protein MdtA	*mdtA*	251788831	Fig. 2(b)	2 aa deletion	200–242
AMP-dependent synthetase and ligase	–	251788831	Fig. S2	2 aa insert	41–63
4-Amino-4-deoxy-L-arabinose transferase*	*arnT*	251791779	Fig. S3	1 aa insert	232–264
HAD-superfamily hydrolase	–	251787900	Fig. S4	1 aa insert	189–214
Hypothetical protein Dd1591-2304	–	251789904	Fig. S5	1 aa insert	220–259
Electron transport complex, RnfABCDGE type, C subunit	*rnfC*	251789757	Fig. S6	1 aa insert	313–348
Molybdenum cofactor synthesis domain-containing protein	*moeA*	251790145	Fig. S7	1 aa deletion	50–89
2-Succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylate synthase	*menD*	251790594	Fig. S8	1 aa deletion	264–293
β-D-Galactosidase*	*lacZ*	251790316	Fig. S9	1 aa deletion	577–602

*Homologous protein/part of the sequences corresponding to the region containing the CSIs was found to be missing in *Brenneria* sp. EniD312.

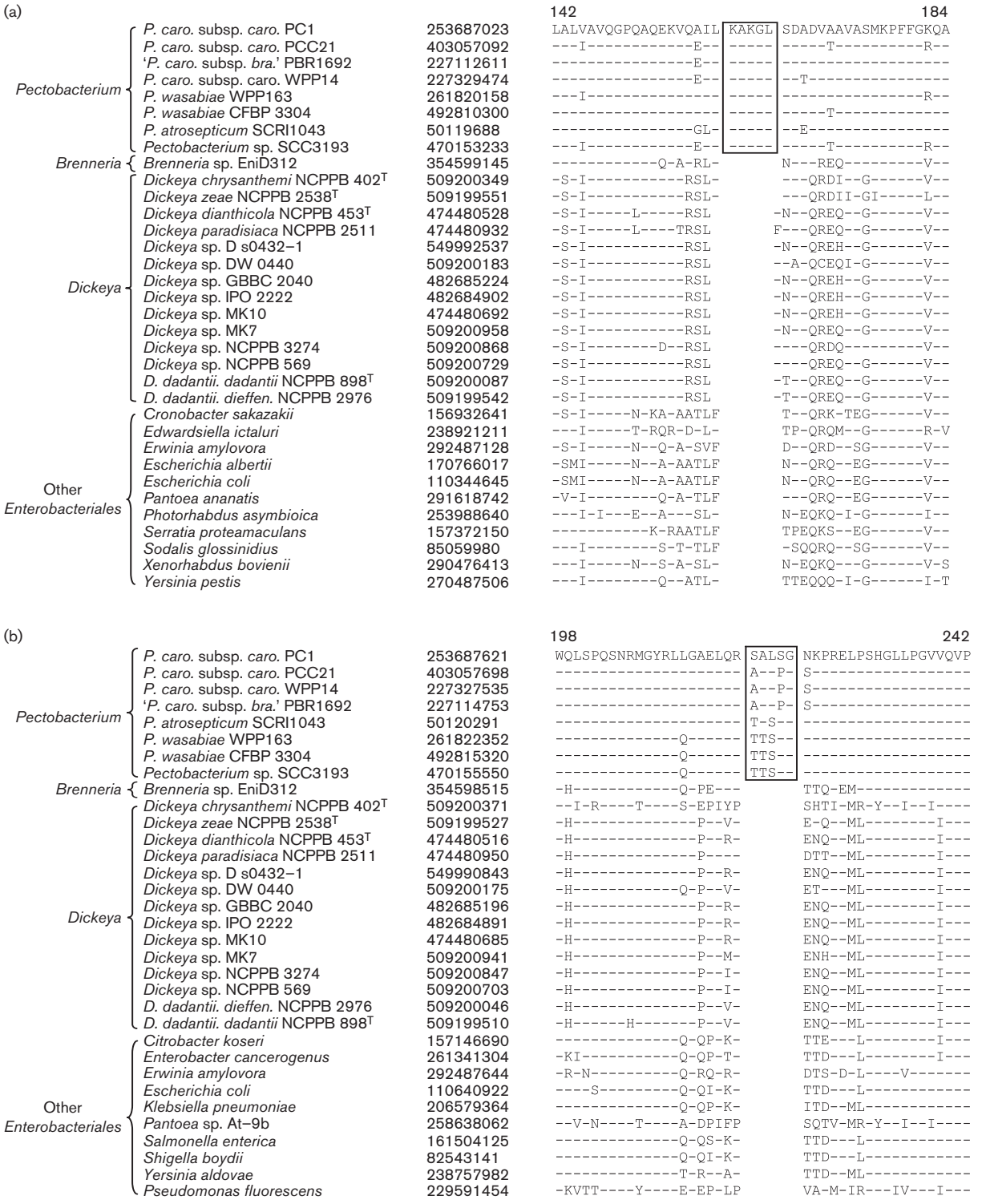


Fig. 3. Excerpts from the sequence alignments of alycine cleavage system T protein (a) and urea amidolyase-like protein (b), each containing 5 aa inserts that are uniquely found in various sequenced species/strains of the genus *Pectobacterium* but not found in other bacteria. The information for other CSIs that are also specific for this genus is presented in Table 3 and Figs S10–S13.

Table 3. CSIs specific for the genus *Pectobacterium* or *Pectobacterium* and *Brenneria*

Protein name	Gene name	GI number	Figure number	Indel size	Indel position
Glycine cleavage system T protein	*gcvT*	253687023	Fig. 3(a)	5 aa insert	142–184
Urea amidolyase related protein	–	253690318	Fig. 3(b)	5 aa insert	198–242
Glycerol-3-phosphate dehydrogenase subunit A	*glpB*	253687621	Fig. S10	8 aa insert	381–421
Sigma E regulatory protein, MucB/RseB	*resB*	253689444	Fig. S11	1 aa insert	168–196
Single-stranded-DNA-specific exonuclease RecJ*	*recJ*	253687048	Fig. S12	5 aa insert	191–239
Phosphoribosyl-formyl-glycinamidine synthase†	*purL*	261820538	Fig. S13	1 aa insert	332–365
Periplasmic serine protease DegS‡	*degS*	253686693	Fig. 4(a)	7 aa insert	166–210
6,7-Dimethyl-8-ribityllumazine synthase‡	*ribH*	253687421	Fig. S14	3 aa insert	52–92
GCN5-like *N*-acetyltransferase‡	-	253689699	Fig. S15	1 aa deletion	593–628

*The CSIs were found to be missing in *Pectobacterium atrosepticum* SCRI1043 and *Pectobacterium wasabiae* CFBP 3304.
†The CSIs were found to be missing in *Pectobacterium atrosepticum* SCRI1043.
‡The CSIs are commonly shared between the genera *Pectobacterium* and *Brenneria*.

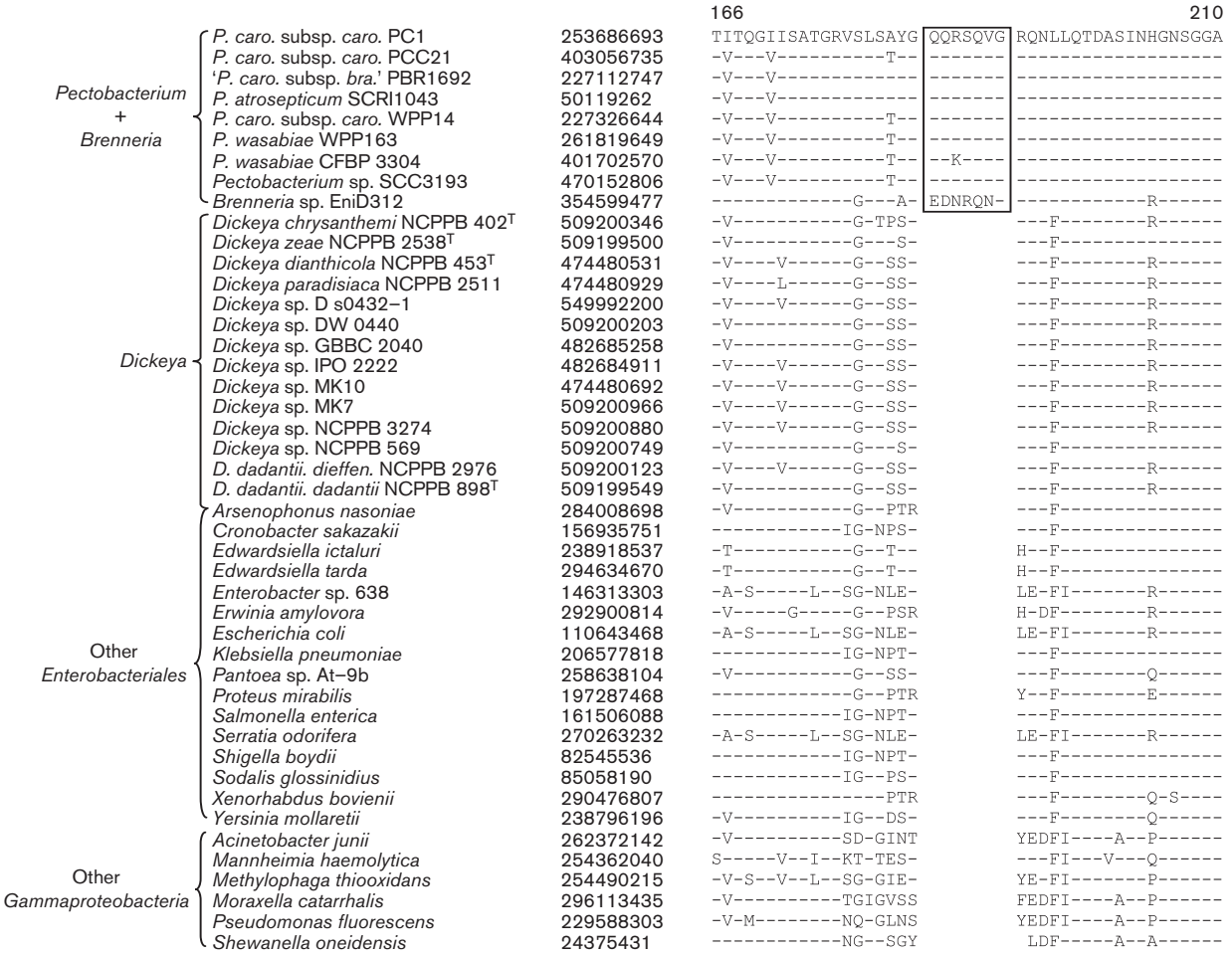


Fig. 4. Partial sequence alignments of the periplasmic serine protease (DegS) protein showing a 7 aa insert that is commonly shared between members of the genus *Pectobacterium* and *Brenneria* sp. EniD312. Information for two other CSIs exhibiting similar specificities is provided in Figs S14 and S15.

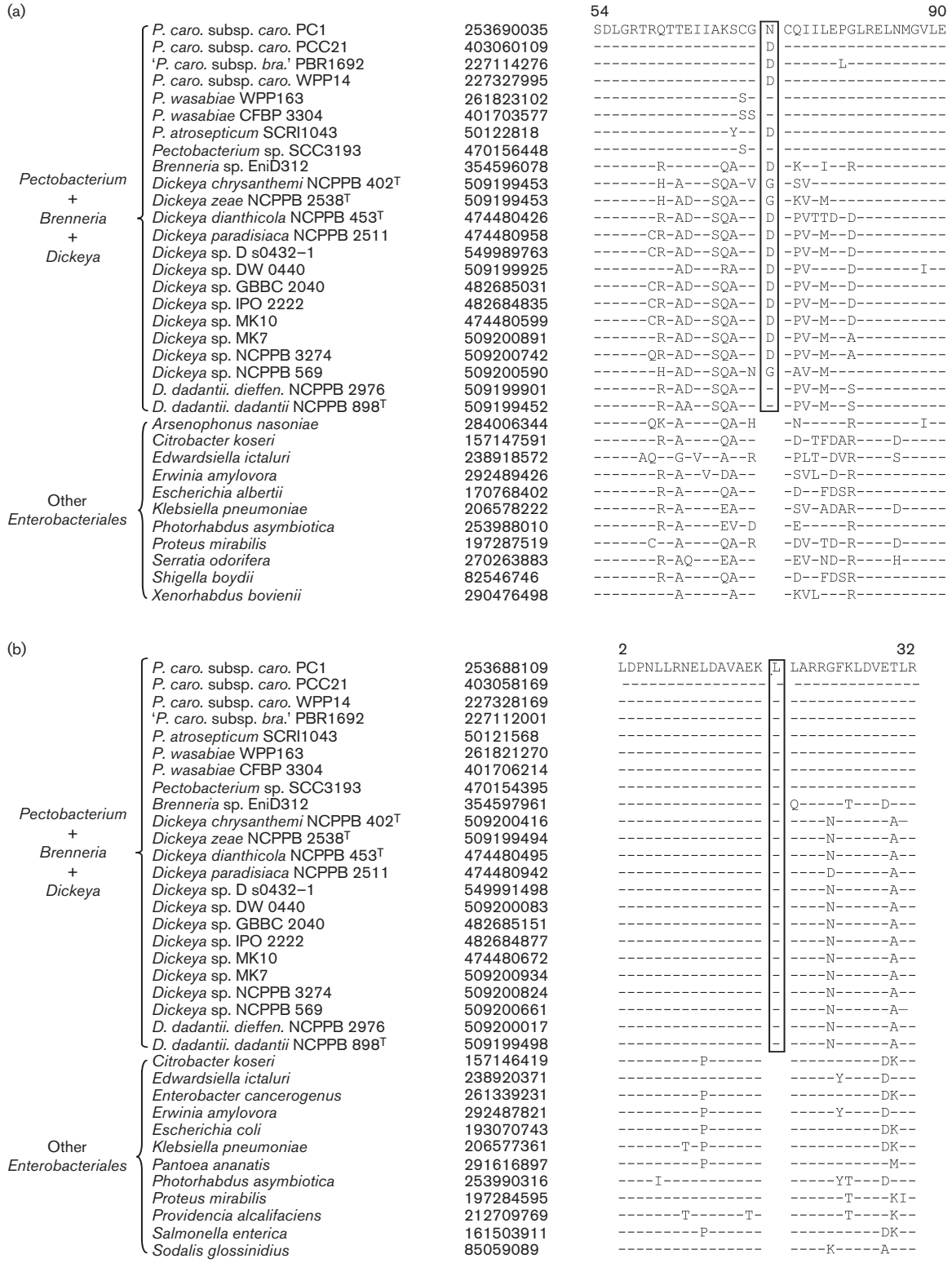


Fig. 5. Partial sequence alignments for the proteins phosphoglycerate mutase (a) and seryl-tRNA synthetase (b), showing two different 1 aa inserts that are uniquely present in all of the genome-sequenced species from the genera *Dickeya*, *Pectobacterium* and *Brenneria*. Information for other CSIs showing similar specificity is provided in Table 4 and Figs S16–S23.

Table 4. CSIs shared by the genera *Dickeya*, *Pectobacterium* and *Brenneria*

Protein name	Gene name	GI number	Figure number	Indel size	Indel position
Phosphoglycerate mutase	*gpmB*	253690035	Fig. 5(a)	1 aa insert	54–90
Seryl-tRNA synthetase	*serS*	253688109	Fig. 5(b)	1 aa insert	2–32
Adenylate cyclase	*cyaA*	253689768	Fig. S16	5 aa insert	11–48
Polyprenyl synthetase	–	253687426	Fig. S17	3 aa insert	268–300
Alkylated DNA repair protein	*alkB*	227329148	Fig. S18	2 aa insert	190–215
Glutamate synthase subunit alpha	*gltB*	227112762	Fig. S19	1 aa insert	821–857
Osmosensitive K channel His kinase sensor*	–	253687609	Fig. S20	1 aa insert	408–438
Cytoplasmic asparaginase I†	*ansA*	253688356	Fig. S21	1 aa insert	304–335
Diguanylate cyclase†	–	253686656	Fig. S22	1 aa insert	259–295
OmpA/MotB domain-containing protein†	–	253688988	Fig. S23	1 aa insert	234–262

*The homologous part of the proteins containing CSIs were found to be missing in *Brenneria* sp. EniD312.
†The CSIs were found to be shared by one or two other gammaproteobacterial species (see respective figures for detail).

supported monophyletic clade within the order *Enterobacteriales*. Within the clade comprising these three genera, species/strains from the genera *Dickeya* and *Pectobacterium* also formed distinct monophyletic clades with *Brenneria* sp. EniD312 forming an outgroup of the *Pectobacterium* clade. Within the genus *Dickeya*, following the recently proposed name changes by Marrero *et al.* (2013), different strains of various *Dickeya* species also formed monophyletic groupings, thus supporting the proposed reclassification.

However, the main objective of this work was to identify molecular markers that are specific for the genera *Dickeya*, *Pectobacterium* and *Brenneria*. Accordingly, this work has identified numerous molecular markers comprising CSIs and CSPs that either are specific for the sequenced members of these three genera or are uniquely shared by some or all of them. A summary of the species specificity of the discovered markers is provided in Fig. 6. Of these molecular signatures, six CSIs and five CSPs are uniquely shared by all sequenced species from these three genera. Four additional CSIs are also largely specific for these three genera, except that they are missing in one of the species or present in an isolated species from some other taxa. The unique shared presence of these molecular markers by species from these three genera provides strong evidence that they shared a common ancestor exclusive of other bacteria and that they form a distinct subgroup within the order *Enterobacteriales*. In addition, 10 CSIs and 17 CSPs discovered during this work were found to be distinctive characteristics of members of the genus *Dickeya*. Several of these CSPs, which are lacking in *D. paradisiaca*, support the deeper branching of this species in comparison with the other *Dickeya* species. Additionally, six CSIs and 19 CSPs are uniquely found in all (or most) *Pectobacterium* species/strains. These molecular signatures provide novel means for identification of species from these groups and for demarcation of these genera in molecular terms. Additionally, this work has also identified three CSIs and one CSP that are uniquely shared by members of the genus *Pectobacterium* and *Brenneria* sp. EniD312. These results provide evidence that *Pectobacterium* and *Brenneria* sp. EniD312 shared a common ancestor exclusive of the genus *Dickeya*. This inference is also supported by the branching of these species in the concatenated protein tree (Fig. 1). However, the inference that *Brenneria* is more closely related to *Pectobacterium* is based only on *Brenneria* sp. EniD312 and it remains to be determined whether other *Brenneria* species also behaved similarly.

In addition to their usefulness for the classification (demarcation) of these genera and for clarifying their evolutionary relationships, the discovered molecular markers also provide novel means for the identification of these bacteria. In addition to their specificity for these taxa, these markers possess a high degree of predictive ability that they

Table 5. CSPs specific for the genera *Dickeya*, *Pectobacterium* and *Brenneria*

Protein name	Accession no.	GI number	Length (aa)
Global regulatory protein	YP_003332283.1	271499258	130
Hypothetical protein Dd586_0697	YP_003332295.1	271499270	140
Hypothetical protein Dd586_1616	YP_003333187.1	271500162	84
Hypothetical protein Dd586_1999	YP_003333560.1	271500535	99
Hypothetical protein Dd586_2255*	YP_003333811.1	271500786	82

*The homologue of this protein is not found in *Brenneria*.

Table 6. CSPs that are uniquely found in *Dickeya* species

Protein name	Accession no.	GI number	Length (aa)
Hypothetical protein Dd586_1497	YP_003333070.1	271500045	52
Hypothetical protein Dd586_1737	YP_003333305.1	271500280	165
Hypothetical protein Dd586_1775	YP_003333343.1	271500318	362
Hypothetical protein Dd586_2539	YP_003334091.1	271501066	130
Hypothetical protein Dd586_2824	YP_003334369.1	271501344	41
Putative lipoprotein	YP_003334921.1	271501895	91
Hypothetical protein Dd586_0422*	YP_003332023.1	271498998	268
Hypothetical protein Dd586_0554*	YP_003332153.1	271499128	57
Hypothetical protein Dd586_1795*	YP_003333363.1	271500338	43
Hypothetical protein Dd586_1801*	YP_003333369.1	271500344	160
Hypothetical protein Dd586_2330*	YP_003333886.1	271500861	91
Hypothetical protein Dd586_2377*	YP_003333933.1	271500908	79
Hypothetical protein Dd586_2418*	YP_003333972.1	271500947	34
Hypothetical protein Dd586_2798*	YP_003334343.1	271501318	69
Hypothetical protein Dd586_3460*	YP_003334996.1	271501970	269
Hypothetical protein Dd586_3464*	YP_003335000.1	271501974	66
Hypothetical protein Dd586_3530*	YP_003335066.1	271502040	51

*The homologues for these proteins were not detected in *Dickeya paradisiaca*.

will also be be found in other members of these groups. This is illustrated by the fact that our initial analysis, which identified these CSIs, was based upon sequence information for only four *Dickeya* species/strains, whose complete genomes were available in the NCBI database (Table 1). The presence of these CSIs in the draft genomes of various other *Dickeya* species/strains (listed in Table S1) was examined only during revision of this manuscript. The fact that all of these CSIs and CSPs are also present in all or most of the other *Dickeya* species provides strong evidence that these molecular markers constitute distinctive characteristics of members of the genus *Dickeya* and that they

Table 7. CSPs that are uniquely found in *Pectobacterium* species

Protein name	Accession no.	GI number	Length (aa)
Hypothetical protein Pecwa_0644	YP_003258075.1	261819969	178
Hypothetical protein Pecwa_0660	YP_003258091.1	261819985	153
Hypothetical protein Pecwa_0689	YP_003258117.1	261820011	171
Hypothetical protein Pecwa_0691	YP_003258119.1	261820013	156
Hypothetical protein Pecwa_0772	YP_003258198.1	261820092	55
Hypothetical protein Pecwa_1061	YP_003258485.1	261820379	156
Hypothetical protein Pecwa_1094	YP_003258515.1	261820409	118
Hypothetical protein Pecwa_1436	YP_003258842.1	261820736	206
Hypothetical protein Pecwa_1592	YP_003258989.1	261820883	95
Hypothetical protein Pecwa_3132	YP_003260481.1	261822375	204
Hypothetical protein Pecwa_2681*	YP_003260044.1	261821938	46
Hypothetical protein Pecwa_2954*	YP_003260310.1	261822204	191
Hypothetical protein Pecwa_3013*	YP_003260366.1	261822260	93
Hypothetical protein Pecwa_3586*	YP_003260929.1	261822823	44
Hypothetical protein Pecwa_3766*	YP_003261108.1	261823002	36
Hypothetical protein Pecwa_3818*	YP_003261159.1	261823053	164
Hypothetical protein Pecwa_0611*	YP_003258042.1	261819936	38
Hypothetical protein Pecwa_0685*	YP_003258113.1	261820007	282
Hypothetical protein Pecwa_1718*	YP_003259112.1	261821006	37
Hypothetical protein Pecwa_0258†	YP_003257719.1	261819613	174

*Homologue of the proteins not found in *Pectobacterium atrosepticum*.
†The CSP is specific to *Pectobacterium* and *Brenneria*.

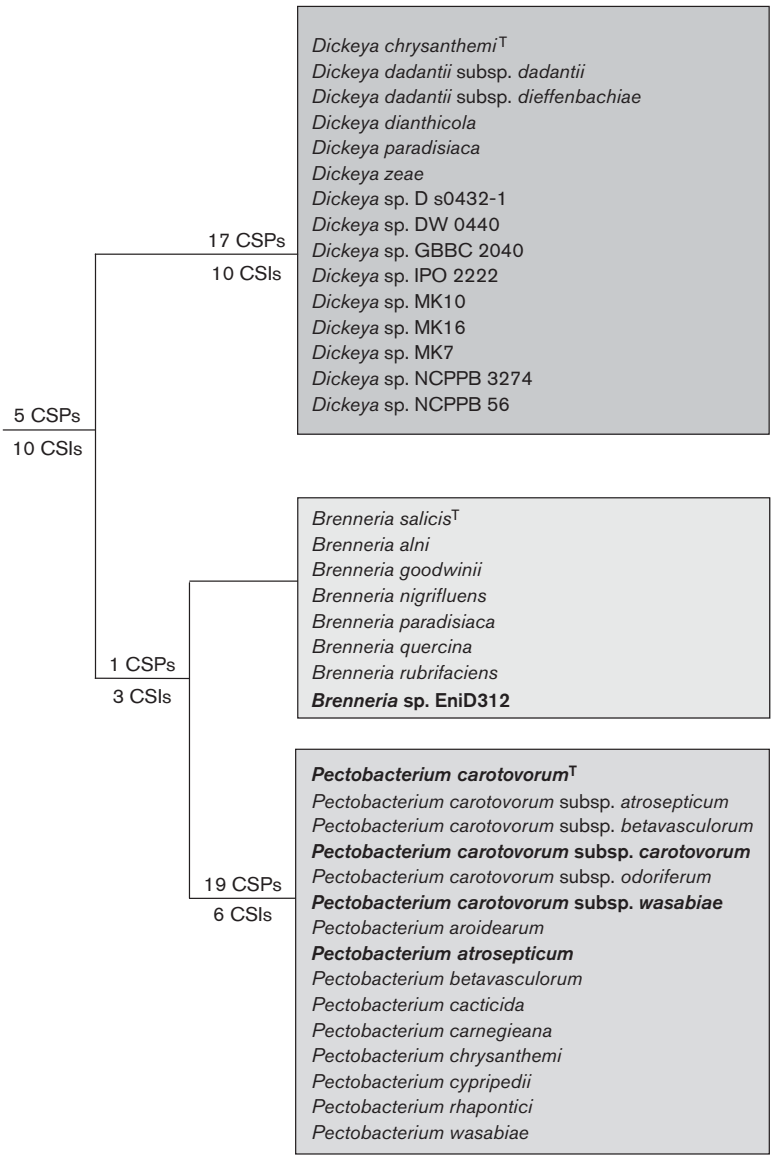


Fig. 6. Summary diagram showing the species distribution pattern of different CSIs and CSPs identified in this work and the evolutionary stages where the genetic changes responsible for them probably occurred. The species/strains whose genomes are sequenced are shown in bold. The type species of the genera are marked by superscript 'T'.

will probably also be found in other species/strains (both known as well as unknown) that are part of this genus.

Members of these three genera are responsible for a broad range of diseases in economically important plants (Hauben *et al.*, 2005; Samson *et al.*, 2005; Charkowski, 2006; Ma *et al.*, 2007; Yishay *et al.*, 2008; Czajkowski *et al.*, 2011; Toth *et al.*, 2011; Costechareyre *et al.*, 2012). Hence, there is a need for developing more rapid, sensitive and specific methods for their identification (Diallo *et al.*, 2009; Van Vaerenbergh *et al.*, 2012). Given their specificity and predictive ability for members of these genera, the molecular markers described here are of great interest in this regard. The primary sequences of the genes/proteins containing many of the described CSIs or CSPs, which are specific for these genera, exhibit high degrees of sequence conservation. Hence, PCR and other molecular probes based upon their gene sequences (including sequence regions flanking the CSIs) should provide novel means for the development of sensitive and specific methods for the identification of both known as well as

novel members of these genera in different settings (Gao & Gupta, 2005; Ahmod *et al.*, 2011).

The cellular functions of the CSIs and CSPs that are specific for these plant-pathogenic genera are presently not known. However, due to the specific presence of these molecular characteristics in members of these genera, the discovered characteristics are expected to play important roles in these bacteria. Our recent work on several CSIs in the Hsp60 (GroEL) and Hsp70 (DnaK) proteins provides evidence that the CSIs such as those identified here are essential for the groups of bacteria where they are found (Singh & Gupta, 2009). Likewise, the CSPs that are limited to a given group of bacteria are also postulated to perform essential functions in the particular groups of bacteria (Fang *et al.*, 2005; Narra *et al.*, 2008; Gao *et al.*, 2009b; Lorenzini *et al.*, 2010). Hence, further studies on understanding the cellular functions of these CSIs and CSPs could lead to the discovery of novel biochemical properties that are specific for these plant-pathogenic bacteria. Furthermore, the conserved indels in protein sequences provide possible means for development of antibacterial agents that can specifically target these groups of plant pathogens (Nandan *et al.*, 2007; Naushad & Gupta, 2013).

Taxonomic implications

At present, no biochemical or molecular marker is known that is specific for members of the genera *Dickeya*, *Pectobacterium* and *Brenneria*. This work describes large numbers of molecular markers that are specific for members of these genera or those that are commonly shared by species from these three genera. The markers that are specific for *Dickeya* or *Pectobacterium* provide novel and more definitive means for demarcation of these taxa in molecular terms. Additionally, the markers identified in this work also provide strong evidence that members of the genera *Dickeya*, *Pectobacterium* and *Brenneria* form a distinct clade within the order *Enterobacteriales* and that this clade should eventually be recognized as a new family-level taxon ('*Pectobacteriaceae*') within this order. The above three genera are presently part of the family *Enterobacteriaceae*, which is the sole family within the order *Enterobacteriales* (Euzéby, 2013). In phylogenetic trees, members of the order *Enterobacteriales* form a number of distinct groups (Gao *et al.*, 2009a; Williams *et al.*, 2010) (Fig. 1). Hence, a formal proposal to place the genera *Dickeya*, *Pectobacterium* and *Brenneria* into a new family cannot be made until more reliable means to divide the order *Enterobacteriales* into a number of distinct families are identified. Nonetheless, based upon the identified signatures, the descriptions of the genera *Dickeya* and *Pectobacterium* are emended to include information for the molecular signatures.

Emended description of the genus *Dickeya* Samson *et al.*, 2005

The morphological and phenotypic characteristics of this genus remain as described by Samson *et al.* (2005). Cells are Gram-negative rods, 0.5–1.0 × 1.0–3.0 μm with rounded ends. They occur mostly alone or in pairs, but sometimes in chains. Cells are usually motile by means of peritrichous flagella. They are facultatively aerobic/anaerobic bacteria that catabolize glucose by a fermentative pathway and reduce nitrates to nitrites. Members of this genus are capable of hydrolysing pectin, produce indole and grow optimally at 36 °C. Catabolize (+)-L-arabinose, *myo*-inositol, (+)-D-malate, malonate, D-mannose, mucate, saccharate and mesotartrate, but do not catabolize (+)-trehalose, methyl α-glucoside, (+)-D-arabitol or sorbitol. Members of this genus cause vascular wilts or soft rots on a range of host plants. The DNA G+C contents of these bacteria range from 53.6 to 59.5 mol%. The type species is *Dickeya chrysanthemi* (Samson *et al.*, 2005). Members of this genus can be distinguished from other bacteria based on CSIs in the following proteins: adenosine deaminase, multidrug resistance protein MdtA, AMP-dependent synthetase and ligase, 4-amino-4-deoxy-L-arabinose transferase, HAD-superfamily hydrolase, hypothetical protein Dd1591-2304, electron transport complex RnfABCDGE type (C subunit), molybdenum cofactor synthesis domain-containing protein, 2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylate synthase and β-D-galactosidase. In addition, the genes for the following CSPs whose GenBank identification numbers are noted here (namely 271498998, 271499128, 271500045, 271500280, 271500318, 271500338, 271500344, 271500861, 271500908, 271500947, 271501066, 271501318, 271501344, 271501895, 271501970, 271501974 and 271502040) are also uniquely found in the sequenced members of the genus *Dickeya*. However, as noted in Table 5, the homologues of some of these CSPs are not detected in the deep branching *D. paradisiaca*.

Emended description of the genus *Pectobacterium* Waldee 1945 (Approved Lists 1980), emend. Hauben *et al.* 1998

The phenotypic characteristics of this genus remain as described by Waldee (1945) and Hauben *et al.* (1998, 2005). Cells are Gram-negative rods which are 0.5–1.0 × 1.0–3.0 μm with rounded ends. They occur mostly alone or in pairs, but chains occur as well. Cells are usually motile by means of peritrichous flagella. Strains are catalase-positive and oxidase-negative. They are facultative anaerobes and use fermentative metabolism to grow on a variety of simple sugars and amino acids as described by Hauben *et al.* (1998). Strains do not possess tryptophan deaminase or urease and hydrolyse aesculin but not starch. They produce acid from *N*-acetylglucosamine as well as a number of other simple sugars. Members of this genus (except *Pectobacterium cypripedii*) possess pectolytic enzymes and cause soft rots, necroses and wilts on food crops and ornamental plants. Members of this genus, whose type species is *Pectobacterium carotovorum* (Jones, 1901), Waldee 1945 (Approved Lists 1980) (Jones, 1901; Waldee, 1945; Skerman *et al.*, 1980; Hauben *et al.*, 1998),

can be distinguished from other bacteria by CSIs in the following proteins: glycine cleavage system T protein, urea amidolyase related protein, glycerol-3-phosphate dehydrogenase subunit A, sigma E regulatory protein, MucB/RseB, single-stranded-DNA-specific exonuclease RecJ and phosphoribosyl-formyl-glycinamidine synthase. In addition, the genes for the following CSPs whose GenBank identification numbers are noted here (namely 261819969, 261819985, 261820011, 261820013, 261820092, 261820379, 261820409, 261820736, 261820883, 261822375, 261821938, 261822204, 261822260, 261822823, 261823002, 261823053, 261819936, 261820007 and 261821006) are also uniquely found in the sequenced *Pectobacterium* species.

ACKNOWLEDGEMENTS

This work was supported by a research grant from the Ontario Ministry of Innovation and Economic Development-Ontario Research Fund. We thank Misbah Sohail for assistance in the analysis of sequence data for identification of CSIs.

REFERENCES

Ahmod, N. Z., Gupta, R. S. & Shah, H. N. (2011). Identification of a *Bacillus anthracis* specific indel in the *yeaC* gene and development of a rapid pyrosequencing assay for distinguishing *B. anthracis* from the *B. cereus* group. *J Microbiol Methods* **87**, 278–285.

Ajawatanawong, P. & Baldauf, S. L. (2013). Evolution of protein indels in plants, animals and fungi. *BMC Evol Biol* **13**, 140.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402.

Bell, K. S., Sebaihia, M., Pritchard, L., Holden, M. T., Hyman, L. J., Holeva, M. C., Thomson, N. R., Bentley, S. D., Churcher, L. J. & other authors (2004). Genome sequence of the enterobacterial phytopathogen *Erwinia carotovora* subsp. *atroseptica* and characterization of virulence factors. *Proc Natl Acad Sci U S A* **101**, 11105–11110.

Bhandari, V. & Gupta, R. S. (2014). Molecular signatures for the phylum (class) *Thermotogae* and a proposal for its division into three orders (*Thermotogales*, *Kosmotogales* ord. nov. and *Petrotogales* ord. nov.) containing four families (*Thermotogaceae*, *Fervidobacteriaceae* fam. nov., *Kosmotogaceae* fam. nov. and *Petrotogaceae* fam. nov.) and a new genus *Pseudothermotoga* gen. nov. with five new combinations. *Antonie van Leeuwenhoek* **105**, 143–168.

Bhandari, V., Naushad, H. S. & Gupta, R. S. (2012). Protein based molecular markers provide reliable means to understand prokaryotic phylogeny and support Darwinian mode of evolution. *Front Cell Infect Microbiol* **2**, 98.

Brady, C. L., Cleenwerck, I., Denman, S., Venter, S. N., Rodríguez-Palenzuela, P., Coutinho, T. A. & De Vos, P. (2012). Proposal to reclassify *Brenneria quercina* (Hildebrand and Schroth 1967) Hauben *et al.* 1999 into a new genus, *Lonsdalea* gen. nov., as *Lonsdalea quercina* comb. nov., descriptions of *Lonsdalea quercina* subsp. *quercina* comb. nov., *Lonsdalea quercina* subsp. *iberica* subsp. nov. and *Lonsdalea quercina* subsp. *britannica* subsp. nov., emendation of the description of the genus *Brenneria*, reclassification of *Dickeya dieffenbachiae* as *Dickeya dadantii* subsp. *dieffenbachiae* comb. nov., and emendation of the description of *Dickeya dadantii*. *Int J Syst Evol Microbiol* **62**, 1592–1602.

Brenner, D. J., McWhorter, A. C., Kai, A., Steigerwalt, A. G. & Farmer, J. J., III (1986). *Enterobacter asburiae* sp. nov., a new species found in clinical specimens, and reassignment of *Erwinia dissolvens* and *Erwinia nimipressuralis* to the genus *Enterobacter* as *Enterobacter dissolvens* comb. nov. and *Enterobacter nimipressuralis* comb. nov. *J Clin Microbiol* **23**, 1114–1120.

Brown, E. W., Davis, R. M., Gouk, C. & van der Zwet, T. (2000). Phylogenetic relationships of necrogenic *Erwinia* and *Brenneria* species as revealed by glyceraldehyde-3-phosphate dehydrogenase gene sequences. *Int J Syst Evol Microbiol* **50**, 2057–2068.

Bull, C. T., De Boer, S. H., Denny, T. P., Firrao, G., Fischer-Le Saux, M., Saddler, G. S., Scortichini, M., Stead, D. E. & Takikawa, Y. (2012) Letter to the Editor: List of new names of plant pathogenic bacteria (2008–2010). *J. Plant Pathol* **94**, 21–27.

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

Charkowski, A. (2006). The soft rot *Erwinia*. In *Plant-Associated Bacteria*, pp. 423–505. Edited by S. Gnanamanickam. Dordrecht: Springer.

Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B. & Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283–1287.

Costechareyre, D., Balmand, S., Condemine, G. & Rahbé, Y. (2012). *Dickeya dadantii*, a plant pathogenic bacterium producing Cyt-like entomotoxins, causes septicemia in the pea aphid *Acyrthosiphon pisum*. *PLoS ONE* **7**, e30702.

Czajkowski, R., Pérombelon, M. C. M., van Veen, J. A. & van der Wolf, J. M. (2011). Control of blackleg and tuber soft rot of potato caused by *Pectobacterium* and *Dickeya* species: a review. *Plant Pathol* **60**, 999–1013.

Denman, S., Brady, C., Kirk, S., Cleenwerck, I., Venter, S., Coutinho, T. & De Vos, P. (2012). *Brenneria goodwinii* sp. nov., associated with acute oak decline in the UK. *Int J Syst Evol Microbiol* **62**, 2451–2456.

Diallo, S., Latour, X., Groboillot, A., Smadja, B., Copin, P., Orange, N., Feuilloley, M. & Chevalier, S. (2009). Simultaneous and selective detection of two major soft rot pathogens of potato: *Pectobacterium atrosepticum* (*Erwinia carotovora* subsp. *atrosepticum*) and *Dickeya* spp. (*Erwinia chrysanthemi*). *Eur J Plant Pathol* **125**, 349–354.

Dutilh, B. E., Snel, B., Ettema, T. J. & Huynen, M. A. (2008). Signature genes as a phylogenomic tool. *Mol Biol Evol* **25**, 1659–1667.

Dye, D. W. (1968). A taxonomic study of the genus *Erwinia* I: the 'amylovora' group. *N Z J Sci* **11**, 590–607.

Euzéby, J. P. (2013). List of bacterial names with standing in nomenclature: a folder available on the Internet. [Last full update 22 November 2013] http://www.bacterio.cict.fr

Fang, G., Rocha, E. & Danchin, A. (2005). How essential are nonessential genes? *Mol Biol Evol* **22**, 2147–2156.

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

Gao, B. & Gupta, R. S. (2007). Phylogenomic analysis of proteins that are distinctive of *Archaea* and its main subgroups and the origin of methanogenesis. *BMC Genomics* **8**, 86.

Gao, B. & Gupta, R. S. (2012a). Microbial systematics in the post-genomics era. *Antonie van Leeuwenhoek* **101**, 45–54.

Gao, B. & Gupta, R. S. (2012b). Phylogenetic framework and molecular signatures for the main clades of the phylum *Actinobacteria*. *Microbiol Mol Biol Rev* **76**, 66–112.

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

Gao, B., Sugiman-Marangos, S., Junop, M. S. & Gupta, R. S. (2009b). Structural and phylogenetic analysis of a conserved actinobacteria-specific protein (ASP1; SCO1997) from *Streptomyces coelicolor*. *BMC Struct Biol* **9**, 40.

Gardan, L., Gouy, C., Christen, R. & Samson, R. (2003). Elevation of three subspecies of *Pectobacterium carotovorum* to species level: *Pectobacterium atrosepticum* sp. nov., *Pectobacterium betavasculorum* sp. nov. *and Pectobacterium wasabiae* sp. nov. *Int J Syst Evol Microbiol* **53**, 381–391.

Gavini, F., Mergaert, J., Beji, A., Mielcarek, C., Izard, D., Kersters, K. & De Ley, J. (1989). Transfer of *Enterobacter agglomerans* (Beijerinck 1888) Ewing and Fife 1972 to *Pantoea* gen. nov. as *Pantoea agglomerans* comb. nov. and description of *Pantoea dispersa* sp. nov. *Int J Syst Bacteriol* **39**, 337–345.

Glasner, J. D., Marquez-Villavicencio, M., Kim, H. S., Jahn, C. E., Ma, B., Biehl, B. S., Rissman, A. I., Mole, B., Yi, X. & other authors (2008). Niche-specificity and the variable fraction of the *Pectobacterium* pan-genome. *Mol Plant Microbe Interact* **21**, 1549–1560.

Glasner, J. D., Yang, C. H., Reverchon, S., Hugouvieux-Cotte-Pattat, N., Condemine, G., Bohin, J. P., Van Gijsegem, F., Yang, S., Franza, T. & other authors (2011). Genome sequence of the plant-pathogenic bacterium *Dickeya dadantii* 3937. *J Bacteriol* **193**, 2076–2077.

Griffiths, E. & Gupta, R. S. (2004). Signature sequences in diverse proteins provide evidence for the late divergence of the Order *Aquificales*. *Int Microbiol* **7**, 41–52.

Griffiths, E. & Gupta, R. S. (2006). Lateral transfers of serine hydroxymethyltransferase (*glyA*) and UDP-N-acetylglucosamine enolpyruvyl transferase (*murA*) genes from free-living *Actinobacteria* to the parasitic chlamydiae. *J Mol Evol* **63**, 283–296.

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

Gupta, R. S. (2000). The phylogeny of *Proteobacteria*: relationships to other eubacterial phyla and eukaryotes. *FEMS Microbiol Rev* **24**, 367–402.

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

Gupta, R. S. (2010). Applications of conserved indels for understanding microbial phylogeny. In *Molecular Phylogeny of Microorganisms*, pp. 135–150. Edited by A. Oren & R. T. Papke. Norwich: Caister Academic Press.

Gupta, R. S. & Gao, B. (2010). Recent advances in understanding microbial systematics. In *Microbial Population Genetics*, pp. 1–14. Edited by J. Xu. Norwich: Caister Academic Press.

Gupta, R. S. & Griffiths, E. (2002). Critical issues in bacterial phylogeny. *Theor Popul Biol* **61**, 423–434.

Gupta, R. S. & Mathews, D. W. (2010). Signature proteins for the major clades of *Cyanobacteria*. *BMC Evol Biol* **10**, 24.

Gupta, R. S. & Mok, A. (2007). Phylogenomics and signature proteins for the alpha proteobacteria and its main groups. *BMC Microbiol* **7**, 106.

Hauben, L., Moore, E. R., Vauterin, L., Steenackers, M., Mergaert, J., Verdonck, L. & Swings, J. (1998). Phylogenetic position of phytopathogens within the *Enterobacteriaceae*. *Syst Appl Microbiol* **21**, 384–397.

Hauben, L., Gijsegem, F. V. & Swings, J. (2005). Genus XXIV. *Pectobacterium*. In *Bergey's Manual of Systematic Bacteriology*, pp. 721–730. Edited by G. M. Garrity. New York: Springer.

Jones, L. R. (1901). A soft rot of carrot and other vegetabels caused by *Bacillus carotovorus*. *Vt Agric Exp Stn Annu Rep* **13**, 299–332.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8**, 275–282.

Koskinen, J. P., Laine, P., Niemi, O., Nykyri, J., Harjunpää, H., Auvinen, P., Paulin, L., Pirhonen, M., Palva, T. & Holm, L. (2012). Genome sequence of *Pectobacterium* sp. strain SCC3193. *J Bacteriol* **194**, 6004.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A. & other authors (2007). CLUSTAL W and CLUSTAL_X version 2.0. *Bioinformatics* **23**, 2947–2948.

Lerat, E., Daubin, V., Ochman, H. & Moran, N. A. (2005). Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol* **3**, e130.

Lorenzini, E., Singer, A., Singh, B., Lam, R., Skarina, T., Chirgadze, N. Y., Savchenko, A. & Gupta, R. S. (2010). Structure and protein–protein interaction studies on *Chlamydia trachomatis* protein CT670 (YscO homolog). *J Bacteriol* **192**, 2746–2756.

Ludwig, W. & Klenk, H.-P. (2005). Overview: a phylogenetic backbone and taxonomic framework for prokaryotic systematics. In *Bergey's Manual of Systematic Bacteriology*, pp. 49–65. Edited by D. J. Brenner, N. R. Krieg, J. T. Staley & G. M. Garrity. Berlin: Springer-Verlag.

Ma, B., Hibbing, M. E., Kim, H. S., Reedy, R. M., Yedidia, I., Breuer, J., Breuer, J., Glasner, J. D., Perna, N. T. & other authors (2007). Host range and molecular phylogenies of the soft rot enterobacterial genera *Pectobacterium* and *Dickeya*. *Phytopathology* **97**, 1150–1163.

Marrero, G., Schneider, K. L., Jenkins, D. M. & Alvarez, A. M. (2013). Phylogeny and classification of *Dickeya* based on multilocus sequence analysis. *Int J Syst Evol Microbiol* **63**, 3524–3539.

Nandan, D., Lopez, M., Ban, F., Huang, M., Li, Y., Reiner, N. E. & Cherkasov, A. (2007). Indel-based targeting of essential proteins in human pathogens that have close host orthologue(s): discovery of selective inhibitors for *Leishmania donovani* elongation factor-1α. *Proteins* **67**, 53–64.

Narra, H. P., Cordes, M. H. & Ochman, H. (2008). Structural features and the persistence of acquired proteins. *Proteomics* **8**, 4772–4781.

Naum, M., Brown, E. W. & Mason-Gamer, R. J. (2008). Is 16S rDNA a reliable phylogenetic marker to characterize relationships below the family level in the Enterobacteriaceae? *J Mol Evol* **66**, 630–642.

Naum, M., Brown, E. W. & Mason-Gamer, R. J. (2011). Is a robust phylogeny of the enterobacterial plant pathogens attainable? *Cladistics* **27**, 80–93.

Naushad, H. S. & Gupta, R. S. (2013). Phylogenomics and molecular signatures for species from the plant pathogen-containing order Xanthomonadales. *PLoS ONE* **8**, e55216.

Nykyri, J., Niemi, O., Koskinen, P., Nokso-Koivisto, J., Pasanen, M., Broberg, M., Plyusnin, I., Törönen, P., Holm, L. & other authors (2012). Revised phylogeny and novel horizontally acquired virulence determinants of the model soft rot phytopathogen *Pectobacterium wasabiae* SCC3193. *PLoS Pathog* **8**, e1003013.

Olsen, G. J. & Woese, C. R. (1993). Ribosomal RNA: a key to phylogeny. *FASEB J* **7**, 113–123.

Oren, A. (2010) Microbial systematics. In *Handbook of Environmental Engineering, Vol. 10: Environmental Biotechnology*, pp. 81–120. Edited by L. K. Wang, V. Ivanov, J.-H. Tay & Y. T. Hung. New York: Springer Science + Business Media.

Park, T. H., Choi, B. S., Choi, A. Y., Choi, I. Y., Heu, S. & Park, B. S. (2012). Genome sequence of *Pectobacterium carotovorum* subsp. *carotovorum* strain PCC21, a pathogen causing soft rot in Chinese cabbage. *J Bacteriol* **194**, 6345–6346.

Parkinson, N., Stead, D., Bew, J., Heeney, J., Tsror Lahkim, L. & Elphinstone, J. (2009). *Dickeya* species relatedness and clade structure determined by comparison of *recA* sequences. *Int J Syst Evol Microbiol* **59**, 2388–2393.

Pritchard, L., Humphris, S., Saddler, G. S., Parkinson, N. M., Bertrand, V. & Elphinstone, J. G., andToth, I. K. (2013a) Detection of phytopathogens of the genus *Dickeya* using a PCR primer prediction pipeline for draft bacterial genome sequences. *Plant Pathol* **62**, 587–596.

Pritchard, L., Humphris, S., Baeyen, S., Maes, M., Van Vaerenbergh, J., Elphinstone, J., Saddler, G., and Toth, I. (2013b) Draft genome sequences of four *Dickeya dianthicola* and four *Dickeya solani* strains. Genome Announc 1, e00087-12.

Rokas, A. & Holland, P. W. (2000). Rare genomic changes as a tool for phylogenetics. *Trends Ecol Evol* **15**, 454–459.

Samson, R., Legendre, J. B., Christen, R., Fischer-Le Saux, M., Achouak, W. & Gardan, L. (2005). Transfer of *Pectobacterium chrysanthemi* (Burkholder *et al.* 1953) Brenner *et al.* 1973 and *Brenneria paradisiaca* to the genus *Dickeya* gen. nov. as *Dickeya chrysanthemi* comb. nov. and *Dickeya paradisiaca* comb. nov. and delineation of four novel species, *Dickeya dadantii* sp. nov., *Dickeya dianthicola* sp. nov., *Dickeya dieffenbachiae* sp. nov. and *Dickeya zeae* sp. nov. *Int J Syst Evol Microbiol* **55**, 1415–1427.

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

Skerman, V. B. D., McGowan, V. & Sneath, P. H. A. (1980). Approved lists of bacterial names. *Int J Syst Bacteriol* **30**, 225–420.

Spröer, C., Mendrock, U., Swiderski, J., Lang, E. & Stackebrandt, E. (1999). The phylogenetic position of *Serratia*, *Buttiauxella* and some other genera of the family *Enterobacteriaceae*. *Int J Syst Bacteriol* **49**, 1433–1438.

Stackebrandt, E. (2006). Defining taxonomic ranks. In *The Prokaryotes*, pp. 29–57. Edited by M. Dworkin, S. Falkow, E. Rosenberg, K.-H. Schleifer & E. Stackebrandt. New York: Springer.

Tamura, K., Dudley, J., Nei, M. & Kumar, S. (2007). MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**, 1596–1599.

Toth, I. K., van der Wolf, J. M., Saddler, G., Lojkowska, E., Helias, V. & Porhonen, M. (2011). *Dickeya* species: an emerging problem for potato production in Europe. *Plant Pathol* **60**, 385–399.

Van Vaerenbergh, J., Baeyen, S., De Vos, P. & Maes, M. (2012). Sequence diversity in the *Dickeya fliC* gene: phylogeny of the *Dickeya* genus and TaqMan® PCR for '*D. solani*', new biovar 3 variant on potato in Europe. *PLoS ONE* **7**, e35738.

Waldee, E. L. (1945). Comparative studies of some peritrichous phytopathogenic bacteria. *Iowa State Coll J Sci* **19**, 435–484.

Whelan, S. & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* **18**, 691–699.

Williams, K. P., Gillespie, J. J., Sobral, B. W., Nordberg, E. K., Snyder, E. E., Shallom, J. M. & Dickerman, A. W. (2010). Phylogeny of gammaproteobacteria. *J Bacteriol* **192**, 2305–2314.

Winslow, C. E., Broadhurst, J., Buchanan, R. E., Krumwiede, C., Rogers, L. A. & Smith, G. H. (1917). The families and genera of the Bacteria: Preliminary report of the Committee of the Society of American Bacteriologists on characterization and classification of bacterial types. *J Bacteriol* **2**, 505–566.

Woese, C. R. (1998). Default taxonomy: Ernst Mayr's view of the microbial world. *Proc Natl Acad Sci U S A* **95**, 11043–11046.

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N. N., Kunin, V., Goodwin, L., Wu, M. & other authors (2009). A phylogeny-driven genomic encyclopaedia of *Bacteria* and *Archaea*. *Nature* **462**, 1056–1060.

Yarza, P., Ludwig, W., Euzéby, J., Amann, R., Schleifer, K. H., Glöckner, F. O. & Rosselló-Móra, R. (2010). Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analyses. *Syst Appl Microbiol* **33**, 291–299.

Yishay, M., Burdman, S., Valverde, A., Luzzatto, T., Ophir, R. & Yedidia, I. (2008). Differential pathogenicity and genetic diversity among *Pectobacterium carotovorum* ssp. *carotovorum* isolates from monocot and dicot hosts support early genomic divergence within this taxon. *Environ Microbiol* **10**, 2746–2759.

Young, J. M. & Park, D. C. (2007). Relationships of plant pathogenic enterobacteria based on partial *atpD*, *carA*, and *recA* as individual and concatenated nucleotide and peptide sequences. *Syst Appl Microbiol* **30**, 343–354.

CHAPTER 3

Molecular Signatures (Conserved Indels) in Protein Sequences that are Specific for the Order *Pasteurellales* and Distinguish Two of Its Main Clades

This Chapter describes the identification of CSIs for the order *Pasteurellales*. These CSIs are the first reported molecular markers for this order. The identification of CSIs for two different clades of this order provided strong evidence for the division of *Pasteurellales* into different families. The division was also supported by phylogenetic trees. My contribution encompassed the performance of comparative genomic analysis and the construction of the phylogenetic trees highlighted in the methods section. In addition, I was involved in data analysis, in writing of the manuscript and the construction of the figures and tables.

*Due to limited space, supplementary figures (1-51) are not included in the chapter but can be accessed along with the rest of the manuscript at:

Naushad,H.S. and Gupta,R.S. (2012). Molecular signatures (conserved indels) in protein sequences that are specific for the order Pasteurellales and distinguish two of its main clades. Antonie Van Leeuwenhoek 101, 105-124.

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

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

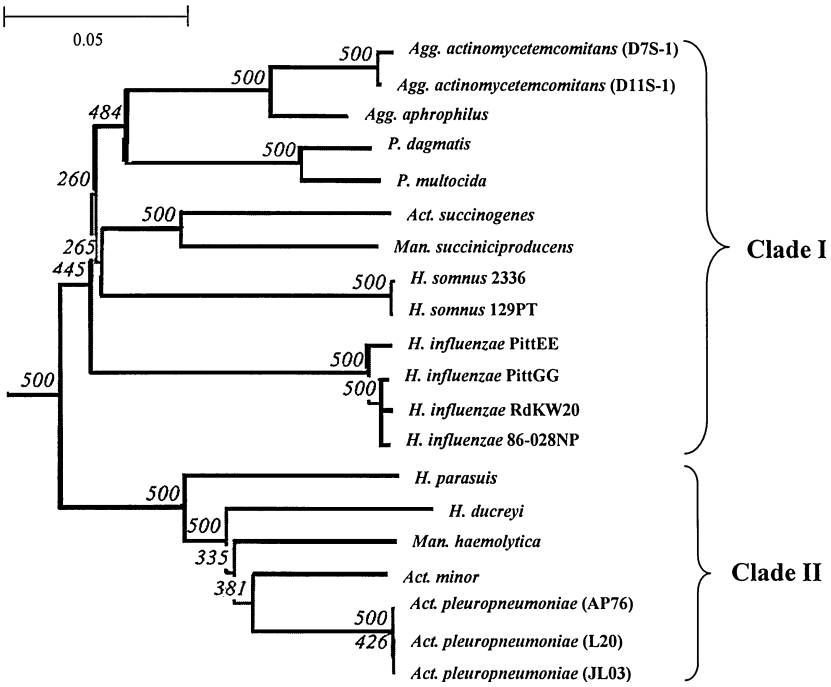


Fig. 1 A neighbor-joining distance tree for the sequenced Pasteurellales based upon concatenated sequences for 10 conserved proteins. The tree was rooted using sequences for other Gammaproteobacteria (viz. Vibrionales or Enterobacteriales) and the *numbers* on the nodes indicate the bootstrap values 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

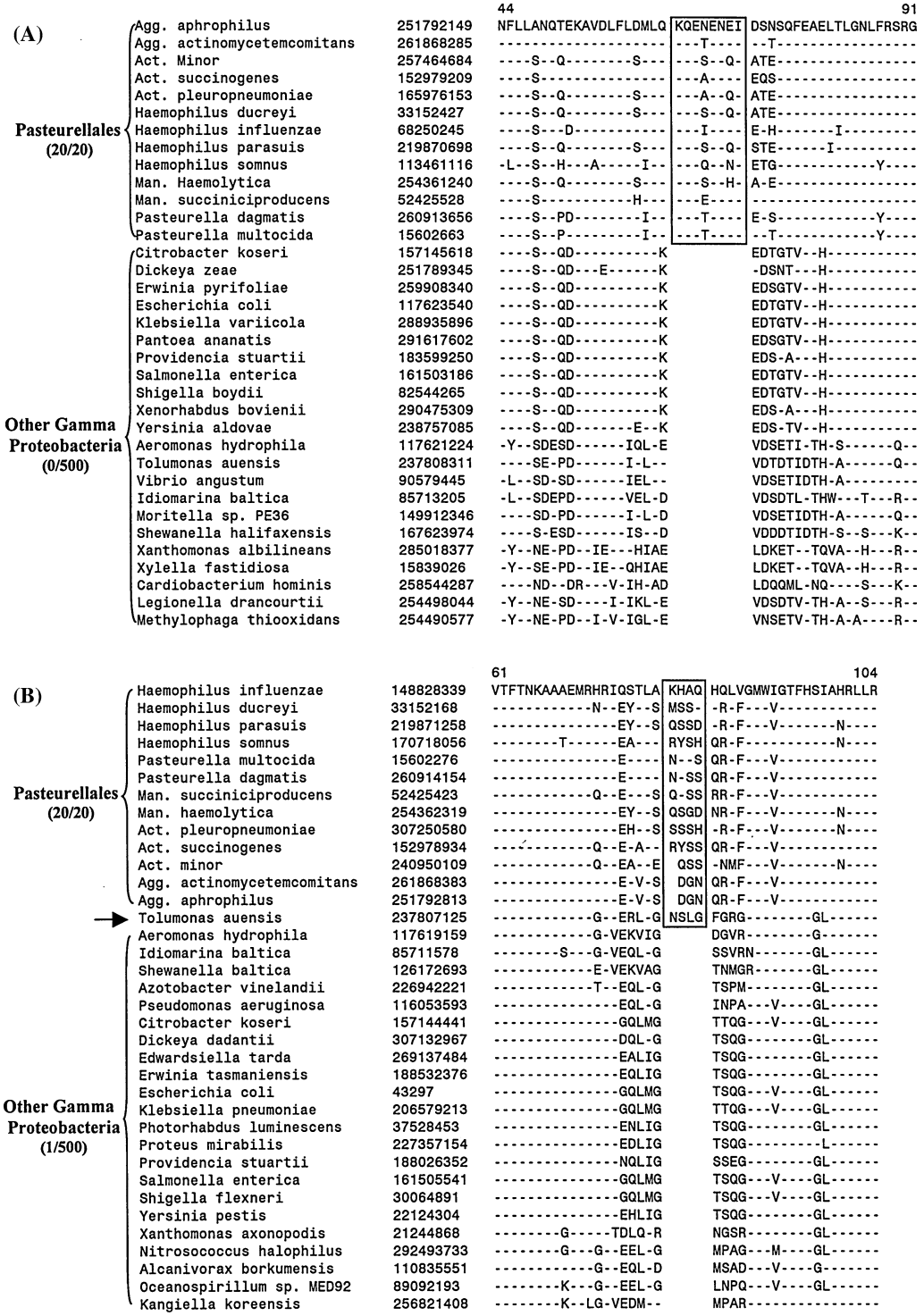
above eight species plus *Agg. aphrophilus* and *Actinobacillus succinogens*) based upon concatenated sequences for different orthologous proteins found in their genomes. Although, these trees provide useful resources for understanding the evolutionary relationships among the indicated *Pasteurellaceae* 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)
Pasteurellales (20/20)
Other Gamma Proteobacteria (0/500)
(B)
Pasteurellales (20/20)
Other Gamma Proteobacteria (1/500)

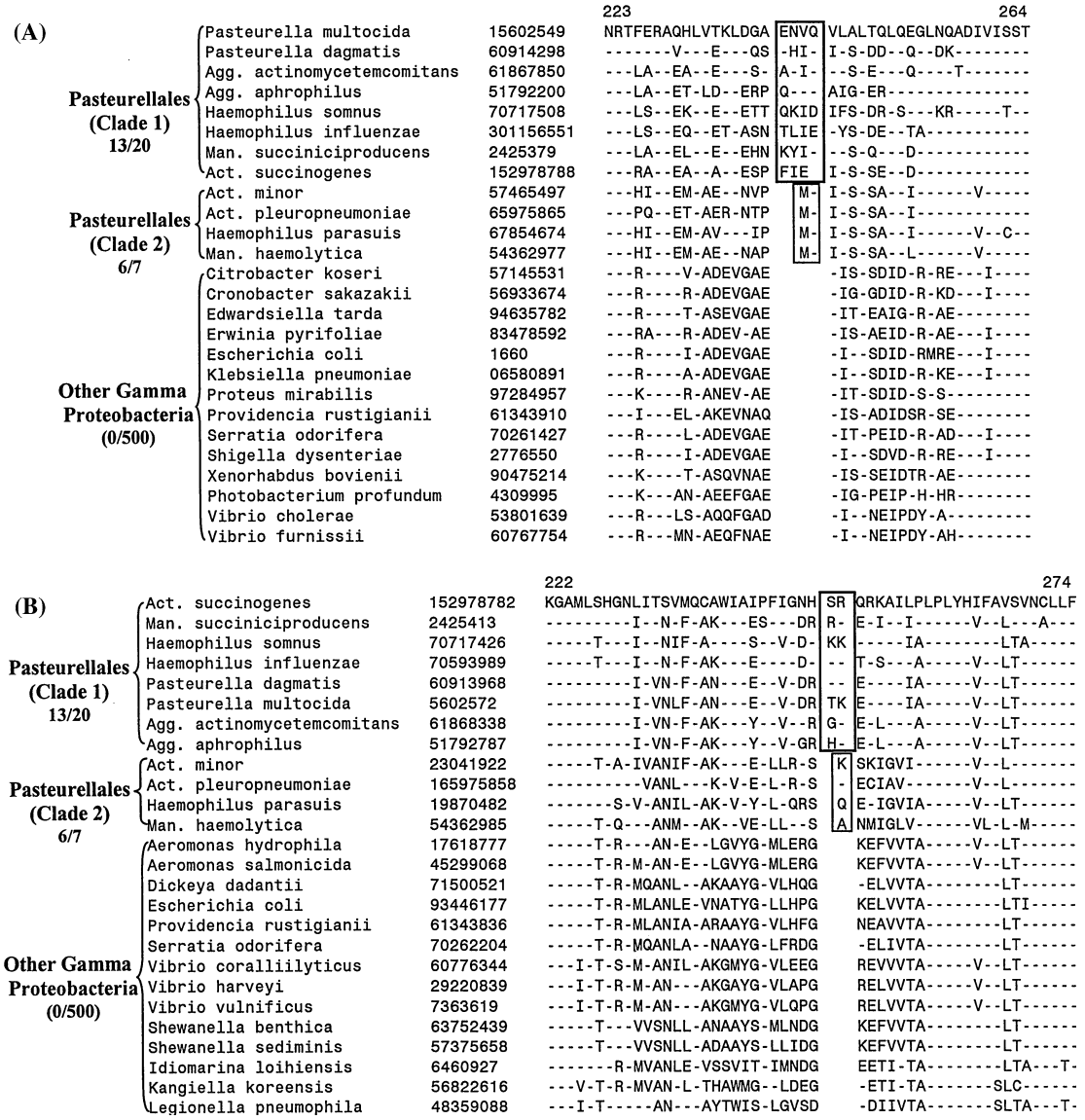


Fig. 3 Partial sequence alignments of **a** glutamyl-tRNA reductase and **b** long-chain-fatty-acid-CoA ligase, each containing two CSIs of different lengths (*boxed*) at the same positions that are specific for the two Pasteurellales clades. The *dashes* in the sequence alignments indicate identity 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*.

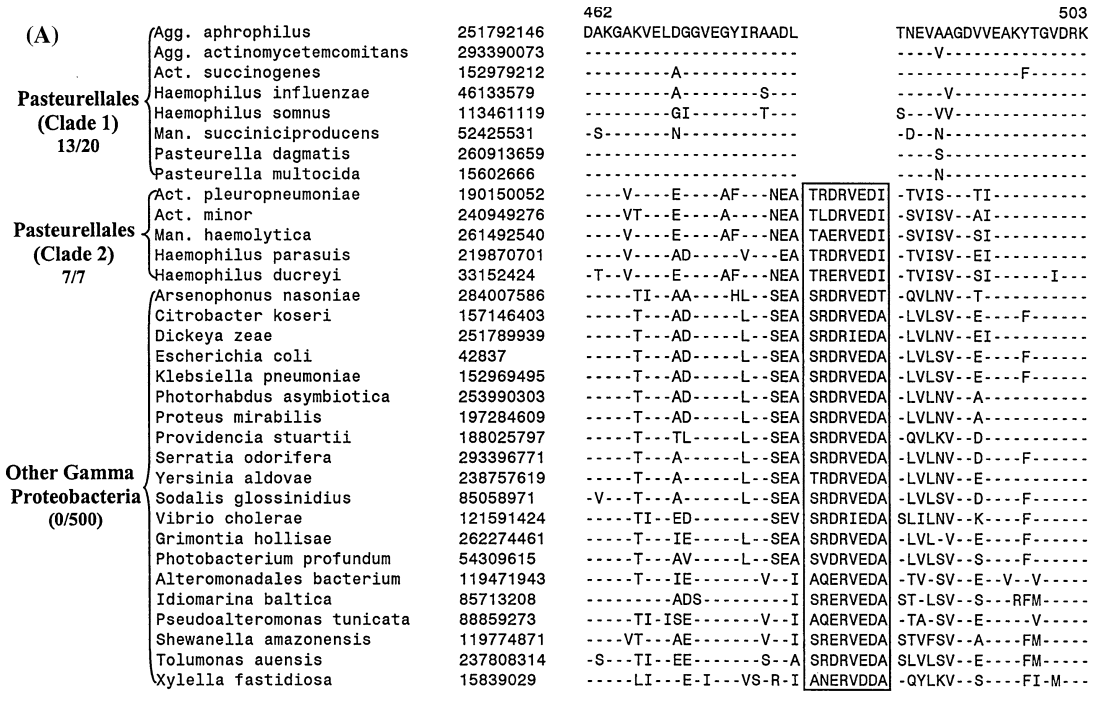
In addition to these CSIs that distinguish both Clades I and II species, we have also identified 11 CSIs in widely distributed proteins that are either uniquely or mainly found in the Clade I species (Table 3A). Two examples of such CSIs are presented in Fig. 4. In the universally distributed ribosomal 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.

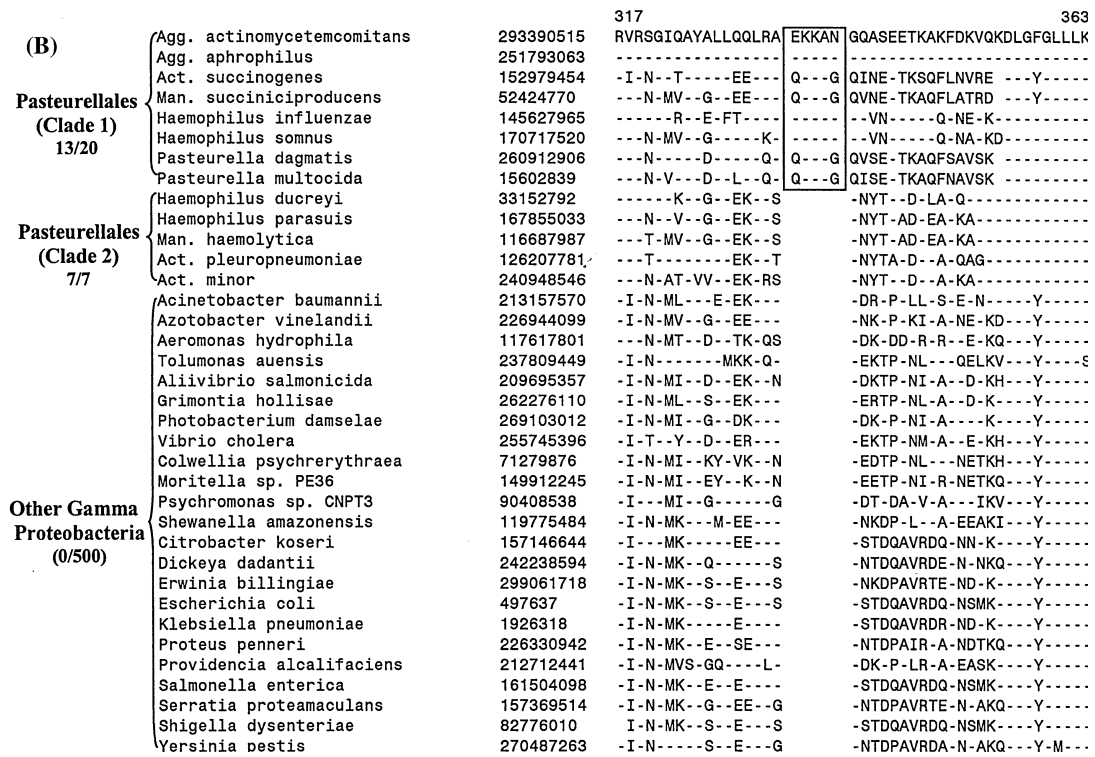
Four other proteins also contain CSIs that are largely specific for the Clade I. Within Clade I, *H. influenzae* shows deepest branching in the phylogenetic tree (Fig. 1). We have identified a 2 aa insert in the protein thiamine-monophosphate kinase that is commonly shared by all Clade I species except *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.

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).

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


(A)
Pasteurellales (Clade 1) 13/20
Pasteurellales (Clade 2) 7/7
Other Gamma Proteobacteria (0/500)
462
503
Agg. aphrophilus 251792146 DAKGAKVELDGGVEGYIRAADL TNEVAAGDVVEAKYTGVDRK
Agg. actinomycetemcomitans 293390073 ----V----
Act. succinogenes 152979212 ----------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)
Pasteurellales (Clade 1) 13/20
Pasteurellales (Clade 2) 7/7
Other Gamma Proteobacteria (0/500)
317
363
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

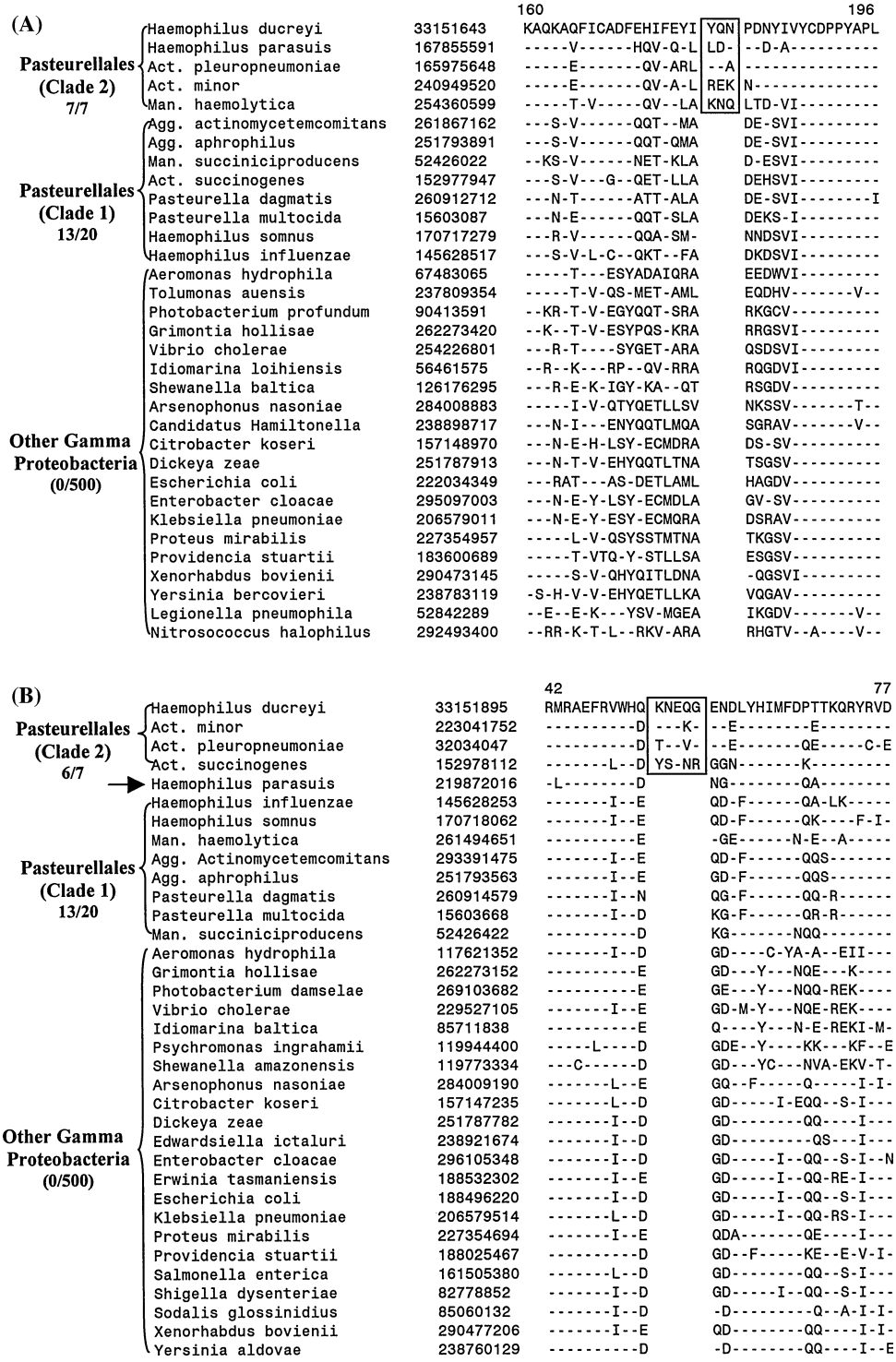

(A)
160
196
Pasteurellales (Clade 2) 7/7
Pasteurellales (Clade 1) 13/20
Other Gamma Proteobacteria (0/500)
Haemophilus ducreyi 33151643 KAQKAQFICADFEHIFEYI YQN PDNYIVYCDPPYAPL
Haemophilus parasuis 167855591 -----V------HQV-Q-L LD- --D-A----------
Act. pleuropneumoniae 165975648 -----E-------QV-ARL --A --------------
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*

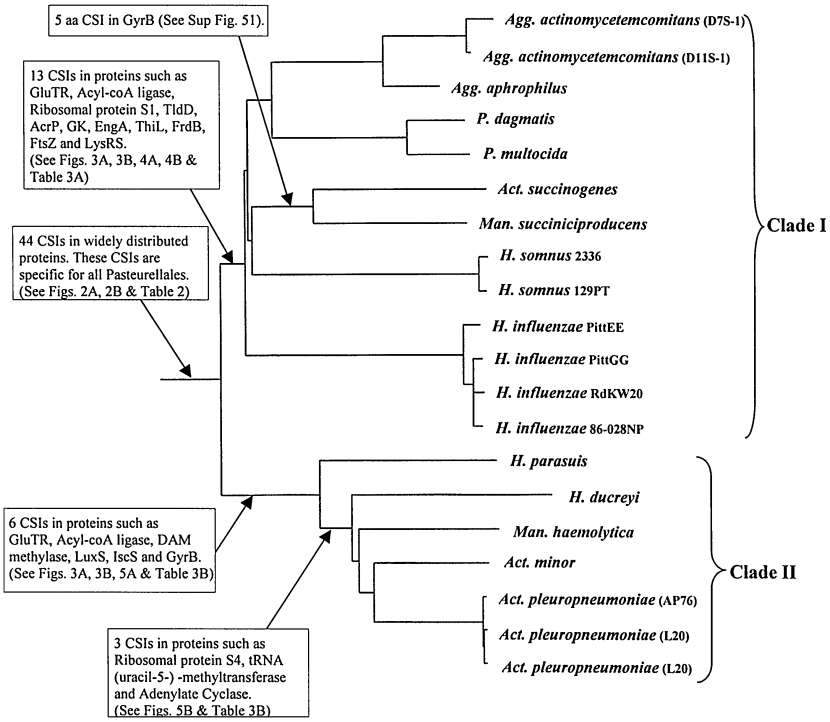


Fig. 6 A summary diagram showing the distribution patterns of various Pasteurellales-specific CSIs indicating the evolutionary relationships among Pasteurellales species. The different clades within this order that are supported by both phylogenetic studies and the identified molecular signatures are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

CHAPTER 4

Phylogenomic and Molecular Demarcation of the Core Members of the Polyphyletic *Pasteurellaceae* genera *Actinobacillus*, *Haemophilus*, and *Pasteurella*

*Due to limited space, supplementary figures and tables are not included in the chapter which are available along with the rest of the manuscript at:

Sohail Naushad, Mobolaji Adeolu, Nisha Goel, Aqeel Al-Dahwi, and Radhey S. Gupta (2015) International Journal of Genomics; Volume 2015 (2015), Article ID 198560, 15 pages

Preface

This chapter highlights the use of CSIs for the identification and rectification of different polyphyletic genera of the family *Pasteurellaceae*. The core members of the *Actinobacillus*, *Haemophilus*, and *Pasteurella* are identified into "*sensu stricto*" clades. The CSIs are also compared with phylogenetic trees to highlight the groupings of organisms within the family. The manuscript is currently under review process. My contribution towards the completion of this chapter encompassed the performance of comparative genomic analysis and the construction of the phylogenetic trees highlighted in the methods section. In addition, I was involved in analyzing the results, preparing the manuscript, and for the preparation of the figures and tables.

ABSTRACT

The family *Pasteurellaceae* contains a number of important human and animal pathogens housed within the genera *Actinobacillus, Aggregatibacter*, *Haemophilus, Mannheimia*, and *Pasteurella*. The genera *Actinobacillus, Haemophilus,* and *Pasteurella* exhibit extensive polyphyletic branching in phylogenetic trees and do not represent coherent clusters of species. In this study, we have utilized molecular signatures identified from comparative analyses of *Pasteurellaceae* genomes in conjunction with core genome based and multilocus sequence based phylogenetic analyses to clarify the phylogenetic and taxonomic boundary of these genera. We have identified large clusters of *Actinobacillus, Haemophilus,* and *Pasteurella* species which represent the "*sensu stricto*" members of these genera. We have identified 3, 7, and 6 unique molecular signatures, in the form of conserved signature indels (CSIs), which are specifically shared by members of the *Actinobacillus sensu stricto, Haemophilus sensu stricto,* and *Pasteurella sensu stricto*, respectively. We have also identified two different sets of 4 molecular signatures that are unique characteristics of the pathogen containing genera *Aggregatibacter* and *Mannheimia*, repectively. Based upon the CSIs identified in this work, it is now possible to demarcate the genera *Actinobacillus sensu stricto, Haemophilus sensu stricto,* and *Pasteurella sensu stricto* on the basis of discrete molecular signatures. The other members of the genera *Actinobacillus, Haemophilus,* and *Pasteurella* that do not fall within the "*sensu stricto*" clades and contain these molecular signatures should be reclassified as other genera. Additionally, the CSIs identified in this work serve as useful diagnostic targets for the development of highly specific diagnostic assays for current and novel members of the genera *Actinobacillus sensu stricto, Haemophilus sensu stricto, Pasteurella sensu stricto*, *Aggregatibacter* and *Mannheimia*.

INTRODUCTION

The family *Pasteurellaceae*, the single constituent family of the order *Pasteurellales*, represents a diverse group of commensal and pathogenic bacteria within the class *Gammaproteobacteria*. The family currently contains 19 genera, some of which are particularly important human and animal pathogens (Parte, 2013; Muehldorfer et al., 2014). The genera *Haemophilus* contains species responsible for human bacteremia, pneumonia, acute bacterial meningitis, and the sexually transmitted disease chancroid (Spinola et al., 2002; Christensen & Bisgaard, 2010; Nørskov-Lauritsen, 2014); *Aggregatibacter* species have been implicated in juvenile periodontitis (Henderson et al., 2010); members of the genera *Mannheimia*, *Pasteurella*, and *Actinobacillus* have been implicated in the causation of shipping fever in cattle, fowl cholera and pleuropneumonia in pigs, respectively (Angen et al., 1999; Bossé et al., 2002; Wilson & Ho, 2013).

The family *Pasteurellaceae* was originally proposed as a higher level taxonomic grouping of the related pathogenic genera *Actinobacillus, Haemophilus,* and *Pasteurella* (Pohl, 1979). Classification of organisms into these three genera was primarily based on DNA G-C content, and a handful of phenotypic traits (Mannheim et al., 1979). The phenotypic traits were later found not to be characteristic of any single genus (Dewhirst et al., 1992). Consequently, the genera *Actinobacillus, Haemophilus,* and *Pasteurella* each exhibit extensive polyphyly in subsequent 16S rRNA based phylogenies (Dewhirst et al., 1992; Olsen et al., 2005). Additional studies based on individual or concatenated gene sets and DNA-DNA/rRNA-DNA hybridization also support the presence of extensive polyphyly within the genera *Actinobacillus, Haemophilus,* and *Pasteurella* (Dewhirst et al., 1993; Christensen et al., 2004; Korczak et al., 2004; Kuhnert & Korczak, 2006; Christensen et al., 2007; Bonaventura et al., 2010; Naushad & Gupta, 2012).

Extensive work has been undertaken to amend the classification of the genera *Actinobacillus, Haemophilus,* and *Pasteurella* (Christensen et al., 2007; Parte, 2013; Wilson & Ho, 2013; Nørskov-Lauritsen, 2014). New genera have been created to house phylogenetically coherent clusters of *Actinobacillus, Haemophilus,* and *Pasteurella*. The species [*Actinobacillus*] *actinomycetemcomitans*, [*Haemophilus*] *aphrophilus*, [*Haemophilus*] *paraphrophilus* and [*Haemophilus*] *segnis* have been transferred to the genus *Aggregatibacter* (Nørskov-Lauritsen & Kilian, 2006); the species [*Haemophilus*] *paragallinarum*, [*Pasteurella*] *gallinarum*, [*Pasteurella*] *avium* and [*Pasteurella*] *volantium* have been transferred to the genus

Avibacterium (Blackall et al., 2005); the species [*Haemophilus*] *somnus* and [*Haemophilus*] *agni* have been transferred to the genus *Histophilus* (Angen et al., 2003); and the species [*Pasteurella*] *haemolytica* and [*Pasteurella*] *granulomatis* have been transferred to the genus *Mannheimia* (Angen et al., 1999). Additionally, some individual species within the genera *Actinobacillus, Haemophilus,* and *Pasteurella* that do not cluster with other members of their genus in phylogenetic trees have been moved or proposed to be moved to novel or neighbouring genera (viz. the transfer of the species [*Haemophilus*] *pleuropneumoniae* to the genus *Actinobacillus* (Pohl et al., 1983), the transfer of the species [*Pasteurella*] *anatis* to the genus *Gallibacterium* (Christensen et al., 2003), the transfer of the species [*Pasteurella*] *trehalosi* to the genus *Bibersteinia* (Blackall et al., 2007), the transfer of the species [*Pasteurella*] *ureae* to the genus *Actinobacillus* (Mutters et al., 1986), and the proposed transfer of the species [*Haemophilus*] *ducreyi* to a novel genus (Christensen & Kuhnert, 2012)). However, despite these changes, the classification of the genera *Actinobacillus, Haemophilus,* and *Pasteurella* is still problematic and each genera continues to contain members which exhibit polyphyletic branching (Kuhnert & Korczak, 2006; Christensen et al., 2007; Bonaventura et al., 2010; Naushad & Gupta, 2012; Nørskov-Lauritsen, 2014).

Multiple studies have attempted to define a core group of species which cluster around the nomenclatural type species of *Actinobacillus, Haemophilus,* or *Pasteurella* as the only true members of these genera (i.e. *sensu stricto*) (Hedegaard et al., 2001; Christensen et al., 2004; Korczak et al., 2004; Nørskov-Lauritsen et al., 2005; Olsen et al., 2005; Cattoir et al., 2006; Kuhnert & Korczak, 2006), but the taxonomy and phylogeny of these bacteria continue to remain inconclusive (Kilian, 2005; Naushad & Gupta, 2012; Nørskov-Lauritsen et al., 2012). Several methods have been employed for the demarcation of these genera, however, no simple method or criteria is available that can clearly delimit these genera. It has been suggested that genome based studies may provide reliable means of clarifying the evolutionary relationships of these bacteria (Nørskov-Lauritsen et al., 2012).

Since the availability of the first complete genome sequence of the *Haemophilus influenzae* (Fleischmann et al., 1995), a large number of genomes for the members of the family *Pasteurellaceae* have become available in public databases (Wattam et al., 2013; NCBI, 2014a). The availability of these genomes provides us with an opportunity to complete comprehensive genome scale phylogenetic analyses of the family *Pasteurellaceae*. These genome sequences

have also been utilized to carry out comparative genomic analyses to identify molecular signatures (viz. Conserved Signature Indels (CSIs) in various proteins), commonly shared by all or closely related subsets of species within the family *Pasteurellaceae*. On the basis of the molecular signatures identified from comparative analyses of *Pasteurellaceae* genomes in conjunction with core genome based and multilocus sequence based phylogenetic analyses, we have identified *sensu stricto* clades of *Actinobacillus, Haemophilus,* and *Pasteurella* that are supported by 3, 7, and 6 unique molecular signatures, respectively. We also report sets of molecular signatures that are unique characteristics of the pathogen containing genera *Aggregatibacter* and *Mannheimia*.

METHODS

Multilocus Sequence Analysis

Multilocus sequence analysis was completed for members of the family *Pasteurellaceae* using widely available nucleotide sequences of the 16S rDNA, *infB* (Translation initiation factor IF-2), *recN* (DNA repair protein), and *rpoB* (DNA-directed RNA polymerase subunit beta) genes which have been used, individually or as part of a set, in a number of previous phylogenetic analyses of the family *Pasteurellaceae* (Hedegaard et al., 2001; Christensen et al., 2004; Korczak et al., 2004; Nørskov-Lauritsen et al., 2005; Kuhnert & Korczak, 2006). Gene sequences for these four genes were obtained for 52 *Pasteurellaceae* strains, representing a large majority of the known *Pasteurellaceae* species, and 2 members of *Vibrio cholerae* from the NCBI nucleotide database (NCBI, 2014b). Species which were missing one of these four genes or which did not have a gene sequence that was at least 50% of the length of the full gene were excluded from the analysis. The four genes were individually aligned using MUSCLE (Edgar, 2004) and manually concatenated to create a combined dataset that contained 10 183 nucleotide long alignments. A maximum-likelihood tree based on 100 bootstrap replicates of this alignment was constructed using MEGA 6.0 (Tamura et al., 2013) while employing maximum composite likelihood substitution model.

Pasteurellaceae Core Genome Phylogenetic Tree

A phylogenetic tree of 76 *Pasteurellaceae* strains, rooted using 7 members of the family *Vibrionaceae*, based on the core genome of the family *Pasteurellaceae* was created for this

study. The core set of *Pasteurellaceae* proteins were identified using the UCLUST algorithm (Edgar, 2010) to identify widely distributed protein families with at least 30% sequence identity and 50% sequence length. Proteins families which were present in less than 50% of the input genomes were excluded from further analysis. Potentially paralogous sequences (additional proteins from the same organism in a single protein family) within the remaining protein families were also excluded from further analysis. Each protein family was individually aligned using MAFFT 7 (Katoh & Standley, 2013). Aligned amino acid positions which contained gaps in more than 50% of organisms were excluded from further analysis. The remaining amino acid positions were concatenated to create a combined dataset that contained 128 080 amino acid long alignments. An approximately maximum-likelihood tree based on this alignment was constructed using FastTree 2 (Price et al., 2010) while employing the Whelan and Goldman substitution model (Whelan & Goldman, 2001).

Identification of Molecular Signatures (CSIs) for different genera of the family *Pasteurellaceae*

The detailed outline of the process of identifying CSIs has been recently published (Gupta, 2014). In brief: Blastp searches were performed on all proteins from the genome of *Haemophilus influenzae* F3047 (Strouts et al., 2012). Ten to fifteen high scoring homologues that were present in *Haemophilus,* other *Pasteurellaceae*, and *Gammaproteobacteria* species were retrieved, and their multiple sequence alignments were constructed using Clustal X 1.83 (Jeanmougin et al., 1998). The alignments were visually inspected to identify any conserved inserts or deletions (indels) that are restricted to the particular clades of the family Pasteurellaceae, which are flanked on each side by at least 5–6 identical/conserved residues in the neighbouring 30–40 amino acids. The selected sequences containing the indels and their flanking conserved regions were further evaluated by detailed Blastp searches to determine species distribution and group specificity. The results of these Blast searches were processed using Sig_Create and Seq_Style to construct signature files (Gupta, 2014). Due to space constraints, the sequence alignment files presented here contain sequence information for a limited number of species within the order *Pasteurellaceae* and a representative selection of outgroup species. However, in each case, all members of the order and outgroups exhibited similar sequence characteristics to the representatives.

RESULTS AND DISCUSSION

Phylogenetic Analysis of the *Pasteurellaceae*

Elucidating an accurate phylogeny of the members of the family *Pasteurellaceae* has been a long standing challenge in *Pasteurellaceae* research (Mannheim et al., 1979; Pohl, 1979; Dewhirst et al., 1992; Christensen et al., 2007; Bonaventura et al., 2010). Early 16S rRNA based studies revealed that the established taxonomy of the family *Pasteurellaceae* was not consistent with their genetically inferred phylogeny (Dewhirst et al., 1992, 1993). This has led to a long series of taxonomic revisions within the family *Pasteurellaceae*; a process which is still taking place today (Angen et al., 1999; Blackall et al., 2005; Christensen et al., 2007; Christensen & Kuhnert, 2012). However, it was subsequently discovered that phylogenetic trees of *Pasteurellaceae* species based on different genes did not completely agree with each other (Christensen et al., 2004; Korczak et al., 2004; Cattoir et al., 2006). In particular, phylogenetic trees based on the 16S rRNA gene, often considered the gold standard in bacterial taxonomy and phylogeny (Stackebrandt & Ebers, 2006; Konstantinidis & Stackebrandt, 2013), disagreed with highly robust multilocus sequence and concatenated protein sequence based phylogenetic trees (Gioia et al., 2006; Kuhnert & Korczak, 2006; Redfield et al., 2006; Bonaventura et al., 2010; Naushad & Gupta, 2012; Wilson & Ho, 2013).

Phylogenetic trees based on concatenated sequences for a large number of unlinked and conserved loci are more reliable and robust than phylogenetic trees based on any single gene or protein (Rokas et al., 2003; Wu et al., 2009). Due to a rapid increase in the availability of genomic sequence data, we are now able to complete genome scale phylogenetic analyses of the family *Pasteurellaceae* which cover a vast majority of the diversity within the family. In this work we have produced a phylogenetic tree for 74 genome sequenced members of the family *Pasteurellaceae* based on 128 080 aligned amino acid positions (Figure 1A). The branching patterns of the core genome phylogenetic tree produced in this work largely agree with a previous genome based phylogenetic tree produced for a limited number of *Pasteurellaceae* species (Bonaventura et al., 2010) and a concatenated protein based phylogenetic tree of the family *Pasteurellaceae* produced by our lab in a previous study (Naushad & Gupta, 2012). Additionally, we have also produced a multilocus sequence based phylogenetic tree using the 16S rDNA, *infB*, *recN*, and *rpoB* genes which are commonly used in the phylogenetic analysis of

the family *Pasteurellaceae* (Figure 1B) (Hedegaard et al., 2001; Christensen et al., 2004; Korczak et al., 2004; Nørskov-Lauritsen et al., 2005; Kuhnert & Korczak, 2006). This tree also showed broadly similar branching patterns to past multilocus sequence based phylogenetic trees (Kuhnert & Korczak, 2006; Christensen et al., 2007) and to our core genome based phylogenetic tree. Both our core genome based and multilocus sequence based phylogenetic trees provide evidence for a division of the *Pasteurellaceae* into at least two higher taxonomic groups (families) which are broadly similar to the two clades of *Pasteurellales* identified in our previous work (Naushad & Gupta, 2012). A similar division of the family *Pasteurellaceae* into two or more large groups is seen in many other robust multilocus or concatenated protein based phylogenetic trees (Gioia et al., 2006; Kuhnert & Korczak, 2006; Redfield et al., 2006; Bonaventura et al., 2010), however, this division is not readily apparent in phylogenies based on the 16S rRNA gene (Wilson & Ho, 2013; Yilmaz et al., 2013).

A majority of the known genera within the family *Pasteurellaceae* form well-defined and coherent clusters in phylogenetic trees (Figure 1) (Kuhnert & Korczak, 2006; Bonaventura et al., 2010; Naushad & Gupta, 2012; Wilson & Ho, 2013; Yilmaz et al., 2013). The genera *Actinobacillus, Haemophilus,* and *Pasteurella*, which were described before the advent of genetic characterization, exhibit polyphyletic branching in all gene and protein based phylogenetic trees, including the core genome based and multilocus sequence based phylogenetic trees created in this work (Figure 1). However, there are large clusters of *Actinobacillus, Haemophilus,* and *Pasteurella* species identifiable in the phylogenetic trees which represent the core or "*sensu stricto*" members of each genera. The clusters of species that represent *Actinobacillus sensu stricto, Haemophilus sensu stricto,* and *Pasteurella sensu stricto* are indicated in Figure 1. Members of each genera which fall outside of the *sensu stricto* clusters, indicated in our phylogenetic trees by the presence of square brackets around their genus name (ex. [*Pasteurella*] *pneumotropica*), are only distantly related to the *sensu stricto* members of their genus and will require reclassification in order to make their taxonomy and phylogeny concordant.

The Usefulness of Conserved Signature Indels as Phylogenetic and Taxonomic Markers

Whole genome sequences are a rich resource for the discovery of molecular signatures which are unique to a group of organisms (Gao et al., 2009; Cutino-Jimenez et al., 2010;

Naushad & Gupta, 2013). One useful class of shared molecular signatures are Conserved Signature Indels (CSIs), which are insertions/deletions uniquely present in protein sequences from a group of evolutionarily related organisms (Gupta, 2010; Gupta, 2014; Naushad et al., 2014). The unique, shared presence of multiple CSIs by a group of related species is most parsimoniously explained by the occurrence of the genetic changes that resulted in these CSIs in a common ancestor of the group, followed by vertical transmission of these CSIs to various descendant species (Gupta, 1998; Rokas & Holland, 2000; Gupta, 2014; Naushad et al., 2014). Hence, these CSIs represent molecular synapomorphies (markers of common evolutionary decent) which can be used to identify and demarcate specific bacterial groups in molecular terms and for understanding their interrelationships independently of phylogenetic trees (Gupta, 1998, 2010; Gupta, 2014; Naushad et al., 2014). CSIs have recently been used to propose important taxonomic changes for a number of bacterial groups (viz. Aquificae, Spirochaetes, Thermotogae, *Xanthomonadales*, and *Borrelia*) at different taxonomic ranks (Gupta & Lali, 2013; Gupta et al., 2013; Naushad & Gupta, 2013; Adeolu & Gupta, 2014; Bhandari & Gupta, 2014). In the present work, we have completed comprehensive comparative analysis of *Pasteurellaceae* genomes (Table 1) in order to identify CSIs that are primarily restricted to the different genera within the family *Pasteurellaceae*. We have identified 3, 7, and 6 unique molecular signatures which are shared by *Actinobacillus sensu stricto, Haemophilus sensu stricto,* and *Pasteurella sensu stricto*, respectively. Information regarding these CSIs and their evolutionary significances are discussed below.

Molecular signatures specific for *Actinobacillus sensu stricto*

The genus *Actinobacillus* was originally defined as a group of growth factor independent host-associated rods which shared phenotypic or biochemical similarity with *Actionbacillus lignieresii*, the type species of the genus (Pohl et al., 1983; Olsen, 1993). However, the original classification scheme for the genus *Actionbacillus* led to the inclusion of a highly heterogeneous and polyphyletic grouping of species within the genus (Dewhirst et al., 1992, 1993; Olsen et al., 2005). An assemblage of *Actionbacillus* species closely related to *Actionbacillus lignieresii* has been recognized as *Actinobacillus sensu stricto* (i.e. the core members of the genus *Actionbacillus*) in both our phylogenetic analysis (Figure 1) and past phylogenetic analyses (Dewhirst et al., 1992, 1993; Olsen et al., 2005; Kuhnert & Korczak, 2006). Differentiation of

Actinobacillus sensu stricto from other *Actionbacillus* species and the modern criteria for placing novel species within the genus *Actionbacillus sensu stricto* is heavily reliant on genetic and genomic criteria, namely, DNA-DNA hybridization values, 16S rRNA sequence similarity, and other single gene sequence comparisons (Olsen et al., 2005; Christensen et al., 2007). There are currently no known discrete characteristics which are unique to *Actionbacillus* that define the genus. In this work, we have completed a comprehensive comparative analysis of *Pasteurellaceae* genomes in order to identify unique, defining molecular signatures for different genera within the family *Pasteurellaceae*. We have identified 3 CSIs which are unique, defining molecular signatures for the sequenced members of *Actionbacillus sensu stricto* (viz. *Actionbacillus capsulatus, A. pleuropneumoniae, A. suis,* and *A. ureae*). An example of a CSI specific for *Actionbacillus sensu stricto* is shown in Figure 2. The CSI consists of a 1 amino acid insertion in a conserved region of a 3'-nucleotidase which is present in all sequenced members of *Actionbacillus sensu stricto* and absent in all other sequenced *Gammaproteobacteria*. Sequence information for 2 other CSIs which are also unique characteristics of the *Actionbacillus sensu stricto* clade are presented in Supplemental Figure 1 - 2 and their characteristics are briefly summarized in Table 2A.

Molecular signatures specific for *Haemophilus sensu stricto*

The classification of novel species into the genus *Haemophilus* was initially based on phenotypic and biochemical properties, most importantly, the dependence of growth on the presence of factor V and factor X in blood (Olsen, 1993; Olsen et al., 2005; Hayashimoto et al., 2007). As with *Actinobacillus*, the classification of *Haemophilus* on the basis of phenotypic and biochemical properties has led to the genus containing an extremely heterogeneous group of species (Dewhirst et al., 1992, 1993; Kilian, 2005; Olsen et al., 2005). Species from the genus *Haemophilus* have undergone a number of transfers and reclassifications (Pohl et al., 1983; Angen et al., 2003; Blackall et al., 2005; Nørskov-Lauritsen & Kilian, 2006; Christensen & Kuhnert, 2012). However, the genus remains highly polyphyletic (Figure 1) (Kuhnert & Korczak, 2006; Bonaventura et al., 2010; Christensen & Kuhnert, 2012).The core members of the genus *Haemophilus* (viz. *Haemophilus sensu stricto*) consists of *Haemophilus influenzae, H. aegyptius*, and *H. haemolyticus* based on 16S rRNA sequence analysis (Dewhirst et al., 1992, 1993; Kilian, 2005; Olsen et al., 2005). However, phylogenetic analysis based on DNA-DNA

hybridization and multilocus sequence analysis suggests that *H. parainfluenzae* and *H. pittmaniae* are also members of *Haemophilus sensu stricto* (Mutters et al., 1989; Nørskov-Lauritsen et al., 2005). Phylogenetic analysis of *rpoB*, *infB*, and concatenated gene sets also suggest that [*Pasteurella*] *pneumotropica* and related isolates are closely related to *Haemophilus sensu stricto* (Christensen et al., 2004; Korczak et al., 2004).

Our comparative analysis of *Pasteurellaceae* genomes has led to the identification of 7 CSIs that are unique characteristics of *Haemophilus sensu stricto* which consists of *Haemophilus influenzae, H. aegyptius, H. haemolyticus, H. parainfluenzae, H. pittmaniae*, and [*Pasteurella*] *pneumotropica* (Figure 1). One example of a CSI specific for the members of *Haemophilus sensu stricto*, shown in Figure 3, consists of a 4 amino acid deletion in a biotin-protein ligase which is uniquely found in homologs from *Haemophilus sensu stricto* and absent in all other sequenced *Gammaproteobacteria*. Sequence information for 6 additional CSIs which are also unique characteristics of *Haemophilus sensu stricto* are presented in Supplemental Figure 3 - 8 and their characteristics are briefly summarized in Table 2B. These CSIs and our phylogenetic trees (Figure 1) suggest that *Haemophilus influenzae, H. aegyptius, H. haemolyticus, H. parainfluenzae, H. pittmaniae*, and [*Pasteurella*] *pneumotropica* share a close evolutionary relationship and should all be considered members of *Haemophilus sensu stricto*. Additionally, these results also suggest that [*Pasteurella*] *pneumotropica* is incorrectly classified as a member of the genus *Pasteurella* and should be reclassified as "*Haemophilus pneumotropica*".

Molecular signatures specific for *Pasteurella sensu stricto*

The genus *Pasteurella* is highly heterogeneous and polyphyletic (Figure 1) (Olsen et al., 2005). Similar to the members of *Actinobacillus*, bacterial isolates were originally classified as members of the genus *Pasteurella* based on growth factor independent growth and phenotypic or biochemical similarity to *Pasteurella multocida*, the type species of the genus (Snipes & Biberstein, 1982; Olsen, 1993). The monophyletic cluster of *Pasteurella* species that branch with *Pasteurella multocida* are considered the core members of the genus (viz. *Pasteurella sensu stricto*) (Korczak et al., 2004; Olsen et al., 2005; Kuhnert & Korczak, 2006; Wilson & Ho, 2013). Our comparative analysis of *Pasteurellaceae* genomes has led to the identification of 6 CSIs which are unique characteristics for the sequenced members of *Pasteurella sensu stricto* (viz. *Pasteurella multocida* and *P. dagmatis*). An example of a CSI uniquely found in the

sequenced members of *Pasteurella sensu stricto*, consisting of a 4 amino acid insertion in a conserved region of Menaquinone-specific isochorismate synthase, is shown in Figure 4. This CSI is only found in the sequenced members of *Pasteurella sensu stricto* and is absent from all other sequenced *Gammaproteobacteria*. Partial sequence alignments for 5 additional CSIs which are also unique characteristics of *Pasteurella sensu stricto* are presented in Supplemental Figures 9 - 13 and their characteristics are briefly summarized in Table 2C.

Molecular signatures specific for the genera *Aggregatibacter* or *Mannheimia*

The genus *Aggregatibacter* was proposed as a novel taxonomic classification for a monophyletic cluster of *Actinobacillus* and *Haemophilus* species which branched distinctly from the "*sensu stricto*" members of their respective clades (Nørskov-Lauritsen & Kilian, 2006). Similarly, the genus *Mannheimia* was proposed as a novel classification for the *Pasteurella Haemolytica* complex which did not branch with *Pasteurella sensu stricto* in phylogenetic trees (Angen et al., 1999). Currently other than branching in phylogenetic trees or relatedness in DNA-DNA hybridization studies, the members of the genera *Aggregatibacter* or *Mannheimia* do not share any single unique or defining biochemical or molecular characteristic that can differentiate them from all other bacteria (Angen et al., 2002; Nørskov-Lauritsen, 2014).

In this study we have identified 4 CSIs that are unique molecular characteristics shared by all sequenced species of the genus *Aggregatibacter* and another 4 CSIs which are uniquely found in all sequenced members of the genus *Mannheimia*. Examples of CSIs specific to the sequenced members of *Aggregatibacter* and *Mannheimia* are shown in Figure 5. A partial sequence alignment of a *nhaC* family sodium:proton antiporter containing a 3 amino acid insertion specific for all sequenced species of the genus *Aggregatibacter* is shown in Figure 5A and a partial sequence alignment of a methyl-galactoside ABC transporter substrate-binding protein containing a 1 amino acid deletion specific for all sequenced species of the genus *Mannheimia* is shown in Figure 5B. In each case, the identified CSIs were only found in the sequenced members of the genera *Aggregatibacter* or *Mannheimia* and were absent from all other sequenced *Gammaproteobacteria*. Partial sequence alignments additional CSIs specific for the genera *Aggregatibacter* or *Mannheimia* are provided in Supplemental Figures 14 - 19 and their characteristics are summarized in Table 2D - 2E. These CSIs are the first discrete molecular characteristics which are unique for the genera *Aggregatibacter* and *Mannheimia* and support

their observed monophyly in phylogenetic trees. Additionally, these CSIs could be useful targets for the development of PCR based diagnostic assays for the genera *Aggregatibacter* and *Mannheimia* which amplify the CSI containing DNA segment using the conserved flanking regions of the CSIs (Ahmod et al., 2011; Wong et al., 2014).

CONCLUSION

The genera *Actinobacillus, Haemophilus,* and *Pasteurella*, within the family *Pasteurellaceae*, are known to exhibit extensive polyphyletic branching. We have utilized molecular signatures and phylogenetic analyses to clarify the taxonomic boundary of these genera. We have been able to identify large clusters of *Actinobacillus, Haemophilus,* and *Pasteurella* species which represent the "*sensu stricto*" members of these genera. We have identified 3, 7, and 6 unique molecular signatures which are specifically shared by the members of the genera *Actinobacillus sensu stricto, Haemophilus sensu stricto,* and *Pasteurella sensu stricto*, respectively. The group specificity of the molecular signatures we have identified in this work are summarized in Figure 6 and their characteristics are briefly summarized in Tables 2. Our comparative genomic analyses have not come across any CSIs that were unique characteristics of all sequenced members of the genera *Actinobacillus, Haemophilus,* or *Pasteurella* as currently defined, suggesting that the members of these genera that do not fall into the "*sensu stricto*" clusters should not be considered members their respective genus.

Examinations of phenotypic and biochemical characteristics do not provide a reliable means of assigning a novel isolate to the genera *Actinobacillus, Haemophilus,* and *Pasteurella* (Christensen et al., 2007). However, based upon the CSIs described in this work, it is now possible to demarcate the genera *Actinobacillus sensu stricto, Haemophilus sensu stricto,* and *Pasteurella sensu stricto* on the basis of the presence or absence of unique molecular signatures. It is important to note that the current analysis of CSIs is limited to the currently available genomic sequence data and may show slight variance as additional bacterial genomes are sequenced. However, earlier work on CSIs for other groups of bacteria provides evidence that the identified CSIs have strong predictive value and will likely be found in other members of these groups as more species are sequenced and novel species are isolated (Gao & Gupta, 2012; Gupta & Lali, 2013; Bhandari & Gupta, 2014; Howard-Azzeh et al., 2014). The conserved nature of the sequence regions that contain these CSIs, in conjunction with their strong predictive

value, makes CSIs promising targets for the development of highly specific diagnostic assays for *Actinobacillus sensu stricto, Haemophilus sensu stricto, Pasteurella sensu stricto*, *Aggregatibacter* and *Mannheimia* (Ahmod et al., 2011; Wong et al., 2014). Additionally, further analysis of these genus specific CSIs should lead to the discovery of their functional role in their respective organisms and may provide important insights into novel distinguishing features of these groups of organisms.

TABLE 1

Genome characteristics of the sequenced *Pasteurellaceae* included in our analyses

Organism name	BioProject	Size (Mb)	Proteins	G-C (%)	References
Actinobacillus pleuropneumoniae L20	CP000569	2.27	2013	41.3	(Foote et al., 2008)
Actinobacillus pleuropneumoniae serovar 3 str. JL03	CP000687	2.24	2036	41.2	(Xu et al., 2008)
Actinobacillus pleuropneumoniae serovar 7 str. AP76	CP001091	2.35	2142	41.2	STHH[b] (2008)
Actinobacillus ureae ATCC 25976[a]	AEVG0	2.30	2475	-	BCM[g] (2011)
Actinobacillus minor 202[a]	ACFT0	2.13	2050	39.3	McGill University (2010)
Actinobacillus minor NM305[a]	ACQL0	2.43	2411	39.3	McGill University (2010)
Actinobacillus succinogenes 130Z	CP000746	2.32	2079	44.9	DOE-JGIB (2007)
Aggregatibacter aphrophilus NJ8700	CP001607	2.31	2219	42.2	Di Bonaventural et al, 2009
Aggregatibacter actinomycetemcomitans D11S-1	CP001733	2.20	2280	44.3	(Chen et al., 2009)
Aggregatibacter actinomycetemcomitans D7S-1	CP003496	2.31	2250	44.3	(Chen et al., 2010)
Aggregatibacter segnis ATCC 33393[a]	AEPS0	1.99	1956	-	BCM[g] (2010)
Gallibacterium anatis UMN179	CP002667	2.69	2500	39.9	Johnson et al, 2011
Haemophilus aegyptius ATCC 11116[a]	AFBC0	1.92	2020	-	BCM[g] (2011)
Haemophilus ducreyi 35000HP	AE017143	1.70	1717	38.2	Ohio State University (2003)
Haemophilus haemolyticus M21621[a]	AFQQ0	2.09	1894	-	(Jordan et al., 2011)
Haemophilus influenzae 10810	FQ312006	1.98	1903	38.1	WTSI[h] (2010)
Haemophilus influenzae F3031	FQ670178	1.99	1770	38.2	(Strouts et al., 2012)
Haemophilus influenzae F3047	FQ670204	2.01	1786	38.2	(Strouts et al., 2012)
Haemophilus influenzae 22.1-21[a]	AAZD0	1.89	2224	38.0	(Hogg et al., 2007)
Haemophilus influenzae 3655	AAZF0	1.88	1929	38.0	(Hogg et al., 2007)
Haemophilus influenzae 6P18H1[a]	ABWW0	1.91	1893	38.2	CGS, ASRI[e] (2008)
Haemophilus influenzae 7P49H1[a]	ABWV0	1.83	1752	37.9	CGS, ASRI[e] (2008)
Haemophilus influenzae NT127[a]	ACSL0	1.87	1809	38.0	BIGSP[c] (2009)
Haemophilus influenzae PittAA[a]	AAZG0	1.88	1981	38.1	(Hogg et al., 2007)
Haemophilus influenzae PittII[a]	AAZI0	1.95	2028	38.0	(Hogg et al., 2007)
Haemophilus influenzae PittHH[a]	AAZH0	1.84	1977	38.0	(Hogg et al., 2007)
Haemophilus influenzae R3021[a]	AAZJ0	1.88	2307	37.9	(Hogg et al., 2007)
Haemophilus influenzae RdAW[a]	ACSM0	1.80	1718	38.0	BIGSP[c] (2009)
Haemophilus influenzae 86-028NP	CP000057	1.91	1792	38.2	(Harrison et al., 2005)
Haemophilus influenzae PittEE	CP000671	1.81	1613	38.0	(Hogg et al., 2007)
Haemophilus influenzae PittGG	CP000672	1.89	1661	38.0	(Hogg et al., 2007)
Haemophilus influenzae Rd KW20	L42023	1.83	1657	38.2	(Fleischmann et al., 1995)
Haemophilus influenzae R2846	CP002276	1.82	1636	38.0	UW-SBRI[d] (2004)
Haemophilus influenzae R2866	CP002277	1.93	1795	38.1	UW-SBRI[d] (2004)
Haemophilus parainfluenzae ATCC 33392 [a]	AEWU0	2.11	2010	-	BCM[g] (2011)
Haemophilus parainfluenzae T3T1	FQ312002	2.09	1975	39.6	WTSI[h] (2011)
Haemophilus parasuis 29755[a]	ABKM0	2.22	2244	39.8	Iowa State University (2008)
Haemophilus parasuis SH0165	CP001321	2.27	2021	40.0	(Yue et al., 2009)
Haemophilus pittmaniae HK 85[a]	AFUV0	2.18	2390	-	J. Craig Venter Institute (2011)
Haemophilus sputorum CCUG13788[a]	AFNK0	2.14	2073	-	Aarhus University Hospital (2011)
Haemophilus parahaemolyticus HK385[a]	AJSW0	1.81	1764	-	J. Craig Venter Institute (2011)
Haemophilus paraphrohaemolyticus HK411[a]	AJMU0	2.02	2025	-	J. Craig Venter Institute (2011)
Haemophilus sp. oral taxon 851 str. F0397[a]	AGRK0	1.84	1809	-	GCG-WU[f] (2012)
Histophilus somni 2336	CP000947	2.26	1980	37.4	DOE-JGIB (2010)
Histophilus somni 129PT	CP000436	2.01	1798	37.2	(Challacombe et al., 2007)
Mannheimia succiniciproducens MBEL55E	AE016827	2.31	2370	42.5	(Hong et al., 2004)
Mannheimia haemolytica PHL213[a]	AASA0	2.57	2695	41.1	(Gioia et al., 2006)
Pasteurella multocida subsp. *multocida* str. *Pm70*	AE004439	2.26	2012	40.4	(May et al., 2001)
Pasteurella dagmatis ATCC 43325[a]	ACZR0	2.25	2053	37.4	BCM[g] (2009)

[a] The genomes of these species/strains are currently under scaffolds/contigs status
[b] Stiftung Tieraerztliche Hochschule Hannover (STHH)
[c] The Broad Institute Genome Sequencing Platform (BIGSP)
[d] University of Washington; Seattle Biomedical Research Institute (UW-SBRI)
[e] Center for Genomic Sciences, Allegheny-Singer Research Institute (CGS, ASRI)
[f] Genome Sequencing Center (GSC) at Washington University (WashU) School of Medicine
[g] Baylor College of Medicine (BCM)
[h] Wellcome Trust Sanger Institute (WTSI)

TABLE 2

Conserved signature indels specific for genera within the family *Pasteurellaceae*

Protein name	Gene Name	GenBank Identifier	Figure no.	Indel Size	Indel Position[a]
(A) CSIs Specific for *Actinobacillus sensu stricto*					
3'-nucleotidase	*surE*	126208128	**Fig. 2**	1 aa ins	367-402
GTP pyrophosphokinase	*relA*	126207889	Sup. Fig 1	1 aa ins	368-412
Anaerobic glycerol-3-phosphate dehydrogenase subunit	*glpA*	491834528	Sup. Fig 2	1 aa ins	359-400
(B) CSIs Specific for *Haemophillus sensu stricto*					
Biotin-protein ligase	*birA*	144979005	**Fig. 3**	6 aa del	138-178
Aspartate ammonia-lyase	*aspA*	145630289	Sup. Fig 3	1 aa ins	34-75
NAD(P) transhydrogenase subunit alpha	*pntA*	145631394	Sup. Fig 4	1 aa del	352-378
Fumarate reductase subunit C	*frdC*	301169552	Sup. Fig 5	3 aa ins	31-89
Hypothetical tRNA/rRNA methyltransferase	-	145636352	Sup. Fig 6	1 aa del	17-58
Gamma-glutamyl kinase	*proB*	145629980	Sup. Fig 7	1 aa ins	197-253
ACP phosphodieterase	*acpD*	68250119	Sup. Fig 8	2 aa del	119-159
(C) CSIs Specific for *Pasteurella sensu stricto*					
Menaquinone-specific isochorismate synthase	*menF*	386834899	**Fig. 4**	4 aa ins	29-86
tRNA s(4)U8 sulfurtransferase	*thiI*	15602400	Sup. Fig 9	2 aa del	412-446
FKBP-type peptidyl-prolyl cis-trans isomerase	*slyD*	378775595	Sup. Fig 10	2 aa del	151-188
Aspartate-semialdehyde dehydrogenase	*asd*	383311492	Sup. Fig 11	1 aa del	173-245
Lactate permease family transporter	*lldP*	492154065	Sup. Fig 12	2 aa ins	390-427
Cell division protein *ftsA*	*ftsA*	492155843	Sup. Fig 13	1 aa ins	357-387
(D) CSIs Specific for *Aggregatibacter*					
nhaC family sodium:proton antiporter	*nhaC*	493769836	**Fig. 5A**	3 aa ins	396-437
Outer membrane protein	*omp*	261866907	Sup. Fig 14	4 aa del	25-64
Multidrug transporter *murJ*	*murJ*	365966332	Sup. Fig 15	1 aa del	190-220
NADH dehydrogenase	*nuoE*	387120244	Sup. Fig 16	1 aa ins	372-412
(E) CSIs Specific for *Mannheimia*					
Methyl-galactoside ABC transporter substrate-binding protein	-	472335016	**Fig. 5B**	1 aa del	33-73
UDP-N-acetylmuramoylalanyl-D-glutamate--2,6-diaminopimelate ligase	*murE*	472333011	Sup. Fig 17	2 aa del	418-473
Glutathione-regulated potassium-efflux protein	*kefC*	472333189	Sup. Fig 18	1 aa ins	504-531
Glycerol-3-phosphate acyltransferase	*plsB*	472334521	Sup. Fig 19	2 aa del	214-252

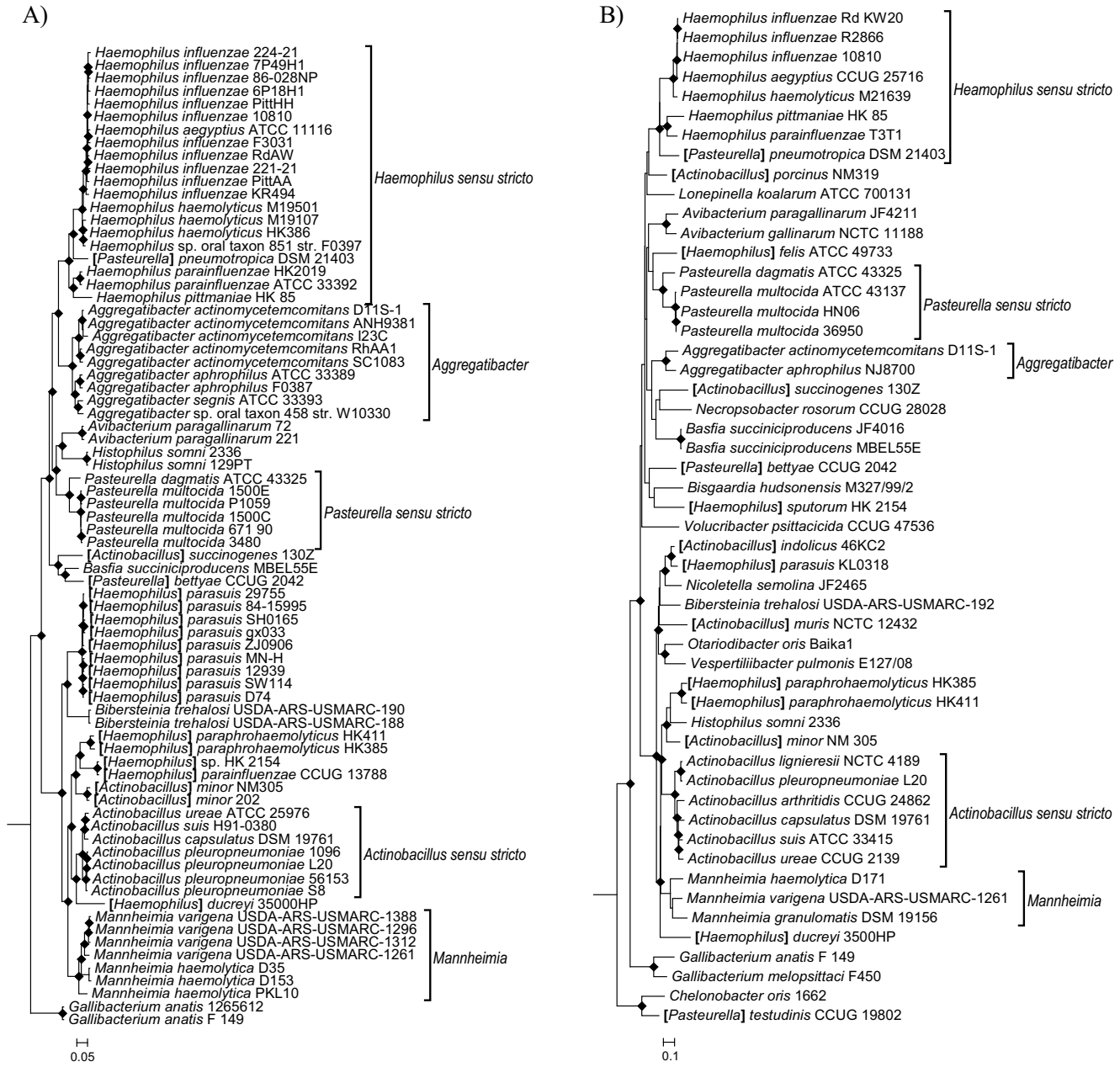


Figure 1(A). A maximum likelihood whole genome phylogenetic tree of sequenced members of the family Pasteurellaceae (B) A maximum likelihood phylogenetic tree based on concatenated nucleotide sequence alignments of the 16S rDNA, infB, recN, and rpoB genes. Both trees are rooted using members of the Vibrionaceae (not shown). Nodes with >80% bootstrap support are indicated by diamond shaped symbols at the node. Clusters of species representing Actinobacillus sensu stricto, Haemophilus sensu stricto, Pasteurella sensu stricto, Aggregatibacter and Mannheimia are indicated by brackets. Members of the genera Actinobacillus, Haemophilus, and Pasteurella which do not fall into their respective "sensu stricto" clades are indicated by the presence of square brackets around their generic name (ex. [Pasteurella] pneumotropica).

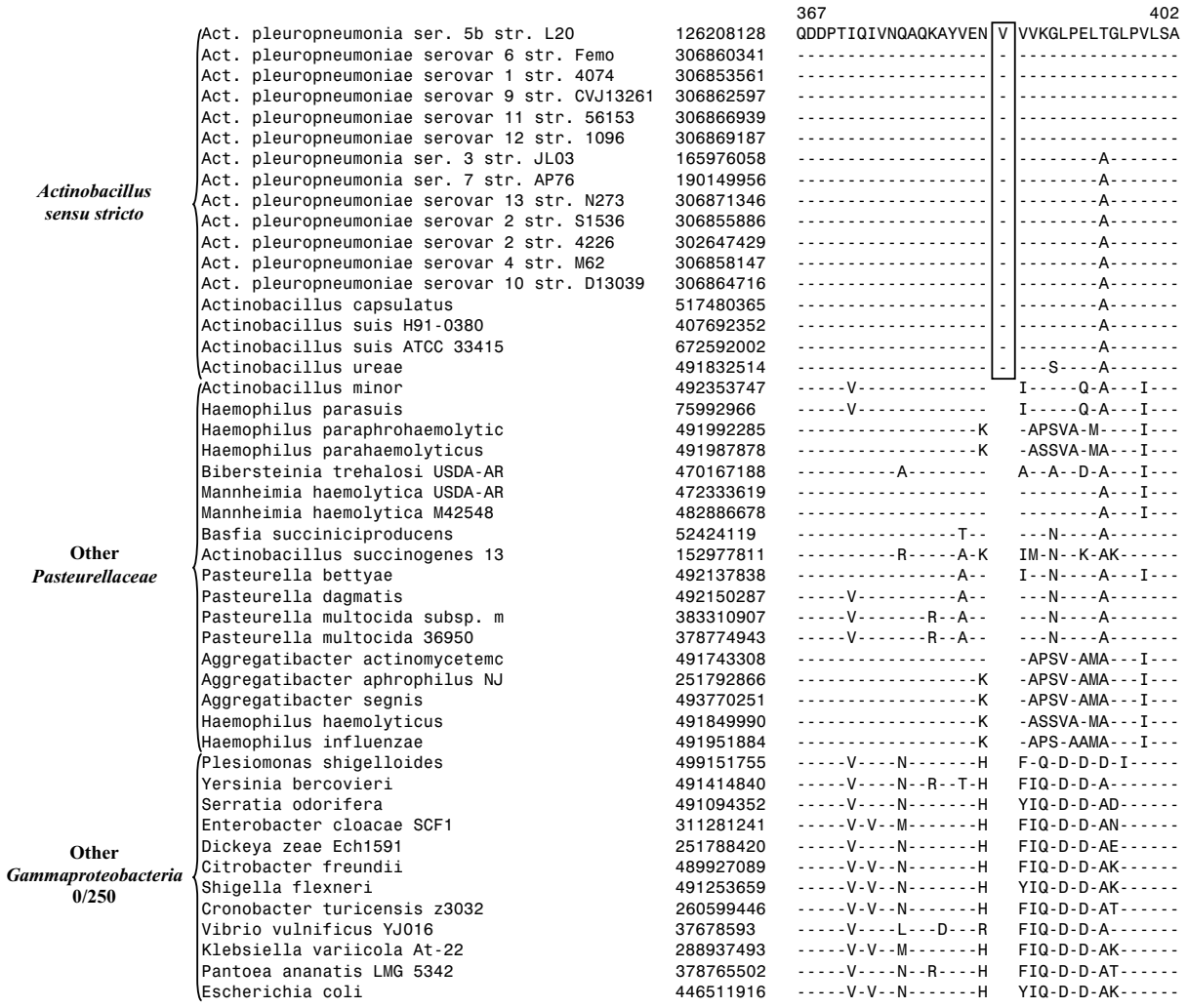


Figure 2. A partial sequence alignment of a 3'-nucleotidase showing a 1 amino acid insertion identified in all members of *Actinobacillus sensu stricto*. This insertion was not found in the homologues from any member of the genus *Actinobacillus* that was not part of the "*sensu stricto*" clade or any other member of the *Gammaproteobacteria*. Sequence information for a representative subset of the family *Pasteurellaceae* and the class *Gammaproteobacteria* are shown, but unless otherwise indicated, similar CSIs were detected in all members of the indicated group and not detected in any other bacterial species in the top 250 BLAST hits. The dashes (-) in the alignments indicate identity with the residue in the top sequence. GenBank identification (GI) numbers for each sequence are indicated in the second column. Sequence information for other CSIs specific to *Actinobacillus sensu stricto* are presented in Supplemental Figures 1 - 2 and their characteristics are summarized in Table 2A.

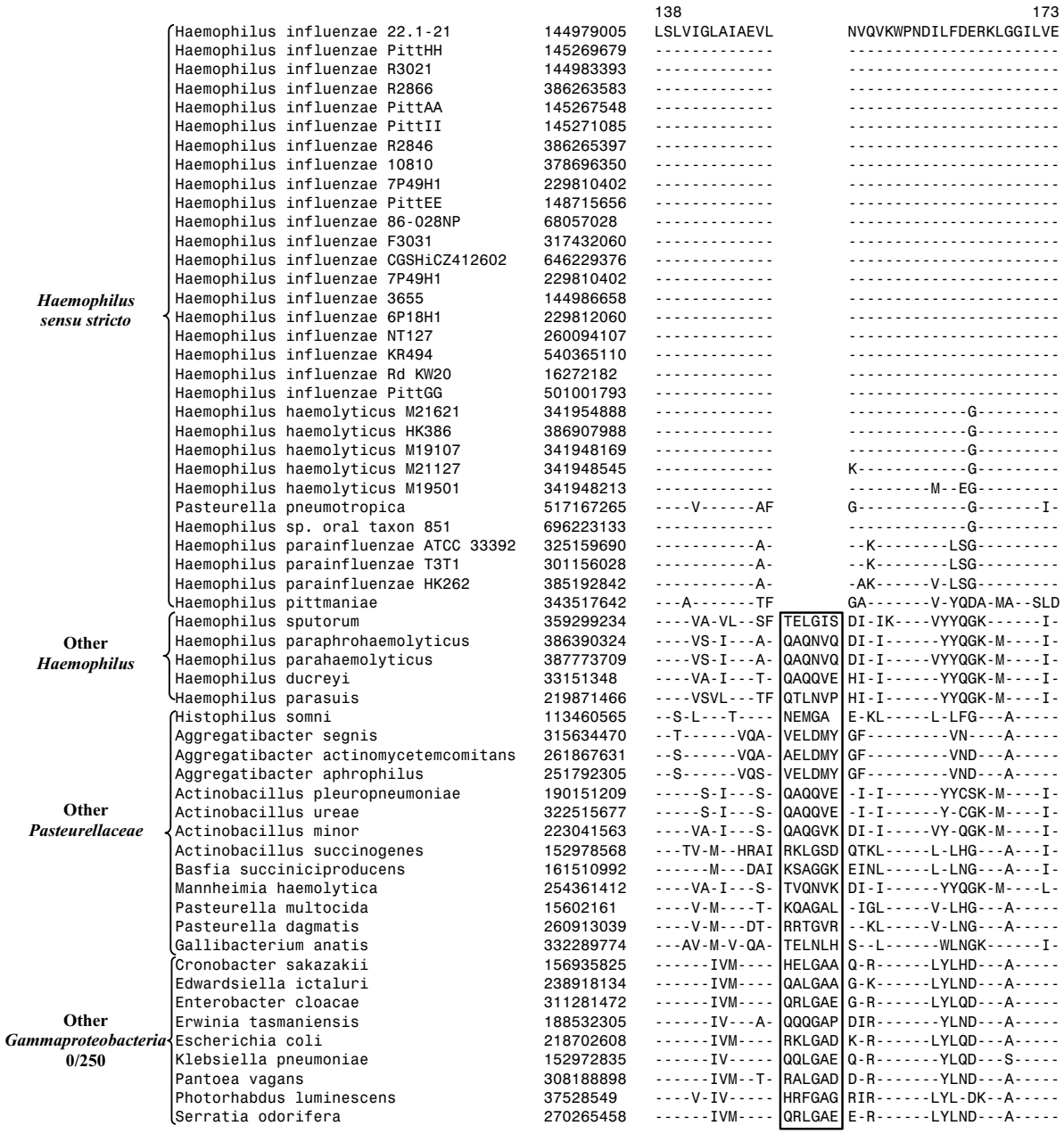


Figure 3 A partial sequence alignment of 1,4-dihydroxy-2-naphthoate octaprenyltransferase showing a 2 amino acid insertion identified in all members of *Haemophilus sensu stricto*. This insertion was not found in the homologues from any member of the genus *Haemophilus* that was not part of the "*sensu stricto*" clade or any other member of the *Gammaproteobacteria*. Sequence information for other CSIs specific to *Haemophilus sensu stricto* are presented in Supplemental Figures 3 - 8 and their characteristics are summarized in Table 2B.

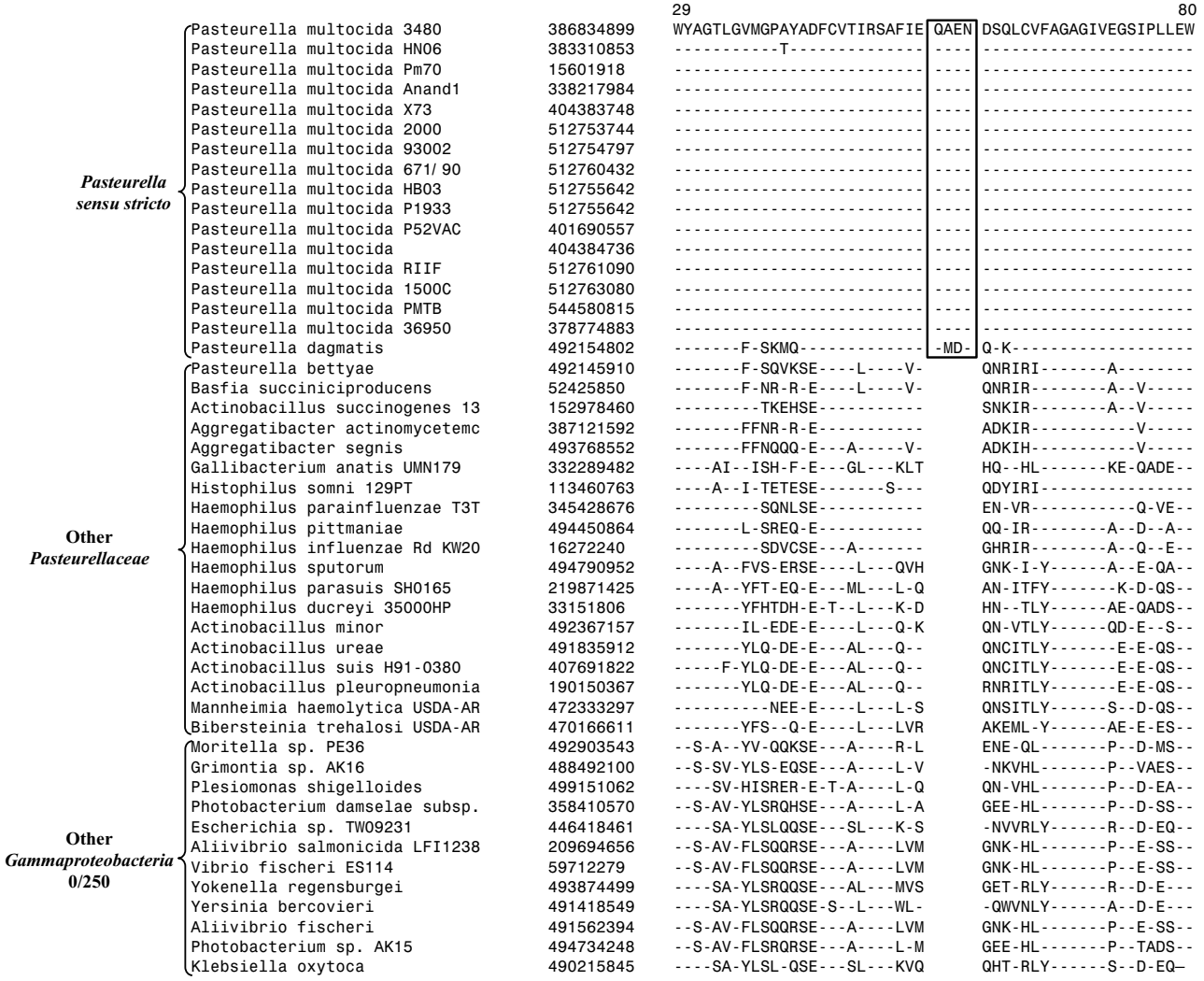


Figure 4. A partial sequence alignment of Menaquinone-specific isochorismate synthase showing a 4 amino acid insertion identified in all members of *Pasteurella sensu stricto*. This insertion was not found in the homologues from any member of the genus *Pasteurella* that was not part of the "*sensu stricto*" clade or any other member of the *Gammaproteobacteria*. Sequence information for other CSIs specific to *Pasteurella sensu stricto* are presented in Supplemental Figures 9 - 13 and their characteristics are summarized in Table 2C.

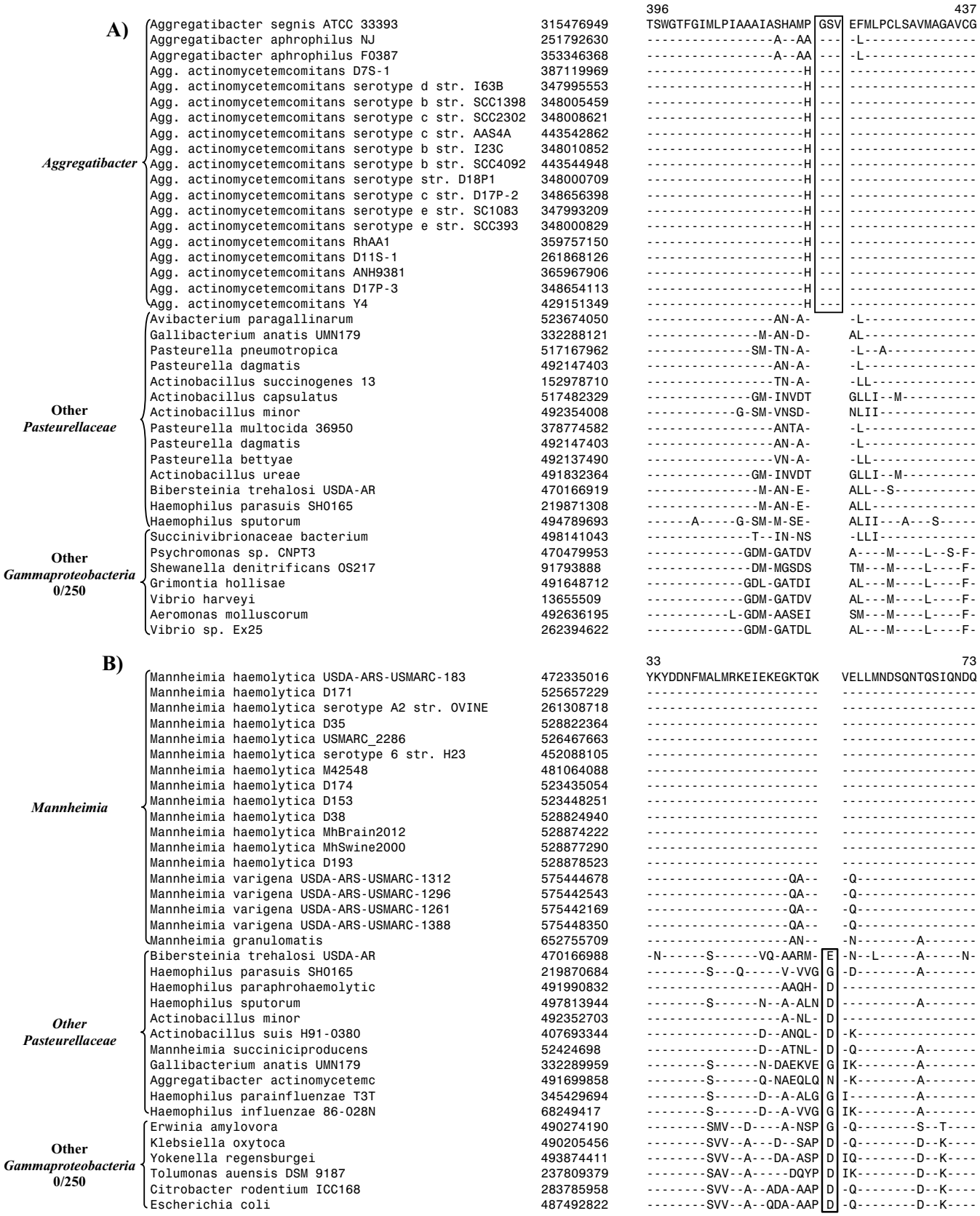

A)
396
437
Aggregatibacter segnis ATCC 33393 315476949 TSWGTFGIMLPIAAAIASHAMP GSV EFMLPCLSAVMAGAVCG
Aggregatibacter aphrophilus NJ 251792630 ----------A--AA --- -L---------
Aggregatibacter aphrophilus F0387 353346368 ----------A--AA --- -L---------
Agg. actinomycetemcomitans D7S-1 387119969 ----------H --- ----------
Agg. actinomycetemcomitans serotype d str. I63B 347995553 ----------H --- ----------
Agg. actinomycetemcomitans serotype b str. SCC1398 348005459 ----------H --- ----------
Agg. actinomycetemcomitans serotype c str. SCC2302 348008621 ----------H --- ----------
Agg. actinomycetemcomitans serotype c str. AAS4A 443542862 ----------H --- ----------
Agg. actinomycetemcomitans serotype b str. I23C 348010852 ----------H --- ----------
Aggregatibacter
Agg. actinomycetemcomitans serotype b str. SCC4092 443544948 ----------H --- ----------
Agg. actinomycetemcomitans serotype str. D18P1 348000709 ----------H --- ----------
Agg. actinomycetemcomitans serotype c str. D17P-2 348656398 ----------H --- ----------
Agg. actinomycetemcomitans serotype e str. SC1083 347993209 ----------H --- ----------
Agg. actinomycetemcomitans serotype e str. SCC393 348000829 ----------H --- ----------
Agg. actinomycetemcomitans RhAA1 359757150 ----------H --- ----------
Agg. actinomycetemcomitans D11S-1 261868126 ----------H --- ----------
Agg. actinomycetemcomitans ANH9381 365967906 ----------H --- ----------
Agg. actinomycetemcomitans D17P-3 348654113 ----------H --- ----------
Agg. actinomycetemcomitans Y4 429151349 ----------H --- ----------
Avibacterium paragallinarum 523674050 ----------AN-A- -L---------
Gallibacterium anatis UMN179 332288121 ----------M-AN-D- AL---------
Pasteurella pneumotropica 517167962 ----------SM-TN-A- -L--A--------
Pasteurella dagmatis 492147403 ----------AN-A- -L---------
Actinobacillus succinogenes 13 152978710 ----------TN-A- -LL--------
Actinobacillus capsulatus 517482329 ----------GM-INVDT GLLI--M-------
Other
Actinobacillus minor 492354008 ----------G-SM-VNSD- NLII----------
Pasteurellaceae
Pasteurella multocida 36950 378774582 ----------ANTA- -L---------
Pasteurella dagmatis 492147403 ----------AN-A- -L---------
Pasteurella bettyae 492137490 ----------VN-A- -LL--------
Actinobacillus ureae 491832364 ----------GM-INVDT GLLI--M-------
Bibersteinia trehalosi USDA-AR 470166919 ----------M-AN-E- ALL--S--------
Haemophilus parasuis SH0165 219871308 ----------M-AN-E- ALL---------
Haemophilus sputorum 494789693 ------A-----G-SM-M-SE- ALII---A---S-----
Succinivibrionaceae bacterium 498141043 ----------T--IN-NS -LLI--------
Other
Psychromonas sp. CNPT3 470479953 ----------GDM-GATDV A----M----L--S-F-
Shewanella denitrificans OS217 91793888 ----------DM-MGSDS TM---M----L----F-
Gammaproteobacteria
Grimontia hollisae 491648712 ----------GDL-GATDI AL---M----L----F-
0/250
Vibrio harveyi 13655509 ----------GDM-GATDV AL---M----L----F-
Aeromonas molluscorum 492636195 ----------L-GDM-AASEI SM---M----L----F-
Vibrio sp. Ex25 262394622 ----------GDM-GATDL AL---M----L----F-
B)
33
73
Mannheimia haemolytica USDA-ARS-USMARC-183 472335016 YKYDDNFMALMRKEIEKEGKTQK VELLMNDSQNTQSIQNDQ
Mannheimia haemolytica D171 525657229 ---------- ----------
Mannheimia haemolytica serotype A2 str. OVINE 261308718 ---------- ----------
Mannheimia haemolytica D35 528822364 ---------- ----------
Mannheimia haemolytica USMARC_2286 526467663 ---------- ----------
Mannheimia haemolytica serotype 6 str. H23 452088105 ---------- ----------
Mannheimia haemolytica M42548 481064088 ---------- ----------
Mannheimia haemolytica D174 523435054 ---------- ----------
Mannheimia haemolytica D153 523448251 ---------- ----------
Mannheimia
Mannheimia haemolytica D38 528824940 ---------- ----------
Mannheimia haemolytica MhBrain2012 528874222 ---------- ----------
Mannheimia haemolytica MhSwine2000 528877290 ---------- ----------
Mannheimia haemolytica D193 528878523 ---------- ----------
Mannheimia varigena USDA-ARS-USMARC-1312 575444678 ----------QA-- -Q---------
Mannheimia varigena USDA-ARS-USMARC-1296 575442543 ----------QA-- -Q---------
Mannheimia varigena USDA-ARS-USMARC-1261 575442169 ----------QA-- -Q---------
Mannheimia varigena USDA-ARS-USMARC-1388 575448350 ----------QA-- -Q---------
Mannheimia granulomatis 652755709 ----------AN-- -N--------A-------
Bibersteinia trehalosi USDA-AR 470166988 -N------S------VQ-AARM- E -N--L-----A-----N-
Haemophilus parasuis SH0165 219870684 --------S---Q-----V-VVG G -D--------A-------
Haemophilus paraphrohaemolytic 491990832 ----------AAQH- D ----------
Haemophilus sputorum 497813944 --------S------N--A-ALN D ----------A-------
Other
Actinobacillus minor 492352703 ----------A-NL- D ----------
Pasteurellaceae
Actinobacillus suis H91-0380 407693344 ----------D--ANQL- D -K---------
Mannheimia succiniciproducens 52424698 ----------D--ATNL- D -Q--------A-------
Gallibacterium anatis UMN179 332289959 --------S------N-DAEKVE G IK--------A-------
Aggregatibacter actinomycetemc 491699858 --------S------Q-NAEQLQ N -K--------A-------
Haemophilus parainfluenzae T3T 345429694 --------S------D--A-ALG G I---------A-------
Haemophilus influenzae 86-028N 68249417 --------S------D--A-VVG G IK--------A-------
Erwinia amylovora 490274190 --------SMV--D----A-NSP G -Q--------S--T----
Klebsiella oxytoca 490205456 --------SVV--A---D--SAP D -Q--------D--K----
Other
Yokenella regensburgei 493874411 --------SVV--A---DA-ASP D IQ---------D--K----
Gammaproteobacteria
Tolumonas auensis DSM 9187 237809379 --------SAV--A-----DQYP D IK--------D--K----
0/250
Citrobacter rodentium ICC168 283785958 --------SVV--A--ADA-AAP D -Q--------D--K----
Escherichia coli 487492822 --------SVV--A--QDA-AAP D -Q--------D--K----

Figure 5. A partial sequence alignment of (A) a *nhaC* family sodium/proton antiporter containing a 3 amino acid insertion specific for all sequenced species of the genus *Aggregatibacter* (B) a methyl-galactoside ABC transporter substrate-binding protein containing a 1 amino acid deletion specific for all sequenced species of the genus *Mannheimia*. In each case, the identified CSIs were only found in the sequenced members of the genera *Aggregatibacter* or *Mannheimia* and were absent from all other all other sequenced *Pasteurellaceae* and *Gammaproteobacteria*. Sequence information for other CSIs specific to *Pasteurella sensu stricto* are presented in Supplemental Figures 14 - 19 and their characteristics are summarized in Table 2D and 2E.

PhD – Sohail Naushad McMaster University - Biochemistry

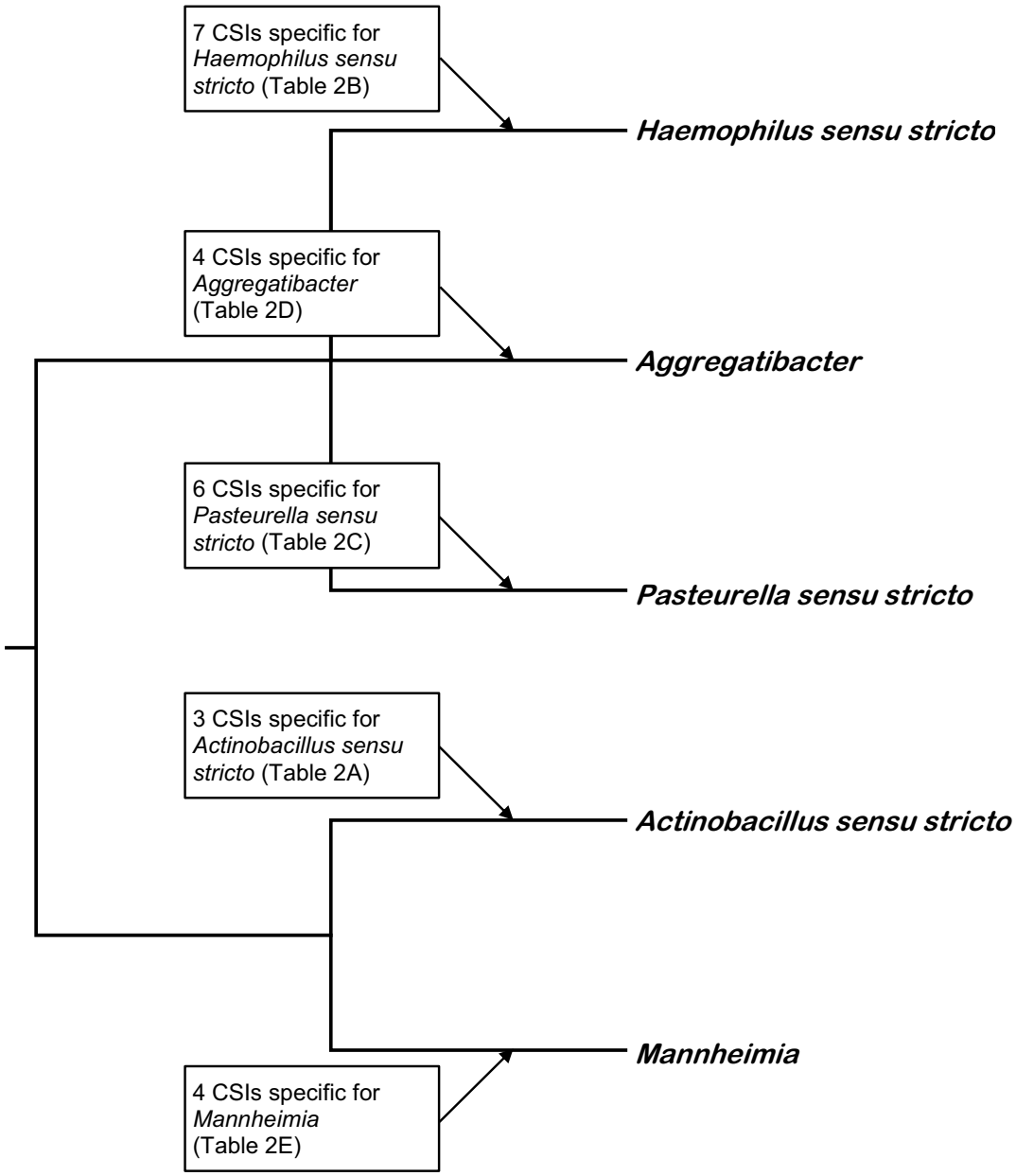


Figure 6. A summary diagram depicting the distribution of identified CSIs for different genera within the family *Pasteurellaceae*.

REFERENCES

Adeolu, M., & Gupta, R. S. (2014). A phylogenomic and molecular marker based proposal for the division of the genus *Borrelia* into two genera: the emended genus *Borrelia* containing only the members of the relapsing fever *Borrelia*, and the genus *Borreliella* gen. nov. containing the members of the Lyme disease *Borrelia* (*Borrelia burgdorferi sensu lato* complex). *Anton Leeuw Int J G, 105*(6), 1049-1072.

Ahmod, N. Z., Gupta, R. S., & Shah, H. N. (2011). Identification of a *Bacillus anthracis* specific indel in the *yeaC* gene and development of a rapid pyrosequencing assay for distinguishing *B. anthracis* from the *B. cereus* group. *J Microbiol Methods, 87*(3), 278-285.

Angen, Ø., Ahrens, P., & Bisgaard, M. (2002). Phenotypic and genotypic characterization of *Mannheimia* (*Pasteurella*) *haemolytica*-like strains isolated from diseased animals in Denmark. *Vet Microbiol, 84*(1), 103-114.

Angen, Ø., Ahrens, P., Kuhnert, P., Christensen, H., & Mutters, R. (2003). Proposal of *Histophilus somni* gen. nov., sp. nov. for the three species incertae sedis '*Haemophilus somnus*','*Haemophilus agni*'and '*Histophilus ovis*'. *Int J Syst Evol Microbiol, 53*(5), 1449-1456.

Angen, Ø., Mutters, R., Caugant, D. A., Olsen, J. E., & Bisgaard, M. (1999). Taxonomic relationships of the [*Pasteurella*] *haemolytica* complex as evaluated by DNA-DNA hybridizations and 16S rRNA sequencing with proposal of *Mannheimia haemolytica* gen. nov., comb, nov., *Mannheimia granulomatis* comb. nov., *Mannheimia glucosida* sp. nov., *Mannheimia ruminalis* sp. nov. and *Mannheimia varigena* sp. nov. *Int J Syst Bacteriol, 49*(1), 67-86.

Bhandari, V., & Gupta, R. S. (2014). Molecular signatures for the phylum (class) *Thermotogae* and a proposal for its division into three orders (*Thermotogales*, *Kosmotogales* ord. nov. and *Petrotogales* ord. nov.) containing four families (*Thermotogaceae*, *Fervidobacteriaceae* fam. nov., *Kosmotogaceae* fam. nov. and *Petrotogaceae* fam. nov.) and a new genus *Pseudothermotoga* gen. nov. with five new combinations. *Antonie Van Leeuwenhoek, 105*(1), 143-168.

Blackall, P., Bojesen, A. M., Christensen, H., & Bisgaard, M. (2007). Reclassification of [*Pasteurella*] *trehalosi* as *Bibersteinia trehalosi* gen. nov., comb. nov. *Int J Syst Evol Microbiol, 57*(4), 666-674.

Blackall, P. J., Christensen, H., Beckenham, T., Blackall, L. L., & Bisgaard, M. (2005). Reclassification of *Pasteurella gallinarum*,[*Haemophilus*] *paragallinarum*, *Pasteurella avium* and *Pasteurella volantium* as *Avibacterium gallinarum* gen. nov., comb. nov., *Avibacterium paragallinarum* comb. nov., *Avibacterium avium* comb. nov. and *Avibacterium volantium* comb. nov. *Int J Syst Evol Microbiol, 55*(1), 353-362.

Bonaventura, M. P. D., Lee, E. K., DeSalle, R., & Planet, P. J. (2010). A whole-genome phylogeny of the family *Pasteurellaceae*. *Mol Phylogen Evol, 54*(3), 950-956.

Bossé, J. T., Janson, H., Sheehan, B. J., Beddek, A. J., Rycroft, A. N., Simon Kroll, J., & Langford, P. R. (2002). *Actinobacillus pleuropneumoniae*: pathobiology and pathogenesis of infection. *Microb Infect, 4*(2), 225-235.

Cattoir, V., Lemenand, O., Avril, J.-L., & Gaillot, O. (2006). The *sodA* gene as a target for phylogenetic dissection of the genus *Haemophilus* and accurate identification of human clinical isolates. *Int J Med Microbiol, 296*(8), 531-540.

Challacombe, J. F., Duncan, A. J., Brettin, T. S., Bruce, D., Chertkov, O., Detter, J. C., Han, C. S., Misra, M., Richardson, P., Tapia, R., et al. (2007). Complete genome sequence of *Haemophilus somnus* (*Histophilus somni*) strain 129Pt and comparison to *Haemophilus ducreyi* 35000HP and *Haemophilus influenzae* Rd. *J Bacteriol, 189*(5), 1890-1898.

Chen, C., Kittichotirat, W., Chen, W., Downey, J. S., Si, Y., & Bumgarner, R. (2010). Genome sequence of naturally competent *Aggregatibacter actinomycetemcomitans* serotype a strain D7S-1. *J Bacteriol, 192*(10), 2643-2644.

Chen, C., Kittichotirat, W., Si, Y., & Bumgarner, R. (2009). Genome sequence of *Aggregatibacter actinomycetemcomitans* serotype c strain D11S-1. *J Bacteriol, 191*(23), 7378-7379.

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 Münch Tierärztl Wschr, 11*(1-2), 20-30.

Christensen, H., Bisgaard, M., Bojesen, A. M., Mutters, R., & Olsen, J. E. (2003). Genetic relationships among avian isolates classified as *Pasteurella haemolytica*, '*Actinobacillus salpingitidis*' or *Pasteurella anatis* with proposal of *Gallibacterium anatis* gen. nov., comb. nov. and description of additional genomospecies within *Gallibacterium* gen. nov. *Int J Syst Evol Microbiol, 53*(1), 275-287.

Christensen, H., & Kuhnert, P. (2012). International Committee on Systematics of Prokaryotes Subcommittee on the taxonomy of *Pasteurellaceae* Minutes of the meetings, 25 August 2011, Elsinore, Denmark. *Int J Syst Evol Microbiol, 62*(1), 257-258.

Christensen, H., Kuhnert, P., Busse, H.-J., Frederiksen, W. C., & Bisgaard, M. (2007). Proposed minimal standards for the description of genera, species and subspecies of the *Pasteurellaceae*. *Int J Syst Evol Microbiol, 57*(1), 166-178.

Christensen, H., Kuhnert, P., Olsen, J. E., & 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*(5), 1601-1609.

Cutino-Jimenez, A. M., Martins-Pinheiro, M., Lima, W. C., Martin-Tornet, A., Morales, O. G., & Menck, C. F. M. (2010). Evolutionary placement of *Xanthomonadales* based on conserved protein signature sequences. *Mol Phylogen Evol, 54*(2), 524-534.

Dewhirst, F. E., Paster, B. J., Olsen, I., & Fraser, G. J. (1992). Phylogeny of 54 representative strains of species in the family *Pasteurellaceae* as determined by comparison of 16S rRNA sequences. *J Bacteriol, 174*(6), 2002-2013.

Dewhirst, F. E., Paster, B. J., Olsen, I., & Fraser, G. J. (1993). Phylogeny of the *Pasteurellaceae* as determined by comparison of 16S ribosomal ribonucleic acid sequences. *Zentralblatt für Bakteriologie, 279*(1), 35-44.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res, 32*(5), 1792-1797.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics, 26*(19), 2460-2461.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-F., Dougherty, B. A., & Merrick, J. M. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science, 269*(5223), 496-512.

Foote, S. J., Bosse, J. T., Bouevitch, A. B., Langford, P. R., Young, N. M., & Nash, J. H. (2008). The complete genome sequence of *Actinobacillus pleuropneumoniae* L20 (serotype 5b). *J Bacteriol, 190*(4), 1495-1496.

Gao, B., & Gupta, R. S. (2012). Phylogenetic framework and molecular signatures for the main clades of the phylum Actinobacteria. *MicrobiolMolBiolRev, 76*(1), 66-112.

Gao, B., Mohan, R., & Gupta, R. S. (2009). Phylogenomics and protein signatures elucidating the evolutionary relationships among the *Gammaproteobacteria*. *Int J Syst Evol Microbiol, 59*(2), 234-247.

Gioia, J., Qin, X., Jiang, H., Clinkenbeard, K., Lo, R., Liu, Y., Fox, G. E., Yerrapragada, S., McLeod, M. P., & McNeill, T. Z. (2006). The genome sequence of *Mannheimia haemolytica* A1: insights into virulence, natural competence, and *Pasteurellaceae* phylogeny. *J Bacteriol, 188*(20), 7257-7266.

Gupta, R. S. (1998). Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. *Microbiol Mol Biol Rev, 62*(4), 1435.

Gupta, R. S. (2010). Applications of conserved indels for understanding microbial phylogeny. In A. Oren & R. T. Papke (Eds.), Molecular phylogeny of microorganisms (pp. 135-150). Norfolk, UK: Caister Academic Press.

Gupta, R. S. (2014). Identification of conserved indels that are useful for classification and evolutionary studies Methods in Microbiology (Vol. 41). Academic Press: http://dx.doi.org/10.1016/bs.mim.2014.05.003.

Gupta, R. S., & Lali, R. (2013). Molecular signatures for the phylum Aquificae and its different clades: proposal for division of the phylum Aquificae into the emended order *Aquificales*, containing the families *Aquificaceae* and *Hydrogenothermaceae*, and a new order *Desulfurobacteriales* ord. nov., containing the family *Desulfurobacteriaceae*. *Antonie Van Leeuwenhoek, 104*(3), 349-368.

Gupta, R. S., Mahmood, S., & Adeolu, M. (2013). A phylogenomic and molecular signature based approach for characterization of the phylum Spirochaetes and its major clades: proposal for a taxonomic revision of the phylum. *Frontiers in microbiology, 4*, 217.

Harrison, A., Dyer, D. W., Gillaspy, A., Ray, W. C., Mungur, R., Carson, M. B., Zhong, H., Gipson, J., Gipson, M., Johnson, L. S., 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*(13), 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*(6), 419-423.

Hedegaard, J., Okkels, H., Bruun, B., Kilian, M., Mortensen, K. K., & Nørskov-Lauritsen, N. (2001). Phylogeny of the genus *Haemophilus* as determined by comparison of partial *infB* sequences. *Microbiology, 147*(9), 2599-2609.

Henderson, B., Ward, J. M., & Ready, D. (2010). *Aggregatibacter* (*Actinobacillus*) *actinomycetemcomitans*: a triple A* periodontopathogen? *Periodontol 2000, 54*(1), 78-105.

Hogg, J. S., Hu, F. Z., Janto, B., Boissy, R., Hayes, J., Keefe, R., Post, J. C., & Ehrlich, G. D. (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*(6), R103.

Hong, S. H., Kim, J. S., Lee, S. Y., In, Y. H., Choi, S. S., Rih, J. K., Kim, C. H., Jeong, H., Hur, C. G., & Kim, J. J. (2004). The genome sequence of the capnophilic rumen bacterium *Mannheimia succiniciproducens*. *Nat Biotechnol, 22*(10), 1275-1281.

Howard-Azzeh, M., Shamseer, L., Schellhorn, H. E., & Gupta, R. S. (2014). Phylogenetic analysis and molecular signatures defining a monophyletic clade of heterocystous cyanobacteria and identifying its closest relatives. *Photosynth Res, 122*(2), 171-185.

Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G., & Gibson, T. J. (1998). Multiple sequence alignment with Clustal X. *Trends Biochem Sci, 23*(10), 403.

Jordan, I. K., Conley, A. B., Antonov, I. V., Arthur, R. A., Cook, E. D., Cooper, G. P., Jones, B. L., Knipe, K. M., Lee, K. J., Liu, X., et al. (2011). Genome sequences for five strains of the emerging pathogen *Haemophilus haemolyticus*. *J Bacteriol, 193*(20), 5879-5880.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol, 30*(4), 772-780.

Kilian, M. (2005). Genus III. *Haemophilus*. In D. J. Brenner, N. R. Krieg, G. M. Garrity & J. T. Staley (Eds.), Bergey's Manual of Systematic Bacteriology (2nd ed., Vol. 2, pp. 883-904). New York: Springer. (Reprinted from: Not in File).

Konstantinidis, K. T., & Stackebrandt, E. (2013). Defining taxonomic ranks The prokaryotes (pp. 229-254): Springer.

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*(4), 1393-1399.

Kuhnert, P., & Korczak, B. M. (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*(9), 2537-2548.

Mannheim, W., Pohl, S., & Holländer, R. (1979). [On the taxonomy of *Actinobacillus*, *Haemophilus*, and *Pasteurella*: DNA base composition, respiratory quinones, and biochemical reactions of representative collection cultures (author's transl)]. *Zentralbl Bakteriol A, 246*(4), 512-540.

May, B. J., Zhang, Q., Li, L. L., Paustian, M. L., Whittam, T. S., & Kapur, V. (2001). Complete genomic sequence of *Pasteurella multocida*, Pm70. *Proc Natl Acad Sci U S A, 98*(6), 3460-3465.

Muehldorfer, K., Speck, S., & Wibbelt, G. (2014). Proposal of *Vespertiliibacter pulmonis* gen. nov., sp. nov. and two genomospecies as new members of the family *Pasteurellaceae* isolated from European bats. *Int J Syst Evol Microbiol*, ijs. 0.062786-062780.

Mutters, R., Mannheim, W., & Bisgaard, M. (1989). Taxonomy of the group. In C. Adlam & J. M. Rutter (Eds.), *Pasturella* and Pasteurellosis (pp. 3-34). London: Academic Press.

Mutters, R., Pohl, S., & Mannheim, W. (1986). Transfer of *Pasteurella ureae* Jones 1962 to the Genus *Actinobacillus* Brumpt 1910: *Actinobacillus ureae* comb. nov. *Int J Syst Bacteriol, 36*(2), 343-344.

Naushad, H. S., & Gupta, R. S. (2012). Molecular signatures (conserved indels) in protein sequences that are specific for the order *Pasteurellales* and distinguish two of its main clades. *Antonie Van Leeuwenhoek, 101*(1), 105-124.

Naushad, H. S., & Gupta, R. S. (2013). Phylogenomics and Molecular Signatures for Species from the Plant Pathogen-Containing Order *Xanthomonadales*. *PLoS One, 8*(2), e55216.

Naushad, H. S., Lee, B., & Gupta, R. S. (2014). Conserved signature indels and signature proteins as novel tools for understanding microbial phylogeny and systematics:

identification of molecular signatures that are specific for the phytopathogenic genera *Dickeya*, *Pectobacterium* and *Brenneria*. *Int J Syst Evol Microbiol, 64*(2), 366-383.

NCBI. (2014a). NCBI Genome Database. http://www.ncbi.nlm.nih.gov/genome/

NCBI. (2014b). NCBI Nucleotide Database. http://www.ncbi.nlm.nih.gov/nuccore/

Nørskov-Lauritsen, N. (2014). Classification, Identification, and Clinical Significance of *Haemophilus* and *Aggregatibacter* Species with Host Specificity for Humans. *Clin Microbiol Rev, 27*(2), 214-240.

Nørskov-Lauritsen, N., Bruun, B., Andersen, C., & Kilian, M. (2012). Identification of haemolytic *Haemophilus* species isolated from human clinical specimens and description of *Haemophilus sputorum* sp. nov. *Int J Med Microbiol, 302*(2), 78-83.

Nørskov-Lauritsen, N., Bruun, B., & Kilian, M. (2005). Multilocus sequence phylogenetic study of the genus *Haemophilus* with description of *Haemophilus pittmaniae* sp. nov. *Int J Syst Evol Microbiol, 55*(1), 449-456.

Nørskov-Lauritsen, N., & Kilian, M. (2006). Reclassification of *Actinobacillus actinomycetemcomitans*, *Haemophilus aphrophilus*, *Haemophilus paraphrophilus* and *Haemophilus segnis* as *Aggregatibacter actinomycetemcomitans* gen. nov., comb. nov., *Aggregatibacter aphrophilus* comb. nov. and *Aggregatibacter segnis* comb. nov., and emended description of *Aggregatibacter aphrophilus* to include V factor-dependent and V factor-independent isolates. *Int J Syst Evol Microbiol, 56*(9), 2135-2146.

Olsen, I. (1993). Recent approaches to the chemotaxonomy of the *Actinobacillus-Haemophilus–Pasteurella* group (family *Paseurellaceae*). *Oral Microbiol Immunol, 8*(6), 327-336.

Olsen, I., Dewhirst, F. E., Paster, B. J., & Busse, H. J. (2005). Family I. *Pasteurellaceae*. In D. J. Brenner, N. R. Krieg, G. M. Garrity & J. T. Staley (Eds.), Bergey's Manual of Systematic Bacteriology (2nd ed., Vol. 2, pp. 851-856). New York: Springer. (Reprinted from: Not in File).

Parte, A. C. (2013). LPSN--list of prokaryotic names with standing in nomenclature. *Nucleic Acids Res, 42*, D613–D616.

Pohl, S. (1979). *Reklassifizierung der gattung Actinobacillus Brumpt 1910, Haemophilus Winslow et al. 1971 und Pasteurella Trevisan 1887 anhand phänotypischer und molekularer daten, insbesondere der DNS-verwandtschaften bei DNS: DNS-hybridisierung in vitro und vorschlag einer neuen familie, Pasteurellaceae.* (Inaug. Diss.), Philipps-Universität, Marburg, Germany.

Pohl, S., Bertschinger, H., Frederiksen, W., & Mannheim, W. (1983). Transfer of *Haemophilus pleuropneumoniae* and the *Pasteurella haemolytica*-like organism causing porcine necrotic pleuropneumonia to the genus *Actinobacillus* (*Actinobacillus pleuropneumoniae* comb. nov.) on the basis of phenotypic and deoxyribonucleic acid relatedness. *Int J Syst Bacteriol, 33*(3), 510-514.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One, 5*(3), e9490.

Redfield, R. J., Findlay, W. A., Bossé, J., Kroll, J. S., Cameron, A. D., & Nash, J. H. (2006). Evolution of competence and DNA uptake specificity in the *Pasteurellaceae*. *BMC Evol Biol, 6*(1), 82.

Rokas, A., & Holland, P. W. H. (2000). Rare genomic changes as a tool for phylogenetics. *Trends Ecol Evol, 15*(11), 454-459.

Rokas, A., Williams, B. L., King, N., & Carroll, S. B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature, 425*(6960), 798-804.

Snipes, K., & Biberstein, E. (1982). *Pasteurella testudinis* sp. nov.: a parasite of desert tortoises (Gopherus agassizi). *Int J Syst Bacteriol, 32*(2), 201-210.

Spinola, S. M., Bauer, M. E., & Munson, R. S. (2002). Immunopathogenesis of *Haemophilus ducreyi* infection (chancroid). *Infect Immun, 70*(4), 1667-1676.

Stackebrandt, E., & Ebers, J. (2006). Taxonomic parameters revisited: tarnished gold standards. *Microbiol Today, 33*(4), 152.

Strouts, F. R., Power, P., Croucher, N. J., Corton, N., Van Tonder, A., Quail, M. A., Langford, P. R., Hudson, M. J., Parkhill, J., & Kroll, J. S. (2012). Lineage-specific virulence determinants of *Haemophilus influenzae* biogroup *aegyptius*. *Emerging Infect Dis, 18*(3), 449-457.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol, 30*(12), 2725-2729.

Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., Gillespie, J. J., Gough, R., Hix, D., & Kenyon, R. (2013). PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res*, gkt1099.

Whelan, S., & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol, 18*(5), 691-699.

Wilson, B. A., & Ho, M. (2013). *Pasteurella multocida*: from zoonosis to cellular microbiology. *Clin Microbiol Rev, 26*(3), 631-655.

Wong, S. Y., Paschos, A., Gupta, R. S., & Schellhorn, H. E. (2014). Insertion/Deletion-Based Approach for the Detection of *Escherichia coli* O157:H7 in Freshwater Environments. *Environ Sci Technol, 48*(19), 11462-11470.

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N. N., Kunin, V., Goodwin, L., Wu, M., & Tindall, B. J. (2009). A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature, 462*(7276), 1056-1060.

Xu, Z., Zhou, Y., Li, L., Zhou, R., Xiao, S., Wan, Y., Zhang, S., Wang, K., Li, W., Li, L., et al. (2008). Genome biology of *Actinobacillus pleuropneumoniae* JL03, an isolate of serotype 3 prevalent in China. *PLoS One, 3*(1), e1450.

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W., & Glöckner, F. O. (2013). The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res*.

Yue, M., Yang, F., Yang, J., Bei, W., Cai, X., Chen, L., Dong, J., Zhou, R., Jin, M., Jin, Q., et al. (2009). Complete genome sequence of *Haemophilus parasuis* SH0165. *J Bacteriol, 191*(4), 1359-1360.

CHAPTER 5

Phylogenomics and Molecular Signatures for Species from the Plant Pathogen-containing Order *Xanthomonadales*

The following Chapter describes the applications of phylogenetic and comparative genomic analyses for the identification of CSIs unique to order *Xanthomonadales*. The members of this order are mostly plant pathogens. The identified CSIs help to demarcate *Xanthomonadales* order from other *Gammaproteobacteria* in clear molecular terms. The Chapter also reviews the impact of LGT on *Xanthomonadales* genome. I was involved in the identification of CSIs, data analysis, writing of the manuscript and construction of the figures and tables.

*Due to limited space, supplementary figures and tables are not included in the chapter but can be accessed along with the rest of the manuscript at:

Naushad,H.S. and Gupta,R.S. (2013). Phylogenomics and molecular signatures for species from the plant pathogen-containing order *Xanthomonadales*. PLoS. One. 8, e55216.

OPEN ACCESS Freely available online PLOS ONE

Phylogenomics and Molecular Signatures for Species from the Plant Pathogen-Containing Order Xanthomonadales

Hafiz Sohail Naushad, Radhey S. Gupta*

Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, Ontario, Canada

Abstract

The species from the order Xanthomonadales, which harbors many important plant pathogens and some human pathogens, are currently distinguished primarily on the basis of their branching in the 16S rRNA tree. No molecular or biochemical characteristic is known that is specific for these bacteria. Phylogenetic and comparative analyses were conducted on 26 sequenced Xanthomonadales genomes to delineate their branching order and to identify molecular signatures consisting of conserved signature indels (CSIs) in protein sequences that are specific for these bacteria. In a phylogenetic tree based upon sequences for 28 proteins, Xanthomonadales species formed a strongly supported clade with *Rhodanobacter* sp. 2APBS1 as its deepest branch. Comparative analyses of protein sequences have identified 13 CSIs in widely distributed proteins such as GlnRS, TypA, MscL, LysRS, LipA, Tgt, LpxA, TolQ, ParE, PolA and TyrB that are unique to all species/strains from this order, but not found in any other bacteria. Fifteen additional CSIs in proteins (viz. CoxD, DnaE, PolA, SucA, AsnB, RecA, PyrG, LigA, MutS and TrmD) are uniquely shared by different Xanthomonadales except *Rhodanobacter* and in a few cases by *Pseudoxanthomonas* species, providing further support for the deep branching of these two genera. Five other CSIs are commonly shared by Xanthomonadales and 1–3 species from the orders *Chromatiales, Methylococcales* and *Cardiobacteriales* suggesting that these deep branching orders of Gammaproteobacteria might be specifically related. Lastly, 7 CSIs in ValRS, CarB, PyrE, GlyS, RnhB, MinD and X001065 are commonly shared by Xanthomonadales and a limited number of Beta- or Gamma-proteobacteria. Our analysis indicates that these CSIs have likely originated independently and they are not due to lateral gene transfers. The Xanthomonadales-specific CSIs reported here provide novel molecular markers for the identification of these important plant and human pathogens and also as potential targets for development of drugs/agents that specifically target these bacteria.

Citation: Naushad HS, Gupta RS (2013) Phylogenomics and Molecular Signatures for Species from the Plant Pathogen-Containing Order Xanthomonadales. PLoS ONE 8(2): e55216. doi:10.1371/journal.pone.0055216

Editor: Celine Brochier-Armanet, Université Claude Bernard - Lyon 1, France

Received May 2, 2012; **Accepted** December 19, 2012; **Published** February 8, 2013

Funding: NSERC (Natural Science and Engineering Research Council of Canada) 249924. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: gupta@mcmaster.ca

Introduction

The Xanthomonadales are gram-negative, non-spore forming, catalase-positive, aerobic, rod shape bacteria [1], which are part of the class Gammaproteobacteria [2]. This order is comprised of two families Xanthomonadaceae and Sinobacteraceae that contain 22 and 6 genera, respectively (http://www.bacterio.cict.fr/classifphyla.html#Proteobacteria). The *Xylella* and *Xanthomonas* species, which are part of the order Xanthomonadales, cause a wide variety of serious diseases in more than 400 agriculturally important plants. Some of the economically important crops that are affected by species from these two genera include tomato, cabbage, pepper, banana, citrus, rice, grapes, peach, plum, almond, coffee and maple [3–9] Additionally, *Xylella fastidiosa* is responsible for causing leaf scorch disease in many landscape and ornamental plants including oak, elm, mulberry, sycamore, maple and oleander [7,9–11]. The diseases caused by these bacteria lead to major crop losses globally and thus they constitute serious agricultural and economic threat. In addition to these important phytopathogens, the Xanthomonadales also harbors the genus *Stenotrophomonas*, whose members (viz. *S. maltophila*) are multidrug resistant opportunistic pathogens, responsible for many hospital-acquired infections in immuno-compromised patients. These latter bacteria are also implicated in respiratory infections in cystic fibrosis patients [12–14].

The species from the order Xanthomonadales and its different families/genera are currently distinguished from other bacteria primarily on the basis of their branching in the 16S rRNA trees [1,4,15]. There is no biochemical, morphological or physiological characteristics known that are uniquely shared by various species from this order. Although Xanthomonadales are an order within the class Gammaproteobacteria, in phylogenetic trees based upon some genes/proteins sequences, these species are observed to branch with other classes of proteobacteria, particularly the Betaproteobacteria [16–20]. However, detailed phylogenetic studies based upon two independent, large datasets of concatenated protein sequences have now established that the species from the order Xanthomonadales are a deep branching clade within the class Gammaproteobacteria [21,22]. Several recently identified molecular signatures that are uniquely shared by Xanthomonadales and all other Gammaproteobacteria also support the placement of this group within the Gammaproteo-

bacteria [21,23]. The anomalous branching of Xanthomonadales in some phylogenetic trees possibly results from the deep branching of Xanthomonadales within the Gammaproteobacteria and also in some cases by lateral gene transfers (LGTs). In particular, extensive work by Menck and coworkers indicate that about 25% of the genes in *Xanthomonas*, which include many genomic islands as well as some genes involved in the biosynthesis of NAD, arginine and cysteine, are acquired by LGTs [16,24–28].

Because Xanthomonadales harbor many major phytopathogens and also some important human pathogens, it is important to understand the evolutionary relationships among these bacteria and identify molecular markers that are specific for either all Xanthomonadales or its different genera. Due to the importance of these bacteria for agriculture and human health, the complete genome sequences for 26 Xanthomonadales species/strains are now available in the NCBI database (see Table 1). In addition, genomes for many other species/strains from this order are currently being sequenced and partial sequence information for them is also available in the databases. These genomes provide valuable resource for discovering molecular and biochemical characteristics that are uniquely shared by these bacteria and which should provide novel means for their identification and also as potential new targets for development of drugs targeting these bacteria. Earlier comparative genomic studies on Xanthomonadales have focused on identifying characteristics that are responsible for the virulence and host specificity of different strains and pathovars of *Xanthomonas* and *Xylella* and on understanding the role of LGTs in their genome evolution [3,4,4,7,8,11,29–34,34–36]. A recent study on DNA repair proteins also identified four conserved indels that were specific for the available Xanthomonadales species [28]. However, thus far no detailed study has been carried out which is aimed at identifying genetic or molecular characteristics that are uniquely shared by either all Xanthomonadales or its different genera.

Using genome sequence data, our recent work has focused on identifying Conserved Signature Indels (inserts or deletions) (CSIs) of defined lengths that are present at specific locations in widely distributed proteins and which are uniquely found in particular groups of organisms [37–40]. The most parsimonious explanation of these CSIs is that they resulted from highly specific genetic changes that first occurred in a common ancestor of the particular groups of species and were then passed on to various descendants [37,40,41]. Further, depending upon the presence or absence of these CSIs in outgroup species, it is possible to infer whether a given CSI is an insert or a deletion and this information can be used to develop rooted phylogenetic relationships independently of phylogenetic trees [21,37,42–45]. Additionally, the shared presence of some CSIs in unrelated groups of bacteria can also identify possible cases of LGTs [46]. In this work, we report detailed phylogenetic and comparative analyses of protein sequences from Xanthomonadales genomes to identify CSIs that are specific for these organisms. These studies have identified 13 CSIs that are specific for all sequenced Xanthomonadales species and many others CSIs that provide information regarding evolutionary relationships among these bacteria. These molecular signatures provide novel and highly specific means for identification of Xanthomonadales species and for different types of studies on these bacteria. We also report here several CSIs that are commonly shared by Xanthomonadales and either Beta- and/or Alpha-proteobacteria. However, our analysis indicates that the shared presence of these CSIs in Xanthomonadales and these other bacterial groups is due to independent occurrence of similar genetic changes and not due to LGTs.

Methods

Phylogenetic Analyses

Phylogenetic analyses were conducted on a concatenated sequence alignment for 28 conserved and widely distributed proteins that have been widely used for phylogenetic studies [21,47,48] and are present in all the Xanthomonadales. These proteins included, alanyl-tRNA synthetase, arginyl-tRNA synthetase, cell division protein FtsY, chaperonin GroEL, dimethyladenosine transferase, DNA gyrase subunit A, DNA gyrase subunit B, DNA polymerase I, DNA-dependent helicase II, elongation factor Tu, histidyl-tRNA synthetase, isoleucyl-tRNA synthetase, methionyl-tRNA synthetase, molecular chaperone DnaK, O-sialoglycoprotein endopeptidase, phenylalanyl-tRNA synthetase subunit alpha, phosphatidate cytidylyltransferase, prolyl-tRNA synthetase, RpoB, RpoC, SecA, SecY, serine hydroxymethyltransferase, seryl-tRNA synthetase, signal recognition particle protein, thioredoxin reductase, tryptophanyl-tRNA synthetase and valyl-tRNA synthetase. For each of these proteins, sequences for all sequenced Xanthomonadales and a number of other Gamma-, Beta- and Alpha-proteobacteria were retrieved by Blastp searches and multiple sequence alignments were created by using the CLUSTAL_X 2.0 [49]. These sequence alignments were concatenated into a single large file and the poorly aligned regions from the alignment were removed using Gblocks 0.91 b program [50]. After removal of poor aligned regions, a total of 14621 aligned positions were present in the final dataset. A neighbor-joining (NJ) tree based on 100 bootstrap replicates was constructed using the JTT matrix-based method [51] in MEGA 5 [52]. A maximum-likelihood tree based upon the same sequence data set was also constructed using the Whelan and Goldman+Freq. model [53] using MEGA5. All positions containing gaps were not considered during these tree constructions.

Identification of Xanthomonadales Specific Conserved Signature Indels (CSIs)

To search for signature sequences in different proteins that are specific for Xanthomonadales or for its subclades, Blastp searches were carried out on each proteins (open reading frame) from the genome of *Xylella fastidiosa* 9a5c against the NCBI nr database [35]. The results of blast searches were examined for high scoring homologs. For those proteins for whom high scoring homologs (E value $<1e^{-20}$) were present in Xanthomonadales and several other bacteria, about 10–15 sequences representing different groups were retrieved and multiple sequence alignments were constructed using the CLUSTAL_X 2.0 program [49]. The sequence alignments were visually inspected to identify any conserved indels that were restricted to Xanthomonadales and which were flanked on both sides by at least 5–6 identical/conserved residues in the neighboring 30–50 amino acids [21,40,54]. The conserved indels, which in addition to Xanthomonadales were also present in few other species, were also retained. The indels that were not flanked on both sides by conserved regions were not further evaluated as they do not provide useful molecular markers [23,37,40]. The species distribution of all indels thus identified (~150) was further examined by detailed Blastp searches against the nr database (500 top hits) on short sequence segments containing the indels and their flanking conserved regions. Based upon detailed Blast searches, many original indels queries were found to be uninformative for this study due to a variety of reasons including their presence in only a single species/strain, lack of sequence conservation, presence of other confounding indels in the same area in other species, lack of specificity of the indels for any particular group and large variation

Table 1. Sequence Characteristics of Xanthomonadales genomes.

Organism	GenBank Accession No.	Size (Mbp)	No. of Proteins	% GC content	Reference
Stenotrophomonas maltophilia K279a	AM743169.1	4.8	4386	66	[12]
Stenotrophomonas maltophilia R551-3	CP001111.1	4.6	4039	66	[12]
Stenotrophomonas sp. SKA14	ACDV00000000	4.9	4469	66	JCVI *
Xanthomonas albilineans GPE PC73	FP565176.1	3.7	3114	63	[32]
Xanthomonas axonopodis pv. citri str. 306	AE008923.1	5.3	4312	64	[56]
Xanthomonas axonopodis pv. citrumelo FL 1195	CP002914.1	5.0	4181	65	[55]
Xanthomonas campestris pv. raphani strain 756C	CP002789.1	4.9	4520	65	[33]
Xanthomonas campestris pv. campestris str. 8004	CP000050.1	5.1	4271	64	[30]
Xanthomonas campestris pv. campestris str. ATCC 33913	AE008922.1	5.1	4179	65	[56]
Xanthomonas campestris pv. campestris str. B100	AM920689.1	5.1	4466	65	[56]
Xanthomonas campestris pv. vesicatoria str. 85-10	AM039952.1	5.4	4487	64	[29]
Xanthomonas vesicatoria ATCC 35937	AEQV00000000	5.5	4927	65	[31]
Xanthomonas gardneri ATCC 19865	AEQX00000000	5.5	5027	65	[31]
Xanthomonas oryzae pv. oryzicola BLS256	CP003057.1	4.8	4474	64	[33]
Xanthomonas oryzae pv. oryzae KACC10331	AE013598.1	5.0	4064	63	[5]
Xanthomonas oryzae pv. oryzae MAFF 311018	AP008229.1	5.0	4372	63	NIAS*
Xanthomonas oryzae pv. oryzae PXO99A	CP000967.1	5.2	4988	63	[6]
Xanthomonas perforans 91-118	AEQW00000000	5.2	4637	63	[31]
Xylella fastidiosa 9a5c	AE003849.1	2.8	2766	52	[35]
Xylella fastidiosa M23	CP000941.1	2.5	2104	51	[10]
Xylella fastidiosa M12	CP001011.1	2.6	2161	51	[10]
Xylella fastidiosa Temecula1	AE009442.1	2.5	2034	51	[34]
Xylella fastidiosa subsp. fastidiosa GB514	CP002165	2.5	2216	51	[57]
Pseudoxanthomonas spadix BD-a59	CP003093.2	3.5	3149	67	[58]
Pseudoxanthomonas suwonensis 11-1	CP002446.1	3.4	3070	70	DOE-JGI*
Rhodanobacter sp. 2APBS1	AGIL00000000	4.0	3800	68	DOE-JGI*

*NIAS = Genome was sequenced by National Institute of Agrobiological Sciences, Japan.
*DOE-JGI = Genome was sequenced by DOE Joint Genome Institute USA.
*JCVI = Genome was sequenced by J. Craig Venter Institute, USA.
doi:10.1371/journal.pone.0055216.t001

in their lengths, etc. Hence, such indels were not further studied. However, for different indels those were specific for Xanthomonadales or present in a limited number of other bacteria, sequence information for them were compiled into signature files that are shown here. Due to space considerations, the signature files shown here contain information for only a limited number of species from other bacteria such as Alpha, Beta and Gammaproteobacteria and different strains of the same species are also not shown. However, unless otherwise noted, all of these CSIs are specific for the indicated groups and they are also present in different strains of the Xanthomonadaceae species for which sequence information is available (Table 1).

Results

Phylogenetic Analysis of Genome Sequenced Xanthomonadales

The genome sequences are now available for 26 Xanthomonadales including 15 for *Xanthomonas* species/strains [5,6,29–33,55,56], 5 for different strains/pathovars of *Xylella fastidiosa* [10,34,35,57], 3 for *Stenotrophomonas* species/strains [12], 2 for *Pseudoxanthomonas* species [58] and for the *Rhodanobacter* sp. 2APBS1. Some characteristics of these genomes are listed in Table 1. Their genome sizes varied from 2.5 Mb to 5.3 Mb and the xylem-inhabiting bacterium *Xylella fastidiosa* had the smallest or most reduced genome. Further, in contrast to other Xanthomonadales species/strains whose mol G+C % was in the range of 61–67%, the *Xylella* strains/pathovars have much lower G+C content. The reduced genome size and the lower G+C mole content of *Xylella* strains/pathovars have likely resulted from their adaptation to the more stable xylem environment [7].

The sequence information from these genomes was also used to examine the evolutionary relationships among the sequenced Xanthomonadales species. Detailed phylogenetic studies on Gammaproteobacteria and other proteobacteria based upon concatenated sequences for different large datasets of protein sequences have been reported previously [19,21,22]. In these trees [4,28,59], species from the order Xanthomonadales formed a monophyletic clade and one of the deepest branching lineages within the Gammaproteobacteria [21,22]. Hence, in the present work, phylogenetic trees based upon concatenated sequences were mainly constructed to clarify the branching order of species within the order Xanthomonadales. The dataset employed in this study included sequence information for only a limited number of other

proteobacteria. Figure 1 shows a NJ distance tree based upon concatenated protein sequences, which was rooted using sequences from Alphaproteobacteria. The branching order of various Xanthomonadales species in the ML tree (Figure S1) is very similar to that seen in the NJ tree. In both ML and NJ tree, the Xanthomonadales species formed a strongly supported clade branching within the other Gammaproteobacteria. This clade was separated from all other Gammaproteobacteria by a long branch. Similar monophyletic grouping and branching of the Xanthomonadales species within the Gammaproteobacteria have been observed in earlier studies [19,21,22]. Among the sequenced Xanthomonadales species, *Rhodanobacter* was found to be the deepest branching species and it was separated from all other Xanthomonadales by a long branch. Interestingly, the sequenced *Xanthomonas* species showed polyphyletic branching in the tree, with *X. albilineans* branching deeply and separately from the other *Xanthomonas* species (Figure 1 and Figure S1). The tree shown in Figure 1 provides a phylogenetic framework for understanding and interpreting the significance of various CSIs observed in this work.

Identification of Conserved Signature Indels that are Specific for Xanthomonadales

Our work has identified 13 CSIs that are uniquely present in all sequenced Xanthomonadales including the deepest branching *Rhodanobacter*. Two examples of these CSIs are shown in Figure 2A & B. In the first case (Figure 2A), an 18 aa insert in highly conserved region of the protein glutaminyl-tRNA synthetase, which plays an essential role in protein synthesis by linking glutamine to its cognate tRNA [60]. The large insert in GlnRS is uniquely shared by all available sequences from Xanthomonadales species but not found in any other bacteria (at least the top 500 blast hits). In the other example shown here (Figure 2B), a 4 aa insert in a GTP-binding elongation factor protein (typA) is commonly shared by all sequenced Xanthomonadales, but again it is not found in any other bacteria. Both these CSIs are present in highly conserved regions of the proteins and their sequences are also highly conserved. Because these CSIs are lacking in other bacteria, they constitute inserts in the Xanthomonadales rather than deletions in other bacteria [38]. The sequence information for other CSIs that are uniquely present in all sequenced species/strains of Xanthomonadales is presented in Figures S2–S12 and a summary of their characteristics is provided in Table 2 (first 13 entries). These CSIs include a 7 aa insert in amino acid/peptide transported protein; 5 aa insert in conserved region of the large-conductance mechanosensitive channel protein; a 3 aa insert in LysRS; 2 aa insert in highly conserved region of the protein lipoyl synthase (LipA); 1 aa inserts in the proteins Tgt, LpxA and TolQ; a 13 aa deletion in alpha-2-macroglobulin domain-containing protein and 1 aa deletions in the ParE, PolA and TyrB proteins. Because these CSIs are present in all sequenced Xanthomonadales but not found in any other bacteria, the most likely explanation is that genetic changes responsible for them first occurred in a common ancestor of the Xanthomonadales and then passed on to various descendants by vertical descent.

In addition to these CSIs that are uniquely found in all Xanthomonadales, we have also come across 6 other CSIs, where in addition to the Xanthomonadales, the identified CSIs are also present in 1–3 other Gammaproteobacteria. These species are generally from some of the other deep branching orders of Gammaproteobacteria such as *Chromatiales*, *Methylococcales* and *Cardiobacteriales*, which branch in the proximity of Xanthomonadales [21,22,28]. One example of such a CSI consisting of a 1 aa deletion in a conserved region of the protein glutaminyl-tRNA synthetase that is commonly shared by various Xanthomonadales and also by a few *Methylococcales* and *Cardiobacteriales* species is presented in Figure 3. Sequence information for others CSIs of this kind is presented in Figures S13–S17 and in Table 2 (last six records). Cutino-Jimenez et al. [28]also reported a CSI in Topoisomerase I that was commonly shared by various *Xanthomonadales*, *Methylococcales*, *Cardiobacteriales*, *Chromatiales*, *Legionellales* and *Thiotrichales*. The information provided by these CSIs could prove useful in establishing a specific relationship of the Xanthomonadales to these other deep branching orders of Gammaproteobacteria.

CSIs Supporting the Deeper Branching of *Rhodanobacter* within the Xanthomonadales

In the phylogenetic tree shown in Figure 1 and Figure S1, *Rhodanobacter* sp. 2APBS1 exhibited the deepest branching amongst the sequenced Xanthomonadales. During our analyses, we have found 15 CSIs that are uniquely shared by all other Xanthomonadales except *Rhodanobacter*, supporting the deeper branching of this species in comparison to other Xanthomonadales. Two examples of such CSIs that are uniquely found in different Xanthomonadales, but not in *Rhodanobacter* are shown in Figure 4A & B. In both these cases 5 aa inserts in highly conserved regions of the proteins uroporphyrinogen decarboxylase (Figure 4A) and in the protein tRNA delta(2)-isopentenylpyrophosphate transferase (Figure 4B) are uniquely shared by different sequenced Xanthomonadales except *Rhodanobacter*. These CSIs are not present in any other bacteria. A summary of the characteristics of different CSIs showing this type of species distribution pattern is presented in Table 3 and the sequence alignments of the corresponding proteins are provided as Figures S18–S30. The proteins in which these CSIs are found include protoheme IX farnesyltransferase (CoxD), DNA polymerase III alpha subunit (DnaE), DEAD box helicase domain-containing protein, ribose-5-phosphate isomerase A (RpiA), DNA polymerase I (PolA), glucose-6-phosphate 1-dehydrogenase (Zwf1), AspRS, 2-oxoglutarate-dehydrogenase E1 component (SucA), coproporphyrinogen III oxidase (CpoX), and TrmD. In a few of these cases, the CSIs under consideration was also not found in one or both of the *Pseudoxanthomonas* species, supporting their deeper branching in comparison to other Xanthomonadales genera (viz. *Xylella*, *Xanthomonas* and *Stenotrophomonas*) (Figures S32–S34). In a recent study, Cutino-Jimenez et al. [28] had reported four CSIs in DNA repair proteins that were indicated to be specific for Xanthomonadales. Our analyses of these CSIs, which were also identified in our work, indicate that they are lacking in either *Rhodanobacter* (4 aa insert in DnaE and 1 aa insert in RecA) or both *Rhodanobacter* and in *P. suwonensis* (5 aa insert in MutS and >50 aa insert in LigA) (Figures S29–S31 and S35). The information for these CSIs is also summarized in Table 3. Based upon the species distribution of these CSIs and the branching positions of *Rhodanobacter* (and *Pseudoxanthomonas*) in phylogenetic trees, the genetic changes responsible for these CSIs likely occurred in common ancestors of other Xanthomonadales species after the divergence of *Rhodanobacter* sp. 2APBS1 and also in some cases that of *Pseudoxanthomonas* species.

In addition to the CSIs discussed above 4 other proteins contains CSIs of different lengths at the same position, which are uniquely shared by all sequenced species/strains of Xanthomonadales except *Rhodanobacter* sp. 2APBS1 and in some cases *Pseudoxanthomonas*. However, these CSIs due to differences in their lengths are also able to distinguish between different genera of *Xanthomonadaceae*. Two examples of such CSIs are presented in Figure 5. In the first case in the protein 5′-nucleotidase (Figure 5A), which catalyzes the hydrolysis of nucleotides to nucleosides, a 13

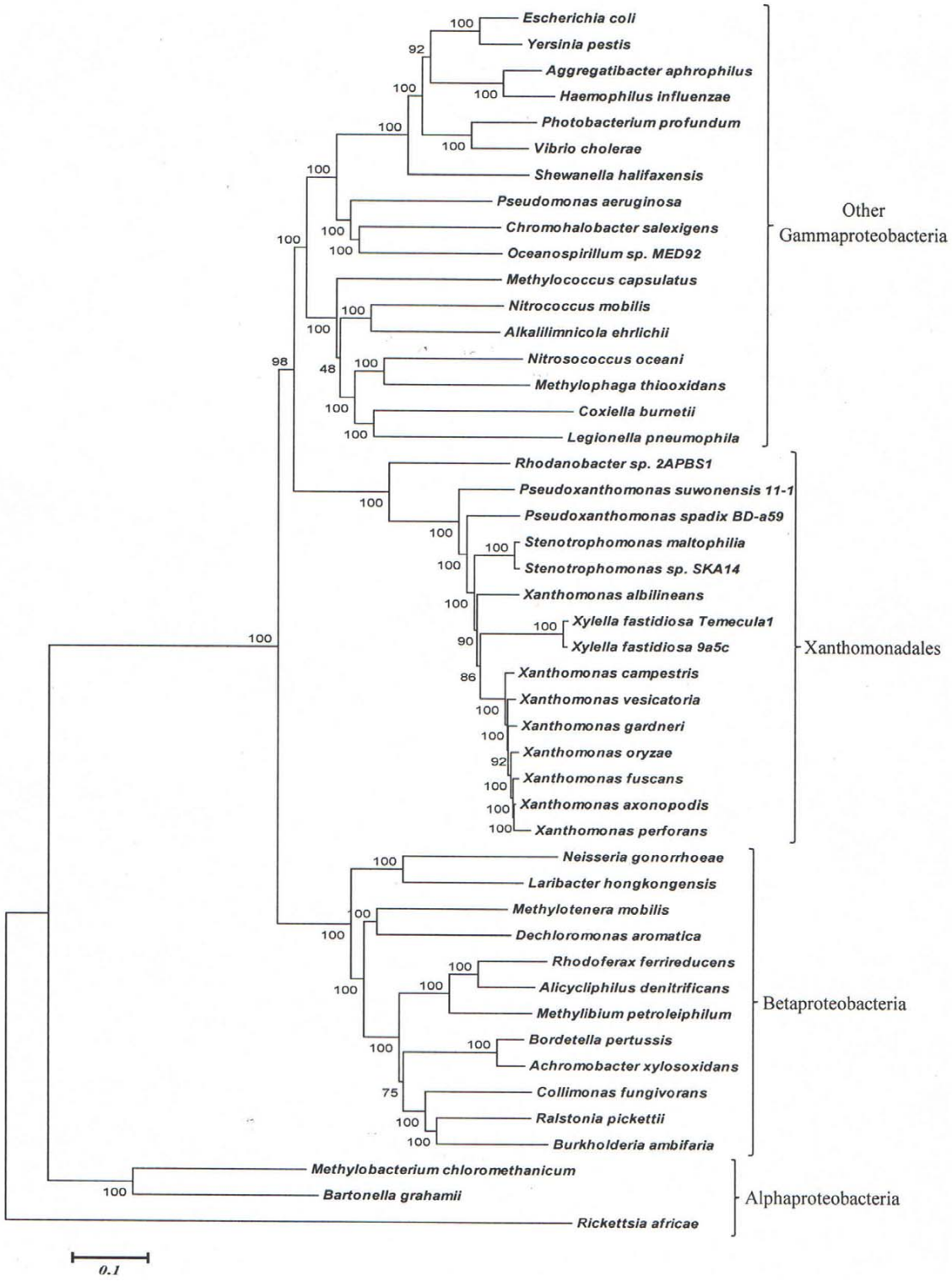


Figure 1. Phylogenetic tree for Xanthomonadales based on concatenated sequences for 28 conserved proteins. The tree shown is a NJ distance tree, however, similar branching was observed in the ML tree (Figure S1). The observed bootstrap scores for various nodes are shown on the branch points. The tree was rooted using sequences from Alphaproteobacteria.
doi:10.1371/journal.pone.0055216.g001

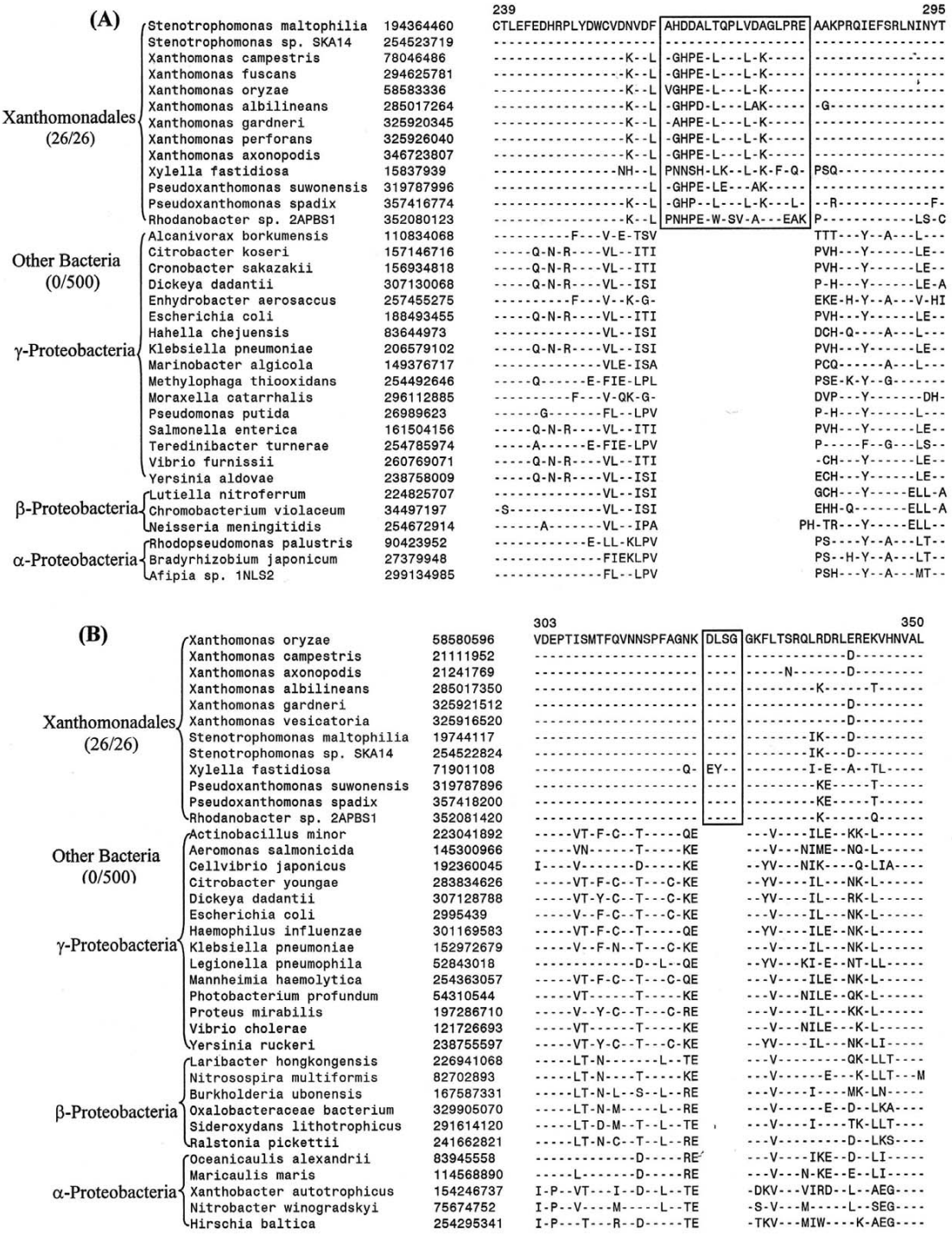


Figure 2. Examples of conserved signature indels (CSIs) that are specific for the order Xanthomonadales. Excerpts are shown from the sequence alignments of (A) Glutaminyl t-RNA synthetase and (B) GTP-binding elongation factor proteins showing two CSIs that are uniquely found in various sequenced Xanthomonadales species, but not found in any other bacteria. Information for other CSIs that are specific for the Xanthomonadales is provided in Figures S2–S12 and Table 2. The dashes in these as well as all other alignments show identity with the amino acid on the top line. The Gene bank identification numbers of various sequences are shown in the second column and the numbers on the top indicate the position of this sequence in the species shown on the top line. The sequence information is shown here for only representative species. However, unless otherwise indicated, these CSIs are highly specific for the indicated group of species.
doi:10.1371/journal.pone.0055216.g002

Table 2. Conserved Signatures Indels that are specific for Xanthomonadales.

Protein Name	Gene Name	GenBank Identifier	Figure No	Indel Size	Indel Position[a]	Exceptions[b]
Glutaminyl-tRNA synthetase	glnS	194364460	Figure 2A	18 aa ins	239–295	None
GTP-binding elongation factor protein	typA	58580596	Figure. 2B	4 aa ins	303–350	None
Amino acid/peptide transporter	–	71275790	Figure S2	7 aa ins	164–213	None
Large-conductance mechanosensitive channel	mscL	294667079	Figure S3	5 aa ins	40–85	None
Lysyl-tRNA synthetase	lysS	194365604	Figure S4	3 aa ins	34–85	None
Lipoyl synthase	lipA	58583575	Figure S5	2 aa ins	156–209	None
Queuine tRNA-ribosyltransferase	tgt	194365393	Figure S6	1 aa ins	289–339	None
Acyl-(acyl-carrier-protein)–UDP-N-acetyl-glucosamine O-acyltransferase	lpxA	71275623	Figure S7	1 aa ins	164–210	None
TolQ protein	tolQ	21232451	Figure S8	1 aa ins	177–217	None
Alpha-2-macroglobulin domain-containing protein	–	194366795	Figure S9	13 aa del	607–661	None
DNA topoisomerase IV subunit B	parE	84624476	Figure S10	1 aa del	282–326	None
DNA polymerase I	polA	194367713	Figure S11	1 aa del	28–65	None
Aromatic amino acid aminotransferase	tyrB	28197970	Figure S12	1 aa del	306–354	None
Glutaminyl-tRNA synthetase	glnS	194364460	Figure 3	1 aa del	77–131	*Methylobacter tundripaludum, Methylomicrobium album BG8,Dichelobacter nodosus*
DNA polymerase III subunit beta	RpoB	194363780	Figure S13	1 aa del	44–81	*Marinomonas* sp. MWYL1, *Thioalkalivibrio sp. HL-EbGR7*
Lipid-A-disaccharide synthase	lpxB	190573490	Figure S14	2 aa ins	317–358	*Cardiobacterium hominis, Allochromatium vinosum, Alteromonadales bacterium*
Carbamoyl phosphate synthase large subunit	carB	166711938	Figure S15	1 aa ins	403–457	*Marinobacter sp. ELB17*
Putative secreted protein	–	188992701	Figure S16	1 aa ins	1285–1318	*Teredinibacter turnerae*
Aminopeptidase P	pepP	294627124	Figure S17	1 aa del	211–246	*Thioalkalivibrio sp. HL-EbGR7,Alkalilimnicola ehrlichii*

[a]The indel position provided indicates the region of the protein containing the CSI.
[b]For details go to respective figures.
doi:10.1371/journal.pone.0055216.t002

aa insert is uniquely shared by all *Xanthomonas* and *Xylella* species, whereas the two *Stenotrophomonas* species have an 11 aa insert in the same position. Because both these CSIs are present at the same position and they are related in sequences, the most likely explanation about their occurrence is that a 13 aa insert was initially introduced in a common ancestor of the *Xanthomonas*, *Xylella* and *Stenotrophomonas* genera and it was followed by a 2 aa deletion in the genus *Stenotrophomonas*. Alternatively, an 11 aa insert was initially introduced in a common ancestor of these three genera followed by another 2 aa insert in a common ancestor of the *Xanthomonas* and *Xylella* genera. Likewise, in a conserved region of the asparagine synthase b protein (AsnB), a 5 aa insert is present in various *Xylella*, *Xanthomonas* and *Pseudoxanthomonas*, whereas the two *Stenotrophomonas* species have a smaller insert (4 aa) in this position (Figure S27). The AsnB protein also contains another CSI in a different position (see Figure S28), where a 1 aa insert is present in *Xylella*, *Xanthomonas* and *Pseudoxanthomonas*, species, whereas the two *Stenotrophomonas* species have a 2 aa insert in the same position. In another example of this kind, in the protein CTP synthetase, a 2 aa insert in a conserved region is uniquely shared by various *Xylella*, *Xanthomonas* and *Stenotrophomonas* species/strains, whereas the two *Pseudoxanthomonas* species contain a 1 aa insert in this position (Figure 5B). These CSIs, in addition to supporting the deeper branching of *Rhodanobacter* in comparison to other Xanthomonadales, also serve to differentiate *Stenotrophomonas* and *Pseudoxanthomonas* species from other genera of *Xanthomonadaceae*.

CSIs that are Commonly Shared by Xanthomonadales and Some Alpha- and Beta-proteobacteria

In addition to the above proteins that contained CSIs, which were highly specific for Xanthomonadales species (or 1–2 closely related species), our analyses have also identified 7 other CSIs, which in addition to various Xanthomonadales are also shared by some Betaproteobacteria and/or Alphaproteobacteria. Two examples of these CSIs are shown in Figures 6 and 7. In the protein valyl-tRNA synthetase, which plays an essential role in protein synthesis, a 13 aa insert in a highly conserved region is present in all sequenced Xanthomonadales, except *Rhodanobacter* (Figure 6). Interestingly, a very similar CSI is also present in several species belonging to the class Alphaproteobacteria (e.g. *Ahrensia sp. R2A130*, *Labrenzia alexandrii*, *Rhodobacter capsulatus*, *Sagittula stellata* etc.) whereas other Alphaproteobacteria do not contain this insert. In the other example shown here (Figure 7), in the protein carbamoyl phosphate synthase large subunit (CarB), a 1 aa insert in a conserved region is commonly shared by various Xanthomonadales and a subgroup of Betaproteobacteria (mainly *Burkholderiales*), but not by any other bacterial groups. The shared presence of similar CSIs by different Xanthomonadales and species from these other classes of proteobacteria could result from a variety of possibilities including lateral transfers of genes for these proteins between these two groups of bacteria or alternatively by independent occurrence of similar genetic changes in these lineages.

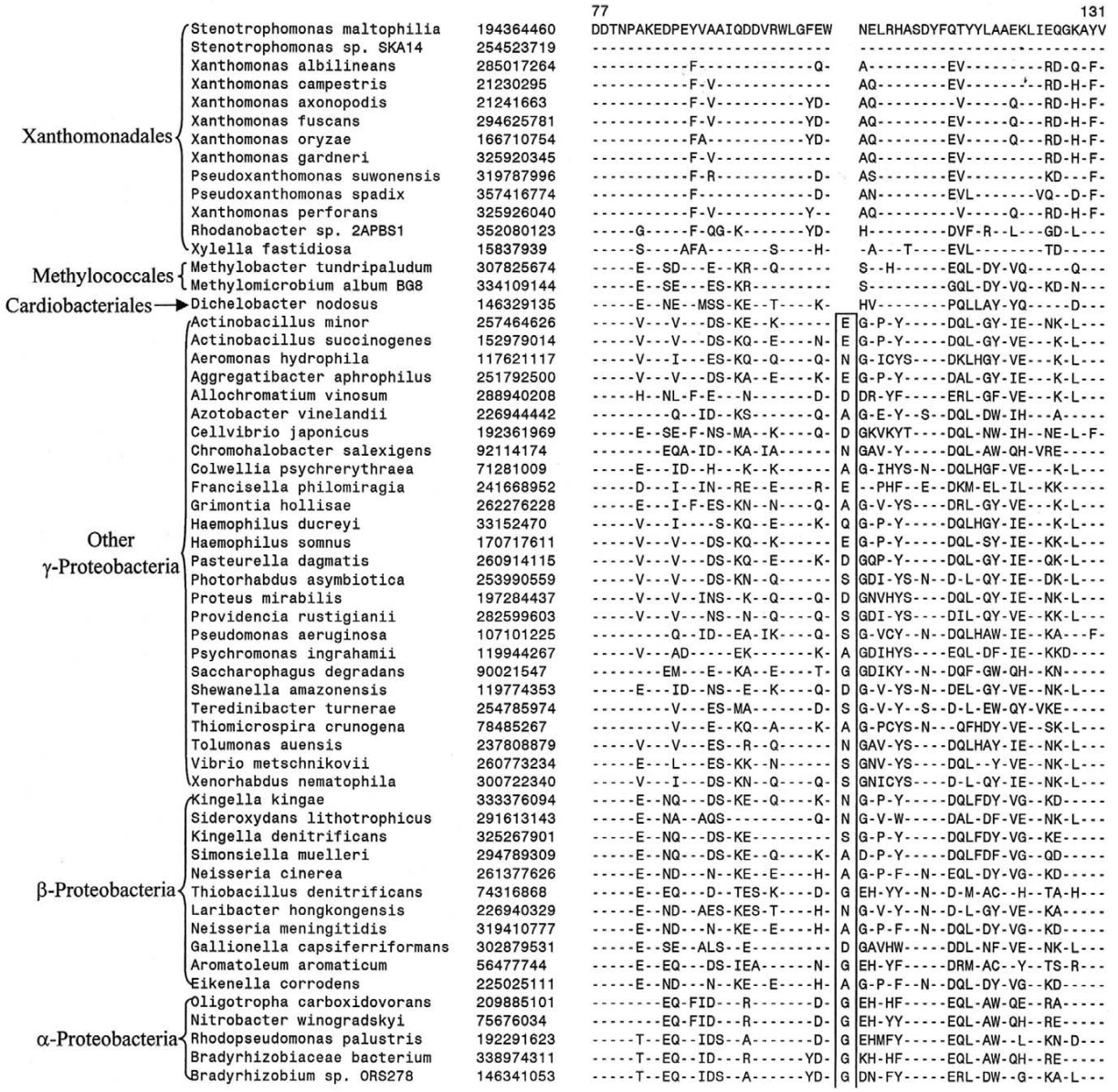


Figure 3. Partial sequence alignment of glutaminyl t-RNA synthetase showing a CSI that is specifically present in various sequenced Xanthomonadales and some other Gammaproteobacteria. This CSI as well as a few other CSIs identified in this work (see Table 2 and Figures S13–S17) suggest a possible relationship of Xanthomonadales to these deep branching orders of Gammaproteobacteria.
doi:10.1371/journal.pone.0055216.g003

To distinguish between these possibilities, phylogenetic trees for the ValRS and CarB sequences for the same species as shown in Figures 6 and Figure 7 were constructed. In the tree based upon ValRS sequences, which is shown Figure 8, all of the Alphaproteobacteria species (both containing and lacking the insert) formed a strongly supported clade that branched distinctly from the Xanthomonadales. The Xanthomonadales species in this tree branched in between the clades consisting of Betaproteobacteria and the other Gammaproteobacteria, but that is not surprising in view of phylogenetic position within the Gammaproteobacteria. If the shared presence of the CSI in the Xanthomonadales and the CSI-containing Alphaproteobacteria was due to LGTs, then the Alphaproteobacteria containing this CSI should have branched with the Xanthomonadales, which is not observed here. Similarly, in the tree based upon CarB sequences (Figure S36), all of the Betaproteobacteria branched together and no association was observed between the insert containing Betaproteobacteria and the Xanthomonadales. These results do not support the possibility that LGT was responsible for the shared presence of CSIs in these two groups. Instead in the phylogenetic trees shown in Figure 8 and Figure S36, the clades comprising of the inserts containing Alphaproteobacteria or Betaproteobacteria formed distinct subclades within the rest of the Alpha- or Beta-proteobacteria. Thus, it is likely that the genetic changes responsible for these CSI

Table 3. CSIs that are specific for Xanthomonadales except *Rhodanobacter* sp. 2APBS1.

Protein Name	Gene Name	GenBank Identifier	Figure No	Indel Size	Indel Position[a]	Specificity within Xanthomonadales
Uroporphyrinogen decarboxylase	hemE	294625972	Figure 4A	5 aa ins	295–340	All except *Rhodanobacter* sp. 2APBS1
tRNA delta(2)-isopentenylpyrophosphate transferase	miaA	194365248	Figure 4B	5 aa ins	219–256	All except *Rhodanobacter* sp. 2APBS1
Protoheme IX farnesyltransferase	coxD	15837961	Figure S18	4 aa ins	150–192	All except *Rhodanobacter* sp. 2APBS1
DNA polymerase III subunit alpha	dnaE	21242159	Figure S19	1 aa ins	583–638	All except *Rhodanobacter* sp. 2APBS1
DEAD box helicase domain-containing protein	–	194364258	Figure S20	1 aa ins	155–200	All except *Rhodanobacter* sp. 2APBS1
Ribose-5-phosphate isomerase A	rpiA	194367055	Figure S21	1 aa ins	127–169	All except *Rhodanobacter* sp. 2APBS1
DNA polymerase I	polA	21244827	Figure S22	1 aa ins	136–180	All except *Rhodanobacter* sp. 2APBS1
Aspartyl-tRNA synthetase	aspS	194366904	Figure S23	4 aa del	343–391	All except *Rhodanobacter* sp. 2APBS1
2-oxoglutarate-dehydrogenase E1 component	sucA	194366403	Figure S24	1 aa del	782–830	All except *Rhodanobacter* sp. 2APBS1
Coproporphyrinogen III oxidase	cpoX	194367710	Figure S25	1 aa del	166–215	All except *Rhodanobacter* sp. 2APBS1
2-oxoglutarate-dehydrogenase E1 component	sucA	194366403	Figure S26	1 aa del	106–164	All except *Rhodanobacter* sp. 2APBS1
Asparagine synthase b protein	asnB	285018780	Figure S27	4–5 aa ins	404–445	All except *Rhodanobacter* sp. 2APBS1
Asparagine synthase b protein	asnB	194365058	Figure S28	1–2 aa ins	96–132	All except *Rhodanobacter* sp. 2APBS1
DNA polymerase III subunit alpha [b]	dnaE	77747494	Figure S29	4 aa ins	522–576	All except *Rhodanobacter* sp. 2APBS1
DNA repair protein RecA[b]	recA	15836728	Figure S30	2 aa ins	172–238	All except *Rhodanobacter* sp. 2APBS1
5′-nucleotidase	–	21231001	Figure 5A	11–13 aa ins	123–188	All except *Rhodanobacter* sp. 2APBS1 & *Pseudoxanthomonas suwonensis*
CTP synthetase	pyrG	194365226	Figure 5B	2 aa ins	253–291	All except *Rhodanobacter* sp. 2APBS1 *Pseudoxanthomonas suwonensis* & *Pseudoxanthomonas suwonensis*
DNA mismatch repair protein MutS [b]	mutS	15838317	Figure S31	5 aa ins	765–806	All except *Rhodanobacter* sp. 2APBS1 & *Pseudoxanthomonas suwonensis*
DNA polymerase III subunit alpha	dnaE	194365029	Figure S32	2 aa del	65–120	All except *Rhodanobacter* sp. 2APBS1 & *Pseudoxanthomonas suwonensis*
tRNA (guanine-N(1)-)-methyltransferase	trmD	194364933	Figure S33	2 aa ins	140–200	All except *Rhodanobacter* sp. 2APBS1 & *Pseudoxanthomonas suwonensis*
Glucose-6-phosphate 1-dehydrogenase	zwf	190573773	Figure S34	4 aa del	290–334	All except *Rhodanobacter* sp. 2APBS1 & *Pseudoxanthomonas spadix BD-a59*
DNA ligase NAD dependent [b]	ligA	77747612	Figure S35	57–65 aa ins	461–583	All except *Rhodanobacter* sp. 2APBS1 & *Pseudoxanthomonas suwonensis*

[a]The indel position provided indicates the region of the protein containing the CSI.
[b]These CSIs have been previously described [28].
doi:10.1371/journal.pone.0055216.t003

occurred independently in the common ancestors of these subclades of species.

Besides these two proteins that contained CSIs, which were commonly shared by Xanthomonadales and either some Alpha- or Beta-proteobacteria, five other proteins were identified that contained CSIs showing similar species distributions. These included: two CSIs consisting of 1 aa conserved deletions in a hypothetical protein XOO1065 and the protein orotate phosphoribosyltransferase (PyrE) that are commonly shared by various Xanthomonadales and some Betaproteobacteria (Figures S37 and S38); two CSIs consisting of 1 aa and 2 aa inserts in the proteins putative ribonuclease HII (RnhB) and glycyl-tRNA synthetase subunit beta (GlyS) that are also commonly shared by various Xanthomonadales and some Betaproteobacteria (Figures S39 and S40); a 1 aa deletion in a conserved region in the septum site-determining protein MinD that is commonly shared by Xanthomonadales and some Alpha- and Beta-proteobacteria (Figure S41). The phylogenetic trees based upon the sequences of these proteins are shown in Figures S42 to S46. In all of these trees, the proteobacterial groups which contained similar CSIs as found in the Xanthomonadales did not branch with the Xanthomonadales. These results provide evidence that the CSIs in these other bacterial groups have originated independently and their shared presence is not due to LGTs from Xanthomonadales.

Discussion

The Xanthomonadales species harbor many major plant pathogens [3,4,9] as well as some important human pathogens. However, these bacteria are presently distinguished from other bacteria solely on the basis of their branching in phylogenetic trees (primarily 16S rRNA) and no molecular or biochemical characteristic that is uniquely shared by various species from this group of bacteria is currently known [1]. This paper reports detailed phylogenetic and comparative genomic analyses of sequenced Xanthomonadales species to identify molecular markers that are specific for these bacteria and which are also helpful in understanding their evolutionary relationships. We report here for the first time 13 molecular signatures consisting of conserved indels in widely distributed proteins that are distinctive characteristics of all sequenced Xanthomonadales species, but they are not found in any other bacteria. In view of their Xanthomonadales-specificity, the most parsimonious explanation to account for these CSIs is that the rare genetic changes responsible for them occurred

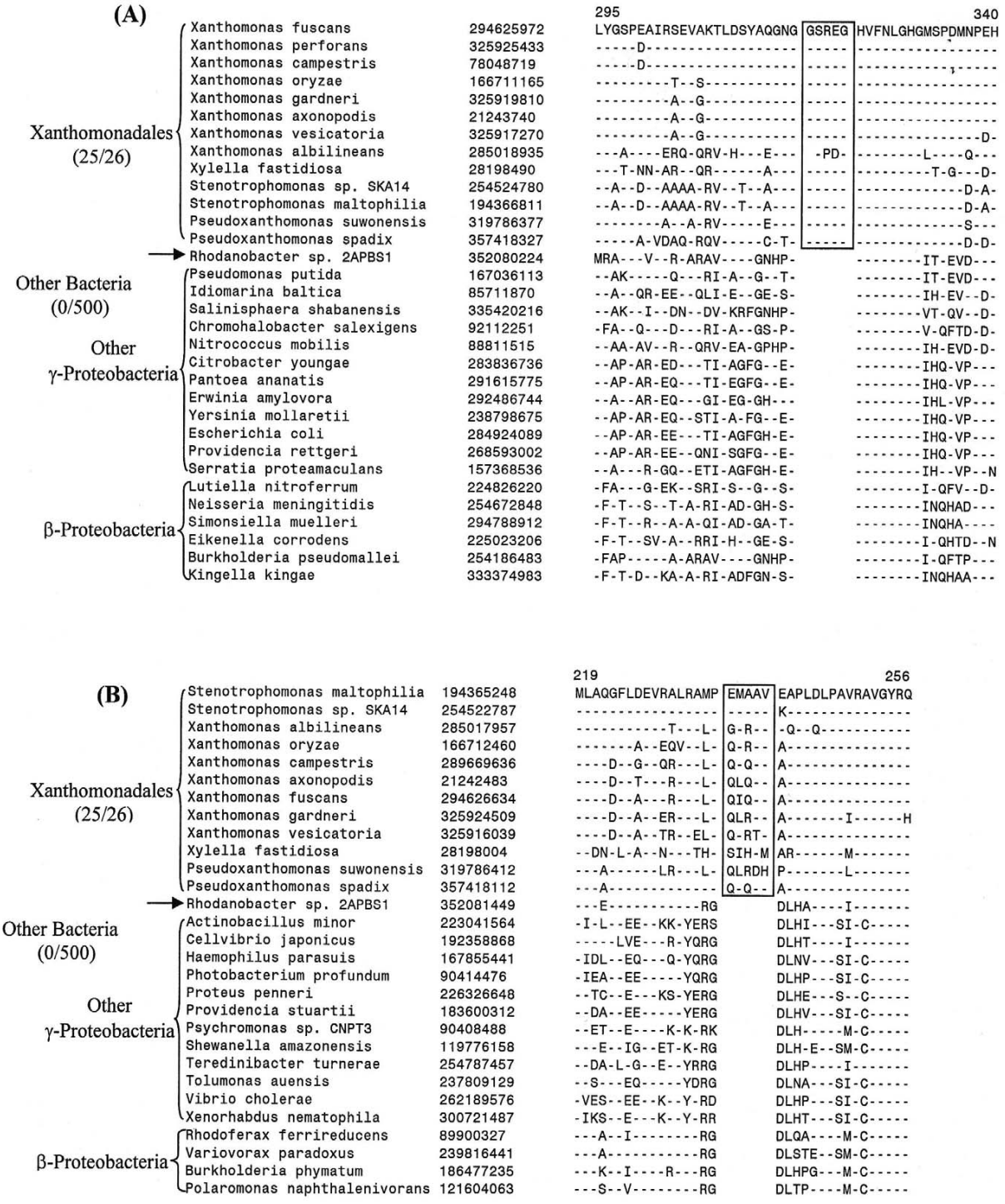


Figure 4. Examples of CSIs those are present in various Xanthomonadales species except *Rhodanobacter* sp. 2APBS1. Excerpts are shown from the sequence alignments of (A) uroporphyrinogen decarboxylase (HemE) and (B) tRNA delta(2)-isopentenylpyrophosphate transferase (MiaA) proteins showing two conserved signature indels (boxed) that are specifically found in various sequenced Xanthomonadales species, except *Rhodanobacter* sp. 2APBS1. These CSIs were likely introduced in these genes in a common ancestor of the Xanthomonadales after branching of *Rhodanobacter*. Information for CSIs in other proteins showing similar species specificities is provided in Figures S18–S30 and Table 3.
doi:10.1371/journal.pone.0055216.g004

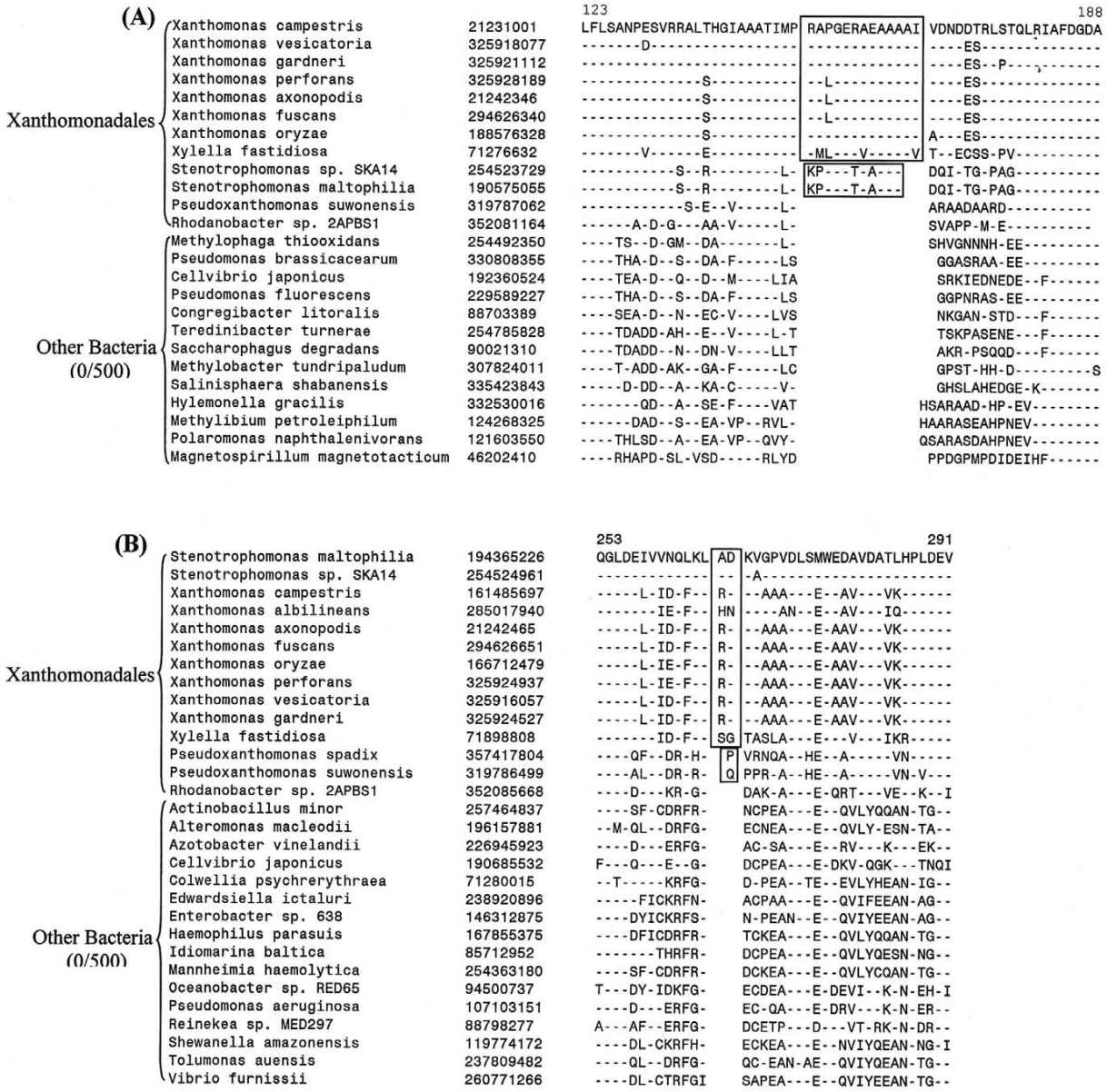


Figure 5. Example of CSIs those are able to distinguish two different clades of Xanthomonadales. Partial sequence alignments are shown of the proteins (A) 5′-nucleotidase and (B) CTP synthetase showing two CSI, which due to their different lengths are able to distinguish between two different clades of Xanthomonadales. In (A), a 13 aa insert is present in all of the *Xanthomonas* and *Xylella* species, whereas the two *Stenotrophomonas* spp. contain an 11 aa insert in this position. Similarly, in (B), all of the *Xanthomonas*, *Xylella* and *Stenotrophomonas* species have a 2 aa insert, whereas the two Pseudoxanthomonas spp. contain a 1 aa insert in this position. Different possibilities to account for these CSIs are discussed in the text.

doi:10.1371/journal.pone.0055216.g005

only once in a common ancestor of the Xanthomonadales and were then passed on to various descendent species vertically as shown in Figure 9 [37,54,61]. Further, the absence of these CSIs in all other bacteria strongly indicates that the genes for these proteins have not been laterally transferred from Xanthomonadales to other bacterial groups or vice versa. Thus, these molecular signatures (or synapomorphies) provide novel means for the identification and circumscription of species from the order Xanthomonadales in clear molecular terms.

We also report in this work detailed phylogenetic analyses of (sequenced) Xanthomonadales species based upon concatenated sequences for 28 widely distributed proteins. Earlier phylogenetic studies on Xanthomonadales are mainly based upon 16S rRNA or single genes such as Gyrase B and most of them cover only the genus *Xanthomonas* [4,15,59,62,63]. Among a number of novel relationships seen in this tree, these trees showed that *Rhodanobacter* sp. 2APBS1 formed the deepest branch within the Xanthomonadales and it was separated from all other species by a long branch.

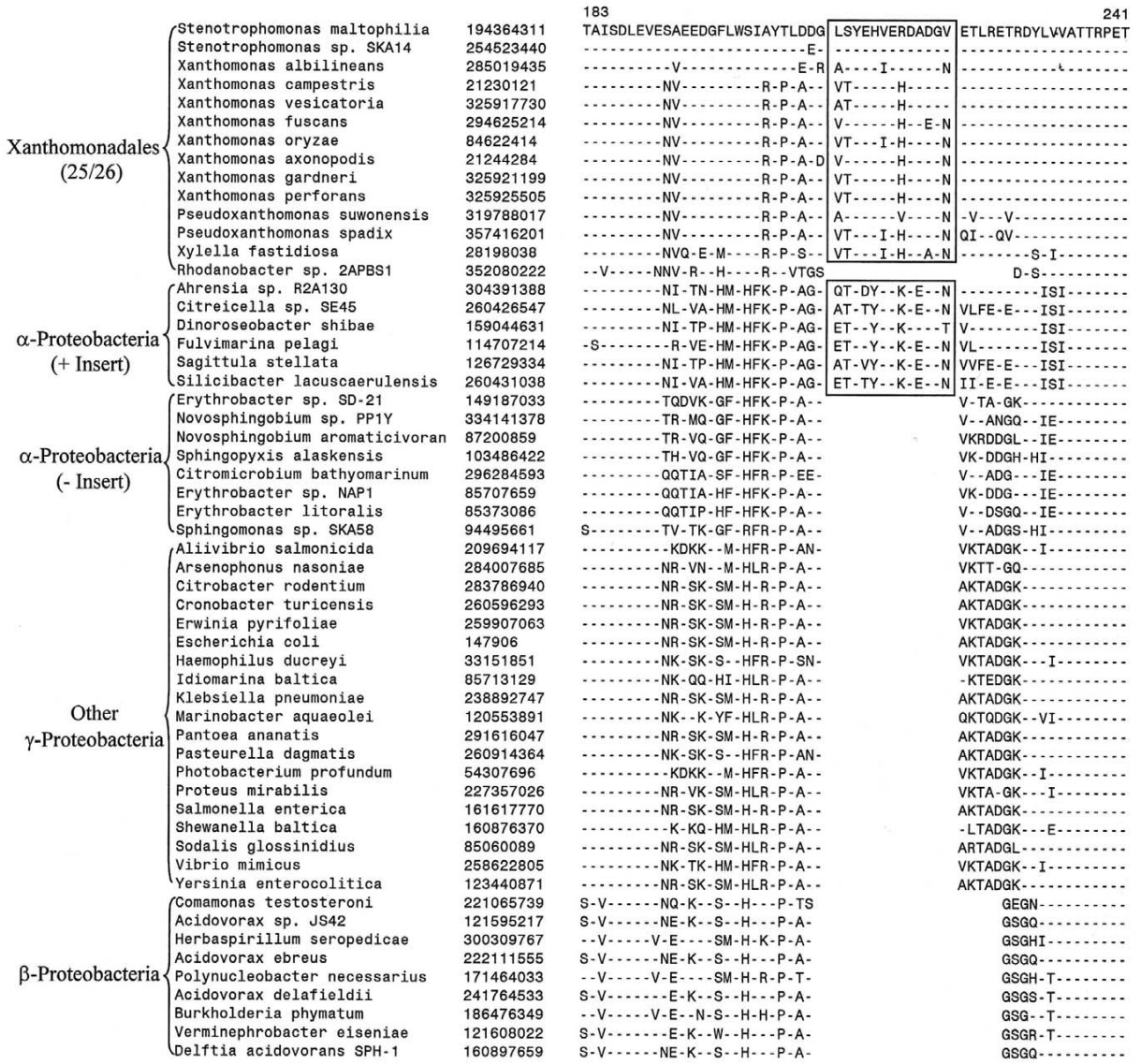


Figure 6. Partial sequence alignments of valyl t-RNA synthetase showing a 13 aa insert that is commonly shared by various Xanthomonadales and a subgroup of Alphaproteobacteria. Other Alpha- and Gamma-proteobacteria do not contain this insert.
doi:10.1371/journal.pone.0055216.g006

The branching of *Pseudoxanthomonas* and then other *Xanthomonadaceae* genera followed it. Importantly, our analyses have also identified 15 CSIs that are uniquely present in all other Xanthomonadales, except *Rhodanobacter* and in a few cases also by the *Pseudoxanthomonas* species. The genetic changes responsible for theses CSIs were likely introduced in a common ancestor of the other Xanthomonadales after the branching of *Rhodanobacter* and also in some cases *Pseudoxanthomonas* (Figure 9) and they provide independent evidence for the deep branching of these lineages with respect to other genera within this order.

Xanthomonadales species are indicated to have undergone extensive LGTs with other prokaryotic taxa particularly Alpha, Beta and some orders of Gamma- proteobacteria and in some cases with Archaea as well [16,24–27]. In the present work, we have also identified several examples where a given CSI, in addition to being shared by all or most Xanthomonadales, was also present in some species from other groups of bacteria, most commonly from Alpha-, Beta- and Gamma- proteobacteria. Of these CSIs, five were present only in 1–3 species from other deep branching orders of Gammaproteobacteria and their possible significance is discussed below. Seven other CSIs were commonly shared by various Xanthomonadales and also several Betaproteobacteria and/or both Alpha- and Beta- proteobacteria. The shared presence of these CSIs between Xanthomonadales and these other proteobacteria could result from a number of possibilities including later transfer of the corresponding genes between these groups of bacteria or independent occurrence of similar genetic changes in these groups. However, phylogenetic trees based upon these protein sequences showed that Xanthomonadales species and the Alpha- and/or Beta- proteobacteria containing similar

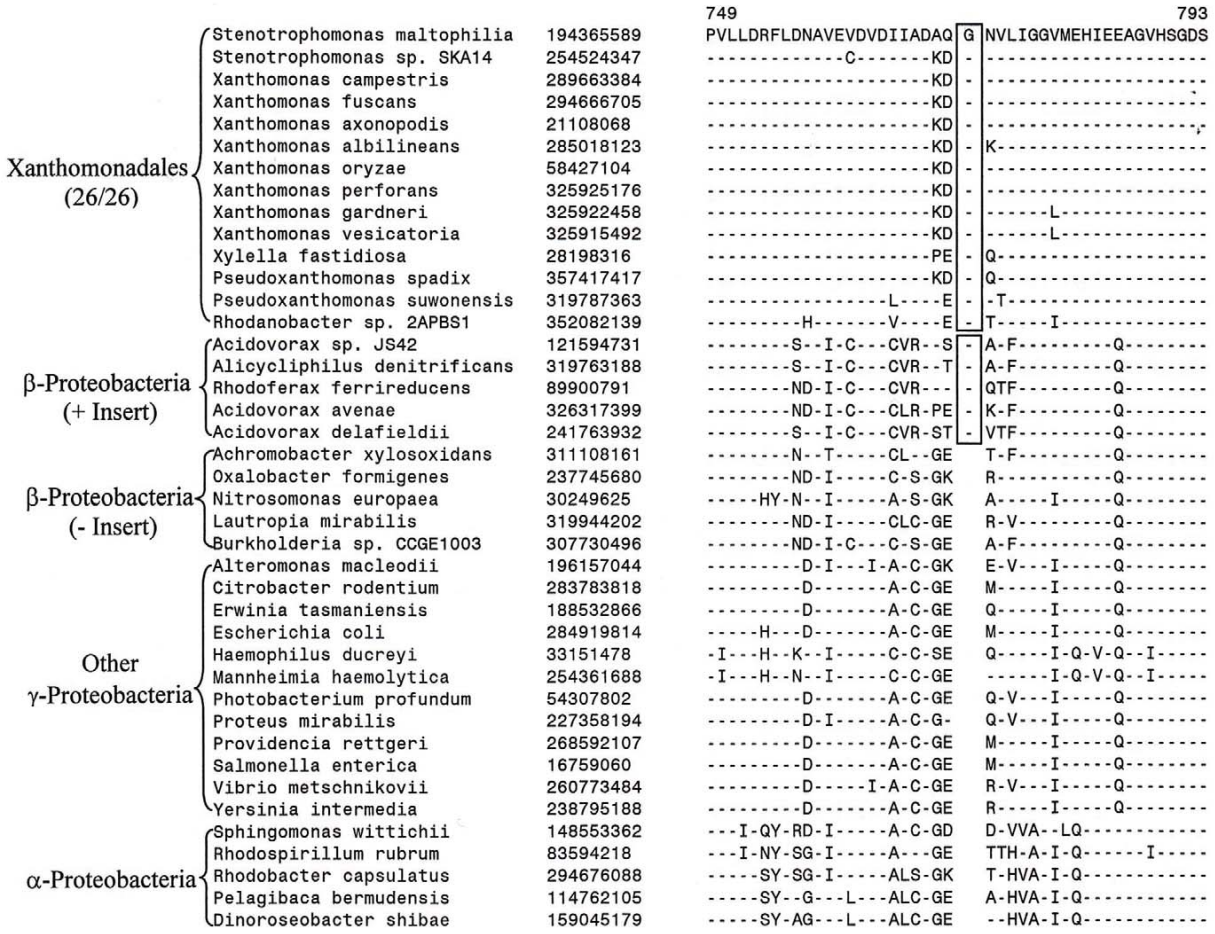


Figure 7. Partial sequence alignment of carbamoyl phosphate synthase showing a 1 aa insert that is commonly shared by Xanthomonadales and a subgroup of Betaproteobacteria. The distinct branching of these two groups in a phylogenetic tree based upon CarB sequence (Figure S36) provides evidence that this shared CSIs is not a result of LGT.
doi:10.1371/journal.pone.0055216.g007

CSIs branched separately from each other, indicating that the presence of similar CSIs in these groups of bacteria was not due to LGTs. Therefore, genetic changes leading to similar CSIs in these groups likely occurred independently due to similar functional requirements for these CSIs. Although in our work we have not come across many examples of LGTs between Xanthomonadales and other groups of bacteria, our analyses is based only on proteins that contain conserved indels. Such genes/proteins represent only a small fraction of the total genes that are found in various genomes. Because most of these proteins are involved in essential functions, they are less prone to LGTs. In contrast, extensive work that Menck and coworkers have carried out on identification of cases of LGTs is primarily on species from the genus *Xanthomonas* [16,24–27], which have thus far not studied in detail.

Xanthomonadales is one of the deepest branching orders within the Class Gammaproteobacteria. Some of the other orders that branch in its proximity include *Chromatiales*, *Methylococcales*, *Cardiobacteriales*, *Legionelalles* and *Thiotrichales*. However, the relationship of Xanthomonadales to these other orders is presently not understood. In the present work, we also identified six other CSIs (Table 2, last six entries), which in addition to various Xanthomonadales were also uniquely shared by 1 or 2 species from these orders of Gammaproteobacteria. The shared presence of these CSIs by Xanthomonadales and some of these other orders of Gammaproteobacteria suggests that either these orders are closely related or that similar genetic changes have occurred in them independently. However, further information from additional species from these orders will be necessary to establish whether the Xanthomonadales and some of these other orders of Gammaproteobacteria are specifically related and form a higher taxonomic clade within the Gammaproteobacteria.

The focus of the present study was on identifying molecular signatures that are specific for either the entire Xanthomonadales order or some of its deep branching lineages. Thus far, we have not carried out careful analyses of various signature sequences that are specific for specific genera viz. *Xylella, Xanthomonas*, *Stenotrophomonas* and *Pseudoxanthomonas* and such studies will be part of our future work. Nonetheless, based upon the identified molecular signatures it is now possible to identify and circumscribe species from the order Xanthomonadales from all other bacteria in clear molecular terms based upon large numbers of discrete molecular characteristics. Based upon our earlier work on CSIs for other groups/phyla of bacteria, most of these CSIs have degree of

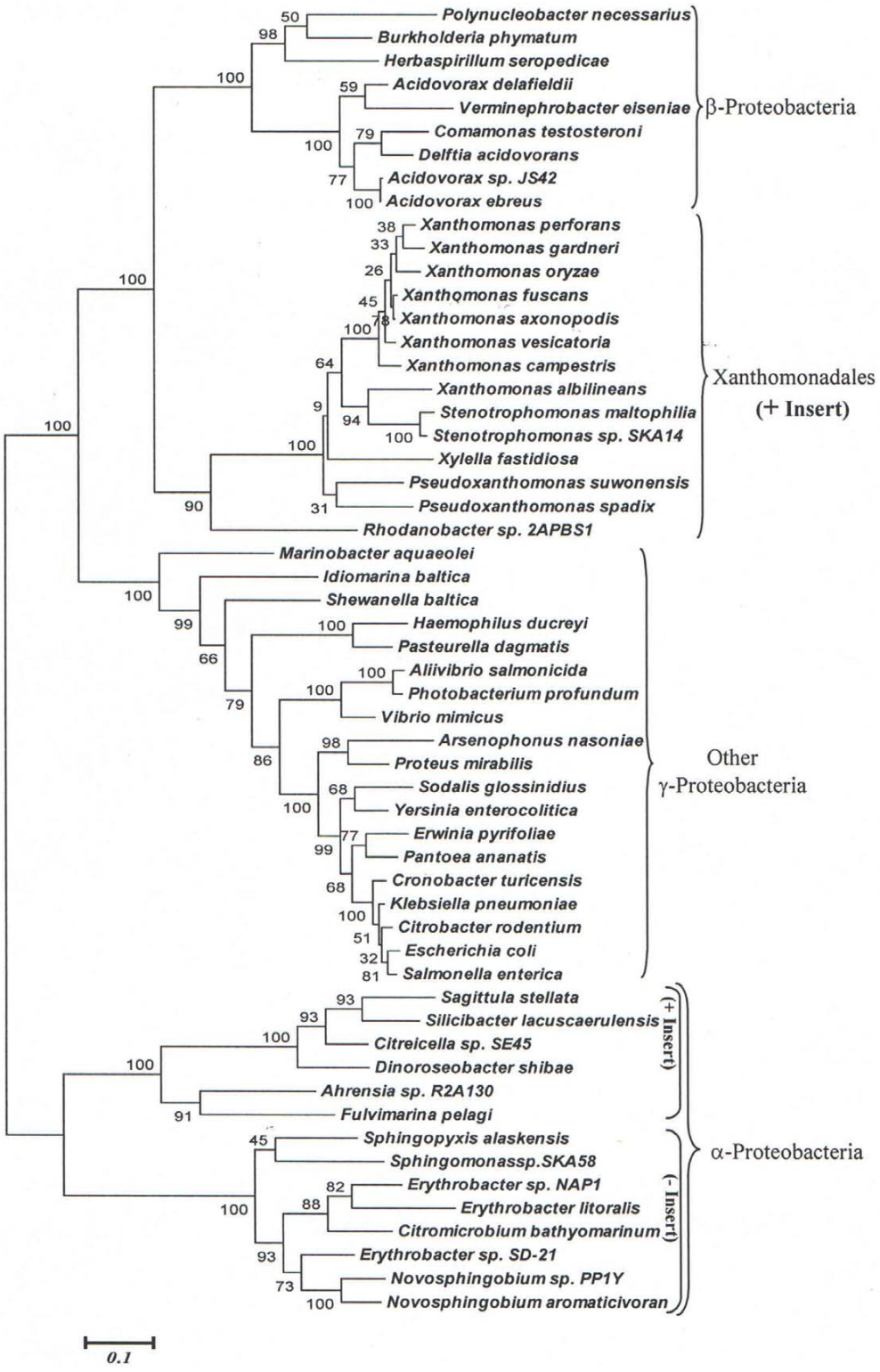


Figure 8. Phylogenetic tree based upon valyl t-RNA synthetase sequences. The distinct branching of Xanthomonadales and the Alphaproteobacteria containing this insert suggests that the shared presence of this CSIs in these two groups is not due to a LGT.
doi:10.1371/journal.pone.0055216.g008

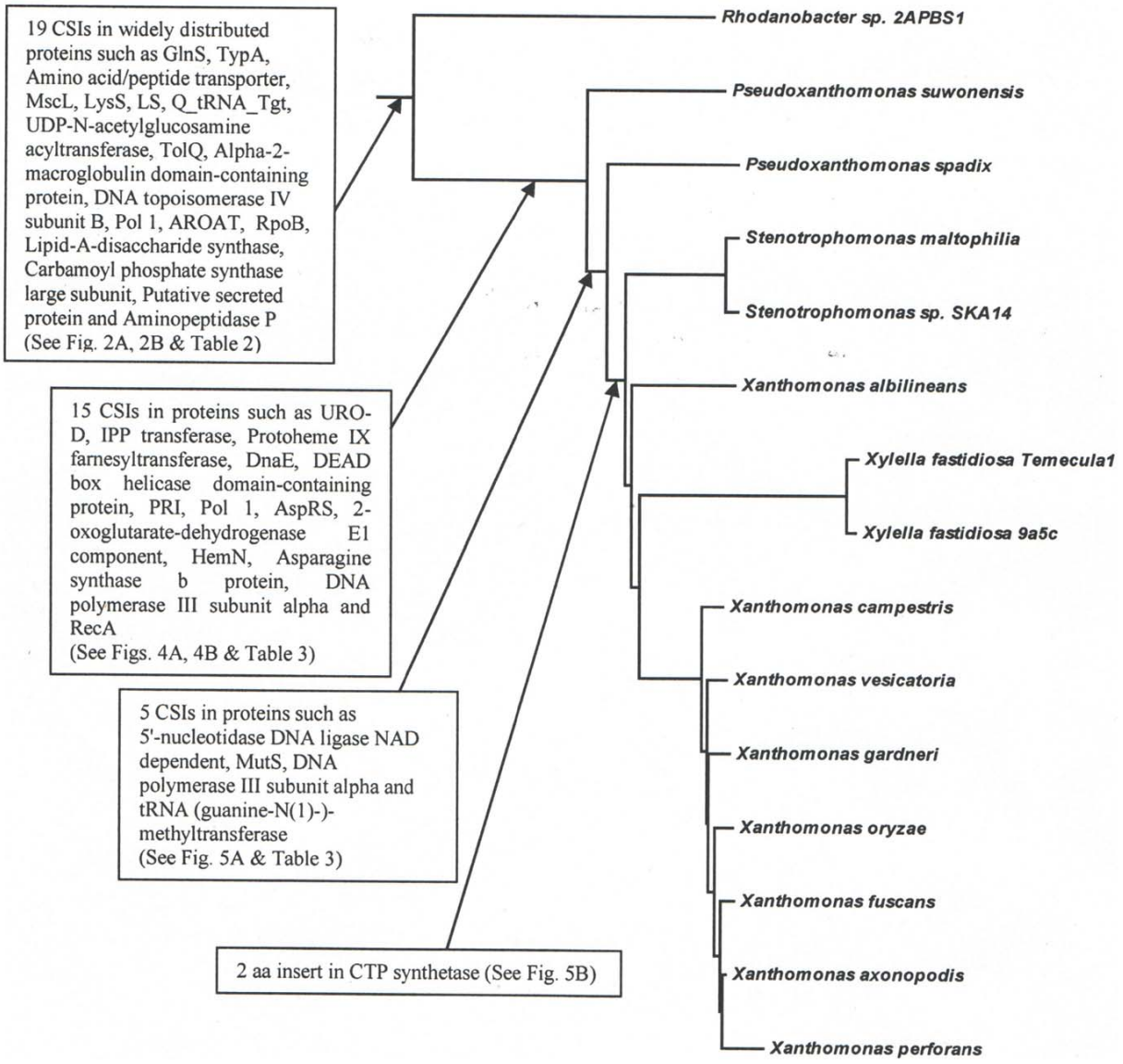


Figure 9. A summary diagram showing the species specificity of various CSIs identified in this work and the evolutionary stages where the genetic changes responsible for them were likely introduced.
doi:10.1371/journal.pone.0055216.g009

predictive ability [21,64–66] and thus they are useful in identifying both known as well as unknown species belonging to these clades (viz. Xanthomonadales) in different environments. Xanthomonadales harbor many important plant pathogens that cause a variety of diseases in economically important crops and plants [3–9]. In addition, they also contain Stenotrophomonas, which are opportunistic human pathogens [12–14]. Thus, novel methods for sensitive and specific identification of species from this order in different settings are of much importance. Most of the Xanthomonadales-specific CSIs discovered in this work are present in highly conserved regions of the genes/proteins. Hence, based upon these gene sequences degenerate PCR primers (based upon either flanking conserved regions or the indel region and a flanking conserved region) could be readily designed to examine the presence or absence of gene sequences containing these CSIs in any given sample [64,67]. Thus, molecular probes based upon these CSIs and/or their flanking regions should provide novel and specific means for the detection of new as well as existing Xanthomonadales species in different environments. The Xanthomonadales-specific CSIs, in addition to their usefulness for evolutionary and diagnostic studies, also provide novel and useful tools for genetic and biochemical investigations and possible means for identification of agents that specifically target these plant pathogenic bacteria.

Supporting Information

Figure S1 A maximum-likelihood tree based upon concatenated sequences for 28 conserved proteins. The tree shows the branching of Xanthomonadales group with Gammaproteobacteria. The tree was rooted using Alphaproteobacteria. The numbers on the nodes indicate statistical support for the nodes.
(PDF)

Figure S2 Partial sequence alignment of a conserved region in the amino acid/peptide transporter showing a 7 aa insert that is specific for all Xanthomonadales.
(PDF)

Figure S3 Partial sequence alignment of large-conductance mechanosensitive channel protein showing the presence of a 5 aa insert that is commonly shared by Xanthomonadales.
(PDF)

Figure S4 Partial sequence alignment of lysyl-tRNA synthetase showing a 3 aa insert that is uniquely shared by all members of Xanthomonadales.
(PDF)

Figure S5 Partial sequence alignment of lipoyl synthase showing a 2 aa insert that is commonly shared by Xanthomonadales.
(PDF)

Figure S6 Partial sequence alignment of a conserved region in the queuine tRNA-ribosyltransferase showing a 1 aa insert that is specific for Xanthomonadales.
(PDF)

Figure S7 Partial sequence alignment of a conserved region in the acyl-(acyl-carrier-protein)–UDP-N-acetylglucosamine O-acyltransferase showing a 1 aa insert that is commonly shared by Xanthomonadales.
(PDF)

Figure S8 Partial sequence alignment of a conserved region in the TolQ protein showing a 1 aa insert that is commonly shared by Xanthomonadales.
(PDF)

Figure S9 Partial sequence alignment of a conserved region in alpha-2-macroglobulin domain-containing protein showing a 13 aa deletion that is uniquely present in Xanthomonadales.
(PDF)

Figure S10 Partial sequence alignment of a conserved region of DNA topoisomerase IV subunit B showing a 1 aa deletion that is commonly specifically found in Xanthomonadales.
(PDF)

Figure S11 Partial sequence alignment of DNA polymerase I showing a 1 aa deletion that is uniquely shared by all members of Xanthomonadales.
(PDF)

Figure S12 Partial sequence alignment of a conserved region of aromatic amino acid aminotransferase showing a 1 aa deletion that is uniquely shared by Xanthomonadales.
(PDF)

Figure S13 Partial sequence alignment of a conserved region of DNA polymerase III subunit beta showing a 1 aa deletion that is present in Xanthomonadales. The CSI has also been found to be shared by *Marinomonas sp. MWYL1* and *Thioalkalivibrio sp. HL-EbGR7*.
(PDF)

Figure S14 Partial sequence alignment of a conserved region of lipid-A-disaccharide synthase a 2 aa insert that is present in Xanthomonadales. The CSI has also been found to be shared by *Cardiobacterium hominis*, *Allochromatium vinosum* and *Alteromonadales bacterium*.
(PDF)

Figure S15 Partial sequence alignment of carbamoyl phosphate synthase large subunit a 1 aa insert that is present in Xanthomonadales. The CSI has also been found to be shared by *Marinobacter* sp. ELB17.
(PDF)

Figure S16 Partial sequence alignment of putative secreted protein showing a 1 aa insert that is present in Xanthomonadales. The CSI has also been found to be shared by *Teredinibacter turnerae*.
(PDF)

Figure S17 Partial sequence alignment of a conserved region of aminopeptidase P, showing a 1 aa deletion that is commonly shared by Xanthomonadales. The CSI has also been found to be shared by *Thioalkalivibrio* sp. HL-EbGR7 and *Alkalilimnicola ehrlichii*.
(PDF)

Figure S18 Partial sequence alignment of protoheme IX farnesyltransferase showing a 4 aa insert that is uniquely shared by subclade of Xanthomonadales after the divergence of *Rhodanobacter* sp. 2APBS1.
(PDF)

Figure S19 Partial sequence alignment of DNA polymerase III subunit alpha, showing a 1 aa insert that is uniquely present in Xanthomonadales except *Rhodanobacter* sp. 2APBS1.
(PDF)

Figure S20 Partial sequence alignment of a conserved region in the DEAD box helicase domain-containing protein showing a 1 aa insert that is specific for Xanthomonadales except *Rhodanobacter* sp. 2APBS1.
(PDF)

Figure S21 Partial sequence alignment of a conserved region in ribose-5-phosphate isomerase A, showing a 1 aa insert that is commonly shared by Xanthomonadales except *Rhodanobacter* sp. 2APBS1.
(PDF)

Figure S22 Partial sequence alignment of a conserved region in DNA polymerase I, showing a 1 aa insert that is uniquely shared by a subclade of Xanthomonadales except *Rhodanobacter* sp. 2APBS1.
(PDF)

Figure S23 Partial sequence alignment of aspartyl-tRNA synthetase, showing a 4 aa deletion that is commonly shared by all Xanthomonadales except *Rhodanobacter* sp. 2APBS1.
(PDF)

Figure S24 Partial sequence alignment of a conserved region of 2-oxoglutarate-dehydrogenase E1 component, showing a 1 aa deletion that is unique to Xanthomonadales except *Rhodanobacter* sp. 2APBS1.
(PDF)

Figure S25 Partial sequence alignment of a conserved region of coproporphyrinogen III oxidase, showing a 1 aa deletion that is unique to Xanthomonadales except *Rhodanobacter* sp. 2APBS1.
(PDF)

Figure S26 Partial sequence alignment of a conserved region in 2-oxoglutarate-dehydrogenase E1 component, showing a 1 aa deletion that is uniquely shared by Xanthomonadales except *Rhodanobacter* sp. 2APBS1.
(PDF)

Figure S27 Partial sequence alignment of a conserved region of asparagine synthase b protein that is showing a 4–5 aa insert, unique to Xanthomonadales except *Rhodanobacter* sp. 2APBS1.
(PDF)

Figure S28 Partial sequence alignment of a conserved region of Asparagine synthase b protein, showing a 1–2 aa insert that is uniquely shared by Xanthomonadales except *Rhodanobacter* sp. 2APBS1. While genus Stenotrophomonas can be differentiated from other Xanthomonadales because of having 1 aa insert instead of 2 aa.
(PDF)

Figure S29 Partial sequence alignment of a conserved region DNA polymerase III subunit alpha showing a 4 aa insert that is commonly shared by Xanthomonadales except *Rhodanobacter* sp. 2APBS1. This CSI was previously identified by [28] as all Xanthomonadales specific signature.
(PDF)

Figure S30 Partial sequence alignment of a conserved region of RecA showing a 2 aa insert that is commonly shared by Xanthomonadales except *Rhodanobacter* sp. 2APBS1. This CSI was previously identified by [28] as all Xanthomonadales specific signature.
(PDF)

Figure S31 Partial sequence alignment of a conserved region of MutS showing a 5 aa insert that is commonly shared by Xanthomonadales except *Pseudoxanthomonas suwonensis* and *Rhodanobacter* sp. 2APBS1. This CSI was previously identified by [28] as all Xanthomonadales specific signature.
(PDF)

Figure S32 Partial sequence alignment of a conserved region of DNA polymerase III subunit alpha, showing a 2 aa deletion that is commonly shared by all Xanthomonadales except *Pseudoxanthomonas suwonensis* and *Rhodanobacter* sp. 2APBS1. These two have only 1 aa deletion at the same position.
(PDF)

Figure S33 Partial sequence alignment of tRNA (guanine-N(1)-)-methyltransferase, showing a 2 aa insert that is commonly present in all members of Xanthomonadales except *Pseudoxanthomonas suwonensis* and *Rhodanobacter* sp. 2APBS1.
(PDF)

Figure S34 Partial sequence alignment of a conserved region in glucose-6-phosphate 1-dehydrogenase, showing a 4 aa deletion that is uniquely present in all Xanthomonadales except *Pseudoxanthomonas spadix* BD-a59 and *Rhodanobacter* sp. 2APBS1 which has 3 aa insert.
(PDF)

Figure S35 Partial sequence alignment of a conserved region of DNA ligase NAD dependent, showing a 57–65 aa insert that is commonly shared by all Xanthomonadales except *Pseudoxanthomonas suwonensis* and *Rhodanobacter* sp. 2APBS1. Both these species do not contain the insert of same length. This CSI was previously identified by [28] as all Xanthomonadales specific signature.
(PDF)

Figure S36 A Neighbor-joining tree based upon carbamoyl phosphate synthase large subunit sequence. The Tree is showing the distinct branching of Xanthomonadales from various β-Proteobacteria with and without insert.
(PDF)

Figure S37 Partial sequence alignment of a conserved region of Hypothetical protein XOO1065, showing a 1 aa deletion that is present in Xanthomonadales. The deletion has also been found to be shared by few species from β-Proteobacteria but not in all of them.
(PDF)

Figure S38 Partial sequence alignment of a conserved region of orotate phosphoribosyltransferase, showing a 1 aa deletion that is present in all Xanthomonadales. The deletion has also been found to be shared by species from β-Proteobacteria.
(PDF)

Figure S39 Partial sequence alignment of a conserved region of Putative ribonuclease HII, showing a 1 aa insert that is present in Xanthomonadales. The CSI has also been found to be shared by few species from β-Proteobacteria but is not present in all.
(PDF)

Figure S40 Partial sequence alignment of a conserved region of glycyl-tRNA synthetase subunit beta, showing a 2 aa insert that is present in Xanthomonadales. The insert has also been found to be shared by some β-Proteobacteria.
(PDF)

Figure S41 Partial sequence alignment of a conserved region of the septum-site determining protein MinD, showing a 1 aa deletion that is present in Xanthomonadales. This CSI is also present in some species from β-Proteobacteria.
(PDF)

Figure S42 A Neighbor-joining tree based upon sequences from hypothetical protein X001065. The Tree is showing the Xanthomonadales and various β-Proteobacteria that share the 1 aa deletion. Species representing some other Gammaproteobacteria are also shown in tree.
(PDF)

Figure S43 A Neighbor-joining tree for Proteobacterial species based upon orotate phosphoribosyltransferase sequences. The Xanthomonadales and different β-Proteobacteria that contain the 1 aa deletion in this protein do not branch together in this tree suggesting that the deletion in these two

groups have likely occurred independently. The tree shows only representative species from other Gammaproteobacteria and Alphaproteobacteria and it was rooted using sequences from Epsilonproteobacteria.
(PDF)

Figure S44 A Neighbor-joining tree based upon sequences from putative ribonuclease HII. The Tree is showing the Xanthomonadales and various β-Proteobacteria with insert. The tree also shows representative species from other Gammaproteobacteria and Alphaproteobacteria.
(PDF)

Figure S45 A maximum-likelihood tree based upon sequences from glycyl-tRNA synthetase subunit beta. The Tree shows the branching of Xanthomonadales separately from the other insert containing Betaproteobacteria. The species distribution of this insert could be explained by either the independent occurrence of a similar genetic event in the Betaproteobacteria and the Xanthomonadales, or that this insert was introduced in a common ancestor of the Beta- and Gammaproteobacteria, followed by its loss from other Gammaproteobacteria after the divergence of deep-branching Xanthomonadales.
(PDF)

Figure S46 A Neighbor-joining tree based upon sequences from septum site-determining protein MinD protein. The Tree is showing the branching of Xanthomonadales distinctly from the other insert containing Betaproteobacteria.
(PDF)

Author Contributions

Conceived and designed the experiments: RSG. Performed the experiments: RSG HSN. Analyzed the data: HSN RSG. Contributed reagents/materials/analysis tools: RSG. Wrote the paper: RSG HSN.

References

1. Saddler GS, Bradbury J.S (2005) "Order Xanthomonadales" Brenner DJ, Krieg NR, Staley JT editors Bergey's Manual of Systematic Bacteriology, Vol. 2, 2nd edition, New York: Springer.
2. Brenner DJ, Krieg NR, Staley JT (2005) *Bergey's Manual of Systematic Bacteriology* (2005) 2nd edition, Vol. 2, The Proteobacteria, New York: Springer.
3. Chatterjee S, Almeida RP, Lindow S (2008) Living in two worlds: the plant and insect lifestyles of Xylella fastidiosa. Annu Rev Phytopathol 46: 243–271.
4. Ryan RP, Vorholter FJ, Potnis N, Jones JB, Van Sluys MA, et al. (2011) Pathogenomics of Xanthomonas: understanding bacterium-plant interactions. Nat Rev Microbiol 9: 344–355.
5. Lee BM, Park YJ, Park DS, Kang HW, Kim JG, et al. (2005) The genome sequence of Xanthomonas oryzae pathovar oryzae KACC10331, the bacterial blight pathogen of rice. Nucleic Acids Res 33: 577–586.
6. Salzberg SL, Sommer DD, Schatz MC, Phillippy AM, Rabinowicz PD, et al. (2008) Genome sequence and rapid evolution of the rice pathogen Xanthomonas oryzae pv. oryzae PXO99A. BMC Genomics 9: 204.
7. Van Sluys MA, Monteiro-Vitorello CB, Camargo LE, Menck CF, da Silva AC, et al. (2002) Comparative genomic analysis of plant-associated bacteria. Annu Rev Phytopathol 40: 169–189.
8. Bhattacharyya A, Stilwagen S, Ivanova N, D'Souza M, Bernal A, et al. (2002) Whole-genome comparative analysis of three phytopathogenic Xylella fastidiosa strains. Proc Natl Acad Sci U S A 99: 12403–12408.
9. Purcell AH, Hopkins DL (1996) Fastidious xylem-limited bacterial plant pathogens. Annu Rev Phytopathol 34: 131–151.
10. Chen J, Xie G, Han S, Chertkov O, Sims D, et al. (2010) Whole genome sequences of two Xylella fastidiosa strains (M12 and M23) causing almond leaf scorch disease in California. J Bacteriol 192: 4534.
11. Monteiro-Vitorello CB, de Oliveira MC, Zerillo MM, Varani AM, Civerolo E, et al. (2005) *Xylella* and *Xanthomonas* Mobil'omics. OMICS 9: 146–159.
12. Crossman LC, Gould VC, Dow JM, Vernikos GS, Okazaki A, et al. (2008) The complete genome, comparative and functional analysis of Stenotrophomonas maltophilia reveals an organism heavily shielded by drug resistance determinants. Genome Biol 9: R74.
13. Looney WJ, Narita M, Muhlemann K (2009) Stenotrophomonas maltophilia: an emerging opportunist human pathogen. Lancet Infect Dis 9: 312–323.
14. Waters V, Yau Y, Prasad S, Lu A, Atenafu E, et al. (2011) Stenotrophomonas maltophilia in Cystic Fibrosis: Serologic Response and Effect on Lung Disease. Am J Respir Crit Care Med 183: 635–640.
15. Yarza P, Ludwig W, Euzeby J, Amann R, Schleifer KH, et al. (2010) Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analyses. Syst Appl Microbiol 33: 291–299.
16. Martins-Pinheiro M, Galhardo RS, Lage C, Lima-Bessa KM, Aires KA, et al. (2004) Different patterns of evolution for duplicated DNA repair genes in bacteria of the Xanthomonadales group. BMC Evol Biol 4: 29.
17. Comas I, Moya A, Azad RK, Lawrence JG, Gonzalez-Candelas F (2006) The evolutionary origin of Xanthomonadales genomes and the nature of the horizontal gene transfer process. Mol Biol Evol 23: 2049–2057.
18. Dutilh BE, Huynen MA, Bruno WJ, Snel B (2004) The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. J Mol Evol 58: 527–539.
19. Gupta RS, Sneath PHA (2007) Application of the character compatibility approach to generalized molecular sequence data: Branching order of the proteobacterial subdivisions. J Mol Evol 64: 90–100.
20. Schneider A, Dessimoz C, Gonnet GH (2007) OMA Browser–exploring orthologous relations across 352 complete genomes. Bioinformatics 23: 2180–2182.
21. 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.
22. Williams KP, Gillespie JJ, Sobral BW, Nordberg EK, Snyder EE, et al. (2010) Phylogeny of gammaproteobacteria. J Bacteriol 192: 2305–2314.
23. Gupta RS (2000) The phylogeny of *Proteobacteria*: relationships to other eubacterial phyla and eukaryotes. FEMS Microbiol Rev 24: 367–402.
24. Lima WC, Van Sluys MA, Menck CF (2005) Non-gamma-proteobacteria gene islands contribute to the Xanthomonas genome. OMICS 9: 160–172.
25. Lima WC, Paquola AC, Varani AM, Van Sluys MA, Menck CF (2008) Laterally transferred genomic islands in Xanthomonadales related to pathogenicity and primary metabolism. FEMS Microbiol Lett 281: 87–97.
26. Lima WC, Menck CF (2008) Replacement of the arginine biosynthesis operon in Xanthomonadales by lateral gene transfer. J Mol Evol 66: 266–275.
27. Lima WC, Varani AM, Menck CF (2009) NAD biosynthesis evolution in bacteria: lateral gene transfer of kynurenine pathway in Xanthomonadales and Flavobacteriales. Mol Biol Evol 26: 399–406.
28. Cutino-Jimenez AM, Martins-Pinheiro M, Lima WC, Martin-Tornet A, Morales OG, et al. (2010) Evolutionary placement of Xanthomonadales based on conserved protein signature sequences. Mol Phylogenet Evol 54: 524–534.
29. Thieme F, Koebnik R, Bekel T, Berger C, Boch J, et al. (2005) Insights into genome plasticity and pathogenicity of the plant pathogenic bacterium Xanthomonas campestris pv. vesicatoria revealed by the complete genome sequence. J Bacteriol 187: 7254–7266.
30. Qian W, Jia Y, Ren SX, He YQ, Feng JX, et al. (2005) Comparative and functional genomic analyses of the pathogenicity of phytopathogen Xanthomonas campestris pv. campestris. Genome Res 15: 757–767.
31. Potnis N, Krasileva K, Chow V, Almeida NF, Patil PB, et al. (2011) Comparative genomics reveals diversity among xanthomonads infecting tomato and pepper. BMC Genomics 12: 146.
32. Pieretti I, Royer M, Barbe V, Carrere S, Koebnik R, et al. (2009) The complete genome sequence of Xanthomonas albilineans provides new insights into the reductive genome evolution of the xylem-limited Xanthomonadaceae. BMC Genomics 10: 616.
33. Bogdanove AJ, Koebnik R, Lu H, Furutani A, Angiuoli SV, et al. (2011) Two new complete genome sequences offer insight into host and tissue specificity of plant pathogenic Xanthomonas spp. J Bacteriol 193: 5450–5464.
34. Van Sluys MA, de Oliveira MC, Monteiro-Vitorello CB, Miyaki CY, Furlan LR, et al. (2003) Comparative analyses of the complete genome sequences of Pierce's disease and citrus variegated chlorosis strains of Xylella fastidiosa. J Bacteriol 185: 1018–1026.
35. Simpson AJ, Reinach FC, Arruda P, Abreu FA, Acencio M, et al. (2000) The genome sequence of the plant pathogen Xylella fastidiosa. The Xylella fastidiosa Consortium of the Organization for Nucleotide Sequencing and Analysis. Nature 406: 151–159.
36. Doddapaneni H, Yao J, Lin H, Walker MA, Civerolo EL (2006) Analysis of the genome-wide variations among multiple strains of the plant pathogenic bacterium Xylella fastidiosa. BMC Genomics 7: 225.
37. 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.
38. Gupta RS, Griffiths E (2002) Critical issues in bacterial phylogeny. Theor Popul Biol 61: 423–434.
39. 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.

40. Gupta RS (2009) Protein signatures (molecular synapomorphies) that are distinctive characteristics of the major cyanobacterial clades. Int J Syst Evol Microbiol 59: 2510–2526.
41. Rokas A, Holland PW (2000) Rare genomic changes as a tool for phylogenetics. Trends Ecol Evol 15: 454–459.
42. Baldauf SL, Palmer JD (1993) Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. Proc Natl Acad Sci USA 90: 11558–11562.
43. Rivera MC, Lake JA (1992) Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science 257: 74–76.
44. Gupta RS (2010) Molecular signatures for the main phyla of photosynthetic bacteria and their subgroups. Photosynth Res 104: 357–372.
45. Gupta RS (2001) The branching order and phylogenetic placement of species from completed bacterial genomes, based on conserved indels found in various proteins. Int Microbiol 4: 187–202.
46. Griffiths E, Gupta RS (2006) Lateral transfers of serine hydroxymethyl transferase (*glyA*) and UDP-N-acetylglucosamine enolpyruvyl transferase (*murA*) genes from free-living *Actinobacteria* to the parasitic chlamydiae. J Mol Evol 63: 283–296.
47. Harris JK, Kelley ST, Spiegelman GB, Pace NR (2003) The genetic core of the universal ancestor. Genome Res 13: 407–412.
48. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, et al. (2006) Toward automatic reconstruction of a highly resolved tree of life. Science 311: 1283–1287.
49. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ (1998) Multiple sequence alignment with Clustal x. Trends Biochem Sci 23: 403–405.
50. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17: 540–552.
51. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Computer applications in the biosciences : CABIOS 8: 275–282.
52. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596–1599.
53. Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18: 691–699.
54. Naushad HS, Gupta RS (2012) Molecular signatures (conserved indels) in protein sequences that are specific for the order Pasteurellales and distinguish two of its main clades. Antonie van Leeuwenhoek 101: 105–124.
55. Jalan N, Aritua V, Kumar D, Yu F, Jones JB, et al. (2011) Comparative genomic analysis of Xanthomonas axonopodis pv. citrumelo F1, which causes citrus bacterial spot disease, and related strains provides insights into virulence and host specificity. J Bacteriol 193: 6342–6357.
56. da Silva AC, Ferro JA, Reinach FC, Farah CS, Furlan LR, et al. (2002) Comparison of the genomes of two Xanthomonas pathogens with differing host specificities. Nature 417: 459–463.
57. Schreiber HL, Koirala M, Lara A, Ojeda M, Dowd SE, et al. (2010) Unraveling the First Xylella fastidiosa Subsp. Fastidiosa Genome from Texas. Southwestern Entomologist 35: 479–483.
58. Lee SH, Jin HM, Lee HJ, Kim JM, Jeon CO (2012) Complete Genome Sequence of the BTEX-Degrading Bacterium Pseudoxanthomonas spadix BD-a59. J Bacteriol 194: 544.
59. Parkinson N, Cowie C, Heeney J, Stead D (2009) Phylogenetic structure of Xanthomonas determined by comparison of gyrB sequences. Int J Syst Evol Microbiol 59: 264–274.
60. Handy J, Doolittle RF (1999) An attempt to pinpoint the phylogenetic introduction of glutaminyl-tRNA synthetase among bacteria. Journal of Molecular Evolution 49: 709–715.
61. Gupta RS, Mathews DW (2010) Signature proteins for the major clades of *Cyanobacteria*. BMC Evol Biol 10: 24.
62. Parkinson N, Aritua V, Heeney J, Cowie C, Bew J, et al. (2007) Phylogenetic analysis of Xanthomonas species by comparison of partial gyrase B gene sequences. Int J Syst Evol Microbiol 57: 2881–2887.
63. Young JM, Park DC, Shearman HM, Fargier E (2008) A multilocus sequence analysis of the genus Xanthomonas. Syst Appl Microbiol 31: 366–377.
64. 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.
65. Gupta RS (2011) Origin of diderm (Gram-negative) bacteria: antibiotic selection pressure rather than endosymbiosis likely led to the evolution of bacterial cells with two membranes. Antonie van Leeuwenhoek 100: 171–182.
66. Gao B, Gupta RS (2012) Microbial systematics in the post-genomics era. Antonie van Leeuwenhoek 101: 45–54.
67. 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.

CHAPTER 6

A Phylogenomic and Molecular Marker Based Taxonomic Framework for the Order *Xanthomonadales*: Proposal to Transfer of the Families *Algiphilaceae* and *Solimonadaceae* to the Order *Nevskiales* ord. nov. and to Create a New Family Within the Order *Xanthomonadales*, the Family *Rhodanobacteraceae* fam. nov., Containing the Genus *Rhodanobacter* and its Closest Relatives

The work presented in this Chapter highlights the utility of CSIs for the bacterial systematics and taxonomy. The CSIs identified in our earlier work, described in Chapter 5, were rechecked for their reliability and predictability. Based upon our analyses on all known and 5 *de novo* sequenced *Xanthomonadales* genomes, we have proposed a complete taxonomic revision of this order. My contribution towards the completion of this chapter included the culturing of bacteria, DNA extraction, sequencing and assembling of 5 new *Xanthomonadales* genomes, followed by the examination of the genomes for the detection of CSIs. I was also involved in the construction of phylogenetic trees, writing of the manuscript and construction of the figures and tables provided.

*Due to limited space, supplementary figures and tables are not included in the chapter but can be accessed along with the rest of the manuscript at:

Naushad,S., Adeolu,M., Wong,S., Sohail,M., Schellhorn,H.E., and Gupta,R.S. (2014). A phylogenomic and molecular marker based taxonomic framework for the order Xanthomonadales: proposal to transfer the families Algiphilaceae and Solimonadaceae to the order Nevskiales ord. nov. and to create a new family within the order Xanthomonadales, the family Rhodanobacteraceae fam. nov., containing the genus Rhodanobacter and its closest relatives. Antonie Van Leeuwenhoek.

Antonie van Leeuwenhoek (2015) 107:467–485
DOI 10.1007/s10482-014-0344-8

ORIGINAL PAPER

A phylogenomic and molecular marker based taxonomic framework for the order *Xanthomonadales*: proposal to transfer the families *Algiphilaceae* and *Solimonadaceae* to the order *Nevskiales* ord. nov. and to create a new family within the order *Xanthomonadales*, the family *Rhodanobacteraceae* fam. nov., containing the genus *Rhodanobacter* and its closest relatives

Sohail Naushad · Mobolaji Adeolu · Shirley Wong · Misbah Sohail · Herbert E. Schellhorn · Radhey S. Gupta

Received: 8 September 2014 / Accepted: 28 November 2014 / Published online: 7 December 2014

Abstract The current taxonomy of the order *Xanthomonadales* is highly problematic and no comprehensive phylogenomic studies have been completed that include the most divergent members within the order. In this work, we have completed a phylogenomic analysis of a wide range of genomes, five of which were sequenced for the first time for this work, representing the vast majority of the diversity within the order *Xanthomonadales*. Using comparative genomic techniques, we have identified a large number of conserved signature inserts/deletions (CSIs) that are specifically found in different groups of related organisms, at different taxonomic levels, within the order. Our phylogenetic analyses do not support a monophyletic grouping of the members of the order *Xanthomonadales* and no CSIs were identified which are uniquely shared by all sequenced species within this order. However, our work has identified 10 CSIs which are specific to all members of the family *Xanthomonadaceae* and an additional 10 and 11 CSIs that are specific to one of two phylogenetically well-defined clades within the family *Xanthomonadaceae*. On the basis of the identified CSIs and the results of phylogenomic analyses, we propose a new taxonomic framework for the order *Xanthomonadales*. In this proposal, the families *Algiphilaceae* and *Solimonadaceae* (*Nevskiaceae*), which do not branch with the other members of the order *Xanthomonadales*, are transferred into the order *Nevskiales* ord. nov. The remaining members of the order *Xanthomonadales* are divided into two families: the family *Xanthomonadaceae,* containing the genus *Xanthomonas* and its closest relatives, and a new family, *Rhodanobacteraceae* fam. nov., containing the genus *Rhodanobacter* and its closest relatives. Additionally, we have also emended descriptions of the order *Lysobacterales*, the family *Lysobacteraceae*, and the family *Nevskiaceae* to indicate that they are earlier synonyms of the order *Xanthomonadales*, the family *Xanthomonadaceae*, and the family *Solimonadaceae*, respectively.

Electronic supplementary material The online version of this article (doi:10.1007/s10482-014-0344-8) contains supplementary material, which is available to authorized users.

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

S. Wong · H. E. Schellhorn
Department of Biology, McMaster University, Hamilton, ON L8S 4K1, Canada

Keywords *Xanthomonadales* · *Lysobacterales* · *Lysobacteraceae* · *Rhodanobacteraceae* · *Nevskiales* · *Nevskiaceae* · *Salinisphaeraceae* · Phylogenetic trees · Conserved signature indels · Molecular signatures

Introduction

The order *Xanthomonadales* is an early diverging group of bacteria within the class *Gammaproteobacteria* (Cutino-Jimenez et al. 2010; Williams et al. 2010; Naushad and Gupta 2013). The order *Xanthomonadales* currently contains 5 families (viz. *Algiphilaceae, Nevskiaceae, Sinobacteraceae*, *Solimonadaceae*, and *Xanthomonadaceae*) which contain 30 genera encompassing a large number of species that possess a diverse range of phenotypic and biochemical characteristics (Saddler and Bradbury 2005a; Parte 2013). The members of this order include a number of major plant pathogens that have significant economic and agricultural impact. Members of the genera *Xylella* and *Xanthomonas*, in particular, are major phytopathogens which cause a wide variety of serious diseases in more than 400 agriculturally important plants including tomatoes, bananas, citrus plants, rice, and coffee plants (da Silva et al. 2002; Van Sluys et al. 2003; Lee et al. 2005; Chatterjee et al. 2008; Salzberg et al. 2008; Ryan et al. 2011). The order also contains the genus *Stenotrophomonas* which harbours a number of increasingly important multidrug resistant opportunistic pathogens that are responsible for hospital-acquired infections in immunodeficient patients (Crossman et al. 2008; Looney et al. 2009). Despite the important plant and human pathogens present within this order, the taxonomy of this group is highly problematic and no comprehensive phylogenetic studies have been completed that focus specifically on the interrelationships of the different members within the order *Xanthomonadales* (Gao et al. 2009; Cutino-Jimenez et al. 2010; Williams et al. 2010; Naushad and Gupta 2013; Tindall 2014b).

The current taxonomy of the order *Xanthomonadales* is largely based on 16S rRNA sequence analysis (Saddler and Bradbury 2005a; Gutierrez et al. 2012; Losey et al. 2013). However, the 16S rRNA gene sequence has shown limited ability to resolve the branching and relationships of organisms within the order *Xanthomonadales* (Zhou et al. 2008; Cutino-Jimenez et al. 2010; Yilmaz et al. 2013). Phylogenetic trees based on the 16S rRNA gene sequence often do not resolve a monophyletic cluster of all *Xanthomonadales*; the most divergent members of the order often branch separately from the majority of the species within the group (Yilmaz et al. 2013). Apart from the 16S rRNA sequence, no biochemical, morphological or physiological characteristics are known which distinguish the order *Xanthomonadales* from all other bacteria or the families and major phylogenetic clusters within the order from each other (Saddler and Bradbury 2005a; Gutierrez et al. 2012; Losey et al. 2013). Thus, it is of interest to identify shared characteristics that can clearly elucidate the evolutionary relationships within this highly diverse group of organisms and form the basis for a coherent taxonomic framework of the order.

Whole genome sequences for members of the order *Xanthomonadales* provide a rich resource for the discovery of molecular characteristics which are unique to evolutionarily related organisms (Gao et al. 2009; Cutino-Jimenez et al. 2010; Naushad and Gupta 2013). One useful type of shared molecular characteristic that has been a focus of recent research are Conserved Signature Indels (CSIs), which are insertions/deletions uniquely present in protein sequences from a group of evolutionarily related organisms (Gupta 2010; Gao and Gupta 2012b; Gupta and Lali 2013; Gupta et al. 2013; Gupta 2014). Due to the specificity of CSIs for particular groups of bacteria, they represent molecular synapomorphies (markers of common evolutionary decent) which can be used to identify and demarcate specific bacterial groups in clear molecular terms (Gupta 1998, 2010). We have previously carried out comparative genomic analysis of a limited number of members from the order *Xanthomonadales* in which we identified a large number of CSIs in diverse proteins that were uniquely present in all analyzed members of the order or a subgroup of the *Xanthomonadales* (Naushad and Gupta 2013). In this work, we have extended these studies, by carrying out detailed phylogenomic and comparative genomic analyses on a greatly expanded dataset on members of the order *Xanthomonadales* which includes 43 genomes from the NCBI, JGI, and EzBioCloud genome databases and 5 additional *Xanthomonadales* genomes, which we have sequenced *de novo,* representing 2 families, 20 genera, and 42 named species. Our analyses have identified no phylogenetic

support for a monophyletic grouping of all sequenced members the order *Xanthomonadales* and no CSIs were identified which are uniquely shared by all sequenced species within the order *Xanthomonadales* suggesting that the order *Xanthomonadales* does not represent a single monophyletic lineage. Additionally, we have identified 31 CSIs which are either specific to the family *Xanthomonadaceae* or to one of its subgroups which demarcate these groups in molecular terms. On the basis of the identified CSIs and the results of phylogenomic analyses, we propose a new taxonomic framework for the order *Xanthomonadales*. In this proposal, the families *Algiphilaceae* and *Solimonadaceae* (*Nevskiaceae*), which do not branch with the other members of the order *Xanthomonadales*, are transferred into the order *Nevskiales* ord. nov, along with their closest evolutionary neighbour, the family *Salinisphaeraceae* fam. nov. The remaining members of the order *Xanthomonadales* are divided into two families: the family *Xanthomonadaceae,* containing the genus *Xanthomonas* and its closest relatives, and a new family, *Rhodanobacteraceae* fam. nov., containing the genus *Rhodanobacter* and its closest relatives. Lastly, the descriptions of the order *Lysobacterales*, the family *Lysobacteraceae*, and the family *Nevskiaceae* are emended to indicate that they are earlier synonyms of the order *Xanthomonadales*, the family *Xanthomonadaceae*, and the families *Solimonadaceae* and *Sinobacteraceae*, respectively.

Methods

DNA extraction and genome sequencing

Five *Xanthomonadales* isolates were sequenced *de novo* in this study; *Dyella japonica* DSM 16301^T (Genbank accession number JPLA00000000), *Luteibacter rhizovicinus* DSM 16549^T (JPLB00000000), *Thermomonas brevis* DSM 15422^T (JPLC00000000), *Xanthomonas hyacinthi* DSM 19077^T (JPLD00000000), and *Xanthomonas pisi* DSM 18956^T (JPLE00000000). The isolates were obtained from the German Collection of Microorganisms and Cell Cultures (Leibniz-Institut DSMZ). The isolates were grown for 24 h under the growth conditions described in Supplemental Table 1. Genomic DNA was extracted using a CTAB based DNA extraction methodology (Wilson 1987) with specific modifications for *Xanthomonadales* described by Jaufeerally-Fakim and Dookun (2000). The DNA samples were diluted to 0.2 ng/µl and standard Illumina multiplex libraries were generated using the Nextera XT DNA Sample Prep Kit. The fragment size distribution of each library was verified using the Agilent High Sensitivity DNA Kit. Sequencing was performed using an Illumina HiSeq 2000 and 150 bp paired end reads were generated. Genomes were assembled using CLC Genomics Workbench 7.0.4 with default *de novo* assembly parameters and trimmed for contamination using the UniVec vector database (Build 8.0) (Table 1).

Phylogenetic sequence analysis

Phylogenetic analysis was performed on a concatenated sequence alignment of 15 highly conserved housekeeping proteins (viz. dimethyladenosine transferase, alanyl-tRNA synthetase, arginyl-tRNA synthetase, chaperone protein DnaK, signal recognition particle-docking protein FtsY, chaperonin GroL, DNA gyrase subunit A, DNA gyrase subunit B, ATP-dependent DNA helicase UvrD, valyl-tRNA synthetase, Isoleucyl-tRNA synthetase, DNA polymerase I, SecA, RpoB, and RpoC) which have been widely used for phylogenetic analysis (Kyrpides et al. 1999; Charlebois and Doolittle 2004; Ciccarelli et al. 2006). Sequences for these proteins were obtained from the NCBI and JGI-IMG genome databases for strains of all *Xanthomonadales* and a representative selection of outgroup *Gammaproteobacteria* (which included members from the orders *Aeromonadales, Alteromonadales, Cardiobacteriales, Chromatiales,* "Enterobacteriales"*, Legionellales, Methylococcales, Oceanospirillales, Pasteurellales, Pseudomonadales,* "Salinisphaerales"*, Thiotrichales*, and "Vibrionales") and *Betaproteobacteria*. Sequences for these proteins were also obtained from the five *Xanthomonadales* genomes which we have sequenced in this work and the genome of *Riemerella anatipestifer*, which was used to root the tree. Multiple sequence alignments for these proteins were created using Clustal_X 1.83 (Jeanmougin et al. 1998) and concatenated into a single alignment file. Poorly aligned regions from this alignment file were removed using Gblocks 0.92 (Castresana 2000). The resulting alignment, which contained 6995 aligned amino acids, was used for phylogenetic analysis. The maximum-likelihood tree based on 100 bootstrap replicates of this alignment

Table 1 Characteristics of the *Xanthomonadales* genomes used for phylogenetic analysis

Organism	Accession #	Genome size (Mb)	G−C %	Genome source
Arenimonas composti TR7-09	AUFF01	3.16	70.8	DOE-JGI
Arenimonas oryziterrae DSM 21050	ATVD01	3.09	65.6	DOE-JGI
Dyella ginsengisoli LA-4	AMSF01	4.55	67.7	Shanghai Jiao Tong University
Frateuria aurantia DSM 6220	CP003350	3.60	63.4	DOE-JGI
Hydrocarboniphaga effusa AP103	AKGD01	5.19	65.2	Chonbuk National University
Ignatzschineria larvae DSM 13226	AZOD01	2.46	40.4	DOE-JGI
Luteimonas mephitis DSM 12574	AULN01	3.42	68.5	DOE-JGI
Lysobacter antibioticus HS124	CAQP01	5.14	69.0	OARDC
Lysobacter defluvii DSM 18482	AUHT01	2.72	70.3	DOE-JGI
Nevskia ramosa DSM 11499	ATVI01	4.52	64.4	DOE-JGI
Pseudoxanthomonas sp. GW2	ALIP01	3.35	71.4	Shanghai Jiao Tong University
Pseudoxanthomonas spadix BD-a59	CP003093	3.45	67.7	Lee et al. (2012)
Pseudoxanthomonas suwonensis 11-1	CP002446	3.42	70.2	DOE-JGI
Rhodanobacter denitrificans 2APBS1	CP003470	4.23	67.5	Kostka et al. (2012)
Rhodanobacter fulvus Jip2	AJXU01	3.88	65.6	Im et al. (2004)
Rhodanobacter sp. 115	AJXS01	4.24	64.7	Kostka et al. (2012)
Rhodanobacter spathiphylli B39	AJXT01	3.91	66.5	De Clercq et al. (2006)
Rhodanobacter thiooxydans LCS2	AJXW01	4.09	67.2	Lee et al. (2007)
Rudaea cellulosilytica DSM 22992	ARJQ01	4.34	63.6	DOE-JGI
Silanimonas lenta DSM 16282	AUBD01	2.65	71.1	DOE-JGI
Singularimonas variicoloris DSM 15731	ARNM01	4.12	69.1	DOE-JGI
Solimonas flavus DSM 18980	AUFV01	4.46	68.9	DOE-JGI
Stenotrophomonas maltophilia K279a	AM743169	4.85	66.3	JCV
Stenotrophomonas sp. SKA14	ACDV01	5.02	66.4	Crossman et al. (2008)
Wohlfahrtiimonas chitiniclastica DSM 18708	AQXD01	1.99	44.1	DOE-JGI
Xanthomonas albilineans GPE PC73	FP565176	3.85	62.9	Pieretti et al. (2009)
Xanthomonas arboricola MAFF 301420	BAVC01	5.00	65.3	NIFTS
Xanthomonas axonopodis 12-2	AJJO01	5.27	64.4	Kasetsart University
Xanthomonas campestris 8004	CP000050	5.15	65.0	Qian et al. (2005)
Xanthomonas citri Aw12879	CP003778	5.40	64.7	Jalan et al. (2013)
Xanthomonas fragariae LMG 25863	AJRZ01	4.18	62.2	ILVO
Xanthomonas fuscans 4834-R	FO681494	5.09	64.7	Darrasse et al. (2013)
Xanthomonas gardneri ATCC 19865	AEQX01	5.53	63.7	University of Florida
Xanthomonas oryzae KACC 10331	AE013598	4.94	63.7	Lee et al. (2005)
Xanthomonas perforans 91-118	AEQW01	5.26	65.0	University of Florida
Xanthomonas sacchari NCPPB 4393	AGDB01	4.90	69.0	Studholme et al. (2011)
Xanthomonas translucens ART-Xtg29	ANGG01	4.10	68.6	ART
Xanthomonas vasicola NCPPB 1326	AKBK01	4.95	63.3	Studholme et al. (2011)
Xanthomonas vesicatoria ATCC 35937	AEQV01	5.53	64.1	University of Florida
Xylella fastidiosa 9a5c	AE003849	2.73	52.6	Meidanis et al. (2002)
Xylella fastidiosa Ann-1	AAAM04	2.73	52.0	DOE-JGI
Xylella fastidiosa M12	CP000941	2.48	51.9	Chen et al. (2010)
Xylella fastidiosa Temecula 1	AE009442	2.52	51.8	Van Sluys et al. (2003)

Genomic information was collected from: http://www.ncbi.nlm.nih.gov/genomes/

DOE-JGI Genome sequenced by the United States Department of Energy Joint Genome Institute, *OARDC* genome sequenced by the Ohio Agricultural Research and Development Center, *JCV* genome sequenced by the J. Craig Venter Institute, *NIFTS* genome sequenced by the National Agriculture and Food Research Organization Institute of Fruit Tree Science, *ILVO* genome sequenced by the Institute for Agricultural and Fisheries Research, *ART* genome sequenced by the Research Station Agroscope Reckenholz-Tänikon

was constructed using MEGA 5.2 (Tamura et al. 2011) employing the Whelan and Goldman substitution model.

A 16S rRNA gene sequence based phylogenetic tree was also created based on 197 sequences that included representative strains of all cultured *Xanthomonadales* genera. 16S rRNA gene sequences larger than 1,200 bp were obtained for all strains used in our concatenated protein based phylogenetic tree and all type strains classified under the order *Xanthomonadales* in the Ribosomal Database Project (Cole et al. 2014). A maximum-likelihood tree based on these sequences was created using 100 bootstrap replicates of the 16S rRNA sequence alignments in MEGA 5.2 (Tamura et al. 2011) employing the General Time-Reversible (Tavaré 1986) substitution model.

Identification and assessment of specificity of conserved signature indels

Identification of CSIs that are commonly shared by members of the *Xanthomonadaceae* was carried out as described by Naushad and Gupta (2013). Briefly, for the identification of CSIs, BLASTp searches were performed on each protein in the genome of *Rhodanobacter fulvus* Jip2. These searches were performed using the default BLAST parameters against all available sequences in the GenBank non-redundant database. For those proteins for which high scoring homologs (E values $< 1e^{-20}$) were present in other species from the *Xanthomonadales*, multiple sequence alignments were created using the Clustal_X 1.83 program (Jeanmougin et al. 1998). These alignments were visually inspected for the presence of insertions or deletions that were flanked on both sides by at least 5–6 conserved amino acid residues in the neighbouring 30–40 amino acids. Indels that were not flanked by conserved regions were not further considered, as they do not provide useful molecular markers. To assess the specificity of the indels we identified here and to reassess the specificity of the indels identified in our previous work, we carried out detailed BLASTp and tBLASTn searches against both the NCBI and JGI-IMG genome databases using as query short sequence segments containing the indel and the flanking conserved regions (60–100 amino acids long). Local tBLASTn searches were also completed on the indel containing regions for genomes of *Xanthomonadales* organisms missing from the NCBI and JGI-IMG genome databases. To ensure that the identified signatures are only present in *Xanthomonadales* homologues, the 250 BLAST hits with the highest similarity to the query sequence were examined for the presence or absence of these CSIs. Signature files were created and formatted using the programs Sig_Create and Sig_Style (accessible from Gleans.net) as described by Gupta (2014). In this work, we report the results of CSIs that are specific for different groups within the *Xanthomonadales* and where similar CSIs were not observed in any other bacteria in the top 250 BLAST hits. Due to space constraints, the sequence alignment files presented here contain sequence information for a limited number of species within the order *Xanthomonadales* and a representative selection of outgroup species. However, in each case, all members of the order and outgroups exhibited similar sequence characteristics to the representatives.

Results

Phylogenetic analysis

The current understanding of the evolutionary relationships of the *Xanthomonadales* is based largely on analyses of the 16S rRNA gene (Saddler and Bradbury 2005a; Gutierrez et al. 2012; Losey et al. 2013). In past studies, the 16S rRNA gene sequence has shown limited ability to resolve some of the phylogenetic relationships of organisms within the order *Xanthomonadales* (Zhou et al. 2008; Yilmaz et al. 2013). Phylogenetic trees based on multiple conserved genes/proteins have been shown to provide greater resolving power than those based on any single gene or protein (Rokas et al. 2003; Wu et al. 2009). Thus, we have constructed a highly resolved phylogenetic tree of the *Xanthomonadales* based on a concatenated set of 15 housekeeping and ribosomal proteins (Fig. 1). In this concatenated protein based phylogenetic tree a majority of the members of the *Xanthomonadales* formed a well-supported monophyletic clade which branched as an outgroup of the other members of the *Gammaproteobacteria*. The members of the order *Xanthomonadales* formed two distinct and well-supported main monophyletic clades: one clade consisting of members of the family *Xanthomonadaceae* and another clade consisting of the family *Solimonadaceae* (including the genera *Nevskia* and *Hydrocarboniphaga*) and the

species *Salinisphaera shabanensis*, a member of the family "Salinisphaeraceae". The *Xanthomonadaceae* clade contained two smaller clades that were well-supported by bootstrap analysis. The first of these clades, contained the genera *Xanthomonas*, *Xylella*, *Stenotrophomonas*, *Lutimonas*, *Lysobacter* and their relatives (Clade 1) while the second clade contained the genera *Rudaea*, *Dylella*, *Luteibacter*, *Rhodanobacter* and *Frateuria* (Clade 2). Two members of the *Xanthomonadales*, *Wohlfahrtiimonas chitiniclastica* and *Ignatzschineria larvae*, branched separately from the rest of the order, in a well-supported clade with members of the order *Cardiobacteriales*, another early diverging group within the class *Gammaproteobacteria*.

We have also produced a phylogenetic tree based on the 16S rRNA gene which contains representative species of *Xanthomonadales* that encompass all of the currently named genera (Fig. 2). The 16S rRNA based phylogenetic tree exhibited broadly similar branching to our concatenated protein based phylogenetic tree. In the 16S rRNA gene tree, the families *Xanthomonadaceae* and *Solimonadaceae* did not form a monophyletic clade and were separated by a large number of organisms. The family *Xanthomonadaceae* was divided into two well-supported clades which were analogous to the clades found in our concatenated protein based phylogenetic tree (Clades 1 and 2). In the 16S rRNA gene tree, the family *Solimonadaceae* (including the genera *Nevskia*, *Hydrocarboniphaga*, and *Alkanibacter*) branched with the family *Algiphilaceae*, another disparate group within the *Xanthomonadales*, and the genus *Steroidobacter,* which is currently recognized as a member of the *Xanthomonadaceae*. As in our concatenated protein based phylogenetic tree, *Solimonadaceae* and the other disparate members of the *Xanthomonadales* showed an association with the members of the genus *Salinisphaera*, the sole members of the family "Salinisphaeraceae". Additionally, in the 16S rRNA based phylogenetic tree, the genera *Wohlfahrtiimonas* and *Ignatzschineria*, which branched with the order *Cardiobacteriales* in our concatenated protein based phylogenetic tree, formed a weakly supported monophyletic group with the other members of the family *Xanthomonadaceae*. However, *Wohlfahrtiimonas* and *Ignatzschineria* were well separated from the other *Xanthomonadaceae* by a long branch.

Conserved signature indels

CSIs that are restricted to a group of related species are a novel class of molecular marker with high utility for evolutionary studies (Gupta 1998; Rokas and Holland 2000; Gupta 2010; Gao and Gupta 2012a; Gupta 2014). Recently, CSIs have been used to define novel taxonomic groups and to propose important taxonomic changes for various groups of bacteria (viz. Spirochaetes, Aquificae, *Neisseriales*, and *Bacillus*) at different taxonomic ranks (Adeolu and Gupta 2013; Bhandari et al. 2013; Gupta and Lali 2013; Gupta et al. 2013). We have recently reported a comparative genomic analysis on a limited number of members of the order *Xanthomonadales* in which we identified a large number of CSIs in diverse proteins that were uniquely present in all available members of the order or different phylogenetic groups within the order and absent in homologs from all other bacterial groups (Naushad and Gupta 2013). However, the genomes analyzed in our previous study were all from members of one family within the *Xanthomonadales*, the family *Xanthomonadaceae*, and did not include any of the more divergent species within the order whose phylogenetic placement is less clear. In this work, we have reassessed the specificity of these previously identified CSIs for a large number of additional *Xanthomonadales*, including five strains which we have sequenced, *de novo*, covering a vast majority of the diversity within the order and thereby have identified 31 CSIs which are either specific to the family *Xanthomonadaceae* or to one of its subgroups and absent in all other sequenced bacterial groups.

Of the 31 CSIs described in this work, none were present in all members of the order *Xanthomonadales*. All of the CSIs identified in our previous study of the *Xanthomonadales* (Naushad and Gupta 2013) were found to be specific to only the family *Xanthomonadaceae* or one of its subgroups. Of the 31 CSIs identified, 10 were uniquely found in all or most members of the *Xanthomonadaceae*, except *Wohlfahrtiimonas* and *Ignatzschineria*, and absent in organisms from all other sequenced bacterial groups. One example of a CSI uniquely present in members of the *Xanthomonadace* is shown in Fig. 3. In the example, an 18 aa insertion in a conserved region of glutaminyl t-RNA synthetase is uniquely present in all members of the *Xanthomonadace*, except

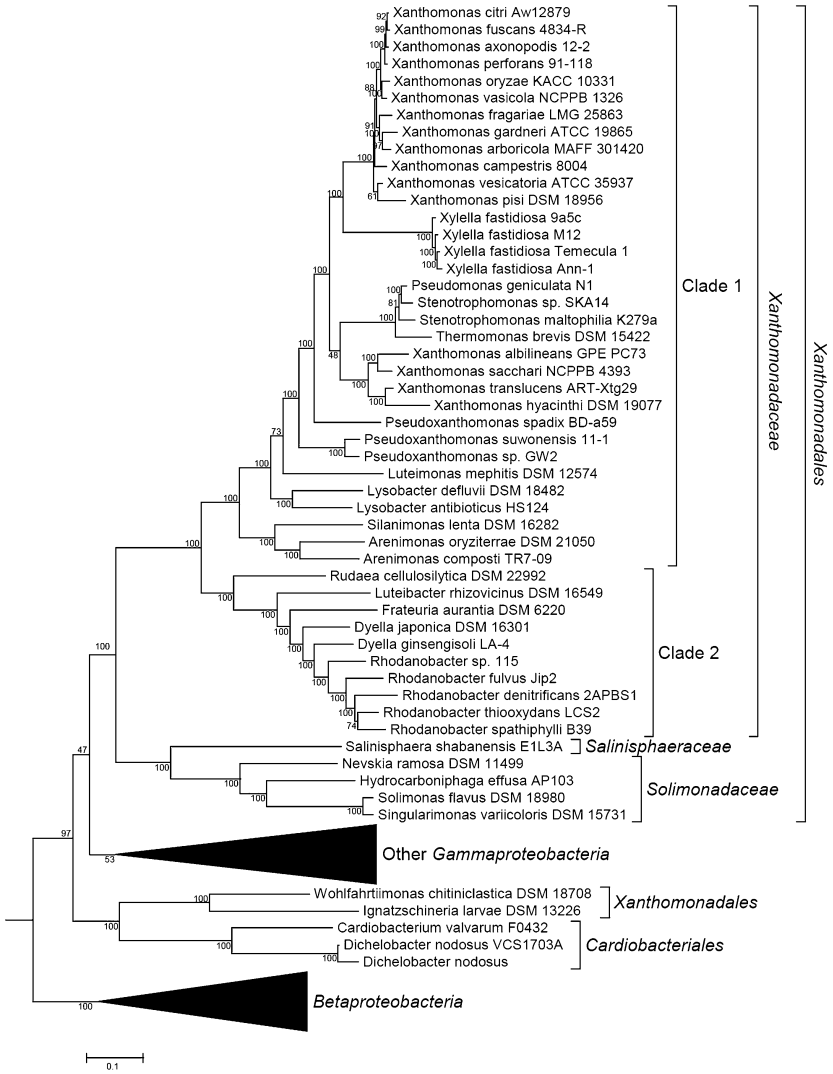


Fig. 1 A maximum-likelihood phylogenetic tree of the order *Xanthomonadales*, other *Gammaproteobacteria*, and *Betaproteobacteria* based on the concatenated amino acid sequences of 25 conserved proteins. Bootstrap values are shown at branch nodes. The major groups within the order *Xanthomonadales* as well as the related taxa, *Salinisphaeraceae* and *Cardiobacteriales*, are indicated

Wohlfahrtiimonas and *Ignatzschineria*, but it is not found in sequences from any other bacterial group. Sequence information for the 9 other CSIs specific for all members of the *Xanthomonadace*, except *Wohlfahrtiimonas* and *Ignatzschineria* are presented in Supplemental Figs. 1–9 and a summary of all 10 *Xanthomonadace* specific CSIs is presented in Table 2A. Our analyses have also identified 10 CSIs which were found to be unique molecular characteristics of most members of Clade 1 of the *Xanthomonadaceae*. Two examples of such CSIs are presented in Fig. 4. One CSI, a 4 aa insert in DNA polymerase III subunit alpha, is present in all members of Clade 1 of the *Xanthomonadaceae* except *Silanimonas lenta* (Fig. 4a), while the other CSI, a 4 aa insert in the protein protoheme IX farnesyltransferase, is uniquely present in all members of Clade 1 of the *Xanthomonadaceae* except the early branching genera *Arenimonas* and *Silanimonas* (Fig. 4b). Sequence information for the other identified CSIs specific to

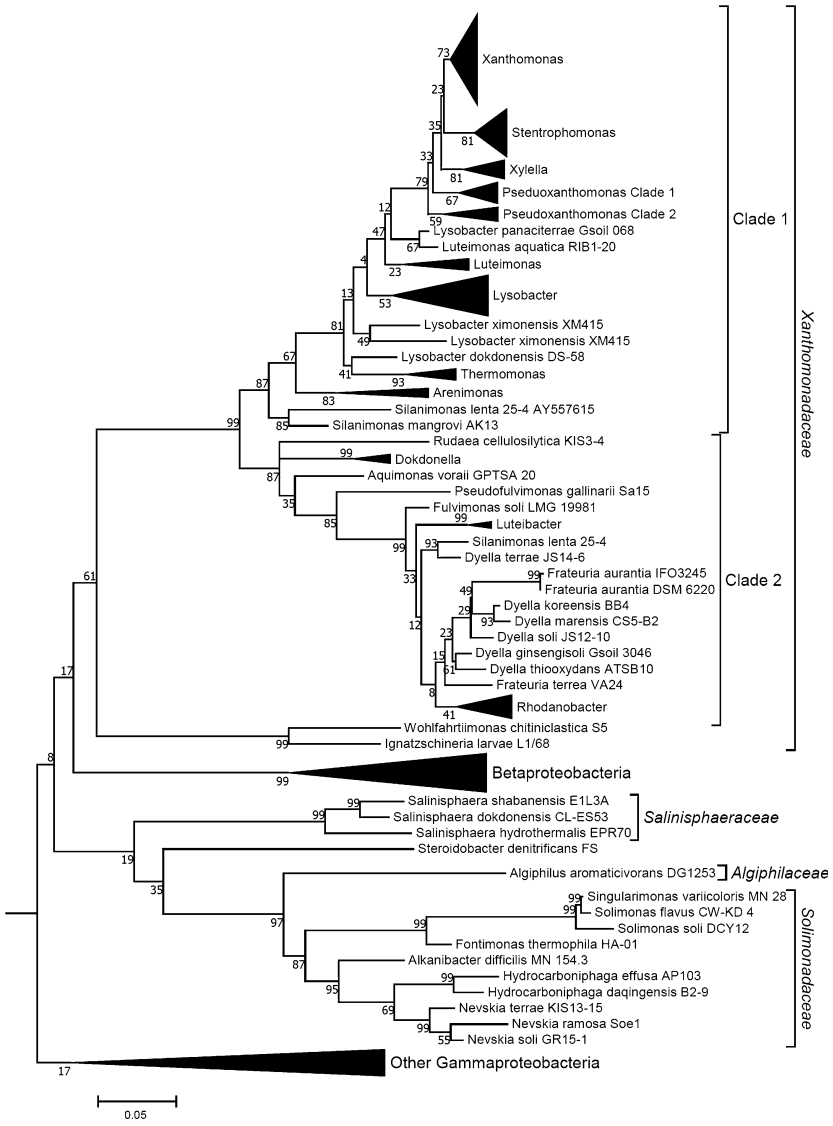


Fig. 2 A maximum-likelihood tree based on the 16S rRNA gene sequences of representative strains of all named *Xanthomonadales* species. Bootstrap values are shown at branch nodes. The major groups within the order *Xanthomonadales* as well as the related taxon, *Salinisphaeraceae*, are indicated

the members of Clade 1 of the *Xanthomonadaceae* are presented in Supplemental Figs. 10–17 and summarized in Table 2B.

Our analyses have also identified 11 CSIs that were specifically found in diverse proteins from members of Clade 2 of the *Xanthomonadaceae*, 7 of which were uniquely found in all members of Clade 2 of the *Xanthomonadaceae* except the early branching genus *Rudaea*. An example of a CSI specifically found in all members of Clade 2 of the *Xanthomonadaceae* is shown in Fig. 5a. In this CSI a 1 aa insert in the protein uridylyltransferase is shown to be found in all members of Clade 2 of the *Xanthomonadaceae* and absent in all other *Xanthomonadales* and all other bacterial groups. Another CSI, a 4 aa insert in the protein CDP-diacylglycerol–glycerol-3-phosphate 3-phosphatidyltransferase, specifically found in all members of Clade 2 of the *Xanthomonadaceae* except *Rudaea cellulosilytica* is shown in Fig. 5b. Sequence information for the other identified CSIs

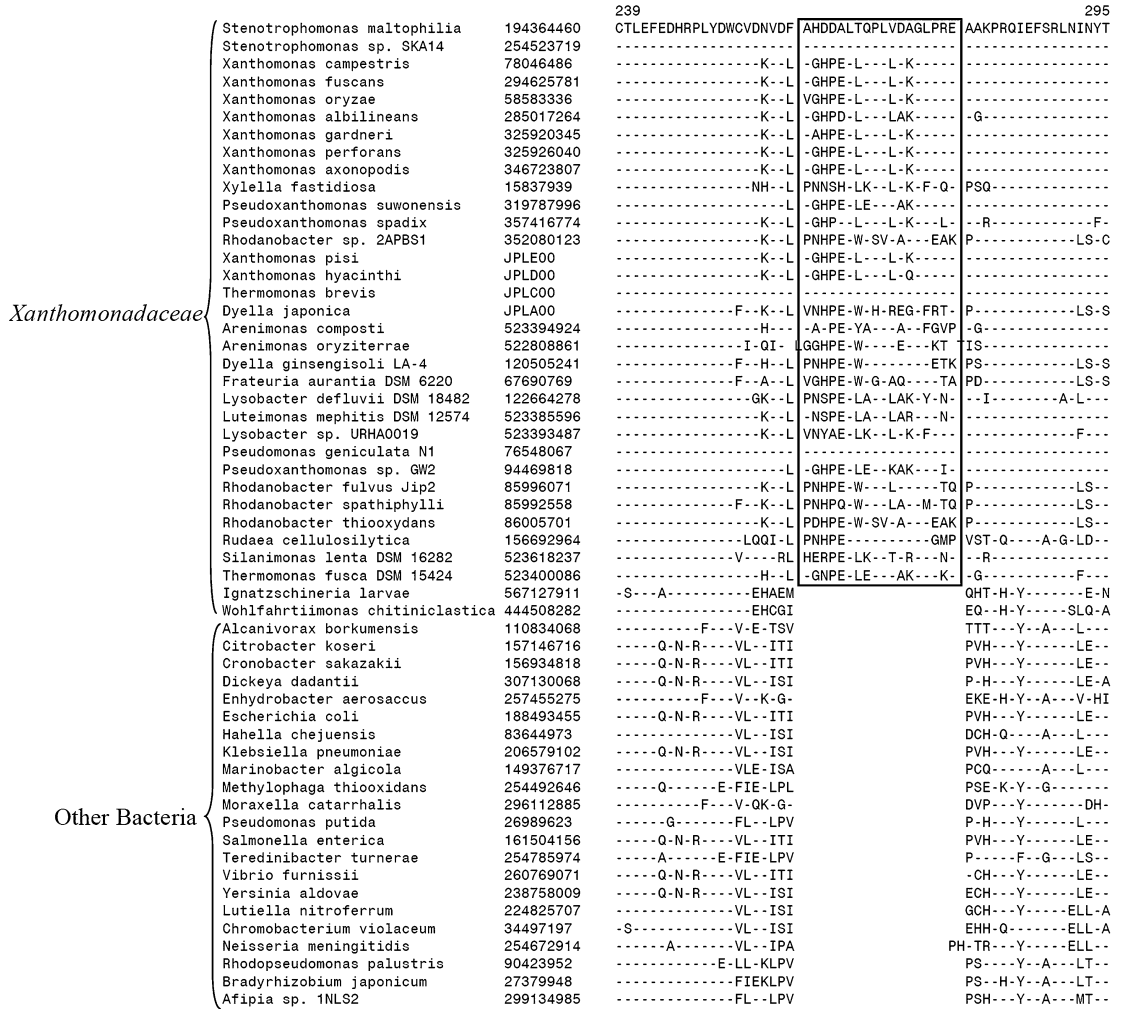


Fig. 3 A partial sequence alignment of the protein Glutaminyl t-RNA synthetase, showing a CSI (*boxed*) that is uniquely present in all members of the order *Xanthomonadales*. Sequence information for only representative *Xanthomonadales* and a limited number other bacteria is shown here. However, unless otherwise indicated, similar CSIs were present in all members of the indicated group and not detected in any other bacterial species in the top 250 BLAST hits. The *dashes* (-) in the alignments indicate identity with the residue in the top sequence. GenBank identification (GI) numbers for each sequence are indicated in the second column. Sequence information for 10 other CSIs that are specific for all sequenced *Xanthomonadales* is provided in Supplemental Figs. 1–9 and Table 2A

specific to the members of Clade 2 of the *Xanthomonadaceae* are presented in Supplemental Figs. 18–26 and summarized in Table 2C and D. Our analyses have not identified any CSIs uniquely found in all of the disparate members of the *Xanthomonadales* or uniquely shared by the genera *Wohlfahrtiimonas* and *Ignatzschineria* and the rest of the *Xanthomonadaceae*.

Discussion

The current phylogeny of the order *Xanthomonadales* is based largely on the analysis of 16S rRNA gene sequences (Saddler and Bradbury 2005a; Gutierrez et al. 2012; Losey et al. 2013). However, the 16S rRNA gene based phylogenies exhibit limited support for a single monophyletic clade consisting of all

Table 2 Conserved signature indels that are specific for different groups of *Xanthomonadales*

Protein name	GI number	Figure number	Indel size	Indel position
A: CSIs Specific for *Xanthomonadales* (*Lysobacterales*)				
Glutaminyl t-RNA synthetase	194364460	Figure 3	18 aa ins	239–295
GTP-binding protein	58580596	Sup. Fig. 1	4 aa ins	303–350
Queuine tRNA-ribosyltransferase	194365393	Sup. Fig. 2	1 aa ins	289–339
Lipoyl synthase	58583575	Sup. Fig. 3	2 aa ins	156–209
Lysyl-tRNA synthetase	194365604	Sup. Fig. 4	3 aa ins	34–85
Dihydroorotate dehydrogenase	71275790	Sup. Fig. 5	7 aa ins	164–213
Carbamoyl phosphate synthase large subunit	166711938	Sup. Fig. 6	1 aa ins	403–457
Aspartate aminotransferase	28197970	Sup. Fig. 7	1 aa del	316–354
DNA polymerase I	194367713	Sup. Fig. 8	1 aa del	28–65
DNA topoisomerase IV subunit B	84624476	Sup. Fig. 9	1 aa del	282–326
B: CSIs Specific for *Xanthomonadaceae* (*Lysobacteriaceae*)				
DNA polymerase III subunit alpha	77747494	Figure 4a	4 aa ins	522–576
Uroporphyrinogen decarboxylase	294625972	Sup. Fig. 10	5 aa ins	295–340
DNA polymerase I	21244827	Sup. Fig. 11	1 aa ins	136–180
Coproporphyrinogen III oxidase	194367710	Sup. Fig. 12	1 aa del	166–215
tRNA isopentenyltransferase	194365248	Sup. Fig. 13	5 aa ins	219–256
Protoheme IX farnesyltransferase	15837961	Figure 4b	4 aa ins	150–192
Ribose-5-phosphate isomerase A	194367055	Sup. Fig. 14	1 aa ins	127–169
Aspartyl-tRNA synthetase	194366904	Sup. Fig. 15	4 aa del	343–391
2-oxoglutarate dehydrogenase E1	194366403	Sup. Fig. 16	1 aa del	782–830
Asparagine synthetase B	194365058	Sup. Fig. 17	2 aa ins	98–132
C: CSIs Specific for *Rhodanobacteriaceae*				
Uridylyltransferase	495713257	Figure 5a	1 aa ins	272–310
Xanthomonadin exporter protein	383315419	Sup. Fig. 18	1 aa del	171–196
Signal peptidase	494142978	Sup. Fig. 19	24 aa ins	111–165
Tryptophan synthase subunit alpha	383316227	Sup. Fig. 20	1 aa del	121–157
D: CSIs Specific for all *Rhodanobacteriaceae* except *Rudaea*				
CDP-diacylglycerol–glycerol-3-phosphate 3-phosphatidyltransferase	469817908	Figure 5b	4 aa ins	63–120
Protease *tldD*	495491439	Sup. Fig. 21	1 aa del	75–126
S-adenosylmethionine decarboxylase	383315616	Sup. Fig. 22	2 aa del	71–123
DEAD/DEAH box helicase	494777343	Sup. Fig. 23	1 aa ins	720–756
F0F1 ATP synthase subunit gamma	495082201	Sup. Fig. 24	17 aa ins	177–230
Proline aminopeptidase P II	469819587	Sup. Fig. 25	1 aa del	135–178
Glycosyl transferase	469816683	Sup. Fig. 26	2 aa del	101–140

members of the *Xanthomonadales* (Yilmaz et al. 2013; Fig. 2). The current taxonomy of the order *Xanthomonadales* is not concordant with 16S rRNA gene based phylogenies of the members of the order and the nomenclature of the order *Xanthomonadales* and a majority of the family names within this order are problematic and not in accordance with the International Code of Nomenclature of Bacteria (Oren 2010; Yilmaz et al. 2013; Tindall 2014a, b). However, apart from the 16S rRNA gene, no reliable morphological, biochemical, or molecular characteristics are known that are specifically shared by all members of this order or its distinct subgroups and can be used for their demarcation and classification (Saddler and

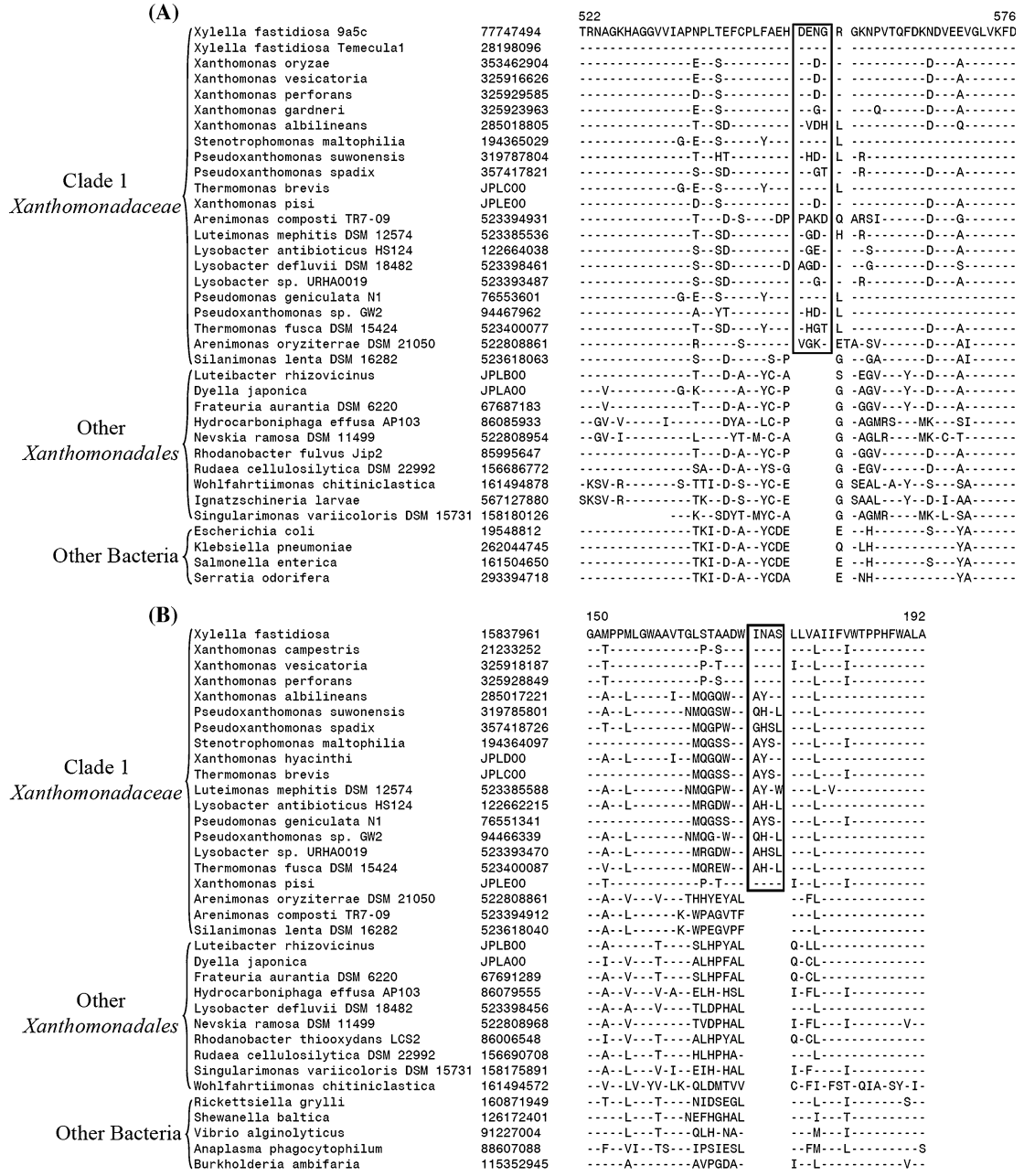
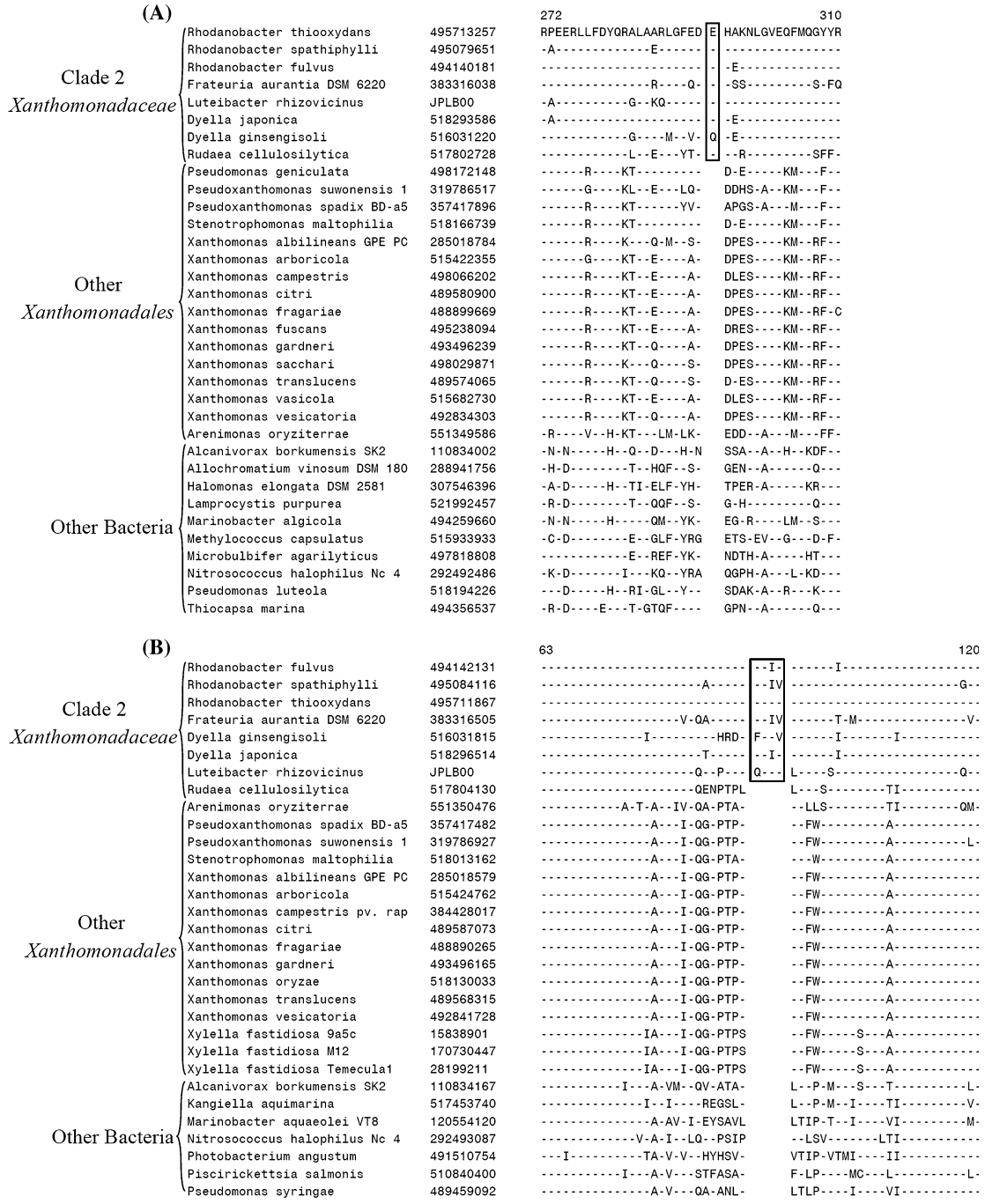


Fig. 4 Partial sequence alignments of **a** DNA polymerase III subunit alpha showing a 4 amino acid insertion (*boxed*) identified in all members of Clade 1 of the *Xanthomonadaceae* except *Silanimonas lenta* **b** the protein Protoheme IX farnesyltransferase showing a 4 amino acid insertion (*boxed*) identified in all members of Clade 1 of the *Xanthomonadaceae* except the genera *Arenimonas* and *Silanimonas*. Due to space constraints, sequence information for only representative *Xanthomonadales* and a limited number other bacteria is shown here, but similar CSIs were present in all members of the indicated group and not detected in any other bacterial species in the top 250 BLAST hits. Sequence information for other CSIs showing similar group specificities are presented in Supplemental Figs. 10–17 and summarized in Tables 2B


(A)
272
310
Clade 2 Xanthomonadaceae
Other Xanthomonadales
Other Bacteria
Rhodanobacter thiooxydans 495713257 RPEERLLFDYQRALAARLGFED E HAKNLGVEQFMQGYYR
Rhodanobacter spathiphylli 495079651 -A----------E------ - ----------
Rhodanobacter fulvus 494140181 ---------- - -E----------
Frateuria aurantia DSM 6220 383316038 ----------R----Q- - -SS---------S-FQ
Luteibacter rhizovicinus JPLB00 -A---------G--KQ----- - ----------
Dyella japonica 518293586 -A---------- - -E----------
Dyella ginsengisoli 516031220 ----------G----M--V- Q -E----------
Rudaea cellulosilytica 517802728 ----------L--E---YT- - --R---------SFF-
Pseudomonas geniculata 498172148 ------R----KT--------- D-E-----KM---F--
Pseudoxanthomonas suwonensis 1 319786517 ------G----KL--E---LQ- DDHS-A--KM---F--
Pseudoxanthomonas spadix BD-a5 357417896 ------R----KT------YV- APGS-A---M---F--
Stenotrophomonas maltophilia 518166739 ------R----KT--------- D-E-----KM---F--
Xanthomonas albilineans GPE PC 285018784 ------R----K---Q-M--S- DPES----KM--RF--
Xanthomonas arboricola 515422355 ------G----KT--E----A- DPES----KM--RF--
Xanthomonas campestris 498066202 ------R----KT--E----A- DLES----KM--RF--
Xanthomonas citri 489580900 ------R----KT--E----A- DPES----KM--RF--
Xanthomonas fragariae 488899669 ------R----KT--E----A- DPES----KM--RF-C
Xanthomonas fuscans 495238094 ------R----KT--E----A- DRES----KM--RF--
Xanthomonas gardneri 493496239 ------R----KT--Q----A- DPES----KM--RF--
Xanthomonas sacchari 498029871 ------R----K---Q----S- DPES----KM--RF--
Xanthomonas translucens 489574065 ------R----KT--Q----S- D-ES----KM--RF--
Xanthomonas vasicola 515682730 ------R----KT--E----A- DLES----KM--RF--
Xanthomonas vesicatoria 492834303 ------R----KT--Q----A- DPES----KM--RF--
Arenimonas oryziterrae 551349586 -R----V--H-KT---LM-LK- EDD--A---M---FF-
Alcanivorax borkumensis SK2 110834002 -N-N-----H--Q--D---H-N SSA--A--H--KDF--
Allochromatium vinosum DSM 180 288941756 -H-D--------T--HQF--S- GEN--A------Q---
Halomonas elongata DSM 2581 307546396 -A-D-----H--TI-ELF-YH- TPER-A-----KR---
Lamprocystis purpurea 521992457 -R-D--------T--QQF--S- G-H---------Q---
Marinobacter algicola 494259660 -N-N-----H-----QM--YK- EG-R----LM--S---
Methylococcus capsulatus 515933933 -C-D--------E--GLF-YRG ETS-EV--G---D-F-
Microbulbifer agarilyticus 497818808 ----------E--REF-YK- NDTH-A-----HT---
Nitrosococcus halophilus Nc 4 292492486 -K-D-------I---KQ--YRA QGPH-A---L-KD---
Pseudomonas luteola 518194226 ---D-----H--RI-GL--Y-- SDAK-A--R---K---
Thiocapsa marina 494356537 -R-D----E---T-GTQF---- GPN--A------Q---
(B)
63
120
Clade 2 Xanthomonadaceae
Other Xanthomonadales
Other Bacteria
Rhodanobacter fulvus 494142131 ---------- --I- ------I----------
Rhodanobacter spathiphylli 495084116 ----------A----- --IV ----------G--
Rhodanobacter thiooxydans 495711867 ---------- ---- ----------
Frateuria aurantia DSM 6220 383316505 ----------V-QA----- --IV ------T-M----------V-
Dyella ginsengisoli 516031815 ----------I---------HRD- F--V ------I-------I----------
Dyella japonica 518296514 ----------T----- --I- ------I----------
Luteibacter rhizovicinus JPLB00 ----------Q--P--- Q--- L----S----------Q--
Rudaea cellulosilytica 517804130 ----------QENPTPL L---S--------TI----------
Arenimonas oryziterrae 551350476 ----------A-T-A--IV-QA-PTA- --LLS--------TI--------QM-
Pseudoxanthomonas spadix BD-a5 357417482 ----------A---I-QG-PTP- --FW---------A----------
Pseudoxanthomonas suwonensis 1 319786927 ----------A---I-QG-PTP- --FW---------A---------L-
Stenotrophomonas maltophilia 518013162 ----------A---I-QG-PTA- ---W---------A----------
Xanthomonas albilineans GPE PC 285018579 ----------A---I-QG-PTP- --FW---------A----------
Xanthomonas arboricola 515424762 ----------A---I-QG-PTP- --FW---------A----------
Xanthomonas campestris pv. rap 384428017 ----------A---I-QG-PTP- --FW---------A----------
Xanthomonas citri 489587073 ----------A---I-QG-PTP- --FW---------A----------
Xanthomonas fragariae 488890265 ----------A---I-QG-PTP- --FW---------A----------
Xanthomonas gardneri 493496165 ----------A---I-QG-PTP- --FW---------A----------
Xanthomonas oryzae 518130033 ----------A---I-QG-PTP- --FW---------A----------
Xanthomonas translucens 489568315 ----------A---I-QG-PTP- --FW---------A----------
Xanthomonas vesicatoria 492841728 ----------A---I-QG-PTP- --FW---------A----------
Xylella fastidiosa 9a5c 15838901 ----------IA---I-QG-PTPS --FW-----S---A----------
Xylella fastidiosa M12 170730447 ----------IA---I-QG-PTPS --FW-----S---A----------
Xylella fastidiosa Temecula1 28199211 ----------IA---I-QG-PTPS --FW-----S---A----------
Alcanivorax borkumensis SK2 110834167 ----------I---A-VM--QV-ATA- L--P-M---S---T---------L-
Kangiella aquimarina 517453740 ----------I--I----REGSL- L--P-M--I----TI---------V-
Marinobacter aquaeolei VT8 120554120 ----------A-AV-I-EYSAVL LTIP-T--I----VI---------M-
Nitrosococcus halophilus Nc 4 292493087 ----------V-A-I--LQ--PSIP --LSV-------LTI----------
Photobacterium angustum 491510754 ---I----------TA-V-V--HYHSV- VTIP-VTMI----II----------
Piscirickettsia salmonis 510840400 ----------I---A-V---STFASA- F-LP----MC---L----------L-
Pseudomonas syringae 489459092 ----------A-V---QA-ANL- LTLP----I----VI----------

◀**Fig. 5** Partial sequence alignments of **a** the protein Uridylyltransferase showing a 1 amino acid insertion (*boxed*) identified in all members of Clade 2 of the *Xanthomonadaceae* **b** the protein CDP-diacylglycerol–glycerol-3-phosphate 3-phosphatidyltransferase showing a 4 amino acid insertion (*boxed*) identified in all members of Clade 2 of the *Xanthomonadaceae* except the early branching genus *Rudaea*. Sequence information for only representative *Xanthomonadales* and a limited number other bacteria is shown here, but similar CSIs were not detected in any other bacterial species in the top 250 BLAST hits. Sequence information for the other CSIs specific to the Clade 2 *Xanthomonadaceae* are presented in Supplemental Figs. 18–26 and summarized in Table 2C and D

Bradbury 2005a; Gutierrez et al. 2012; Losey et al. 2013). In this work, we have completed a robust phylogenetic analysis of the order *Xanthomonadales* and have utilized comparative genomic techniques to identify large numbers of novel molecular markers of common evolutionary decent (CSIs) shared by subgroups within the *Xanthomonadales*. The CSIs identified in this work both supplement gene based phylogenies and demarcate the groups within the *Xanthomonadales* in more definitive molecular terms. A summary diagram of the identified CSIs and the species in which they are found is shown in Fig. 6.

The results of phylogenetic analyses presented here do not support a monophyletic grouping of the members of the order *Xanthomonadales* and no CSI was identified that is uniquely shared by all members of this order. In phylogenetic trees, members of the order *Xanthomonadales* formed two main clades, one grouping together most of the members from the family *Xanthomonadaceae*, whereas the other clade was comprised of members from the families *Algiphilaceae* (containing the genus *Algiphilus*) and *Solimonadaceae* (containing the genera *Fontimonas*, *Singularimonas* and *Solimonas*) and genera related to these two families (viz. *Alkanibacter*, *Hydrocarboniphaga*, *Nevskia* and *Steroidobacter*). The lack of any identified CSIs or a consistent phylogenetic relationship between these two clades suggests that they may represent distinct evolutionary lineages within the *Gammaproteobacteria*. Additionally, in both our concatenated protein tree and the 16S rRNA gene tree, members of the clade containing *Algiphilaceae* and *Solimonadaceae* families consistently grouped with the members of the genus *Salinisphaera*, the sole members of the family "Salinisphaeraceae", suggesting that the species from these groups may share a common ancestor exclusive of the *Xanthomonadaceae* and other *Gammaproteobacteria*. The genera *Wohlfahrtiimonas* and *Ignatzschineria* branch distinctly from the *Xanthomonadales* in a clade with members of the order *Cardiobacteriales* in our concatenated protein based phylogenetic tree and show limited phylogenetic association with the other members of the *Xanthomonadales* in our 16S rRNA tree. Due to this inconsistent branching, further research will be required to accurately assess the phylogenetic placement of the genera *Wohlfahrtiimonas* and *Ignatzschineria,* but the available data suggests that they do not belong to the order *Xanthomonadales* sensu stricto.

Our work has identified 10 CSIs that support a monophyletic grouping of a majority of the members of the order *Xanthomonadales* that are currently part of the family *Xanthomonadaceae*. These CSIs were initially identified in our earlier comparative genomic study (Naushad and Gupta 2013) and the sequence information for them was updated in the present work for a large number of additional *Xanthomonadales*, including 5 genomes which were sequenced, *de novo*, for this study. Our earlier work identified 13 CSIs which were specific to this group (Naushad and Gupta 2013). Of these, all but 3 CSIs were found to be still specific to the whole group, while the remaining three CSIs were found to be specific for subsets of this large group. The observed specificity of the previously identified CSIs for a distinct bacterial group despite a large increase in the number of analyzed genomes strongly indicate that they constitute reliable molecular characteristics with predictive ability for distinguishing and demarcation of evolutionarily related bacterial groups.

Our work also provides strong molecular and phylogenetic support for the existence of two distinct clades within the *Xanthomonadaceae*: One clade consists of the genera *Xanthomonas*, *Xylella*, *Stenotrophomonas*, *Lutimonas*, *Lysobacter* and their relatives (Clade 1), whereas the other clade groups together members of the genera *Rudaea*, *Dylella*, *Lutibacter*, *Rhodanobacter* and their relatives (Clade 2). The members of these monophyletic clades branch distinctly from each other with strong bootstrap support in both the concatenated protein tree as well as in the 16S rRNA gene trees. Importantly, Clade 1 and Clade 2 are also supported by 10 and 11 identified CSIs, respectively, which serve to clearly distinguish them from each other and every other bacterial group

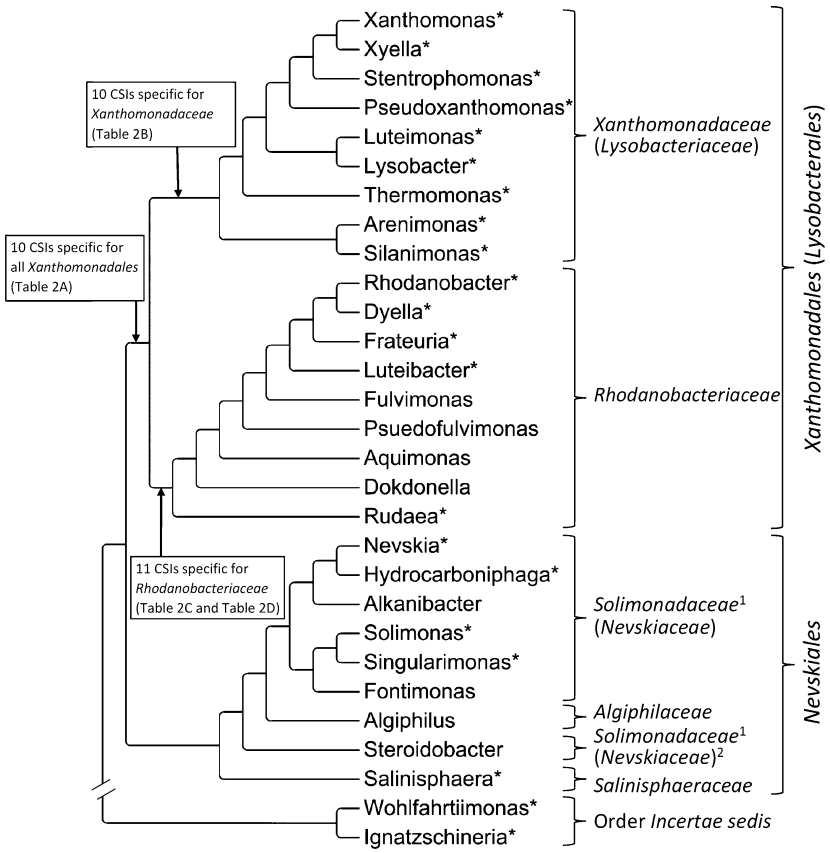


Fig. 6 A summary of the evolutionary relationships of the *Xanthomonadales* genera based upon phylogenetic analyses and the identified CSIs. Genera with genome sequenced members are indicated with *asterisks* (*). The distribution of the identified CSIs and the proposed reclassification of taxonomic groups are indicated. The genera *Wohlfahrtiimonas* and *Ignatzschineria* do not branch with the members of the order *Xanthomonadales* and hence are regarded as order *incertae sedis*. The families *Sinobacteraceae* and *Solimonadaceae* are synonymous; however, only the name of the family *Solimonadaceae* is shown here ([1]). The placement of the genus *Steroidobacter* within the family *Solimonadaceae* (*Nevskiaceae*) is tentative until a more detailed phylogenetic analysis can be completed for this genus ([2])

in molecular terms. This evidence suggests that the Clade 1 and Clade 2 represent two phylogenetically and molecularly distinguishable evolutionary lineages.

Taxonomic implications

Based on the branching of the members of the order *Xanthomonadales* in the concatenated protein and 16S rRNA gene trees and the large number of identified molecular markers (CSIs) that are specific for this group of bacteria, the following main inferences regarding the phylogeny of the *Xanthomonadales* can be derived.

(1) The order presently designated as *Xanthomonadales* contains 2 highly divergent phylogenetic groups, one made up of the members of the family *Xanthomonadaceae* and the other made up of the members of the families *Algiphilaceae*, *Solimonadaceae*, and "Salinisphaeraceae"

(2) The family presently designated as *Xanthomonadaceae*, which harbours a majority of the members from the order *Xanthomonadales*, also contains 2 distinct and distinguishable phylogenetic groups, one consisting of the genera *Xanthomonas*, *Xylella*, *Stenotrophomonas*, *Lutimonas*, *Lysobacter* and their relatives (Clade 1) and another clade consisting of the genera *Rudaea*, *Dylella*, *Lutibacter*, *Rhodanobacter* and their relatives (Clade 2)

Thus, the current taxonomy of the order *Xanthomonadales* does not accurately reflect the evolutionary histories of its members which exhibit enormous genetic diversity. In order to alleviate the taxonomic incongruences within the order *Xanthomonadales*, we

propose that the families *Algiphilaceae* (containing the genus *Algiphilus*) and *Solimonadaceae* (or *Nevskiaceae*) (containing the genera *Fontimonas*, *Singularimonas* and *Solimonas*) and genera related to these two families (viz. *Alkanibacter*, *Hydrocarboniphaga*, *Nevskia* and *Steroidobacter*), which do not branch with the other members of the order *Xanthomonadales*, be placed within a novel order, *Nevskiales* ord. nov., along with their closest evolutionary relatives, the members of the family *Salinisphaeraceae* fam. nov. Further, to recognize the presence of two distinct groups within the family presently designated as *Xanthomonadaceae,* the members of this family should be divided into two families: the family *Xanthomonadaceae* (containing the genera *Arenimonas*, *Luteimonas*, *Lysobacter*, *Metallibacterium*, *Panacagrimonas*, *Pseudoxanthomonas*, *Silanimonas*, *Stenotrophomonas*, *Thermomonas*, *Xanthomonas*, and *Xylella*) and a novel family, *Rhodanobacteraceae* fam. nov. (containing the genera *Aquimonas*, *Chiayiivirga*, *Dokdonella*, *Dyella*, *Frateuria*, *Fulvimonas*, *Luteibacter*, *Pseudofulvimonas*, *Rhodanobacter*, and *Rudaea*). The remaining two genera, *Wohlfahrtiimonas* and *Ignatzschineria*, whose taxonomic affiliation to the above two orders is not supported should be regarded as order *incertae sedis*.

Additionally, the present proposal also serves to help rectify several problems associated with the nomenclature of the order *Xanthomonadales*, the family *Xanthomonadaceae*, and the family *Solimonadaceae* (Tindall 2014b). It has been noted previously (Oren 2010; Tindall 2014a, b) that the names of these taxa are later synonyms for the order *Lysobacterales*, the family *Lysobacteraceae*, and the family *Nevskiaceae*, respectively. In recognition of these nomenclatural concerns, we are providing emended descriptions of the order *Lysobacterales*, the family *Lysobacteraceae*, and the family *Nevskiaceae*, which indicate that they are earlier synonyms of the order *Xanthomonadales*, the family *Xanthomonadaceae*, and the family *Solimonadaceae*, respectively. Descriptions of *Rhodanobacteraceae* fam. nov., *Nevskiales* ord. nov., and *Salinisphaeraceae* fam. nov. and emended descriptions of the order *Lysobacterales* (*Xanthomonadales*), the family *Lysobacteraceae* (*Xanthomonadaceae*), and the family *Nevskiaceae* (*Solimonadaceae*) are provided below.

Emended description of the order *Lysobacterales* Christensen and Cook (1978) (Approved Lists 1980)

Synonym: *Xanthomonadales* Saddler and Bradbury (2005a, b).

The order contains two families, *Lysobacteraceae* and *Rhodanobacteraceae*. Organisms are rods, 0.2–1.8 μm in diameter and 0.8–70 μm in length. Cells are both motile and non-motile. Organisms are aerobic, or facultatively anaerobic. Organisms are chemoorganotrophic and non-spore-forming. Organisms within this order may be either positive or negative in both oxidase and catalase tests. The G + C content of the DNA is 42–75 (mol%).The type genus of the order is *Lysobacter* Christensen and Cook (1978) (Approved Lists 1980) (Skerman et al. 1980) emend. Park et al. (2008).

Organisms from this order are distinguished from all other bacteria examined to date by 10 conserved signature indels in Glutaminyl t-RNA synthetase, GTP-binding protein, Queuine tRNA-ribosyltransferase, Lipoyl synthase, Lysyl-tRNA synthetase, Dihydroorotate dehydrogenase, Carbamoyl phosphate synthase large subunit, Aspartate aminotransferase, DNA polymerase I, and DNA topoisomerase IV subunit B (Tables 2A).

Emended description of the family *Lysobacteraceae* Christensen and Cook (1978) (Approved Lists 1980)

Synonym: *Xanthomonadaceae* Saddler and Bradbury (2005a, b).

The family contains twelve genera, *Arenimonas*, *Luteimonas*, *Lysobacter*, *Metallibacterium*, *Panacagrimonas*, *Pseudoxanthomonas*, *Silanimonas*, *Stenotrophomonas*, *Thermomonas*, *Xanthomonas* and *Xylella*. Organisms are rods, 0.2–1.8 μm in diameter and 0.8–70 μm in length. Cells are both motile and non-motile. Organisms are aerobic, or facultatively anaerobic. Organisms are chemoorganotrophic and non-spore-forming. Organisms within this family may be either positive or negative in both oxidase and catalase tests. The G+C content of the DNA is 42–70 (mol%). The type genus of the family is *Lysobacter*

Christensen and Cook (1978) (Approved Lists 1980) emend. Park et al. (2008).

Organisms from this order are distinguished from all other bacteria examined to date by 10 conserved signature indels in DNA polymerase III subunit alpha, Uroporphyrinogen decarboxylase, DNA polymerase I, Coproporphyrinogen III oxidase, tRNA isopentenyltransferase, Protoheme IX farnesyltransferase, Ribose-5-phosphate isomerase A, Aspartyl-tRNA synthetase, 2-oxoglutarate dehydrogenase E1, and Asparagine synthetase B (Tables 2B).

Description of *Rhodanobacteraceae* fam. nov

Rhodanobacteraceae (Rho.da.no.bac.ter.a,ce'ae N.L. masc. n. *Rhodanobacter* type genus of the family; *-aceae* ending to denote a family; N.L. fem. pl. n. *Rhodanobacteraceae* the family whose nomenclatural type is the genus *Rhodanobacter*).

The family contains nine genera, *Aquimonas*, *Dokdonella*, *Dyella*, *Frateuria*, *Fulvimonas*, *Luteibacter*, *Pseudofulvimonas*, *Rhodanobacter* and *Rudaea*. Organisms are rods, 0.3–0.5 μm in diameter and 1–4.5 μm in length. Cells are both motile and non-motile. Organisms are aerobic, chemoorganotrophic, and non-spore-forming. Organisms within this family may be either positive or negative in both oxidase and catalase tests. The G + C content of the DNA is 62–75 (mol%). The type genus of the family is *Rhodanobacter* Nalin et al. (1999).

Organisms from this order are distinguished from all other bacteria examined to date by 11 conserved signature indels in Uridylyltransferase, a xanthomonadin exporter protein, a signal peptidase, Tryptophan synthase subunit alpha, CDP-diacylglycerol–glycerol-3-phosphate 3-phosphatidyltransferase, Protease *tldD*, S-adenosylmethionine decarboxylase, DEAD/DEAH box helicase, F0F1 ATP synthase subunit gamma, Proline aminopeptidase P II, and Glycosyl transferase (Table 2C, D).

Description of *Nevskiales* ord. nov

Nevskiales (Nev.ski.a'les. N.L. fem. n. *Nevskia* type genus of the order; *-ales* ending to denote an order; N.L. fem. pl. n. *Nevskiales* the order whose nomenclatural type is the genus *Nevskia*).

The order contains three families, *Algiphilaceae, Salinisphaeraceae,* and *Nevskiaceae*. Organisms are rods and cocci, 0.6–1.3 μm in diameter and 0.4–2 μm in length. Cells are non-motile or motile by means of a one or more polar flagella. Organisms are aerobic, or facultatively anaerobic. Organisms are chemoorganotrophic and non-spore-forming. Oxidase and catalase positive. The G+C content of the DNA is 60–68 (mol%). The type genus of the order is *Nevskia* Famintzin 1892 (Approved Lists 1980).

Emended Description of the family *Nevskiaceae* Henrici and Johnson 1935 (Approved Lists 1980)

Synonyms: *Sinobacteraceae* Zhou et al. (2008), *Solimonadaceae* Losey et al. (2013).

The family contains six genera, *Alkanibacter*, *Fontimonas*, *Hydrocarboniphaga*, *Nevskia*, *Solimonas* and *Steroidobacter*.[1] Organisms are rods, 0.6–0.85 μm in diameter and 0.9–2 μm in length. Cells are non-motile or motile by means of a single polar flagellum. Organisms are aerobic, or facultatively anaerobic. Organisms are chemoorganotrophic and non-spore-forming. Oxidase and catalase positive. The G + C content of the DNA is 60–65 (mol%). The type genus of the family is *Nevskia* Famintzin 1892 (Approved Lists 1980).

Description of *Salinisphaeraceae* fam. nov

Salinisphaeraceae (Sa.li.ni.sphae.ra.ce'ae. N.L. fem. n. *Salinisphaera* type genus of the family; *-aceae* ending to denote a family; N.L. fem. pl. n. *Salinisphaeraceae* the family whose nomenclatural type is the genus *Salinisphaera*).

The family contains one genus, *Salinisphaera,* which is also the type genus of the family. The description of the family is the same as that of the

[1] The genus *Steroidobacter* does not branch monophyletically with the other members of the family *Nevskiaceae* in 16S rRNA gene based phylogenies. However, *Steroidobacter* is clearly distinct from the order *Xanthomonadales* and family *Xanthomonadaceae* in which it was previously placed. Its placement within the family *Nevskiaceae* is tentative until more detailed phylogenetic analysis can be completed for this genus.

genus *Salinisphaera* Antunes et al. (2003) emend. Shimane et al. (2013).

Acknowledgments We thank Professor Iain Sutcliffe for valuable comments and suggestions for improvement of this manuscript. This work was supported by a research grant from the Natural Science and Engineering Research Council of Canada to RSG.

References

Adeolu M, Gupta RS (2013) Phylogenomics and molecular signatures for the order *Neisseriales*: proposal for division of the order *Neisseriales* into the emended family *Neisseriaceae* and *Chromobacteriaceae* fam. nov. Anton Leeuw Int J G 104(1):1–24

Antunes A, Eder W, Fareleira P, Santos H, Huber R (2003) *Salinisphaera shabanensis* gen. nov., sp. nov., a novel, moderately halophilic bacterium from the brine–seawater interface of the Shaban Deep, Red Sea. Extremophiles 7(1):29–34

Bhandari V, Ahmod NZ, Shah HN, Gupta RS (2013) Molecular signatures for *Bacillus* species: demarcation of the *Bacillus subtilis* and *Bacillus cereus* clades in molecular terms and proposal to limit the placement of new species into the genus *Bacillus*. Int J Syst Evol Microbiol 63(7):2712–2726

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17(4):540–552

Charlebois RL, Doolittle WF (2004) Computing prokaryotic gene ubiquity: Rescuing the core from extinction. Genome Res 14(12):2469–2477

Chatterjee S, Almeida RPP, Lindow S (2008) Living in two worlds: the plant and insect lifestyles of *Xylella fastidiosa*. Annu Rev Phytopathol 46:243–271

Chen J, Xie G, Han S, Chertkov O, Sims D, Civerolo EL (2010) Whole genome sequences of two *Xylella fastidiosa* strains (M12 and M23) causing almond leaf scorch disease in California. J Bacteriol 192(17):4534

Christensen P, Cook FD (1978) *Lysobacter*, a new genus of nonfruiting, gliding bacteria with a high base ratio. Int J Syst Bacteriol 28(3):367–393

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(5765):1283–1287

Cole J, Wang Q, Fish J, Chai B, McGarrell D, Sun Y, Brown C, Porras-Alfaro A, Kuske C, Tiedje J (2014) Ribosomal database project: data and tools for high throughput rRNA analysis. Nucleic Acids Res 42(1):D633

Crossman LC, Gould VC, Dow JM, Vernikos GS, Okazaki A, Sebaihia M, Saunders D, Arrowsmith C, Carver T, Peters N (2008) The complete genome, comparative and functional analysis of *Stenotrophomonas maltophilia* reveals an organism heavily shielded by drug resistance determinants. Genome Biol 9(4):R74

Cutino-Jimenez AM, Martins-Pinheiro M, Lima WC, Martin-Tornet A, Morales OG, Menck CFM (2010) Evolutionary placement of *Xanthomonadales* based on conserved protein signature sequences. Mol Phylogen Evol 54(2):524–534

da Silva AR, Ferro JA, Reinach F, Farah C, Furlan L, Quaggio R, Monteiro-Vitorello C, Van Sluys M, Almeida N, Alves L (2002) Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. Nature 417(6887):459–463

Darrasse A, Carrere S, Barbe V, Boureau T, Arrieta-Ortiz ML, Bonneau S, Briand M, Brin C, Cociancich S, Durand K et al (2013) Genome sequence of *Xanthomonas fuscans* subsp. *fuscans* strain 4834-R reveals that flagellar motility is not a general feature of xanthomonads. BMC Genomics 14:761

De Clercq D, Van Trappen S, Cleenwerck I, Ceustermans A, Swings J, Coosemans J, Ryckeboer J (2006) *Rhodanobacter spathiphylli* sp. nov., a gammaproteobacterium isolated from the roots of Spathiphyllum plants grown in a compost-amended potting mix. Int J Syst Evol Microbiol 56(Pt 8):1755–1759

Gao B, Gupta RS (2012a) Microbial systematics in the post-genomics era. Anton Leeuw Int J G 101(1):45–54

Gao B, Gupta RS (2012b) Phylogenetic framework and molecular signatures for the main clades of the phylum Actinobacteria. Microbiol Mol Biol Rev 76(1):66–112

Gao B, Mohan R, Gupta RS (2009) Phylogenomics and protein signatures elucidating the evolutionary relationships among the *Gammaproteobacteria*. Int J Syst Evol Microbiol 59(2):234–247

Gupta RS (1998) Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. Microbiol Mol Biol Rev 62(4):1435

Gupta RS (2010) Applications of conserved indels for understanding microbial phylogeny. In: Oren A, Papke RT (eds) Molecular phylogeny of microorganisms. Caister Academic Press, Norfolk, pp 135–150

Gupta RS (2014) Identification of conserved indels that are useful for classification and evolutionary studies Methods in Microbiology, vol 41. Academic Press: 10.1016/bs.mim.2014.05.003

Gupta RS, Lali R (2013) Molecular signatures for the phylum Aquificae and its different clades: proposal for division of the phylum Aquificae into the emended order *Aquificales*, containing the families *Aquificaceae* and *Hydrogenothermaceae*, and a new order *Desulfurobacteriales* ord. nov., containing the family *Desulfurobacteriaceae*. Anton Leeuw Int J G 104(3):349–368

Gupta RS, Mahmood S, Adeolu M (2013) A phylogenomic and molecular signature based approach for characterization of the phylum Spirochaetes and its major clades: proposal for a taxonomic revision of the phylum. Frontiers in microbiology 4:217

Gutierrez T, Green DH, Whitman WB, Nichols PD, Semple KT, Aitken MD (2012) *Algiphilus aromaticivorans* gen. nov., sp. nov., an aromatic hydrocarbon-degrading bacterium isolated from a culture of the marine dinoflagellate *Lingulodinium polyedrum*, and proposal of *Algiphilaceae* fam. nov. Int J Syst Evol Microbiol 62(11):2743–2749

Henrici AT, Johnson DE (1935) Studies of Freshwater Bacteria: II. Stalked Bacteria, a New Order of Schizomycetes. J Bacteriol 30(1):61–93

Im WT, Lee ST, Yokota A (2004) *Rhodanobacter fulvus* sp. nov., a beta-galactosidase-producing gammaproteobacterium. J Gen Appl Microbiol 50(3):143–147

Jalan N, Kumar D, Yu F, Jones JB, Graham JH, Wang N (2013) Complete genome sequence of *Xanthomonas citri* subsp. citri Strain Aw12879, a restricted-host-range citrus canker-causing bacterium. Genome Announc 1(3):e00235-13

Jaufeerally-Fakim Y, Dookun A (2000) Extraction of high quality DNA from polysaccharides-secreting xanthomonads. Sci Technol Res J Univ Maurit 6:33–40

Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ (1998) Multiple sequence alignment with clustal X. Trends Biochem Sci 23(10):403

Kostka JE, Green SJ, Rishishwar L, Prakash O, Katz LS, Marino-Ramirez L, Jordan IK, Munk C, Ivanova N, Mikhailova N et al (2012) Genome sequences for six *Rhodanobacter* strains, isolated from soils and the terrestrial subsurface, with variable denitrification capabilities. J Bacteriol 194(16):4461–4462

Kyrpides N, Overbeek R, Ouzounis C (1999) Universal protein families and the functional content of the last universal common ancestor. J Mol Evol 49(4):413–423

Lee B-M, Park Y-J, Park D-S, Kang H-W, Kim J-G, Song E-S, Park I-C, Yoon U-H, Hahn J-H, Koo B-S (2005) The genome sequence of *Xanthomonas oryzae* pathovar oryzae KACC10331, the bacterial blight pathogen of rice. Nucleic Acids Res 33(2):577–586

Lee CS, Kim KK, Aslam Z, Lee ST (2007) *Rhodanobacter thiooxydans* sp. nov., isolated from a biofilm on sulfur particles used in an autotrophic denitrification process. Int J Syst Evol Microbiol 57(Pt 8):1775–1779

Lee SH, Jin HM, Lee HJ, Kim JM, Jeon CO (2012) Complete genome sequence of the BTEX-degrading bacterium *Pseudoxanthomonas spadix* BD-a59. J Bacteriol 194(2):544

Looney WJ, Narita M, Mühlemann K (2009) *Stenotrophomonas maltophilia*: an emerging opportunist human pathogen. Lancet Infect Dis 9(5):312–323

Losey NA, Stevenson BS, Verbarg S, Rudd S, Moore ER, Lawson PA (2013) *Fontimonas thermophila* gen. nov., sp. nov., a moderately thermophilic bacterium isolated from a freshwater hot spring, and proposal of *Solimonadaceae* fam. nov. to replace *Sinobacteraceae* Zhou et al. 2008. Int J Syst Evol Microbiol 63(1):254–259

Meidanis J, Braga MD, Verjovski-Almeida S (2002) Whole-genome analysis of transporters in the plant pathogen *Xylella fastidiosa*. Microbiol Mol Biol Rev 66(2):272–299

Nalin R, Simonet P, Vogel TM, Normand P (1999) *Rhodanobacter lindaniclasticus* gen. nov., sp. nov., a lindane-degrading bacterium. Int J Syst Bacteriol 49(1):19–23

Naushad HS, Gupta RS (2013) Phylogenomics and molecular signatures for species from the plant pathogen-containing order *Xanthomonadales*. PLoS ONE 8(2):e55216

Oren A (2010) The phyla of prokaryotes—cultured and uncultured. In: Oren A, Papke RT (eds) Molecular phylogeny of microorganisms. Caister Academic Press, Norfolk, pp 85–107

Park JH, Kim R, Aslam Z, Jeon CO, Chung YR (2008) *Lysobacter capsici* sp. nov., with antimicrobial activity, isolated from the rhizosphere of pepper, and emended description of the genus *Lysobacter*. Int J Syst Evol Microbiol 58(2):387–392

Parte AC (2013) LPSN–list of prokaryotic names with standing in nomenclature. Nucleic Acids Res 42:D613–D616

Pieretti I, Royer M, Barbe V, Carrere S, Koebnik R, Cociancich S, Couloux A, Darrasse A, Gouzy J, Jacques MA et al (2009) The complete genome sequence of *Xanthomonas albilineans* provides new insights into the reductive genome evolution of the xylem-limited *Xanthomonadaceae*. BMC Genomics 10:616

Qian W, Jia Y, Ren SX, He YQ, Feng JX, Lu LF, Sun Q, Ying G, Tang DJ, Tang H et al (2005) Comparative and functional genomic analyses of the pathogenicity of phytopathogen *Xanthomonas campestris* pv. *campestris*. Genome Res 15(6):757–767

Rokas A, Holland PWH (2000) Rare genomic changes as a tool for phylogenetics. Trends Ecol Evol 15(11):454–459

Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425(6960):798–804

Ryan RP, Vorhölter F-J, Potnis N, Jones JB, Van Sluys M-A, Bogdanove AJ, Dow JM (2011) Pathogenomics of *Xanthomonas*: understanding bacterium–plant interactions. Nat Rev Microbiol 9(5):344–355

Saddler GS, Bradbury JF (2005a) Order III. *Xanthomonadales* ord. nov. In: Brenner DJ, Krieg NR, Staley JT, Garrity GM, Boone, Vos P, Goodfellow M, Rainey FA, Schleifer K-H (eds) Bergey's manual of systematic bacteriology. Springer, Austin, pp 63–122

Saddler GS, Bradbury JF (2005b) *Xanthomonadaceae* fam. nov Validation of publication of new names and new combinations previously effectively published outside the IJSEM, List no 106. Int J Syst Evol Microbiol 55:2235–2238

Salzberg SL, Sommer DD, Schatz MC, Phillippy AM, Rabinowicz PD, Tsuge S, Furutani A, Ochiai H, Delcher AL, Kelley D (2008) Genome sequence and rapid evolution of the rice pathogen *Xanthomonas oryzae* pv. *oryzae* PXO99A. BMC Genomics 9(1):204

Shimane Y, Tsuruwaka Y, Miyazaki M, Mori K, Minegishi H, Echigo A, Ohta Y, Maruyama T, Grant WD, Hatada Y (2013) *Salinisphaera japonica* sp. nov., a moderately halophilic bacterium isolated from the surface of a deep-sea fish, *Malacocottus gibber*, and emended description of the genus *Salinisphaera*. Int J Syst Evol Microbiol 63(6):2180–2185

Skerman VBD, McGowan V, Sneath PHA (1980) Approved lists of bacterial names. Int J Syst Bacteriol 30(1):225–420

Studholme DJ, Wasukira A, Paszkiewicz K, Aritua V, Thwaites R, Smith J, Grant M (2011) Draft genome sequences of *Xanthomonas sacchari* and two banana-associated xanthomonads reveal insights into the *Xanthomonas* group 1 clade. Genes 2(4):1050–1065

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739

Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. In: Miura RM (ed) Lectures on mathematics in the life sciences, 17th edn. American Mathematical Society, Providence, pp 57–86

Tindall B (2014a) Names at the rank of class, subclass and order, their typification and current status: supplementary information to Opinion 79. Judicial commission of the international committee on systematics of prokaryotes. Int J Syst Evol Microbiol 64(10):3599–3602

Tindall BJ (2014b) The family name *Solimonadaceae Losey* et al. 2013 is illegitimate, proposals to create the names '*Sinobacter soli*' comb. nov. and '*Sinobacter variicoloris*' contravene the Code, the family name *Xanthomonadaceae* Saddler and Bradbury 2005 and the order name *Xanthomonadales* Saddler and Bradbury 2005 are illegitimate and notes on the application of the family names *Solibacteraceae* Zhou et al. 2008, Nevskiaceae Henrici and Johnson 1935 (Approved Lists 1980) and *Lysobacteraceae* Christensen and Cook 1978 (Approved Lists 1980) and order name *Lysobacteriales* Christensen and Cook 1978 (Approved Lists 1980) with respect to the classification of the corresponding type genera *Solibacter* Zhou et al. 2008 *Nevskia* Famintzin 1892 (Approved Lists 1980) and *Lysobacter* Christensen and Cook 1978 (Approved Lists 1980) and importance of accurately expressing the link between a taxonomic name, its authors and the corresponding description/circumscription/emendation. Int J Syst Evol Microbiol 64(1):293–297

Van Sluys MA, de Oliveira MC, Monteiro-Vitorello CB, Miyaki CY, Furlan LR, Camargo LE, da Silva AC, Moon DH, Takita MA, Lemos EG et al (2003) Comparative analyses of the complete genome sequences of Pierce's disease and citrus variegated chlorosis strains of *Xylella fastidiosa*. J Bacteriol 185(3):1018–1026

Williams KP, Gillespie JJ, Sobral BW, Nordberg EK, Snyder EE, Shallom JM, Dickerman AW (2010) Phylogeny of *gammaproteobacteria*. J Bacteriol 192(9):2305–2314

Wilson K (1987) Preparation of genomic DNA from bacteria. In: Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, Struhl K (eds) Current protocols in molecular biology. Wiley, New York, pp 2.4.1–2.4.2

Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature 462(7276):1056–1060

Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO (2013) The SILVA and "all-species living tree project (LTP)" taxonomic frameworks. Nucleic Acids Res. doi:10.1093/nar/gkt1209

Zhou Y, Zhang Y-Q, Zhi X-Y, Wang X, Dong J, Chen Y, Lai R, Li W-J (2008) Description of *Sinobacter flavus* gen. nov., sp. nov., and proposal of *Sinobacteraceae* fam. nov. Int J Syst Evol Microbiol 58(1):184–189

CHAPTER 7

Protein Based Molecular Markers Provide Reliable Means to Understand Prokaryotic Phylogeny and Support a Predominantly Darwinian Mode of Evolution

The following chapter is a review of comparative genomic analyses work performed in Dr. R. S. Gupta's lab and its use for elucidation of prokaryotic relationships. Using CSIs and CSPs, the chapter supports the view that bacterial relationships can be observed in a tree-like pattern and that lateral gene transfer events have only a limited effect on masking prokaryotic relationships. Using previously published data, I was involved in data analysis, the preparation of the manuscript and construction of the figures and tables.

*Due to limited space, supplementary figures and tables are not included in the chapter but can be accessed along with the rest of the manuscript at:

Bhandari, V., Naushad, H. S., and Gupta, R. S. (2012) Protein based molecular markers provide reliable means to understand prokaryotic phylogeny and support Darwinian mode of evolution. Front Cell Infect Microbiol 2, 98

frontiers in
CELLULAR AND INFECTION MICROBIOLOGY

REVIEW ARTICLE
published: 26 July 2012
doi: 10.3389/fcimb.2012.00098

Protein based molecular markers provide reliable means to understand prokaryotic phylogeny and support Darwinian mode of evolution

Vaibhav Bhandari, Hafiz S. Naushad and Radhey S. Gupta*

Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, ON, Canada

Edited by:
Yousef Abu Kwaik, University of Louisville School of Medicine, USA

Reviewed by:
Srinand Sreevastan, University of Minnesota, USA
Andrey P. Anisimov, State Research Center for Applied Microbiology and Biotechnology, Russia

***Correspondence:**
Radhey S. Gupta, Department of Biochemistry and Biomedical Sciences, McMaster University, 1200 Main Street West, Health Sciences Center, Hamilton, ON L8N 3Z5, Canada.
e-mail: gupta@mcmaster.ca

The analyses of genome sequences have led to the proposal that lateral gene transfers (LGTs) among prokaryotes are so widespread that they disguise the interrelationships among these organisms. This has led to questioning of whether the Darwinian model of evolution is applicable to prokaryotic organisms. In this review, we discuss the usefulness of taxon-specific molecular markers such as conserved signature indels (CSIs) and conserved signature proteins (CSPs) for understanding the evolutionary relationships among prokaryotes and to assess the influence of LGTs on prokaryotic evolution. The analyses of genomic sequences have identified large numbers of CSIs and CSPs that are unique properties of different groups of prokaryotes ranging from phylum to genus levels. The species distribution patterns of these molecular signatures strongly support a tree-like vertical inheritance of the genes containing these molecular signatures that is consistent with phylogenetic trees. Recent detailed studies in this regard on the Thermotogae and Archaea, which are reviewed here, have identified large numbers of CSIs and CSPs that are specific for the species from these two taxa and a number of their major clades. The genetic changes responsible for these CSIs (and CSPs) initially likely occurred in the common ancestors of these taxa and then vertically transferred to various descendants. Although some CSIs and CSPs in unrelated groups of prokaryotes were identified, their small numbers and random occurrence has no apparent influence on the consistent tree-like branching pattern emerging from other markers. These results provide evidence that although LGT is an important evolutionary force, it does not mask the tree-like branching pattern of prokaryotes or understanding of their evolutionary relationships. The identified CSIs and CSPs also provide novel and highly specific means for identification of different groups of microbes and for taxonomical and biochemical studies.

Keywords: conserved indels, signature proteins, phylogenetic trees, lateral gene transfers, Thermotogae, Archaea, Crenarchaeota, RpoB signatures

INTRODUCTION

The understanding of prokaryotic relationships is one of the most important goals of evolutionary sciences. These relationships have been difficult to understand due to the simplicity and antiquity of prokaryotic organisms and disagreements in viewpoints among evolutionary biologists regarding the importance of different factors when grouping prokaryotes. Although earlier studies in this regard were based on morphology or physiology (Cowan, 1965; Buchanan and Gibbons, 1974; Stanier et al., 1976), the field itself has evolved to account for new information brought about by technological or informational breakthroughs, viz. molecular data, DNA hybridization and 16S rRNA (Zuckerkandl and Pauling, 1965; Woese and Fox, 1977; Woese, 1987). The most recent breakthrough involves rapid and easily available sequencing of entire genomic sequences (Fleischmann et al., 1995; Iguchi et al., 2009; NCBI genomic database, 2012). This has allowed determination of evolutionary relationships among different organisms based upon large numbers of different gene/protein sequences using a variety of approaches (Gupta, 1998; Haggerty et al., 2009; Puigbo et al., 2009; Blair and Murphy, 2011).

The comparative genomic analyses have revealed that phylogenetic relationships deducted based upon different genes and protein sequences are not congruent and lateral gene transfer (LGT) among different taxa is indicated as the main factor responsible for this lack of concordance (Gogarten et al., 2002; Bapteste and Boucher, 2008; Dagan et al., 2008; Puigbo et al., 2009; Swithers et al., 2009; Andam and Gogarten, 2011). This has led to questioning of whether the Darwinian model of evolution involving vertical inheritance of genes from parents to progenies (Darwin, 1859) is applicable to the prokaryotes (Doolittle, 1999; Pennisi, 1999; Gogarten et al., 2002; Dagan and Martin, 2006; Doolittle and Bapteste, 2007; Dagan et al., 2008; Bapteste et al., 2009; Williams et al., 2011). Multiple mechanisms are known to contribute to the evolution of an organism's genomes including genes that are acquired vertically from the parent organism,

evolution of new genes by gene duplication and divergence, gain of new genes by means of LGTs, as well as gene losses in various lineages (Bapteste et al., 2009; Ragan and Beiko, 2009; Treangen and Rocha, 2011; Williams et al., 2011). LGT, in particular, is being increasingly thought to have an overbearing influence on prokaryotic genome composition. Although rRNAs, ribosomal proteins and other genes involved in the information transfer processes are considered less prone to LGTs due to their involvement in complex gene networks (Jain et al., 1999; Sorek et al., 2007), recent studies indicate that no single gene/protein is completely immune to this process (Yap et al., 1999; Doolittle and Bapteste, 2007; Dagan et al., 2008). Some recent studies have estimated that over time most genes ($81 \pm 15\%$) have undergone at least one LGT event (Doolittle, 1999; Dagan and Martin, 2007; Doolittle and Bapteste, 2007; Dagan et al., 2008). These studies in large part form the basis of the hypothesis that LGTs have led to abolishment of all signals that can be used for determination of prokaryotic evolutionary relationships and a call for uprooting the tree of life (Martin, 1999; Pennisi, 1999; Doolittle, 2000; Gogarten et al., 2002; Delsuc et al., 2005; Bapteste et al., 2009).

Although the importance of LGTs in genome evolution is widely accepted, there is considerable disagreement concerning the prevalence of LGTs and their impact on prokaryotic evolutionary relationships. While some authors have indicated that LGT is so profuse that its influence disguises the Darwinian mode of evolution involving vertical inheritance of genes (Gogarten et al., 2002; Bapteste et al., 2005b, 2009; Doolittle and Bapteste, 2007; Koonin, 2007), others have inferred that the incidences of LGTs are either very minimal or limited and those genes that are laterally transferred have little impact on prokaryotic phylogeny (Wolf et al., 2002; Kurland et al., 2003; Dutilh et al., 2004; Beiko et al., 2005; Kunin et al., 2005; Kurland, 2005; Galtier, 2007; Puigbo et al., 2009; Gao and Gupta, 2012a). However, there are no standardized methods to assess LGTs and the methods used to infer LGTs are varied and based upon large numbers of often poorly supported assumptions (Koski and Golding, 2001; Koski et al., 2001; Ragan, 2001; Beiko et al., 2005; Boto, 2010). Thus, the prevalence of LGTs differ greatly among different studies and often similar datasets have led to dissimilar conclusions (Koski et al., 2001; Ragan, 2001; Wang, 2001; Lerat et al., 2003; Susko et al., 2006; Zhaxybayeva et al., 2007; Marri and Golding, 2008; Roettger et al., 2009). Therefore, prior to concluding that in view of LGTs the Darwinian mode of evolution is not a suitable model for prokaryotes, reliability of the incidences of LGTs and their overall impact on the evolutionary relationships should be critically examined.

Despite the prevalence of LGTs, phylogenetic trees based upon 16S rRNA as well as numerous single genes as well multi-gene analyses strongly support the existence of large numbers of distinct phyla of bacteria (Ludwig and Klenk, 2005). Additionally, these trees also clearly delineate many discrete taxonomic clades within these phyla (Woese, 1987; Ludwig and Klenk, 2005; Ciccarelli et al., 2006; Wu et al., 2009; Gao and Gupta, 2012a). In a recent detailed study Puigbo et al. (2009) reported construction of phylogenetic trees for 6901 prokaryotic genes. Although there were significant topological differences among these trees, a consistent phylogenetic signal was observed in most of these trees, indicating that the LGT events, which were of random nature, did not obscure the central trend resulting from the vertical transfer of genes. The fact that similar prokaryotic clades at different taxonomic levels (ranging from phyla to genera) are consistently identified in phylogenetic trees based upon different gene/protein sequences strongly indicates that the distinctness of the prokaryotic taxa and their evolutionary relationships are in large part discernible and they have not been obliterated by LGTs (Woese, 1987; Daubin et al., 2002; Kurland et al., 2003; Lerat et al., 2003; Beiko et al., 2005; Kurland, 2005; Ludwig and Klenk, 2005; Ciccarelli et al., 2006; Ragan and Beiko, 2009; Wu et al., 2009; Boto, 2010; Yarza et al., 2010; Gupta, 2010b; Gao and Gupta, 2012a). To account for the above observations and the occurrences of LGTs, it has been suggested that the prokaryotic evolution has both tree-like (at intermediate phylogenetic depths) and non-tree (or net-like) (at the base and tips) characteristics (Dagan et al., 2008; Puigbo et al., 2009, 2010; Swithers et al., 2009; Boto, 2010; Beiko, 2011; Dagan, 2011; Kloesges et al., 2011; Popa et al., 2011).

The availability of genome sequences is also enabling development of novel and independent sequence based approaches for determining the evolutionary relationships among organisms and to assess the impact of LGTs on these relationships. In this review, we provide a summary of our recent work in this area based upon two different types of molecular markers that we have used successfully for understanding the evolutionary relationships among prokaryotes. Based upon these markers it is now possible to identify different prokaryotic taxa ranging from phyla to genera in clear molecular terms and the evolutionary relationships among them can also be reliably deducted (Gupta and Griffiths, 2002; Gupta, 2009, 2010a; Gao and Gupta, 2012b). The relationships revealed by these new approaches strongly support a tree-like branching pattern among prokaryotes and the observed incidences of LGTs, which exhibit no specific pattern or statistical significance, apparently have no major impact on the derived relationships. It is contended that these molecular markers provide valuable means for developing a reliable phylogeny and taxonomy of the prokaryotic organisms.

USEFULNESS OF CONSERVED SIGNATURE INDELS (CSIs) AND CONSERVED SIGNATURE PROTEINS (CSPs) FOR UNDERSTANDING EVOLUTIONARY RELATIONSHIPS AMONG PROKARYOTES

Of the two kinds of molecular markers that we are using for studying prokaryotic evolution, the *c*onserved *s*ignature *i*ndels (inserts or deletions), or CSIs, in protein sequences comprises an important category (Gupta, 1998, 2010a; Griffiths and Gupta, 2001). The CSIs that provide useful molecular markers for evolutionary studies are generally of the same lengths and they are flanked on both sides by conserved regions to ensure that the observed changes are not caused by alignment artifacts (Gupta, 1998; Gupta and Griffiths, 2002; Jordan and Goldman, 2012). When such CSIs are present in the same position in a given protein in a group of related species, their presence is most parsimoniously explained by postulating that the genetic change leading to the CSI occurred in a common ancestor of this group

and then this gene with the indel was vertically transmitted to its progeny (Rivera and Lake, 1992; Baldauf and Palmer, 1993; Gupta, 1998, 2000b; Rokas and Holland, 2000; Cutino-Jimenez et al., 2010). The CSIs that are uniquely shared by organisms of one taxa provide molecular tools for identifying the species from this taxa and consolidating the relationships among bacteria of that taxa by delimiting it in molecular terms (Gupta, 2004). Additionally, depending upon the presence or absence of a given CSI in the outgroup species, it can be determined whether the indel represents an insert or a deletion and based upon this a rooted relationship among the species of interest can be derived. Our earlier work in this regard has led to identification of large numbers of CSIs that are specific for different groups of microbes at various phylogenetic levels (**Table 1**; Gupta and Griffiths, 2006; Gupta, 2009; Gupta and Bhandari, 2011; Gupta and Shami, 2011; Gao and Gupta, 2012b).

The second kind of molecular markers that we have usefully employed in our systematic and evolutionary studies are whole proteins that are uniquely found in particular groups or subgroups of bacteria (Gupta, 2006; Gupta and Griffiths, 2006; Gupta and Mok, 2007; Gao and Gupta, 2012b). Comparative analyses of genomic sequences have indicated that many conserved proteins are uniquely present in all species from particular groups, at different phylogenetic depths (Daubin and Ochman, 2004; Lerat et al., 2005; Gupta, 2006; Gupta and Griffiths, 2006; Gupta and Mok, 2007; Dutilh et al., 2008; Gao and Gupta, 2012b). Because of their unique presence in species from particular phylogenetic clades of species, it is likely that the genes for these CSPs originated once in a common ancestor of these groups and then vertically acquired by all its descendants. Because of their taxa specificity these CSPs again provide valuable molecular markers for identifying different groups of species in molecular terms and for evolutionary studies (Gao and Gupta, 2007; Gupta and Mathews, 2010; Gupta, 2010b). However, when a CSP (or CSI) is confined to certain species/strains, then based upon this information alone, it is often difficult to determine whether these species form a clade in the phylogenetic sense or not. Hence, to understand the evolutionary significance of these signatures, such studies are generally performed in conjunction with phylogenetic analysis, which provides a reference point for evaluating the significance of various CSIs and CSPs (Gao and Gupta, 2007; Gupta and Mathews, 2010; Gupta, 2010b).

Molecular markers in the form of CSIs and CSPs have proven useful for examining or consolidating prokaryotic relationships at domain, phylum as well as intra-phylum levels. **Table 1** provides a summary of some bacterial and archaeal taxa for which CSIs and CSPs have been identified (Gupta, 2010a). Two recent detailed studies based upon CSIs and CSPs have focused upon understanding evolutionary relationships within the phylum Thermotogae and the domain Archaea (Gao and Gupta, 2007; Gupta and Bhandari, 2011; Gupta and Shami, 2011). To illustrate the usefulness of these molecular markers for elucidation of prokaryotic evolutionary relationships, and to assess the influence of LGTs on the derived inferences, results for these two taxonomic groups are reviewed here.

MOLECULAR MARKERS FOR THE THERMOTOGAE

The species of the phylum Thermotogae are a group of hyperthermophilic, anaerobic, gram-negative bacteria recognized by a distinctive toga-like sheath structure and their ability to grow at high temperatures (Huber et al., 1986). The approximately 90 species of this phylum are currently divided into nine Genera within a single family termed the Thermotogaceae (Euzeby, 2011; NCBI Taxonomy, 2012). The Thermotogae species, prospectively, are important tools for industrial and biotechnological applications due to the ecological niche they inhabit and the thermo-stable proteins that they harbor (Conners et al., 2006). With the publication of the genome for *T. maritima*, the first species from this phylum (Nelson et al., 1999), the Thermotogae were brought to the forefront of LGT debate. This was due to the fact that based upon Blast searches it was determined that for about 25% of the genes from *T. maritima* genome, the closest blast hits were from archaeal species rather than any bacteria, leading to the inference that Thermotogae species have incurred high degree of LGTs with the archaeal organisms (Nelson et al., 1999). Upon revisiting this issue, Zhaxybayeva et al. (2009) found that for only about 11% of the Thermotogae proteins Archaea were the closest hits, but that the Thermotogae proteins exhibited maximal similarity (42–48% of genes) to the Firmicutes. Based upon these observations, the Thermotogae species genomes were proposed to be a chimera composed of different bacterial and archaeal sources (Zhaxybayeva et al., 2009). However, these estimates for LGTs have been questioned in other studies which indicate that much less (6–7%) of the Thermotogae genome has been laterally transferred (Garcia-Vallve et al., 2000; Ochman et al., 2000). Further, in view of the fact that Thermotogae species branch in proximity of the Firmicutes phylum (Gupta, 2001; Griffiths and Gupta, 2004b), the observation that a preponderance of the top hits for the Thermotogae species are from Firmicutes is an expected results, and it does not indicate that these genes have been laterally transferred (Zhaxybayeva et al., 2009; Andam and Gogarten, 2011).

Apart from their unique protein toga, the species of the phylum Thermotogae are assigned to this group and divided into its different genera primarily on the basis of their branching in the 16S rRNA trees (Reysenbach, 2001; Huber and Hannig, 2006; Zhaxybayeva et al., 2009; Yarza et al., 2010). Until recently, no unique molecular or biochemical characteristics were known that could distinguish the species of this phylum from other bacteria. For identification of molecular markers that could possibly define this phylum and its sub-taxa, a genome wide analysis was performed on protein sequences from 12 Thermotogae spp. whose genomes were available (Gupta and Bhandari, 2011). The protein sequences from these 12 species as well as species representing other bacteria phyla were aligned and examined for the presence of CSIs that were uniquely present in Thermotogae species or those that were commonly shared with some other bacteria. The analysis identified numerous CSIs specific for all Thermotogae. An example of a CSI consisting of a 3 aa long insert in the ribosomal protein L7 that is exclusively present in all sequenced Thermotogae species, including two recently sequenced species, is shown in **Figure 1A**. The unique presence of this CSI of the same length, at the same position in

Table 1 | Overview of the CSIs and CSPs that have been identified for some major prokaryotic taxa.

Taxonomic group	Number of CSPs/CSIs	References
Archaea	*Archaeal Kingdom specific*: 16 CSPs *Subgroups*: Thaumarchaeota—6 CSIs/201 CSPs, Euryarchaeota—6 CSPs, Thermoacidophiles—77 CSPs, Halophiles—127 CSPs, Methanogens—31 CSPs, Thermococcus-Pyrococcus clade—141 CSPs	Gao and Gupta, 2007; Gupta and Shami, 2011
Crenarchaeota	*Phylum specific*: 6 CSIs, 13 CSPs *Subgroups*: Sulfolobales—3 CSIs/151 CSPs, Thermoproteales—5 CSIs/25 CSPs, Desulfurococcales—4CSPs, Sulfolobales-Desulfurococcales clade—2 CSIs/18 CSPs	Gupta and Shami, 2011
Thaumarchaeota	>200 CSPs	Gupta and Shami, 2011
Thermotogae	*Phylum specific*: 18 CSIs *Subgroups*: Thermotoga genus—13 CSIs, Thermosipho genus—7 CSIs, Thermosipho-Fervidobacterium clade—13 CSIs, Thermotoga-Thermosipho-Fervidobacterium clade—5 CSIs, Petrotoga-Kosmotoga clade—4 CSIs	Gupta and Bhandari, 2011
Cyanobacteria	*Phylum specific*: 39 CSPs/10 CSIs *Subgroups*: Cyanobacterial Clade A—14 CSPs/1 CSI, Other Cyanobacteria (outside clade A)—5 CSPs/4 CSIs, Cyanobacterial Clade C—60 CSPs, Nostocales—65 CSPs, Chroococcales—8 CSPs, *Synechococcus*—14 CSPs, *Prochlorococcus*—19 CSPs, Low B/A type *Prochlorococcus*—67 CSPs	Gupta, 2009; Gupta and Mathews, 2010
Chlamydiae	*Phylum specific*: 59 CSPs/8 CSIs *Subgroups*: Chlamydiaceae—79 CSPs, Chlamydophila—20 CSPs, Chlamydia—20 CSPs	Gupta and Griffiths, 2006
Bacteroidetes, chlorobi and fibrobacteres	*Phylum specific*: 1 CSP/2 CSIs *Subgroup specific*: Bacteroidetes—27 CSPs/2 CSIs, Chlorobi—51 CSPs/2 CSIs, Bacteroidetes and Chlorobi clade—5 CSPs/3CSIs	Gupta, 2004
Actinobacteria	*Phylum specific*: 24 CSPs/4 CSIs *Subgroup specific*: CMN group—13 CSPs, *Mycobacterium* and *Nocardia*—14 CSIs, *Mycobacterium*—24 CSPs, *Micrococcineae*—24 CSPs, Corynebacteriales—4 CSPs/2 CSIs, Bifidobacteriales—14 CSPs/1 CSI	Gao and Gupta, 2005, 2012b; Gao et al., 2006
Deinococcus-thermus	*Phylum specific*: 65 CSPs/8 CSIs *Subgroup specific*: Deinococci—206 SPs	Griffiths and Gupta, 2004a, 2007a
Aquificae	*Phylum specific*: 10 CSPs/5 CSIs	Griffiths and Gupta, 2006b, 2004b
α-proteobacteria	*Class specific*: 6 CSPs/13 CSIs *Subgroups*: Rickettsiales—3 CSPs/2 CSIs, Rickettsiaceae—4 CSPs/5 CSIs, Anaplasmataceae—5 CSPs/2 CSIs, Rhodobacterales-Caulobacter-Rhizobiales clade—2 CSIs, Rhodobacterales-Caulobacter clade—1 CSI, Rhizobiales—6 CSPs/1CSI, Bradyrhizobiaceae—62 CSPs/2CSIs	Gupta and Mok, 2007
γ-proteobacteria	*Class specific*: 4 CSPs/1 CSI *Subgroups*: 20 CSPs, 2 CSIs for various subgroup combinations of subgroups	Gao et al., 2009
ε-proteobacteria	*Class specific*: 49 CSPs/4 CSIs *Subgroups*: *Wolinella-Helicobacter* clade—11 CSPs/2 CSIs, *Campylobacter* genus—18 CSPs/1 CSI	Gupta, 2006
Pasteurellales	*Order specific*: 44 CSIs *Subgroups*: Pasteurellales Clade I—13 CSIs, Pasteurellales Clade II—9 CSIs	Naushad and Gupta, 2012
Clostridia sensu stricto	*Genus specific*: 10 CSPs/3 CSIs	Gupta and Gao, 2009

The table provides general information regarding the number of CSIs and CSPs identified for many taxonomic groups on which genomic studies have been conducted. Further details can be obtained from the corresponding studies.

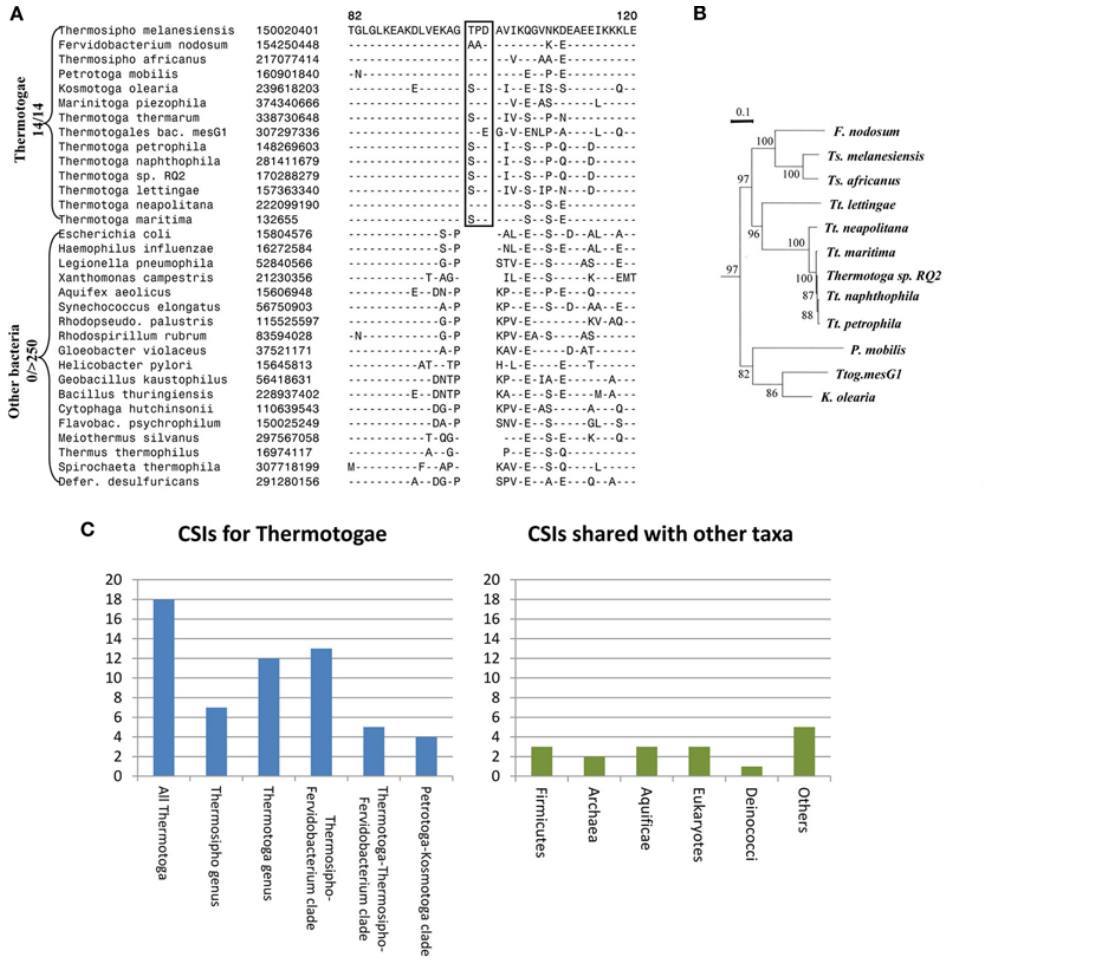


FIGURE 1 | Evolutionary relationships among Thermotogae species based upon CSIs and a Phylogenetic Tree. (A) Partial sequence alignment for the ribosomal protein L7 showing a 3 aa CSI (boxed) that is specific for all detected species of the Thermotogae phylum. The dashes in the alignment (–) indicate amino acid identity with the corresponding residue in the top line; **(B)** A maximum likelihood tree for the 12 sequenced Thermotogae species based upon concatenated sequences for 12 conserved proteins. **(C)** A summary diagram showing the species specificities of different CSIs identified for the Thermotogae group of species. The left panel highlights the CSIs that are specific for the entire Thermotogae phylum or its sub-groups, whereas the right panel indicates the CSIs that were also present in some non-Thermotogae organisms. **Figures 1A,B** modified from Gupta and Bhandari (2011).

this universally distributed protein, in different species from the phylum Thermotogae indicates that the genetic change leading to this CSI occurred once in the common ancestor of the Thermotogae species. In addition to this CSI, this study also identified 17 other CSIs in other important proteins such as DNA recombination protein RecA, DNA polymerase I and tryptophanyl-tRNA synthetase that are also specific for the species from the phylum Thermotogae (Gupta and Bhandari, 2011).

In addition to the large numbers of CSIs that were uniquely present in all Thermotogae species, this study also identified many CSIs that were specific for different sub-groups within the phylum Thermotogae (Gupta and Bhandari, 2011). These included 13 CSIs that were specific for the species of the genus *Thermotoga* and seven others that distinguished species of the genus *Thermosipho* from all others. However, it was observed that the species *Thermotoga lettingae* shared only 1 of 13 CSIs that were otherwise commonly present in other species of this genus. This suggests that *T. lettingae*, which is distantly related to all other *Thermotoga* species, should be assigned to a separate genus. Besides these CSIs that were specific for the species of these two genera, 13 CSIs supported a specific relationships among species of the *Fervidobacterium* and *Thermosipho* genera; 5 CSIs were shared by species from the genus *Thermotoga* and those from the *Fervidobacterium-Thermosipho* clade; and 4 CSIs supported a grouping of the *Petrotoga* and *Kosmotoga* genera along with the species *Thermotogales bacterium MesG1.Ag.4.2* (**Figure 1C**, left panel; Gupta and Bhandari, 2011). Importantly, all of the

relationships indicated by various CSIs were also independently observed in a phylogenetic tree for the Thermotogae species based upon concatenated sequences for 12 conserved proteins (**Figure 1B**).

The CSIs identified in the above study independently and strongly supported different nodes observed in the phylogenetic tree for Thermotogae species all the way from phylum to genus level. If the hypothesis that LGT events have abolished the ability to discern prokaryotic relationships was correct, then it should have been difficult to identify discrete molecular markers supporting distant relationships among these species. At the very least, the Thermotogae species would have shown relationships with species of other prokaryotic groups such as Firmicutes or Archaea as frequently as they did with one another. In this study, in addition to the CSIs that were specific for the Thermotogae species (**Figure 1C**, left panel), several CSIs were also identified that the Thermotogae shared with species from other prokaryotic or eukaryotic organisms (**Figure 1C**, right panel). However, such CSIs, suggesting possible LGT between Thermotogae and other taxa, were far outweighed by CSIs supporting the monophyletic, tree-like relationships among the species of the phylum (left panel) (Gupta and Bhandari, 2011). Assuming that all the CSIs that the Thermotogae shared with other groups are due to LGT, less than 20% (16 of 85) of all Thermotogae genes containing these CSIs have incurred LGTs (Gupta and Bhandari, 2011). Moreover, these presumed LGT events are of random nature and in no case do the Thermotogae species share more than a total of 3 CSIs with any particular phyla of species. Additionally, in most of these cases only a few species from these other taxa contained the indels that were present in most or all Thermotogae species (Gupta and Bhandari, 2011). Thus, these other CSIs, although they are present in a few isolated species from other taxa, are also largely specific for the Thermotogae species and they do not affect the ability of other CSIs to clearly discriminate Thermotogae species from all other bacteria or to deduce the evolutionary relationships amongst species from this phylum.

The shared presence of similar CSI in unrelated taxa can result from two different possibilities, either the gene with the CSI was laterally transferred among the two groups or that independent CSIs owing to two separate genetic events are responsible for these CSIs. After identification of such CSIs, tree-making approaches can be used to test if the presence of the indel in the two groups is due to LGT. Previously, in our work, a number of CSIs in the GlyA and MurA proteins that were commonly shared by the Chlamydiae and a subgroup of Actinobacteria were shown to be due to lateral transfer of genes from Actinobacteria to a common ancestor of the Chlamydiae (Griffiths and Gupta, 2006a). Recently, the shared presence of several CSIs in the bacteriochlorophyll biosynthesis proteins by unrelated phyla of photosynthetic prokaryotes has also been shown to be due to LGTs (Raymond et al., 2002; Gupta, 2012). However, in many other instances phylogenetic analyses have not supported LGT as the possible reason for the presence of a related CSI in unrelated taxa. In these cases, similar CSIs have originated independently in these lineages due to their presumed similar functions in these particular taxa.

MOLECULAR MARKERS FOR THE ARCHAEA AND ITS SUB-GROUPS

Archaea are widely recognized as the third domain of life. They generally inhabit extreme environments such as those of extreme temperature, pH or salinity, where little to no other life exists (Woese et al., 1990). However, recent studies indicate that archaeal species are widespread in the environment and they play a major role in the carbon and nitrogen cycles (Pace, 1997; Herndl et al., 2005; Leininger et al., 2006). Some archaeal species have been found to be commensal organisms residing in human colons (Oxley et al., 2010). The Archaea are generally divided into two main phyla, the Crenarchaeota and Euryarchaeota, based on 16S rRNA data and other phylogenetic data (Woese et al., 1990; Gribaldo and Brochier-Armanet, 2006). The Crenarchaeotes are described as thermophiles with sulfur-reducing capabilities while the Euryarchaeotes are metabolically and morphologically quite diverse (Gribaldo and Brochier-Armanet, 2006; Gupta and Shami, 2011). The mesophilic Crenarchaeota have been recently placed into a separate phylum called the Thaumarchaeota (Brochier-Armanet et al., 2008; Gupta and Shami, 2011).

Despite the importance of Archaea in different environments and in understanding of the evolutionary history of life on earth (Woese et al., 1990; Gupta, 2000a), until recently, very few molecular characteristics were known that are uniquely shared by all Archaea. Additionally, as the higher taxonomic groups within Archaea are described primarily based upon 16S rRNA trees, the characteristics that are unique to different phyla, classes, orders and families of the Archaea have scarcely been elucidated (Boone et al., 2001). The utilization of archaeal genomes for discovery of CSPs as well as CSIs has provided significant information in the form of molecular markers that are distinctive characteristics of Archaea and its taxonomic sub-groups. In 2007, a comprehensive analysis was performed on available archaeal genomes to search for CSPs that were unique to either all Archaea or many of its sub-groups (Gao and Gupta, 2007). Over 1400 such proteins distinctive of Archaea or its main taxa were discovered (**Figure 2**). In the analysis, sixteen proteins specific to all or most Archaea were identified that were not present in any bacterial or eukaryotic organism. Numerous proteins whose homologs were limited to the Crenarchaeota, Euryarchaeota and other sub-groups such as the Thermococci, Thermoplasmata, and Halobacteriales were also detected (**Figure 2**). Significantly, this study also identified 31 proteins that were commonly shared by all methanogenic bacteria (Gao and Gupta, 2007). In the 16S rRNA and other phylogenetic trees, the methanogenic Archaea do not form a monophyletic lineage, but instead are split into a number of distinct clusters separated by non-methanogenic Archaea (Burggraf et al., 1991; Brochier et al., 2004; Bapteste et al., 2005a; Gao and Gupta, 2007). Because most of the proteins that are commonly shared by various methanogens are generally involved in functions related to methanogenesis and their genes are clustered into a few large operons in genomes (Harms et al., 1995; Tersteegen and Hedderich, 1999; Grabarse et al., 2001; Gao and Gupta, 2007), it is likely that the genes for these proteins have been laterally acquired by different Archaea. This could provide a plausible explanation for the observed discrepancy in the branching of methanogenic Archaea in phylogenetic trees and

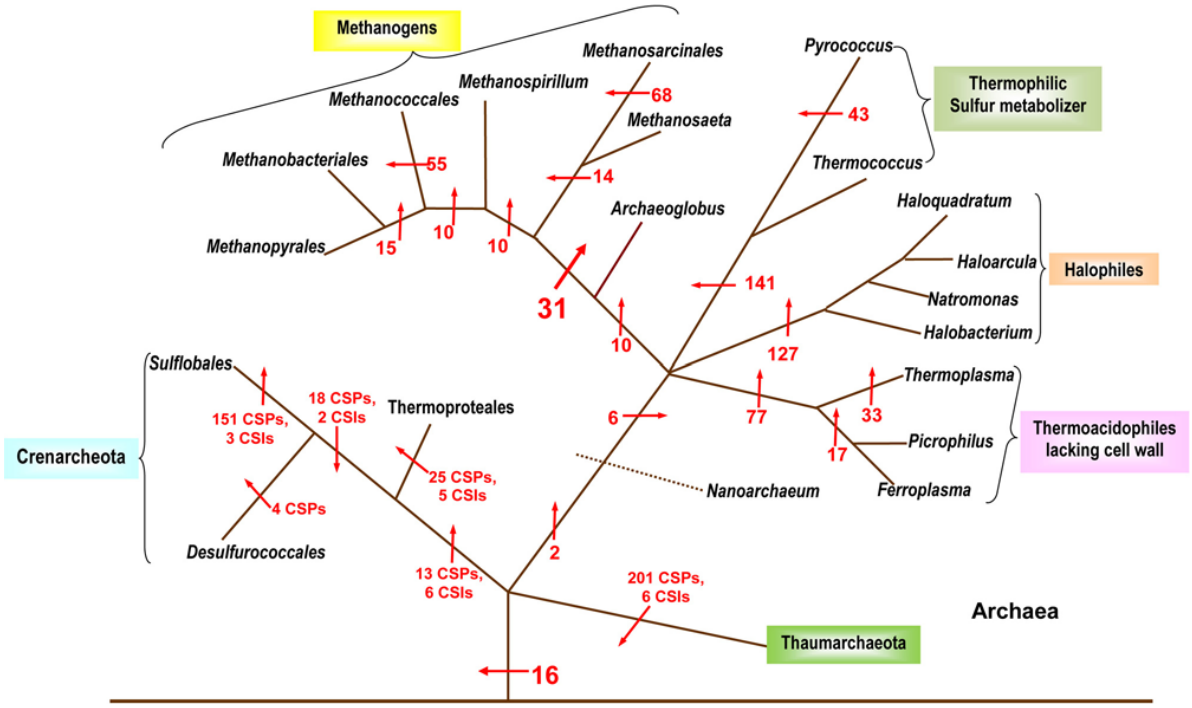


FIGURE 2 | A summary diagram showing the various molecular markers that have been identified for the Archaeal kingdom and its subgroups. The arrows indicate the suggested evolutionary stages where the proteins unique for a particular taxa are proposed to have been introduced. The numbers beside the arrows indicate the number of CSIs and CSPs specific for the various taxa (these numbers indicate CSPs unless otherwise noted). The branching pattern shown is based solely upon the distribution patterns of CSPs and CSIs. Modified from Gao and Gupta (2007) and Gupta and Shami (2011).

their unique sharing of genes for these proteins (Gao and Gupta, 2007).

A recent analysis has further added to the catalogue of molecular signatures for the archaeal organisms (Gupta and Shami, 2011). The focus of this study was on identifying CSIs and CSPs that were specific for the Crenarchaeota and Thaumarchaeota phyla (Gupta and Shami, 2011). Six CSIs and 13 CSPs specific for all species of the phylum Crenarchaeota were identified along with numerous markers for its different orders: the Sulfolobales (151 CSPs, 3 CSIs), Thermoproteales (25 CSPs, 5 CSIs) and the Desulfurococcales (4 CSPs). The study also described the markers (18 CSPs and 2 CSIs) indicative of a close relationship among the Sulfolobales and the Desulfurococcales. The discriminative ability of CSPs is highlighted by the results of blast searches on some CSPs that are specific for the Crenarchaeota or its main groups (Sulfolobales, Thermoproteales, Desulfurococcales and Acidilobales) that are shown in **Table 2**. In these cases, BLASTP searches were carried out on these proteins and the results for all species for whom the observed *E*-values were significant are shown. From the results presented in **Table 2**, it is evident that the first 2 CSPs are specific for the Crenarchaeota phylum, the next two are uniquely found in various species belonging to the orders Desulfurococcales, Acidilobales and Sulfolobales, whereas the last 5 CSPs are distinctive characteristics of species belonging to either the Desulfurococcales (and Acidilobales), the Sulfolobales, or the Thermoproteales orders.

In this study, more than 200 CSPs for various members of the newly defined Thaumarchaeota phylum were also identified (Gupta and Shami, 2011). The Thaumarchaeota are composed of several organisms previously included in the Crenarchaeota (Brochier-Armanet et al., 2008). The two phyla appear as sister groups in phylogenetic analysis and they also share 3 CSIs and 10 CSPs with each other (Gupta and Shami, 2011). Nevertheless, the two groups can be phylogenetically differentiated and numerous markers have been identified for each group that helps to define them molecularly as individual taxa (Gupta and Shami, 2011). A summary diagram depicting the various molecular markers specific for the archaeal species is shown in **Figure 2**. It should be noted that CSIs were only identified for the Thaumarchaeota and the Crenarchaeota and no detailed analysis to identify CSIs has thus far been carried out on the Euryarchaeota.

The two studies noted above have identified numerous CSIs and CSPs for the Archaea, its main phyla (Euryarchaeota, Crenarchaeota, Thaumarchaeota) and a number of its subphylum level taxa (Sulfolobales, Thermococcales, Halobacteriales, etc.; Gao and Gupta, 2007; Gupta and Shami, 2011). Except for the methanogens, the distribution patterns of the identified CSIs and CSPs are also strongly supported by the phylogenetic

Table 2 | A series of proteins specific for the Crenarchaeota and its sub-groups.

	Protein accession #	NP_147640	NP_147284	BAA81469	NP_147588	YP_001041009	YP_254810	YP_254922	NP_559041	NP_559897
	Protein length	262 aa	143 aa	98 aa	228 aa	127 aa	228 aa	270 aa	626 aa	113 aa
Desulfurococcales	*Aeropyrum pernix*	0.0	9e–98	5e–64	7e–161	7e–22	–	–	–	–
	Hyperthermus butylicus	3e–46	9e–43	1e–20	1e–23	3e–25	–	–	–	–
	Ignicoccus hospitalis	3e–41	–	5e–27	4e–19	3e–25	–	–	–	–
	Desulfurococcus kamchatkensis	7e–46	1e–21	2e–20	5e–17	7e–32	–	–	–	–
	Staphylothermus marinus	4e–56	1e–25	3e–21	3e–21	2e–85	–	–	–	–
Acidilobales	*Acidilobus saccharovorans*	9e–56	4e–36	4e–21	1e–46	1e–19	–	–	–	–
Sulfolobales	*Sulfolobus tokodaii*	4e–40	2e–29	3e–20	7e–26	–	1e–77	1e–80	–	–
	Sulfolobus islandicus	4e–42	6e–30	1e–25	1e–15	–	7e–50	8e–65	–	–
	Sulfolobus acidocaldarius	7e–34	3e–23	4e–22	4e–24	–	2e–162	0.0	–	–
	Sulfolobus solfataricus	1e–41	7e–30	5e–26	8e–15	–	5e–50	8e–64	–	–
	Metallosphaera sedula	3e–31	3e–33	3e–20	1e–22	–	4e–39	8e–60	–	–
Thermoproteales	*Pyrobaculum aerophilum*	9e–18	3e–11	–	–	–	–	–	0.0	2e–73
	Pyrobaculum islandicum	3e–18	3e–11	–	–	–	–	–	0.0	6e–54
	Pyrobaculum arsenaticum	1e–18	1e–10	–	–	–	–	–	0.0	2e–63
	Pyrobaculum calidifontis	6e–22	7e–11	–	–	–	–	–	0.0	1e–60
	Thermofilum pendens	1e–35	5e–30	–	–	–	–	–	1e–42	3e–10
	Caldivirga maquilingensis	1e–17	4e–8	–	–	–	–	–	1e–87	2e–22
	Thermoproteus neutrophilus	2e–19	7e–11	–	–	–	–	–	0.0	5e–61
	Thermoproteus tenax	3e–15	6e–10	–	–	–	–	–	0.0	4e–46
Top non-Crenarchaeota hit		*Brucella melitensis* (2e–1)	*Desulfobacterium autotrophicum* (8e–1)	*Aromatoleum aromaticum* (4e–1)	*Serpula lacrymans* (7e–1)	*Clonorchis sinensis* (3e–1)	*Granulicatella elegans* (6e–1)	*Encephalitozoon cuniculi* (7e–1)	*Burkholderia cenocepacia* (9e–1)	*Sordaria macrospora* (1e–1)

Blastp searches were carried out on proteins specific for the Crenarchaeota or its sub-groups and the results for representative species from different sub-groups of the Crenarchaeota are shown with the observed E-values. E-values greater than 1e–3 are considered insignificant hits with lack of homology to the query protein sequence.

The dashes (–) indicate that the homolog for the protein query was not detected in the BlastP searches.

Top non-Crenarchaeota hits indicate detection of species outside the Crenarchaeota that were observed to have the lowest E-value scores.

branching pattern of the archaeal organisms (Gribaldo and Brochier-Armanet, 2006; Gao and Gupta, 2007; Brochier-Armanet et al., 2008; Gupta and Shami, 2011). Considering the specificities of these molecular markers for either all Archaea or different clades of Archaea, these results strongly indicate that LGTs have not obliterated the phylogenetic signal necessary to delineate the evolutionary relationships among this domain of prokaryotes. The discovered CSIs and CSPs also provide novel tools for the identification of different groups of Archaea in various environments.

THE USEFULNESS OF THE CSIs FOR UNDERSTANDING BACTERIAL PHYLOGENY AND TAXONOMY

In addition to the CSIs that are specific for particular prokaryotic taxa, several of the identified CSIs have also proven useful in clarifying the branching order and interrelationships amongst different bacterial phyla (Gupta, 2001, 2011; Gupta and Griffiths, 2002). One example of these kinds of CSIs, which are referred to as the main-line signatures in our work, is shown in **Figure 3A**. In this case, a large ~100 aa insert in the β subunit of RNA polymerase protein (RpoB) is commonly

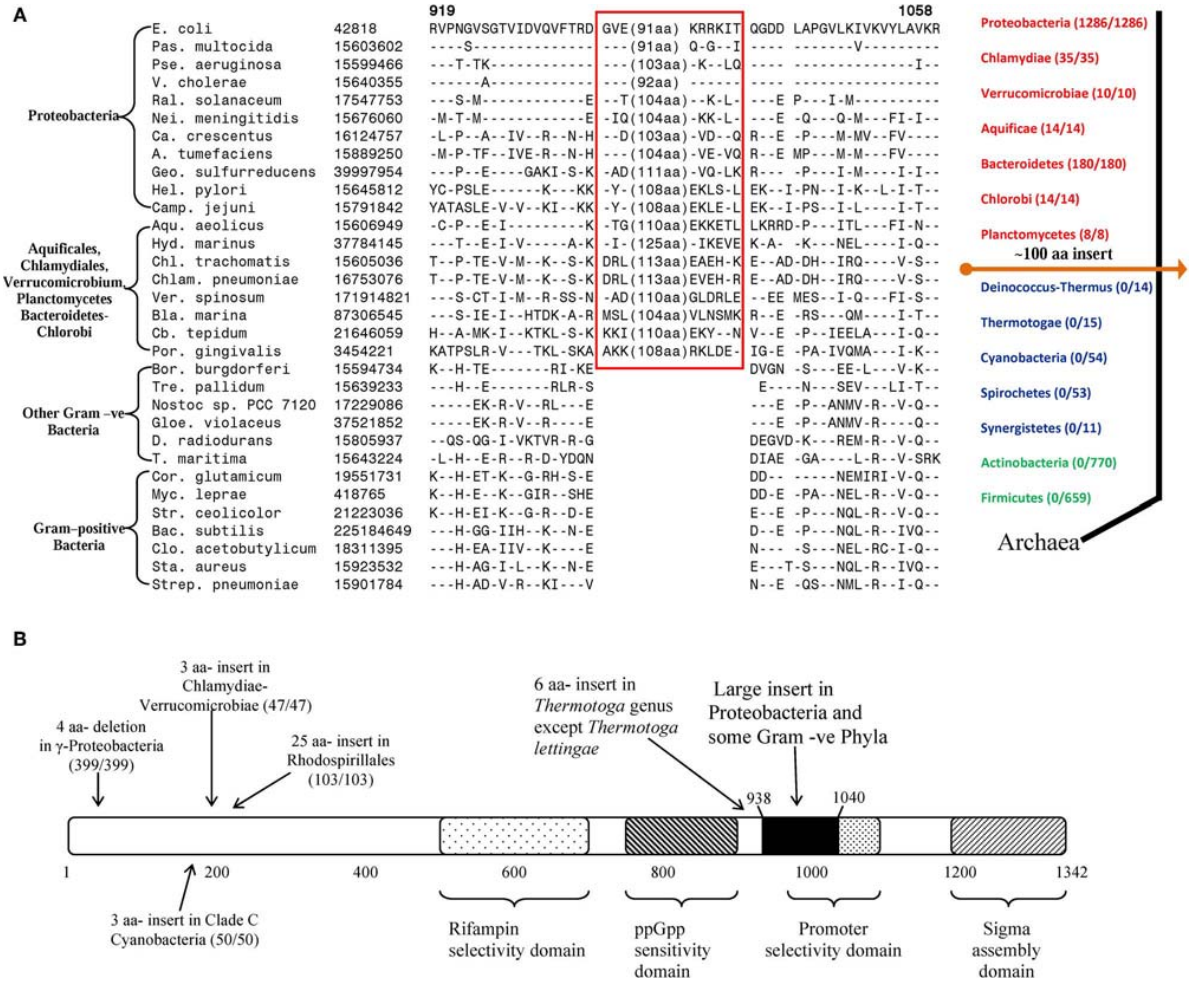


FIGURE 3 | Evolutionary significance of various identified CSIs in the RNA polymerase β subunit. (A) A portion of the RpoB sequence alignment showing a large insert (boxed) that is distinctive characteristic of all Proteobacteria and some Gram-negative phyla (Chlamydiae-Verrucomicrobiae, Aquificales, Planctomycetes, and Bacteroidetes-Chlorobi), but not found in other phyla of bacteria. Due to the large size of the insert, its entire sequence is not shown. Dashes (–) indicate identity with the amino acid on the top line. On the right is a linear representation of prokaryotic relationships based on the presence and absence of this CSI. The numbers in the brackets indicate the species of each phylum, which have been identified to contain the CSI. **(B)** A schematic representation of the sequence for *E. coli* RNA polymerase β subunit (RpoB) showing some functionally important regions and the positions of different lineage-specific inserts that have been identified within this protein. The large insert depicted in **(A)** (≈ 100 aa in *E. coli)* is shown in solid black. The positions of CSIs for different groups are roughly indicated using arrows. The values in the brackets identify the number of organisms in each respective group and the number of these species to harbour the indicated CSI. In all cases no organism outside of the indicated group was identified to contain the indel. The indicated CSIs have been described in earlier work (Griffiths and Gupta, 2004b, 2007b; Gupta and Mok, 2007; Gao et al., 2009; Gupta and Bhandari, 2011; Naushad and Gupta, 2012).

shared by all of the sequenced species belonging to the phyla Proteobacteria (different subclasses), Aquificae, Chlamydiae, Verrucomicrobiae, Bacteroidetes-Chlorobi, and Planctomycetes (Griffiths and Gupta, 2007b). This insert is present in all of the >1500 sequences that are available from species from these phyla. On the other hand, this CSI is not found in any of the >1500 sequences available from various species belonging to the phyla Firmicutes, Actinobacteria, Chloroflexi, Cyanobacteria, Deinococcus-Thermus, Synergistetes, Spirochaetes, etc. This insert is also not found in the archaeal RpoB homologs, thus providing evidence that this indel is an insert in the groups of species where it is found (Griffiths and Gupta, 2004b). Based upon its highly specific species distribution pattern, which argues strongly against the lateral transfer of this gene amongst various phyla, the genetic change responsible for this CSI most likely occurred in a common ancestor of the group of species that contain this CSI, after the divergence of other bacterial phyla that lack this indel as indicated in **Figure 3A** (right panel). A number of other mainline CSIs, which based upon their species distribution patterns have occurred at other important branch points in prokaryotic evolution, have been described in our earlier works (Griffiths and Gupta, 2001, 2004b; Gupta and Griffiths, 2002). Based upon these CSIs, it is possible to determine the branching order of most of the bacterial phyla (Gupta, 1998, 2001, 2003; Griffiths and Gupta, 2004b; see also www.bacterialphylogeny.info).

Within the highly conserved RpoB protein, in addition to the large CSI that is commonly shared by a number of bacterial phyla, several other CSIs have been identified that are specific for different groups/phyla of bacteria. The taxon specificities of these CSIs and their positions within in the RpoB polypeptide are shown in **Figure 3B**. These CSIs include a 4 aa deletion that is commonly and uniquely shared by a number of different orders of the γ-proteobacteria (399/399 species), a 3 aa insert that is specifically present in all of the Chlamydiae-Verrucomicrobiae species (47/47), another 3 aa insert that is a distinctive property of the Clade C cyanobacteria (50/50; Gupta, 2009), a 25 aa insert in various species from the order Rhodospirillales (103/103) and a 6 aa insert in all species from the genus *Thermotoga* except *T. lettingae* (Gupta and Griffiths, 2006; Gupta and Mok, 2007; Griffiths and Gupta, 2007b; Gao et al., 2009; Gupta and Bhandari, 2011). It is highly significant that within a single gene/protein multiple highly specific CSIs are present, each of which is specific for a different group of bacteria and help distinguish these groups from all other bacteria. These CSIs are not present in any species outside of the indicated taxa. The presence of these different taxa-specific characteristics in a single gene/protein strongly indicates that the genetic changes responsible for these CSIs occurred in the gene for this key protein at different stages in the evolution of bacterial domain and that no LGT of the gene for the RpoB protein has occurred among these taxa. Similar to the RpoB protein, multiple CSIs that are specific for different groups of prokaryotes have also been identified in many other important genes/proteins. These observations indicate that strong and consistent phylogenetic signals that are very likely not affected to any significant extent by the LGTs are still present in many conserved and universally distributed genes/proteins and these can be used to trace the evolutionary relationships among prokaryotes.

It is important to point out that virtually all of the higher taxonomic clades (above the Genus rank) within prokaryotes are currently identified solely on the basis of their branching in the 16S rRNA trees. Because the phylogenetic trees are a continuum, based upon them it has proven difficult to clearly define or delimit the boundaries of different taxonomic groups. Additionally, for virtually all of the higher prokaryotic taxa, no molecular, biochemical or physiological characteristics are known that are unique to them. Hence, a very important aspect of microbiology that needs to be understood is that in what respects do species from different main groups of bacteria differ from each other and what, if any, unique molecular, biochemical, structural or physiological characteristics are commonly shared by species from different groups? In this context, the large numbers of CSIs and CSPs for different taxonomic clades of bacteria that are being discovered by comparative genomic analyses provide novel and valuable tools for taxonomic, diagnostic, and biochemical studies (Gupta and Bhandari, 2011; Gao and Gupta, 2012b). In view of the specificities of the discovered CSIs and CSPs for different groups of prokaryotes and their retention by all species from these groups of prokaryotes, it is highly likely that these CSIs and CSPs are involved in functions that are essential for prokaryotes (Galperin and Koonin, 2004; Fang et al., 2005; Singh and Gupta, 2009; Schoeffler et al., 2010). Indeed, recent work on several CSIs have shown that they are essential for the group of organisms where they are found and the deletion or substantial changes in them led to failure of cell growth (Singh and Gupta, 2009; Schoeffler et al., 2010). Hence, further studies on understanding the cellular functions of the different taxa-specific CSIs and CSPs could lead to identification of novel biochemical and other functional characteristics that are specific for these groups of organisms.

It should also be noted that the identified CSIs and CSPs generally constitute robust molecular characteristics that exhibit high degree of predictive ability. Many of these CSIs and CSPs were discovered when the sequence information was available for very few prokaryotic species. However, despite the large increase in the number of sequenced genomes, most of these CSIs and CSPs are still specific for the originally indicated groups of prokaryotes (Gupta, 2009, 2011; Gao and Gupta, 2012b). Additionally, for several Chlamydiae-, Aquificae-, Deinococcus-Thermus- and Actinobacteria- specific degenerate primers based on conserved flanking sequences have been designed and they have been used to amplify the sequence regions predicted to contain the CSIs from large numbers of organisms for whom no sequences were available (Griffiths and Gupta, 2004a,b; Gao and Gupta, 2005; Griffiths et al., 2005). In these studies, in almost all cases the expected inserts or deletions were found to be present in previously un-sequenced organisms from the indicated groups, thus providing evidence that these CSIs and CSPs provide powerful new tools for identification of both known as well as novel species from different groups of prokaryotes.

CONCLUSIONS

There is considerable debate at present concerning the impact of LGTs on understanding prokaryotic phylogeny. While there

is little dispute that LGT plays an important role in microbial evolution, the extreme view taken by some that LGTs are so rampant within the prokaryotes that it totally masks the evolutionary signal from vertical transfer of genes (Doolittle, 2000; Gogarten et al., 2002; Doolittle and Bapteste, 2007; Dagan et al., 2008; Bapteste et al., 2009) is not supported by available evidence. As reviewed here, in phylogenetic trees based upon most gene/protein sequences all of the major groups within prokaryotes (from phylum down to genus level) are generally clearly identified, thus indicating that a strong phylogenetic signal emanating from vertical transfer of genes is maintained throughout prokaryotic evolution (Gupta, 1998, 2000b; Dutilh et al., 2004; Ludwig and Klenk, 2005; Ciccarelli et al., 2006; Puigbo et al., 2009). Most of the differences seen amongst these trees are either at the tips (i.e., species/strains levels) or at the base, i.e., relationships among the higher taxonomic clades such as phyla, class, etc. A recent study indicates that the incidence of LGTs shows linear correlation with the genome sequence and the GC content similarities of the donor and recipient organisms (Kloesges et al., 2011). Hence, while many of the observed inconsistencies between different gene trees at the species/strain levels could be due to LGTs (Puigbo et al., 2009; Kloesges et al., 2011), the differences in branching pattern at the higher taxonomic levels are perhaps in large parts due to loss of the phylogenetic signal and the lack of resolving power of the tree-based phylogenetic approaches (Gupta, 1998; Ludwig and Klenk, 2005; Puigbo et al., 2009).

In this review we have discussed the usefulness of CSIs and CSPs, as novel and important class of molecular markers for understanding the evolutionary relationships among prokaryotes. We have presented compelling evidence that based upon the species distribution patterns of these molecular signatures different prokaryotic taxa from phylum down to the genus levels can be clearly identified. Additionally, based upon these markers it is also possible to reliably deduct the evolutionary relationships amongst different prokaryotic taxa, both within a phylum and among different phyla. The evolutionary relationships deduced based upon these molecular markers generally exhibit high degree of congruency with those indicated by 16S rRNA trees or other gene/protein sequences. The analyses based upon these markers have also been able to clarify some relationships that are not resolved in phylogenetic trees. The species distribution patterns of these markers thus provide strong evidence that different clades of bacteria have evolved in a tree-like manner and that the prokaryotic organisms are not an exception to the Darwinian model of evolution. The relatively small numbers of these CSIs where the indel is also present in some unrelated species, which could be due to LGTs, show no specific pattern or relationship, thus they have minimal or no impact on the strong and consistent tree-like branching pattern that is evident from all other identified CSIs. However, it should be acknowledged that all of the work using CSIs and CSPs on understanding the evolutionary relationships among prokaryotes has thus far been carried out at genus level or higher taxa. Hence, it remains to be seen whether this approach will prove equally useful in clarifying the evolutionary relationships at the species or strain levels or not, where the evolutionary flux and the incidences of LGTs are deemed to be the highest (Daubin et al., 2003; Lerat et al., 2003; Dagan et al., 2008; Puigbo et al., 2009; Kloesges et al., 2011).

The molecular markers such as those described here in addition to their usefulness for understanding prokaryotic phylogeny also provide valuable means to address/clarify a number of important aspects of microbiology. (1) Based upon these markers different prokaryotic taxa can now be identified in clear molecular terms rather than only as phylogenetic entities. (2) Based upon them the boundaries of different taxonomic clades can also be more clearly defined. (3) Due to their high degree of specificity and predictive ability, they provide important diagnostic tools for identifying both known and unknown species belonging to these groups of bacteria. (4) The shared presence of these CSIs by unrelated groups of bacteria provides potential means for identifying novel cases of LGTs. (5) Functional studies on these molecular markers should help in the discovery of novel biochemical or physiological properties that are distinctive characteristics of different groups of prokaryotes.

Lastly, it should be acknowledged that the number of genes which harbor rare genetic changes such as these CSIs is generally small in comparison to the total number of genes that are present in any genome. However, the genes containing these CSIs are involved in different essential functions and they are often are amongst the most conserved proteins found in various organisms. Although, the criticism could be levied that the inferences based upon small numbers of genes/proteins containing these CSIs are not representative of the entire genomes (Dagan and Martin, 2006; Bapteste and Boucher, 2008), it should be emphasized that in a number of studies such as those discussed here, the reported CSIs or CSPs represent analyses of the entire genomes. Based upon these CSIs and/or CSPs, no other significant or consistent relationships or patterns among these organisms, other than those indicated here, can be derived from consideration of all of the gene/protein sequences in these genomes using these approaches. In this context it is also helpful to remember that molecular sequences like all other fossils change and disintegrate over long evolutionary periods of time and they lose their information content at different rates. Hence, a well-preserved fossil is generally considered to be far more informative than hundreds or even thousands of disintegrated fossils. Following this analogy, it is expected that not all genes/proteins will prove equally useful for understanding the evolutionary history of prokaryotes, which spans > 3.5 billion years. Thus, the best we can hope for is to find significant numbers of conserved genes/proteins, which contain consistent and reliable signals such as those described in the present work, whose inferences are generally consistent with all/most other available information.

ACKNOWLEDGMENTS

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

REFERENCES

Andam, C. P., and Gogarten, J. P. (2011). Biased gene transfer in microbial evolution. *Nat. Rev. Microbiol.* 9, 543–555.

Baldauf, S. L., and Palmer, J. D. (1993). Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proc. Natl. Acad. Sci. U.S.A.* 90, 11558–11562.

Bapteste, E., and Boucher, Y. (2008). Lateral gene transfer challenges principles of microbial systematics. *Trends Microbiol.* 16, 200–207.

Bapteste, E., Brochier, C., and Boucher, Y. (2005a). Higher-level classification of the Archaea: evolution of methanogenesis and methanogens. *Archaea* 1, 353–363.

Bapteste, E., Susko, E., Leigh, J., MacLeod, D., Charlebois, R. L., and Doolittle, W. F. (2005b). Do orthologous gene phylogenies really support tree-thinking? *BMC Evol. Biol.* 5, 33.

Bapteste, E., O'Malley, M. A., Beiko, R. G., Ereshefsky, M., Gogarten, J. P., Franklin-Hall, L., Lapointe, F. J., Dupre, J., Dagan, T., Boucher, Y., and Martin, W. (2009). Prokaryotic evolution and the tree of life are two different things. *Biol. Dir.* 4, 34.

Beiko, R. G. (2011). Telling the whole story in a 10,000-genome world. *Biol. Dir.* 6, 34.

Beiko, R. G., Harlow, T. J., and Ragan, M. A. (2005). Highways of gene sharing in prokaryotes. *Proc. Natl. Acad. Sci. U.S.A.* 102, 14332–14337.

Blair, C., and Murphy, R. W. (2011). Recent trends in molecular phylogenetic analysis: where to next? *J. Hered.* 102, 130–138.

Boone, D. R., Castenholz, R. W., and Garrity, G. M. (2001). *Bergey's Manual of Systematic Bacteriology*. New York, NY: Springer, 1–721.

Boto, L. (2010). Horizontal gene transfer in evolution: facts and challenges. *Proc. Biol. Sci.* 277, 819–827.

Brochier, C., Forterre, P., and Gribaldo, S. (2004). Archaeal phylogeny based on proteins of the transcription and translation machineries: tackling the Methanopyrus kandleri paradox. *Genome Biol.* 5, R17.

Brochier-Armanet, C., Boussau, B., Gribaldo, S., and Forterre, P. (2008). Mesophilic Crenarchaeota: proposal for a third Archaeal phylum, the Thaumarchaeota. *Nat. Rev. Microbiol.* 6, 245–252.

Buchanan, R. E., and Gibbons, N. E. (1974). *Bergey's Manual of Deteminative Bacteriology*. Baltimore, MD: Williams and Wilkins.

Burggraf, S., Stetter, K. O., Rouviere, P., and Woese, C. R. (1991). Methanopyrus kandleri: an Archaeal methanogen unrelated to all other known methanogens. *Syst. Appl. Microbiol.* 14, 346–351.

Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B., and Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science* 311, 1283–1287.

Conners, S. B., Mongodin, E. F., Johnson, M. R., Montero, C. I., Nelson, K. E., and Kelly, R. M. (2006). Microbial biochemistry, physiology, and biotechnology of hyperthermophilic Thermotoga species. *FEMS Microbiol. Rev.* 30, 872–905.

Cowan, S. T. (1965). Principles and practice of bacterial taxonomy—a forward look. *J. Gen. Microbiol.* 39, 143–153.

Cutino-Jimenez, A. M., Martins-Pinheiro, M., Lima, W. C., Martin-Tornet, A., Morales, O. G., and Menck, C. F. (2010). Evolutionary placement of Xanthomonadales based on conserved protein signature sequences. *Mol. Phylogenet. Evol.* 54, 524–534.

Dagan, T. (2011). Phylogenomic networks. *Trends Microbiol.* 19, 483–491.

Dagan, T., Artzy-Randrup, Y., and Martin, W. (2008). Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc. Natl. Acad. Sci. U.S.A.* 105, 10039–10044.

Dagan, T., and Martin, W. (2006). The tree of one percent. *Genome Biol.* 7, 118.

Dagan, T., and Martin, W. (2007). Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc. Natl. Acad. Sci. U.S.A.* 104, 870–875.

Darwin, C. (1859). *The Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray.

Daubin, V., Gouy, M., and Perriere, G. (2002). A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res.* 12, 1080–1090.

Daubin, V., and Ochman, H. (2004). Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. *Genome Res.* 14, 1036–1042.

Daubin, V., Moran, N. A., and Ochman, H. (2003). Phylogenetics and the cohesion of bacterial genomes. *Science* 301, 829–832.

Delsuc, F., Brinkmann, H., and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* 6, 361–375.

Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. *Science* 284, 2124–2129.

Doolittle, W. F. (2000). Uprooting the tree of life. *Sci. Am.* 282, 90–95.

Doolittle, W. F., and Bapteste, E. (2007). Pattern pluralism and the Tree of Life hypothesis. *Proc. Natl. Acad. Sci. U.S.A.* 104, 2043–2049.

Dutilh, B. E., Huynen, M. A., Bruno, W. J., and Snel, B. (2004). The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. *J. Mol. Evol.* 58, 527–539.

Dutilh, B. E., Snel, B., Ettema, T. J., and Huynen, M. A. (2008). Signature genes as a phylogenomic tool. *Mol. Biol. Evol.* 25, 1659–1667.

Euzeby, J. P. (2011). List of prokaryotic names with standing in nomenclature. http://www.bacterio.cict.fr/classifphyla.html. (Ref Type: Generic).

Fang, G., Rocha, E., and Danchin, A. (2005). How essential are nonessential genes? *Mol. Biol. Evol.* 22, 2147–2156.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., and Merrick, J. M. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269, 496–512.

Galperin, M. Y., and Koonin, E. V. (2004). 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Res.* 32, 5452–5463.

Galtier, N. (2007). A model of horizontal gene transfer and the bacterial phylogeny problem. *Syst. Biol.* 56, 633–642.

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

Gao, B., and Gupta, R. S. (2007). Phylogenomic analysis of proteins that are distinctive of Archaea and its main subgroups and the origin of methanogenesis. *BMC Genomics* 8, 86.

Gao, B., and Gupta, R. S. (2012a). Microbial systematics in the post-genomics era. *Antonie Van Leeuwenhoek* 101, 45–54.

Gao, B., and Gupta, R. S. (2012b). Phylogenetic framework and molecular signatures for the main clades of the phylum actinobacteria. *Microbiol. Mol. Biol. Rev.* 76, 66–112.

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

Gao, B., Paramanathan, R., and Gupta, R. S. (2006). Signature proteins that are distinctive characteristics of Actinobacteria and their subgroups. *Antonie Van Leeuwenhoek* 90, 69–91.

Garcia-Vallve, S., Romeu, A., and Palau, J. (2000). Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res.* 10, 1719–1725.

Gogarten, J. P., Doolittle, W. F., and Lawrence, J. G. (2002). Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.* 19, 2226–2238.

Grabarse, W., Mahlert, F., Duin, E. C., Goubeaud, M., Shima, S., Thauer, R. K., Lamzin, V., and Ermler, U. (2001). On the mechanism of biological methane formation: structural evidence for conformational changes in methyl-coenzyme M reductase upon substrate binding. *J. Mol. Biol.* 309, 315–330.

Gribaldo, S., and Brochier-Armanet, C. (2006). The origin and evolution of Archaea: a state of the art. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 361, 1007–1022.

Griffiths, E., and Gupta, R. S. (2001). The use of signature sequences in different proteins to determine the relative branching order of bacterial divisions: evidence that Fibrobacter diverged at a similar time to Chlamydia and the Cytophaga-Flavobacterium-Bacteroides division. *Microbiology* 147, 2611–2622.

Griffiths, E., and Gupta, R. S. (2004a). Distinctive protein signatures provide molecular markers and evidence for the monophyletic nature of the deinococcus-thermus phylum. *J. Bacteriol.* 186, 3097–3107.

Griffiths, E., and Gupta, R. S. (2004b). Signature sequences in diverse proteins provide evidence for the late divergence of the Order Aquificales. *Int. Microbiol.* 7, 41–52.

Griffiths, E., and Gupta, R. S. (2006a). Lateral transfers of serine hydroxymethyltransferase (glyA) and UDP-N-acetylglucosamine enolpyruvyl transferase (murA) genes from free-living Actinobacteria to the parasitic chlamydiae. *J. Mol. Evol.* 63, 283–296.

Griffiths, E., and Gupta, R. S. (2006b). Molecular signatures in protein sequences that are characteristics of the phylum Aquificae. *Int. J. Syst. Evol. Microbiol.* 56, 99–107.

Griffiths, E., and Gupta, R. S. (2007a). Identification of signature proteins that are distinctive of the

Deinococcus-Thermus phylum. *Int. Microbiol.* 10, 201–208.

Griffiths, E., and Gupta, R. S. (2007b). Phylogeny and shared conserved inserts in proteins provide evidence that Verrucomicrobia are the closest known free-living relatives of chlamydiae. *Microbiology* 153, 2648–2654.

Griffiths, E., Petrich, A. K., and Gupta, R. S. (2005). Conserved indels in essential proteins that are distinctive characteristics of Chlamydiales and provide novel means for their identification. *Microbiology* 151, 2647–2657.

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

Gupta, R. S. (2000a). The natural evolutionary relationships among prokaryotes. *Crit. Rev. Microbiol.* 26, 111–131.

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

Gupta, R. S. (2001). The branching order and phylogenetic placement of species from completed bacterial genomes, based on conserved indels found in various proteins. *Int. Microbiol.* 4, 187–202.

Gupta, R. S. (2003). Evolutionary relationships among photosynthetic bacteria. *Photosynth. Res.* 76, 173–183.

Gupta, R. S. (2004). The phylogeny and signature sequences characteristics of Fibrobacteres, Chlorobi, and Bacteroidetes. *Crit. Rev. Microbiol.* 30, 123–143.

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

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

Gupta, R. S. (2010a). "Applications of conserved indels for understanding microbial phylogeny," in *Molecular Phylogeny of Microorganisms*, eds A. Oren and R. T. Papke (Norfolk, UK: Caister Academic Press), 135–150.

Gupta, R. S. (2010b). Molecular signatures for the main phyla of photosynthetic bacteria and their subgroups. *Photosynth. Res.* 104, 357–372.

Gupta, R. S. (2011). Origin of diderm (Gram-negative) bacteria: antibiotic selection pressure rather than endosymbiosis likely led to the evolution of bacterial cells with two membranes. *Antonie Van Leeuwenhoek* 100, 171–182.

Gupta, R. S. (2012). Origin and spread of photosynthsis based upon conserved sequence Features in key bacteriochlorophyll biosynthesis proteins. *Mol. Biol. Evol.* PMID: 22628531. [Epub ahead of print].

Gupta, R. S., and Bhandari, V. (2011). Phylogeny and molecular signatures for the phylum Thermotogae and its subgroups. *Antonie Van Leeuwenhoek* 100, 1–34.

Gupta, R. S., and Gao, B. (2009). Phylogenomic analyses of clostridia and identification of novel protein signatures that are specific to the genus Clostridium sensu stricto (cluster I). *Int. J. Syst. Evol. Microbiol.* 59, 285–294.

Gupta, R. S., and Griffiths, E. (2002). Critical issues in bacterial phylogeny. *Theor. Popul. Biol.* 61, 423–434.

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

Gupta, R. S., and Mathews, D. W. (2010). Signature proteins for the major clades of Cyanobacteria. *BMC Evol. Biol.* 10, 24.

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

Gupta, R. S., and Shami, A. (2011). Molecular signatures for the Crenarchaeota and the Thaumarchaeota. *Antonie Van Leeuwenhoek* 99, 133–157.

Haggerty, L. S., Martin, F. J., Fitzpatrick, D. A., and McInerney, J. O. (2009). Gene and genome trees conflict at many levels. *Philos. Trans. R. Soc. Lond. B Biol Sci.* 364, 2209–2219.

Harms, U., Weiss, D. S., Gartner, P., Linder, D., and Thauer, R. K. (1995). The energy conserving N5-methyltetrahydromethanopterin: coenzyme M methyltransferase complex from Methanobacterium thermoautotrophicum is composed of eight different subunits. *Eur. J. Biochem.* 228, 640–648.

Herndl, G. J., Reinthaler, T., Teira, E., van Aken, H., Veth, C., Pernthaler, A., and Pernthaler, J. (2005). Contribution of Archaea to total prokaryotic production in the deep Atlantic Ocean. *Appl. Environ. Microbiol.* 71, 2303–2309.

Huber, R., and Hannig, M. (2006). "Thermotogales," in *The Prokaryotes*, eds M. Dworkin, S. Falkow, E. Rosenberg, K. H. Schleifer, and E. Stackebrandt (New York, NY: Springer), 899–922.

Huber, R., Langworthy, T. A., Konig, H., Thomm, M., Woese, C. R., Sleytr, U. B., and Stetter, K. O. (1986). *Thermotoga maritima* sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90 C*. *Arch. Microbiol.* 144, 324–333.

Iguchi, A., Thomson, N. R., Ogura, Y., Saunders, D., Ooka, T., Henderson, I. R., Harris, D., Asadulghani, M., Kurokawa, K., Dean, P., Kenny, B., Quail, M. A., Thurston, S., Dougan, G., Hayashi, T., Parkhill, J., and Frankel, G. (2009). Complete genome sequence and comparative genome analysis of enteropathogenic *Escherichia coli* O127, H6 strain E2348/69. *J. Bacteriol.* 191, 347–354.

Jain, R., Rivera, M. C., and Lake, J. A. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl. Acad. Sci. U.S.A.* 96, 3801–3806.

Jordan, G., and Goldman, N. (2012). The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol. Biol. Evol.* 29, 1125–1139.

Kloesges, T., Popa, O., Martin, W., and Dagan, T. (2011). Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. *Mol. Biol. Evol.* 28, 1057–1074.

Koonin, E. V. (2007). The Biological Big Bang model for the major transitions in evolution. *Biol. Dir.* 2, 21.

Koski, L. B., and Golding, G. B. (2001). The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.* 52, 540–542.

Koski, L. B., Morton, R. A., and Golding, G. B. (2001). Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol. Biol. Evol.* 18, 404–412.

Kunin, V., Goldovsky, L., Darzentas, N., and Ouzounis, C. A. (2005). The net of life: reconstructing the microbial phylogenetic network. *Genome Res.* 15, 954–959.

Kurland, C. G. (2005). What tangled web: barriers to rampant horizontal gene transfer. *Bioessays* 27, 741–747.

Kurland, C. G., Canback, B., and Berg, O. G. (2003). Horizontal gene transfer: a critical view. *Proc. Natl. Acad. Sci. U.S.A.* 100, 9658–9662.

Leininger, S., Urich, T., Schloter, M., Schwark, L., Qi, J., Nicol, G. W., Prosser, J. I., Schuster, S. C., and Schleper, C. (2006). Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature* 442, 806–809.

Lerat, E., Daubin, V., and Moran, N. A. (2003). From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. *PLoS Biol.* 1:E19. doi: 10.1371/journal.pbio.0000019

Lerat, E., Daubin, V., Ochman, H., and Moran, N. A. (2005). Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol.* 3:e130. doi: 10.1371/journal.pbio.0030130

Ludwig, W., and Klenk, H.-P. (2005). "Overview: a phylogenetic backbone and taxonomic framework for prokaryotic systematics," in *Bergey's Manual of Systematic Bacteriology*, eds D. J. Brenner, N. R. Krieg, J. T. Staley, and G. M. Garrity (Berlin: Springer-Verlag), 49–65.

Marri, P. R., and Golding, G. B. (2008). Gene amelioration demonstrated: the journey of nascent genes in bacteria. *Genome* 51, 164–168.

Martin, W. (1999). Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. *Bioessays* 21, 99–104.

Naushad, H. S., and Gupta, R. S. (2012). Molecular signatures (conserved indels) in protein sequences that are specific for the order Pasteurellales and distinguish two of its main clades. *Antonie Van Leeuwenhoek* 101, 105–124.

NCBI genomic database. (2012). http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi. (Ref Type: Electronic Citation).

NCBI Taxonomy. (2012). http://www.ncbi.nlm.nih.gov/taxonomy. (Ref Type: Electronic Citation).

Nelson, K. E., Clayton, R. A., Gill, S. R., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Nelson, W. C., Ketchum, K. A., McDonald, L., Utterback, T. R., Malek, J. A., Linher, K. D., Garrett, M. M., Stewart, A. M., Cotton, M. D., Pratt, M. S., Phillips, C. A., Richardson, D., Heidelberg, J., Sutton, G. G., Fleischmann, R. D., Eisen, J. A., White, O., Salzberg, S. L., Smith, H. O., Venter, J. C., and Fraser, C. M. (1999). Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* 399, 323–329.

Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* 405, 299–304.

Oxley, A. P., Lanfranconi, M. P., Wurdemann, D., Ott, S., Schreiber, S., McGenity, T. J., Timmis, K. N., and Nogales, B. (2010). Halophilic Archaea in the human intestinal mucosa. *Environ. Microbiol.* 12, 2398–2410.

Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science* 276, 734–740.

Pennisi, E. (1999). Is it time to uproot the tree of life? *Science* 284, 1305–1307.

Popa, O., Hazkani-Covo, E., Landan, G., Martin, W., and Dagan, T. (2011). Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res.* 21, 599–609.

Puigbo, P., Wolf, Y. I., and Koonin, E. V. (2009). Search for a 'Tree of Life' in the thicket of the phylogenetic forest. *J. Biol.* 8, 59.

Puigbo, P., Wolf, Y. I., and Koonin, E. V. (2010). The tree and net components of prokaryote evolution. *Genome Biol. Evol.* 2, 745–756.

Ragan, M. A. (2001). On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol. Lett.* 201, 187–191.

Ragan, M. A., and Beiko, R. G. (2009). Lateral genetic transfer: open issues. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 2241–2251.

Raymond, J., Zhaxybayeva, O., Gogarten, J. P., Gerdes, S. Y., and Blankenship, R. E. (2002). Whole-genome analysis of photosynthetic prokaryotes. *Science* 298, 1616–1620.

Reysenbach, A.-L. (2001). "Phylum BII. Thermotogae phy. nov," in *Bergey's Manual of Systematic Bacteriology* eds G. M. Garrity, D. R. Boone, and R. W. Castenholz (Berlin: Springer), 369–387.

Rivera, M. C., and Lake, J. A. (1992). Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science* 257, 74–76.

Roettger, M., Martin, W., and Dagan, T. (2009). A machine-learning approach reveals that alignment properties alone can accurately predict inference of lateral gene transfer from discordant phylogenies. *Mol. Biol. Evol.* 26, 1931–1939.

Rokas, A., and Holland, P. W. (2000). Rare genomic changes as a tool for phylogenetics. *Trends Ecol. Evol.* 15, 454–459.

Schoeffler, A. J., May, A. P., and Berger, J. M. (2010). A domain insertion in *Escherichia coli* GyrB adopts a novel fold that plays a critical role in gyrase function. *Nucleic Acids Res.* 38, 7830–7844.

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

Sorek, R., Zhu, Y., Creevey, C. J., Francino, M. P., Bork, P., and Rubin, E. M. (2007). Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* 318, 1449–1452.

Stanier, R. Y., Adelberg, E. A., and Ingraham, J. L. (1976). *The Microbial World*. Englewood Cliffs, NJ: Prentice-Hall Inc., 1–871.

Susko, E., Leigh, J., Doolittle, W. F., and Bapteste, E. (2006). Visualizing and assessing phylogenetic congruence of core gene sets: a case study of the gamma-proteobacteria. *Mol. Biol. Evol.* 23, 1019–1030.

Swithers, K. S., Gogarten, J. P., and Fournier, G. P. (2009). Trees in the web of life. *J. Biol.* 8, 54.

Tersteegen, A., and Hedderich, R. (1999). Methanobacterium thermoautotrophicum encodes two multisubunit membrane-bound [NiFe] hydrogenases. Transcription of the operons and sequence analysis of the deduced proteins. *Eur. J. Biochem.* 264, 930–943.

Treangen, T. J., and Rocha, E. P. (2011). Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet.* 7:e1001284. doi: 10.1371/journal.pgen.1001284

Wang, B. (2001). Limitations of compositional approach to identifying horizontally transferred genes. *J. Mol. Evol.* 53, 244–250.

Williams, D., Fournier, G. P., Lapierre, P., Swithers, K. S., Green, A. G., Andam, C. P., and Gogarten, J. P. (2011). A rooted net of life. *Biol. Dir.* 6, 45.

Woese, C. R. (1987). Bacterial evolution. *Microbiol. Rev.* 51, 221–271.

Woese, C. R., and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. U.S.A.* 74, 5088–5090.

Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U.S.A.* 87, 4576–4579.

Wolf, Y. I., Rogozin, I. B., Grishin, N. V., and Koonin, E. V. (2002). Genome trees and the Tree of Life. *Trends Genet.* 18, 472–479.

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N. N., Kunin, V., Goodwin, L., Wu, M., Tindall, B. J., Hooper, S. D., Pati, A., Lykidis, A., Spring, S., Anderson, I. J., D'Haeseleer, P., Zemla, A., Singer, M., Lapidus, A., Nolan, M., Copeland, A., Han, C., Chen, F., Cheng, J. F., Lucas, S., Kerfeld, C., Lang, E., Gronow, S., Chain, P., Bruce, D., Rubin, E. M., Kyrpides, N. C., Klenk, H. P., and Eisen, J. A. (2009). A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462, 1056–1060.

Yap, W. H., Zhang, Z., and Wang, Y. (1999). 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.* 181, 5201–5209.

Yarza, P., Ludwig, W., Euzeby, J., Amann, R., Schleifer, K. H., Glockner, F. O., and Rossello-Mora, R. (2010). Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analyses. *Syst. Appl. Microbiol.* 33, 291–299.

Zhaxybayeva, O., Nesbo, C. L., and Doolittle, W. F. (2007). Systematic overestimation of gene gain through false diagnosis of gene absence. *Genome Biol.* 8, 402.

Zhaxybayeva, O., Swithers, K. S., Lapierre, P., Fournier, G. P., Bickhart, D. M., DeBoy, R. T., Nelson, K. E., Nesbo, C. L., Doolittle, W. F., Gogarten, J. P., and Noll, K. M. (2009). On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales. *Proc. Natl. Acad. Sci. U.S.A.* 106, 5865–5870.

Zuckerkandl, E., and Pauling, L. (1965). Molecules as documents of evolutionary history. *J. Theor. Biol.* 8, 357–366.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Received: 02 April 2012; accepted: 27 June 2012; published online: 2**6** July 2012.*
*Citation: Bhandari V, Naushad HS and Gupta RS (2012) Protein based molecular markers provide reliable means to understand prokaryotic phylogeny and support Darwinian mode of evolution. Front. Cell. Inf. Microbio. **2**:98. doi: 10.3389/fcimb.2012.00098*

CHAPTER 8

Conclusions

Prokaryotic Systematics – Current methods and limitations

Prokaryotic systematics is a scientific discipline that utilizes the differences in kind, diversity and relationships among microbes to determine their classification and taxonomy (Zhi et al., 2012). The goal of prokaryotic taxonomy has been to provide a system for defining bacterial and archaeal relationships that mirrors their evolutionary history leading back to the origin of life (Kampfer, 2012; Oren and Garrity, 2014). Much has been written on what should be considered a phylogenetically meaningful taxonomic classification, yet, there are many questions in the field of taxonomy that need an answer. The criteria, currently in use, for defining prokaryotic groups are subjective and have many drawbacks (Oren and Garrity, 2014). The shortcomings of these methods are discussed in Chapter 1. Though there is a "Bacteriological Code of Nomenclature" that provides rules for the valid publication of new taxa at the family level and below (Lapage et al., 1992; Tindall et al., 2006). No such criteria is established for the taxonomic ranks above family (viz. phylum, class, and order) (Tindall et al., 2010; Kampfer, 2012; Oren and Garrity, 2014).

The development of new methods that can provide criteria, useful in the classification of higher taxonomic ranks has been a key goal in the field of systematics and taxonomy. Most of the taxonomic divisions among prokaryotes are based on phylogenetic trees which are largely dependent on methodology used (Zhi et al., 2012; Ramasamy et al., 2014). The trees are highly dynamic and are dependent on many parameters including the tree making algorithms, alignment quality, models of sequence evolutionary rates, the choice of strains to be included, and the choice of outgroups

species (Zhi et al., 2012; Thiergart et al., 2014; Gupta, 1998; Bhandari et al., 2012; Gao and Gupta, 2012; Gupta, 2014). Thus, the same dataset can often generate dramatically different phylogenetic trees. Additionally, no matter how accurate phylogenetic trees are, they still do not provide unique characteristics for discrimination among different prokaryotic taxa.

Hence, there is a need to develop more discrete and reliable criteria that can characterize prokaryotic taxa in more definitive terms. The new criteria devised for systematics studies should enable identification and circumscription of all the major taxa (at various taxonomic levels) in clear molecular and/or biochemical terms. These criteria should also prove helpful in understanding inter-relationships among different prokaryotic groups. One novel candidate for new systematic criteria are rare genetic changes, in the forms of conserved signature indels (CSIs) and conserved signature proteins (CSPs), which have been used successfully to provide a reliable discriminating criteria for taxonomic studies (Naushad et al., 2014a; Naushad et al., 2014b; Bhandari et al., 2012; Naushad and Gupta, 2013; Adeolu and Gupta, 2014; Sawana et al., 2014; Gupta and Griffiths, 2002; Gupta, 2010). I have used these molecular markers to study phylogenetic relationships among different orders of the class *Gammaproteobacteria.*

Discovery of CSIs and CSPs for the identification and classification of members of several major groups (Orders) within *Gammaproteobacteria*

The class *Gammaproteobacteria*, one of the largest groups of bacteria, encompasses many organisms that are medically, ecologically and scientifically important. The bacteria of this group are very diverse and many are associated with other major forms of life,

having prominent effects on humans, animals and plants. The phylogeny of this class has been difficult to resolve (Gao et al., 2009; Williams et al., 2010; Naushad and Gupta, 2013; Naushad and Gupta, 2012; Naushad et al., 2014a; Gupta, 2000; Ludwig and Klenk, 2005). The classification of different orders within this class is primarily established based on their branching in 16S rRNA gene trees (Brenner et al., 2005). The grouping patterns of species within this class have been further refined by subsequent analyses based on single and multiple gene/protein based phylogenies (Gao et al., 2009; Williams et al., 2010; Ramulu et al., 2014). However, except for branching in phylogenetic trees, no other molecular or biochemical characteristics are known that can distinguish different orders within this class from each other. The focus of my work has been on the identification of molecular markers for different orders of class *Gammaproteobacteria*, with particular emphasis on "Enterobacteriales", *Pasteurellales* and *Xanthomonadales*. Numerous molecular markers in the form of CSIs and CSPs have been identified that are unique to these groups (Naushad and Gupta, 2012; Naushad and Gupta, 2013; Naushad et al., 2014a). These identified signatures have provided important insights into evolutionary relationships of these groups, and have laid the foundation for their taxonomic reappraisal (Naushad et al., 2014b). Additionally, the reliability and predictability of these molecular markers have also been tested against newly sequenced bacterial genomes (Naushad et al., 2014b).

In the preceding Chapters, I have discussed the discovery and utilization of these molecular markers for different orders of the class *Gammaproteobacteria.* The order *Pasteurellales* was the first group for which such studies were performed. These studies,

detailed in Chapters 4 and 5, provided novel insights into the evolutionary relationships of the members of the *Pasteurellales* (Naushad and Gupta, 2012). The main inferences from these studies are briefly summarized as follows; we have identified a large number of CSIs unique to either all *Pasteurellales* or the two distinct clades within *Pasteurellales*, supporting a division of the members *Pasteurellales* into at least two distinct taxonomic units (i.e. two families), multiple CSIs were also identified that are unique characteristics of the "*sensu stricto*" members of the polyphyletic genera *Haemophilus*, *Actinobacillus* and *Pasteurella*. These CSIs should enable demarcation of the monophyletic clades of these genera in molecular terms and assignment of other members from these groups to other genera. This work has also identified several CSIs which are unique to the members of the pathogenic genera *Aggregatibacter* and *Mannheimia* which may serve as useful diagnostic targets.

The second order within the class *Gammaproteobacteria* that we have worked on is *Xanthomonadales*. Most of the members of this group are phytopathogens, which are responsible for diseases in more than 400 plant species. Detailed phylogenomic studies were performed to identify molecular signatures that are specific for all members of the *Xanthomonadales* or its subgroups. Utilizing the CSIs we also investigated the impact of LGT on *Xanthomonadales* genomes (Naushad and Gupta, 2013; Naushad et al., 2014b), and proposed a complete taxonomic revision of this group. These studies, detailed in Chapters 5 and 6, represent the most comprehensive phylogenomic analyses of the order *Xanthomonadales* completed to date. However, the first efforts to use molecular markers to understand the evolutionary relationships of the *Xanthomonadales* were carried out by

Dr. Menck and colleagues in 2010 (Cutino-Jimenez et al., 2010). The CSIs identified in their work were rechecked and updated in our subsequent analyses of the order *Xanthomonadales* (Naushad and Gupta, 2013; Naushad et al., 2014b).

The third group of *Gammaproteobacteria* that we have studied is comprised of several plant pathogenic organisms belonging to the order "Enterobacteriales". This order harbours vast majority of human, animal and plant pathogens (Brenner et al., 2005). We constructed a phylogenetic tree of all sequenced "Enterobacteriales" based upon multiple protein sequences. The tree showed branching of the "Enterobacteriales" species into many distinct clades. We chose to work on a small clade that is composed of *Dickeya, Pectobacterium* and *Brenneria*. The species belonging to these genera are phytopathogens, affecting many important food crops and ornamental plants. We have identified a large number of CSIs and CSPs that are uniquely shared by all members of these three genera which support the proposal that the clade consisting of *Dickeya, Pectobacterium* and *Brenneria* should eventually be recognized as a new family within the order "Enterobacteriales" (viz. "Pectobacteriaceae"). The details of this work and how CSIs and CSPs can be used as novel tools to understand microbial phylogeny and systematics are provided in Chapter 2.

Impact of LGT on Genome Evolution – What to believe?

Lateral gene transfer is a well-established mechanism for prokaryotic evolution (Nyvltova et al., 2015; Suwastika et al., 2014; Ragan et al., 2009; Treangen and Rocha, 2011; Williams et al., 2011). Microbes are known to share genes that provide selective advantages (Ying et al., 2015; Koch, 2014). Some genomic studies have indicated that

LGT events occur so frequently among prokaryotes that they mask parent-to-offspring (Darwinian) inheritance (Koch, 2014; Williams et al., 2011; Bapteste et al., 2009; Kloesges et al., 2011; Gogarten et al., 2002; Doolittle and Bapteste, 2007). However, most of these studies have been focused on traits that provide selective advantage to microbes. These traits, however, may not hold any significance for classification. As Darwin wrote:

> *"…adaptive characters, although of the utmost importance to the welfare of the being, are almost valueless to the systematist."*
>
> *(Darwin, 1859)*

It is of utmost importance to determine the traits that are evolutionarily informative and should be used for classification. The use of rare genetic changes, such as CSIs and CSPs, as evolutionarily informative characters is shown throughout the body of this thesis. In these Chapters we have shown that a large number of CSIs and CSPs have been identified for different groups of the class *Gammaproteobacteria* that show a consistent pattern of vertical inheritance. Chapter 2, in particular, provides arguments about the usefulness of these molecular markers as novel tools for microbial phylogeny and systematics (Naushad et al., 2014a). In Chapter 7 we have reviewed numerous studies performed in our lab in last decade in order to critically re-assess the prevalence of both group specific CSIs and laterally transferred CSIs and found the number of potentially laterally transferred CSIs to be very "minimal". In each of the studies reported in this thesis, the number of group specific CSIs has vastly outnumbered the potential cases of LGT. On the basis of these and other CSI based studies, we concluded that LGT, while

present, is not prevalent enough to obfuscate the tree like pattern of evolution among prokaryotic species.

Conclusions

Prokaryotic systematics is a growing field; techniques are constantly being developed which provide novel means of identifying microbes. It is estimated that most of the higher taxa in different environments will be discovered by the end of this decade (Yarza et al., 2014). One should expect this will lead to the generation of enormous amounts of genomic data, which is already accumulating at incredible pace. The availability of this immense data requires the development of new methods for the determination of prokaryotic taxonomy (Gao and Gupta, 2012; Gupta, 1998; Gupta and Griffiths, 2002). Much advancement has been made involving the utilization of genomic data in prokaryotic systematics (Ramasamy et al., 2014; Chun and Rainey, 2014; Zhi et al., 2012; Kampfer, 2012; Naushad et al., 2014a; Adeolu and Gupta, 2013).

In this thesis, we have utilized a method of discovering novel, evolutionarily significant molecular markers (CSIs and CSPs) from genomic data. Prokaryotic taxa, for the first time, are now being recognized based on these discrete characters (Gao and Gupta, 2007; Gao and Gupta, 2005; Griffiths and Gupta, 2007; Gupta and Bhandari, 2011; Adeolu and Gupta, 2014; Naushad et al., 2014b; Gupta and Griffiths, 2002). Additionally, these markers have been shown to have strong predictive value and will likely be found in other members of the specified taxa as more species are sequenced and novel species are isolated (Naushad et al., 2014b; Gao and Gupta, 2012; Howard-Azzeh

et al., 2014; Gupta and Lali, 2013). The utility of these molecular markers for identification and classification of different prokaryotic taxa have been discussed in the preceding Chapters. The identification of these molecular markers have proved helpful in understanding phylogenetic relationships and also in revising the taxonomy of many prokaryotic groups (Naushad et al., 2014b; Adeolu and Gupta, 2014; Gupta et al., 2014; Gupta and Lali, 2013). The impact of these markers in the field of prokaryotic taxonomy continues to grow and will likely form the foundation for the demarcation of many future prokaryotic groups.

CSIs and CSPs are also useful in the development of novel diagnostic tools. Two different assays based on CSIs identified in our lab have been developed for the identification of *Bacillus anthracis* and *E. coli* O157 (Wong et al., 2014; Ahmod et al., 2011). As we have identified a large number of CSIs and CSPs that are specific for human (*Pasteurellales*) and plant (*Xanthomonadales*, *Dickeya* and *Pectobacterium*) pathogens, similar strategies can be adopted to develop diagnostic assays for these groups of species. CSIs and CSPs also provide useful markers for metagenomic studies. This is particularly relevant as sequencing of the metabiome of patients is becoming increasingly routine in clinical practices (Milshteyn et al., 2014; Ridlon et al., 2014; Gillevet et al., 2010; Ghannoum et al., 2010; Bajaj et al., 2014; Naqvi et al., 2010a; Naqvi et al., 2010b). CSIs and CSPs identified for pathogenic organisms provide a great tool for the identification of infectious disease causing agents in metabiome samples.

Concluding Remarks

The availability of genomic data is enabling the identification of numerous CSIs and CSPs that are specific for different prokaryotic groups at different taxonomic levels. Most of the other approaches in prokaryotic systematics rely on identification of prokaryotes and deducing evolutionary relationships among them based on branching in phylogenetic trees. However, CSIs and CSPs, for the first time, provide quantum criteria for the identification of all prokaryotic taxonomic ranks in definitive molecular terms. CSIs and CSPs are proving helpful in resolving taxonomic issues, which were not resolved by other approaches (Adeolu and Gupta, 2014; Gupta et al., 2014; Adeolu and Gupta, 2013; Gupta et al., 2013; Sawana et al., 2014; Naushad et al., 2014a). The CSIs and CSPs also provide novel and valuable tools for diagnostic and metagenomic analyses. Future studies on these markers should prove helpful in understanding many fundamental aspects related to biology, evolution and adaptation of these microbes.

BIBLIOGRAPHY

Adeolu,M. and Gupta,R.S. (2013). Phylogenomics and molecular signatures for the order *Neisseriales*: proposal for division of the order *Neisseriales* into the emended family *Neisseriaceae* and *Chromobacteriaceae* fam. nov. Antonie Van Leeuwenhoek *104*, 1-24.

Adeolu,M. and Gupta,R.S. (2014). A phylogenomic and molecular marker based proposal for the division of the genus *Borrelia* into two genera: the emended genus *Borrelia* containing only the members of the relapsing fever *Borrelia*, and the genus *Borreliella* gen. nov. containing the members of the Lyme disease *Borrelia* (*Borrelia* burgdorferi sensu lato complex). Antonie Van Leeuwenhoek *105*, 1049-1072.

Ahmod,N.Z., Gupta,R.S., and Shah,H.N. (2011). Identification of a *Bacillus anthracis* specific indel in the yeaC gene and development of a rapid pyrosequencing assay for distinguishing B. *anthracis* from the *B. cereus* group. Journal of Microbiological Methods *87*, 278-285.

Akerborg,O., Sennblad,B., Arvestad,L., and Lagergren,J. (2009). Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. Proc. Natl. Acad. Sci. U. S. A *106*, 5714-5719.

Akiba,T., Koyama,K., Ishikl,Y., KIMURA,S., and FUKUSHIMA,T. (1960). On the mechanism of the development of multiple-drug-resistant clones of *Shigella*. Jpn. J Microbiol. *4*, 219-227.

Alperi,A., Martinez-Murcia,A.J., Ko,W.C., Monera,A., Saavedra,M.J., and Figueras,M.J. (2010). *Aeromonas taiwanensis* sp. nov. and *Aeromonas sanarellii* sp. nov., clinical species from Taiwan. Int J Syst. Evol. Microbiol 60, 2048-2055.

Amann,R.I., Ludwig,W., and Schleifer,K.H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiological Reviews *59*, 143-169.

Andam,C.P. and Gogarten,J.P. (2011). Biased gene transfer in microbial evolution. Nat. Rev. Microbiol *9*, 543-555.

Bajaj,J.S., Heuman,D.M., Hylemon,P.B., Sanyal,A.J., White,M.B., Monteith,P., Noble,N.A., Unser,A.B., Daita,K., Fisher,A.R., Sikaroodi,M., and Gillevet,P.M. (2014). Altered profile of human gut microbiome is associated with cirrhosis and its complications. J Hepatol. *60*, 940-947.

Bapteste,E., O'Malley,M.A., Beiko,R.G., Ereshefsky,M., Gogarten,J.P., Franklin-Hall,L., Lapointe,F.J., Dupre,J., Dagan,T., Boucher,Y., and Martin,W. (2009). Prokaryotic evolution and the tree of life are two different things. Biol. Direct. *4*, 34.

Beiko,R.G., Harlow,T.J., and Ragan,M.A. (2005). Highways of gene sharing in prokaryotes. Proc. Natl. Acad. Sci. U. S. A *102*, 14332-14337.

Bhandari,V. and Gupta,R.S. (2012). Molecular signatures for the phylum Synergistetes and some of its subclades. Antonie Van Leeuwenhoek.

Bhandari,V., Naushad,H.S., and Gupta,R.S. (2012). Protein based molecular markers provide reliable means to understand prokaryotic phylogeny and support Darwinian mode of evolution. Front Cell Infect. Microbiol. *2*, 98.

Boto,L. (2010). Horizontal gene transfer in evolution: facts and challenges. Proc. Biol. Sci. *277*, 819-827.

Boucher,Y., Douady,C.J., Papke,R.T., Walsh,D.A., Boudreau,M.E., Nesbo,C.L., Case,R.J., and Doolittle,W.F. (2003). Lateral gene transfer and the origins of prokaryotic groups. Annu. Rev. Genet. *37*, 283-328.

Boussau,B. and Daubin,V. (2010). Genomes as documents of evolutionary history. Trends Ecol. Evol. *25*, 224-232.

Brady,C., Hunter,G., Kirk,S., Arnold,D., and Denman,S. (2014). *Rahnella victoriana* sp. nov., *Rahnella bruchi* sp. nov., *Rahnella woolbedingensis* sp. nov., classification of *Rahnella* genomospecies 2 and 3 as *Rahnella variigena* sp. nov. and *Rahnella inusitata* sp. nov., respectively and emended description of the genus *Rahnella*. Syst. Appl. Microbiol. *37*, 545-552.

Brenner,D.J., Krieg,N.R., Staley,J.T., and Garrity,G.M. (2005). Bergey's Manual of Systematic Bacteriology "The *Gammaproteobacteria*". (New York: Springer).

Brochier,C., Philippe,H., and Moreira,D. (2000). The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. Trends Genet. *16*, 529-533.

Camelo-Castillo,A., Benitez-Paez,A., Belda-Ferre,P., Cabrera-Rubio,R., and Mira,A. (2014). Streptococcus dentisani sp. nov., a novel member of the mitis group. Int J Syst. Evol. Microbiol. *64*, 60-65.

Chun,J., Grim,C.J., Hasan,N.A., Lee,J.H., Choi,S.Y., Haley,B.J., Taviani,E., Jeon,Y.S., Kim,D.W., Lee,J.H., Brettin,T.S., Bruce,D.C., Challacombe,J.F., Detter,J.C., Han,C.S., Munk,A.C., Chertkov,O., Meincke,L., Saunders,E., Walters,R.A., Huq,A., Nair,G.B., and Colwell,R.R. (2009). Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. Proc. Natl. Acad. Sci. U. S. A *106*, 15442-15447.

Chun,J. and Rainey,F.A. (2014). Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. Int J Syst. Evol. Microbiol. *64*, 316-324.

Ciccarelli,F.D., Doerks,T., von Mering,C., Creevey,C.J., Snel,B., and Bork,P. (2006). Toward automatic reconstruction of a highly resolved tree of life. Science *311*, 1283-1287.

Cole,J.R., Wang,Q., Cardenas,E., Fish,J., Chai,B., Farris,R.J., Kulam-Syed-Mohideen,A.S., McGarrell,D.M., Marsh,T., Garrity,G.M., and Tiedje,J.M. (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. *37*, D141-D145.

Colman,D.R., Thomas,R., Maas,K.R., and Takacs-Vesbach,C.D. (2014). Detection and analysis of elusive members of a novel and diverse archaeal community within a thermal spring streamer consortium. Extremophiles.

Cowan,S.T. (1965). PRINCIPLES AND PRACTICE OF BACTERIAL TAXONOMY--A FORWARD LOOK. J Gen. Microbiol. *39*, 143-153.

Cutino-Jimenez,A.M., Martins-Pinheiro,M., Lima,W.C., Martin-Tornet,A., Morales,O.G., and Menck,C.F. (2010). Evolutionary placement of *Xanthomonadales* based on conserved protein signature sequences. Mol Phylogenet. Evol *54*, 524-534.

Dagan,T. and Martin,W. (2006). The tree of one percent. Genome Biol. *7*, 118.

Darwin,C. (1859). *The Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. (London: John Murray).

Daubin,V., Moran,N.A., and Ochman,H. (2003). Phylogenetics and the cohesion of bacterial genomes. Science *301*, 829-832.

Daubin,V. and Ochman,H. (2004a). Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli. Genome Res. *14*, 1036-1042.

Daubin,V. and Ochman,H. (2004b). Start-up entities in the origin of new genes. Curr. Opin. Genet. Dev. *14*, 616-619.

Davison,J. (1999). Genetic exchange between bacteria in the environment. Plasmid *42*, 73-91.

Deloger,M., El Karoui,M., and Petit,M.A. (2009). A genomic distance based on MUM indicates discontinuity between most bacterial species and genera. J Bacteriol. *191*, 91-99.

Doolittle,W.F. and Bapteste,E. (2007). Pattern pluralism and the Tree of Life hypothesis. Proc. Natl. Acad. Sci. U. S. A *104*, 2043-2049.

Drummond,A.J. and Rambaut,A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. BMC. Evol. Biol. *7*, 214.

Emerson,D., Rentz,J.A., Lilburn,T.G., Davis,R.E., Aldrich,H., Chan,C., and Moyer,C.L. (2007). A novel lineage of proteobacteria involved in formation of marine Fe-oxidizing microbial mat communities. PLoS. One. *2*, e667.

Flavell,R. (1972). Mitochondria and chloroplasts as descendants of prokaryotes. Biochem. Genet. *6*, 275-291.

Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M., and Cowan,S.T. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science *269*, 496-512.

Fournier,P.E. and Raoult,D. (2009). Current knowledge on phylogeny and taxonomy of *Rickettsia* spp. Ann. N. Y. Acad. Sci. *1166*, 1-11.

Fox,G.E., Stackebrandt,E., Hespell,R.B., Gibson,J., Maniloff,J., Dyer,T.A., Wolfe,R.S., Balch,W.E., Tanner,R.S., Magrum,L.J., Zablen,L.B., Blakemore,R., Gupta,R., Bonen,L., Lewis,B.J., Stahl,D.A., Luehrsen,K.R., Chen,K.N., and Woese,C.R. (1980). The phylogeny of prokaryotes. Science *209*, 457-463.

Fox,G.E., Wisotzkey,J.D., and Jurtshuk,P., Jr. (1992). How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. Int J Syst Bacteriol *42*, 166-170.

Freeman,V.J. (1951). Studies on the virulence of bacteriophage-infected strains of *Corynebacterium diphtheriae*. J Bacteriol. *61*, 675-688.

Gao,B. and Gupta,R.S. (2005). Conserved indels in protein sequences that are characteristic of the phylum Actinobacteria. Int J Syst Evol Microbiol *55*, 2401-2412.

Gao,B. and Gupta,R.S. (2007). Phylogenomic analysis of proteins that are distinctive of Archaea and its main subgroups and the origin of methanogenesis. BMC Genomics *8*, 86.

Gao,B. and Gupta,R.S. (2012). Microbial systematics in the post-genomics era. Antonie Van Leeuwenhoek *101*, 45-54.

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

Gao,B., Paramanathan,R., and Gupta,R.S. (2006). Signature proteins that are distinctive characteristics of Actinobacteria and their subgroups. Antonie Van Leeuwenhoek *90*, 69-91.

Gevers,D., Cohan,F.M., Lawrence,J.G., Spratt,B.G., Coenye,T., Feil,E.J., Stackebrandt,E., Van de,P.Y., Vandamme,P., Thompson,F.L., and Swings,J. (2005). Opinion: Re-evaluating prokaryotic species. Nat. Rev. Microbiol. *3*, 733-739.

Ghannoum,M.A., Jurevic,R.J., Mukherjee,P.K., Cui,F., Sikaroodi,M., Naqvi,A., and Gillevet,P.M. (2010). Characterization of the oral fungal microbiome (mycobiome) in healthy individuals. PLoS. Pathog. *6*, e1000713.

Gillevet,P., Sikaroodi,M., Keshavarzian,A., and Mutlu,E.A. (2010). Quantitative assessment of the human gut microbiome using multitag pyrosequencing. Chem. Biodivers. *7*, 1065-1075.

Gogarten,J.P., Doolittle,W.F., and Lawrence,J.G. (2002). Prokaryotic evolution in light of gene transfer. Mol. Biol. Evol. *19*, 2226-2238.

Gogarten,J.P. and Townsend,J.P. (2005). Horizontal gene transfer, genome innovation and evolution. Nat. Rev. Microbiol. *3*, 679-687.

Gomila,M., Prince-Manzano,C., Svensson-Stadler,L., Busquets,A., Erhard,M., Martinez,D.L., Lalucat,J., and Moore,E.R. (2014). Genotypic and phenotypic applications for the differentiation and species-level identification of *achromobacter* for clinical diagnoses. PLoS. One. *9*, e114356.

Goris,J., Konstantinidis,K.T., Klappenbach,J.A., Coenye,T., Vandamme,P., and Tiedje,J.M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst. Evol. Microbiol. *57*, 81-91.

Gray,M.W. (2012). Mitochondrial evolution. Cold Spring Harb. Perspect. Biol. *4*, a011403.

Greisen,K., Loeffelholz,M., Purohit,A., and Leong,D. (1994). PCR primers and probes for the 16S rRNA gene of most species of pathogenic bacteria, including bacteria found in cerebrospinal fluid. Journal of Clinical Microbiology *32*, 335-351.

Griffiths,E. and Gupta,R.S. (2006). Molecular signatures in protein sequences that are characteristics of the phylum Aquificae. Int J Syst Evol Microbiol *56*, 99-107.

Griffiths,E. and Gupta,R.S. (2007). Identification of signature proteins that are distinctive of the Deinococcus-Thermus phylum. Int Microbiol *10*, 201-208.

Gupta,R.S. (1998a). Life's third domain (Archaea): an established fact or an endangered paradigm? Theor Popul Biol *54*, 91-104.

Gupta,R.S. (1998b). Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. Microbiol. Mol. Biol. Rev. *62*, 1435-1491.

Gupta,R.S. (2000). The natural evolutionary relationships among prokaryotes. Crit Rev Microbiol *26*, 111-131.

Gupta,R.S. (2010). Applications of conserved indels for understanding microbial phylogeny. In Molecular phylogeny of microorganisms, A.Oren and R.T.Papke, eds. (Norfolk, U.K.: Caister Academic Press), pp. 135-150.

Gupta,R.S. and Bhandari,V. (2011). Phylogeny and molecular signatures for the phylum Thermotogae and its subgroups. Antonie Van Leeuwenhoek *100*, 1-34.

Gupta,R.S., Bhandari,V., and Naushad,H.S. (2012). Molecular Signatures for the PVC Clade (*Planctomycetes, Verrucomicrobia, Chlamydiae, and Lentisphaerae*) of Bacteria Provide Insights into Their Evolutionary Relationships. Front Microbiol. *3*, 327.

Gupta,R.S., Chen,W.J., Adeolu,M., and Chai,Y. (2013). Molecular signatures for the class Coriobacteriia and its different clades; proposal for division of the class *Coriobacteriia* into the emended order *Coriobacteriales*, containing the emended family *Coriobacteriaceae* and

Atopobiaceae fam. nov., and *Eggerthellales* ord. nov., containing the family *Eggerthellaceae* fam. nov. Int J Syst. Evol. Microbiol. *63*, 3379-3397.

Gupta,R.S. and Griffiths,E. (2002). Critical issues in bacterial phylogeny. Theor Popul Biol *61*, 423-434.

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

Gupta,R.S., Naushad,S., and Baker,S. (2014). Phylogenomic Analyses and Molecular Signatures for the Class *Halobacteria* and its Two Major Clades: A Proposal for Division of the Class *Halobacteria* into an emended order *Halobacteriales* and Two New Orders, *Haloferacales* ord. nov. and *Natrialbales* ord. nov. Int J Syst. Evol. Microbiol.

Gupta,R.S. (2014). Chapter 8 - Identification of Conserved Indels that are Useful for Classification and Evolutionary Studies. In Methods in Microbiology

New Approaches to Prokaryotic Systematics, I.S.a.J.C.Michael Goodfellow, ed. Academic Press), pp. 153-182.

Gupta,R. and Lali,R. (2013). Molecular signatures for the phylum Aquificae and its different clades: proposal for division of the phylum Aquificae into the emended order *Aquificales*, containing the families *Aquificaceae* and *Hydrogenothermaceae*, and a new order *Desulfurobacteriales* ord. nov., containing the family *Desulfurobacteriaceae*. Antonie Van Leeuwenhoek *104*, 349-368.

Haeckel E (1866). *Die Systematische Phylogenie*.

Haley,B.J., Grim,C.J., Hasan,N.A., Choi,S.Y., Chun,J., Brettin,T.S., Bruce,D.C., Challacombe,J.F., Detter,J.C., Han,C.S., Huq,A., and Colwell,R.R. (2010). Comparative genomic analysis reveals evidence of two novel *Vibrio* species closely related to *V. cholerae.* BMC. Microbiol. *10*, 154.

Hall,J.B. (1971). Evolution of the prokaryotes. J Theor. Biol. *30*, 429-454.

Handy,J. and Doolittle,R.F. (1999). An attempt to pinpoint the phylogenetic introduction of glutaminyl-tRNA synthetase among bacteria. J Mol. Evol. *49*, 709-715.

Harel,A., Bromberg,Y., Falkowski,P.G., and Bhattacharya,D. (2014). Evolutionary history of redox metal-binding domains across the tree of life. Proc. Natl. Acad. Sci. U. S. A *111*, 7042-7047.

Harris,J.K., Kelley,S.T., Spiegelman,G.B., and Pace,N.R. (2003). The genetic core of the universal ancestor. Genome Res. *13*, 407-412.

Hauptmann,A.L., Stibal,M., Baelum,J., Sicheritz-Ponten,T., Brunak,S., Bowman,J.S., Hansen,L.H., Jacobsen,C.S., and Blom,N. (2014). Bacterial diversity in snow on North Pole ice floes. Extremophiles. *18*, 945-951.

Helgerson,A.F., Sharma,V., Dow,A.M., Schroeder,R., Post,K., and Cornick,N.A. (2006). Edema disease caused by a clone of *Escherichia coli* O147. J Clin Microbiol *44*, 3074-3077.

Henz,S.R., Huson,D.H., Auch,A.F., Nieselt-Struwe,K., and Schuster,S.C. (2005). Whole-genome prokaryotic phylogeny. Bioinformatics. *21*, 2329-2335.

Hoffmann,M., Monday,S.R., Allard,M.W., Strain,E.A., Whittaker,P., Naum,M., McCarthy,P.J., Lopez,J.V., Fischer,M., and Brown,E.W. (2012). *Vibrio caribbeanicus* sp. nov., isolated from the marine sponge Scleritoderma cyanea. Int J Syst. Evol. Microbiol. *62*, 1736-1743.

Howard-Azzeh,M., Shamseer,L., Schellhorn,H., and Gupta,R. (2014). Phylogenetic analysis and molecular signatures defining a monophyletic clade of heterocystous cyanobacteria and identifying its closest relatives. Photosynth Res *122*, 171-185.

Jain,R., Rivera,M.C., and Lake,J.A. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. Proc. Natl. Acad. Sci. U. S. A *96*, 3801-3806.

Janda,J.M. and Abbott,S.L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin. Microbiol. *45*, 2761-2764.

Jaramillo,V.D., Sukno,S.A., and Thon,M.R. (2015). Identification of horizontally transferred genes in the genus Colletotrichum reveals a steady tempo of bacterial to fungal gene transfer. BMC. Genomics *16*, 2.

Kainth,P. and Gupta,R.S. (2005). Signature proteins that are distinctive of alpha proteobacteria. BMC Genomics *6*, 94.

Kampfer,P. (2012). Systematics of prokaryotes: the state of the art. Antonie Van Leeuwenhoek *101*, 3-11.

Karlin,S., Ladunga,I., and Blaisdell,B.E. (1994). Heterogeneity of genomes: measures and values. Proc. Natl. Acad. Sci. U. S. A *91*, 12837-12841.

Kasting,J.F. and Siefert,J.L. (2002). Life and the evolution of Earth's atmosphere. Science 296, 1066-1068.

Kim,M., Oh,H.S., Park,S.C., and Chun,J. (2014). Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int J Syst. Evol. Microbiol. *64*, 346-351.

Kitahara,K. and Miyazaki,K. (2013). Revisiting bacterial phylogeny: Natural and experimental evidence for horizontal gene transfer of 16S rRNA. Mob. Genet. Elements. *3*, e24210.

Kloesges,T., Popa,O., Martin,W., and Dagan,T. (2011). Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. Mol. Biol. Evol. *28*, 1057-1074.

Knoll,A.H. (1999). A new molecular window on early life. Science *285*, 1025-1026.

Koch,L. (2014). Microbial genetics: Horizontal gene transfer of antibacterial genes. Nat. Rev. Genet. *16*, 5.

Konstantinidis,K.T. and Tiedje,J.M. (2005). Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. U. S. A *102*, 2567-2572.

Koonin,E.V., Makarova,K.S., and Aravind,L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. Annu. Rev. Microbiol. *55*, 709-742.

Koonin,E.V., Puigbo,P., and Wolf,Y.I. (2011). Comparison of phylogenetic trees and search for a central trend in the "forest of life". J Comput. Biol. *18*, 917-924.

Kunin,V., Goldovsky,L., Darzentas,N., and Ouzounis,C.A. (2005). The net of life: reconstructing the microbial phylogenetic network. Genome Res. 15, 954-959.

Kurland,C.G. (2005). What tangled web: barriers to rampant horizontal gene transfer. Bioessays *27*, 741-747.

Kurland,C.G., Canback,B., and Berg,O.G. (2003). Horizontal gene transfer: a critical view. Proc. Natl. Acad. Sci. U. S. A *100*, 9658-9662.

Lapage, Sneath PHA, Lessel EF, Skerman VBD, Seeliger HPR, and Clark WA (1992). International Code of Nomenclature of Bacteria: Bacteriological Code, 1990 Revision. Americal Society of Microbiology.

Lee,S., Oh,J.H., Weon,H.Y., and Ahn,T.Y. (2012). *Flavobacterium cheonhonense* sp. nov., isolated from a freshwater reservoir. J Microbiol. *50*, 562-566.

Linnaeus C (1774). *Systema Naturae*. (Stockholm: Laurentius Salvius).

Loffler,F.E., Yan,J., Ritalahti,K.M., Adrian,L., Edwards,E.A., Konstantinidis,K.T., Muller,J.A., Fullerton,H., Zinder,S.H., and Spormann,A.M. (2013). *Dehalococcoides mccartyi* gen. nov., sp. nov., obligately organohalide-respiring anaerobic bacteria relevant to halogen cycling and bioremediation, belong to a novel bacterial class, *Dehalococcoidia* classis nov., order *Dehalococcoidales* ord. nov. and family *Dehalococcoidaceae* fam. nov., within the phylum *Chloroflexi*. Int J Syst. Evol. Microbiol. *63*, 625-635.

Ludwig,W. and Klenk,H.-P. (2005). Overview: A phylogenetic backbone and taxonomic framework for prokaryotic systamatics. In Bergey's manual of systematic bacteriology, D.J.Brenner, N.R.Krieg, J.T.Staley, and G.M.Garrity, eds. (Berlin: Springer-Verlag), pp. 49-65.

Maiden,M.C., Bygraves,J.A., Feil,E., Morelli,G., Russell,J.E., Urwin,R., Zhang,Q., Zhou,J., Zurth,K., Caugant,D.A., Feavers,I.M., Achtman,M., and Spratt,B.G. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. U. S. A *95*, 3140-3145.

Marchesi,J.R., Sato,T., Weightman,A.J., Martin,T.A., Fry,J.C., Hiom,S.J., and Wade,W.G. (1998). Design and Evaluation of Useful Bacterium-Specific PCR Primers That Amplify Genes Coding for Bacterial 16S rRNA. Applied and Environmental Microbiology *64*, 795-799.

Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J., Chen,Z., Dewell,S.B., Du,L., Fierro,J.M., Gomes,X.V., Godwin,B.C., He,W., Helgesen,S., Ho,C.H., Irzyk,G.P., Jando,S.C., Alenquer,M.L., Jarvie,T.P., Jirage,K.B., Kim,J.B., Knight,J.R., Lanza,J.R., Leamon,J.H., Lefkowitz,S.M., Lei,M., Li,J., Lohman,K.L., Lu,H., Makhijani,V.B., McDade,K.E., McKenna,M.P., Myers,E.W., Nickerson,E., Nobile,J.R., Plant,R., Puc,B.P., Ronan,M.T., Roth,G.T., Sarkis,G.J., Simons,J.F., Simpson,J.W., Srinivasan,M., Tartaro,K.R., Tomasz,A., Vogt,K.A., Volkmer,G.A., Wang,S.H., Wang,Y., Weiner,M.P., Yu,P., Begley,R.F., and Rothberg,J.M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376-380.

Marri,P.R., Bannantine,J.P., and Golding,G.B. (2006). Comparative genomics of metabolic pathways in *Mycobacterium* species: gene duplication, gene decay and lateral gene transfer. FEMS Microbiol. Rev. *30*, 906-925.

Mccarthy,B.J. and Bolton,E.T. (1963). An approach to the measurement of genetic relatedness among organisms. Proc. Natl. Acad. Sci. U. S. A *50*, 156-164.

Meier-Kolthoff,J.P., Auch,A.F., Klenk,H.P., and Goker,M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC. Bioinformatics. *14*, 60.

Migita,L.K. and Doi,R.H. (1970). Formylation of methionyl-transfer RNA from prokaryotes and eukaryotes by *Bacillus* subtilis transformylase. Arch. Biochem. Biophys. *138*, 457-463.

Milshteyn,A., Schneider,J.S., and Brady,S.F. (2014). Mining the Metabiome: Identifying Novel Natural Products from Microbial Communities. Chemistry & biology *21*, 1211-1223.

Naqvi,A., Rangwala,H., Keshavarzian,A., and Gillevet,P. (2010a). Network-based modeling of the human gut microbiome. Chem. Biodivers. *7*, 1040-1050.

Naqvi,A., Rangwala,H., Spear,G., and Gillevet,P. (2010b). Analysis of multitag pyrosequence data from human cervical lavage samples. Chem. Biodivers. *7*, 1076-1085.

Natochin,I., Felitsyn,S.B., Klimova,E.V., and Shakhmatova,E.I. (2012). [K+/Na+ in the animal extracellular fluid at weathering of granitoids and problem of the origin of life]. Zh. Evol. Biokhim. Fiziol. *48*, 409-416.

Naushad,H.S. and Gupta,R.S. (2012). Molecular signatures (conserved indels) in protein sequences that are specific for the order *Pasteurellales* and distinguish two of its main clades. Antonie Van Leeuwenhoek *101*, 105-124.

Naushad,H.S. and Gupta,R.S. (2013). Phylogenomics and molecular signatures for species from the plant pathogen-containing order *Xanthomonadales* . PLoS. One. *8*, e55216.

Naushad,H.S., Lee,B., and Gupta,R.S. (2014a). Conserved signature indels and signature proteins as novel tools for understanding microbial phylogeny and systematics: identification of molecular signatures that are specific for the phytopathogenic genera *Dickeya*, *Pectobacterium* and *Brenneria*. Int J Syst. Evol. Microbiol. *64*, 366-383.

Naushad,S., Adeolu,M., Wong,S., Sohail,M., Schellhorn,H.E., and Gupta,R.S. (2014b). A phylogenomic and molecular marker based taxonomic framework for the order

Xanthomonadales : proposal to transfer the families *Algiphilaceae* and *Solimonadaceae* to the order *Nevskiales* ord. nov. and to create a new family within the order *Xanthomonadales* , the family *Rhodanobacteraceae* fam. nov., containing the genus *Rhodanobacter* and its closest relatives. Antonie Van Leeuwenhoek.

Nelson,K.E., Clayton,R.A., Gill,S.R., Gwinn,M.L., Dodson,R.J., Haft,D.H., Hickey,E.K., Peterson,J.D., Nelson,W.C., Ketchum,K.A., McDonald,L., Utterback,T.R., Malek,J.A., Linher,K.D., Garrett,M.M., Stewart,A.M., Cotton,M.D., Pratt,M.S., Phillips,C.A., Richardson,D., Heidelberg,J., Sutton,G.G., Fleischmann,R.D., Eisen,J.A., White,O., Salzberg,S.L., Smith,H.O., Venter,J.C., and Fraser,C.M. (1999). Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature *399*, 323-329.

Nelson-Sathi,S., Sousa,F.L., Roettger,M., Lozada-Chavez,N., Thiergart,T., Janssen,A., Bryant,D., Landan,G., Schonheit,P., Siebers,B., McInerney,J.O., and Martin,W.F. (2015). Origins of major archaeal clades correspond to gene acquisitions from bacteria. Nature *517*, 77-80.

Nisbet,E.G. and Sleep,N.H. (2001). The habitat and nature of early life. Nature *409*, 1083-1091.

Nishioka,K., Migita,S., Takahashi,M., Kawakami,M., and Fujii,G. (1970). [Symposium: structure, production and function of immune globulin]. Nihon Saikingaku Zasshi *25*, 29-34.

Novoselov,A.A., Serrano,P., Pacheco,M.L., Chaffin,M.S., O'Malley-James,J.T., Moreno,S.C., and Ribeiro,F.B. (2013). From cytoplasm to environment: the inorganic ingredients for the origin of life. Astrobiology. *13*, 294-302.

Nyvltova,E., Stairs,C.W., Hrdy,I., Ridl,J., Mach,J., Paces,J., Roger,A.J., and Tachezy,J. (2015). Lateral gene transfer and gene duplication played a key role in the evolution of *Mastigamoeba balamuthi* hydrogenosomes. Mol. Biol. Evol.

Olsen,G.J., Woese,C.R., and Overbeek,R. (1994). The winds of (evolutionary) change: breathing new life into microbiology. J. Bacteriol. *176*, 1-6.

Oren,A. (2010). "Concepts about Phylogeny of Microorganisms - an Historical Overview". In Molecular Phylogeny of Microorganisms, Aharon Oren and R.Thane Papke, eds. (Norfolk: Caister Academic Press).

Oren,A. and Garrity,G.M. (2014). Then and now: a systematic review of the systematics of prokaryotes in the last 80 years. Antonie Van Leeuwenhoek *106*, 43-56.

Pace,N.R., Sapp,J., and Goldenfeld,N. (2012). Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life. Proc. Natl. Acad. Sci. U. S. A *109*, 1011-1018.

Pace,N.R. (1997). A Molecular View of Microbial Diversity and the Biosphere. Science *276*, 734-740.

Parte,A.C. (2014). LPSN - list of prokaryotic names with standing in nomenclature. Nucleic Acids Res *42*, D613-D616.

Patterson,M., Szollosi,G., Daubin,V., and Tannier,E. (2013). Lateral gene transfer, rearrangement, reconciliation. BMC. Bioinformatics. *14 Suppl 15*, S4.

Prado,S., Dubert,J., and Barja,J.L. (2014). Characterization of pathogenic *vibrios* isolated from bivalve hatcheries in Galicia, NW Atlantic coast of Spain. Description of *Vibrio tubiashii* subsp. *europaensis* subsp. nov. Syst. Appl. Microbiol.

Puigbo,P., Wolf,Y.I., and Koonin,E.V. (2009). Search for a 'Tree of Life' in the thicket of the phylogenetic forest. J Biol. *8*, 59.

Qin,Q.L., Xie,B.B., Zhang,X.Y., Chen,X.L., Zhou,B.C., Zhou,J., Oren,A., and Zhang,Y.Z. (2014). A proposed genus boundary for the prokaryotes based on genomic insights. J Bacteriol. *196*, 2210-2215.

Ragan,M.A., McInerney,J.O., and Lake,J.A. (2009). The network of life: genome beginnings and evolution. Introduction. Philos. Trans. R. Soc. Lond B Biol. Sci. *364*, 2169-2175.

Ramasamy,D., Mishra,A.K., Lagier,J.C., Padhmanabhan,R., Rossi,M., Sentausa,E., Raoult,D., and Fournier,P.E. (2014). A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. Int J Syst. Evol. Microbiol. *64*, 384-391.

Ramulu,H.G., Groussin,M., Talla,E., Planel,R., Daubin,V., and Brochier-Armanet,C. (2014). Ribosomal proteins: toward a next generation standard for prokaryotic systematics? Mol. Phylogenet. Evol. *75*, 103-117.

Raskin,D.M., Seshadri,R., Pukatzki,S.U., and Mekalanos,J.J. (2006). Bacterial genomics and pathogen evolution. Cell *124*, 703-714.

Richter,M. and Rossello-Mora,R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl. Acad. Sci. U. S. A *106*, 19126-19131.

Ridlon,J.M., Kang,D.J., Hylemon,P.B., and Bajaj,J.S. (2014). Bile acids and the gut microbiome. Curr. Opin. Gastroenterol. *30*, 332-338.

Rivera,M.C. and Lake,J.A. (1992). Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science *257*, 74-76.

Ronquist,F. and Huelsenbeck,J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. *19*, 1572-1574.

Sagan,L. (1967). On the origin of mitosing cells. J Theor. Biol. *14*, 255-274.

Sapp,J. (2005). The prokaryote-eukaryote dichotomy: meanings and mythology. Microbiol Mol. Biol. Rev. *69*, 292-305.

Sapp,J. (2009). The New Foundations of Evolution: On the Tree of Life. (London: Oxford University Press).

Sawana,A., Adeolu,M., and Gupta,R.S. (2014). Molecular signatures and phylogenomic analysis of the genus *Burkholderia*: proposal for division of this genus into the emended genus *Burkholderia* containing pathogenic organisms and a new genus *Paraburkholderia* gen. nov. harboring environmental species. Front Genet. *5*, 429.

Schleifer,K.H. (2009). Classification of Bacteria and Archaea: past, present and future. Syst. Appl. Microbiol. *32*, 533-542.

Schopf,J.W. (1978). The evolution of the earliest cells. Sci Am *239*, 110-120.

Sjostrand,J., Tofigh,A., Daubin,V., Arvestad,L., Sennblad,B., and Lagergren,J. (2014). A Bayesian method for analyzing lateral gene transfer. Syst. Biol. *63*, 409-420.

Sorokin,D.Y., Berben,T., Melton,E.D., Overmars,L., Vavourakis,C.D., and Muyzer,G. (2014). Microbial diversity and biogeochemical cycling in soda lakes. Extremophiles. *18*, 791-809.

Stackebrandt E., Murray R.G.E., and Truper H.G. (1988). Proteobacteria classis nov., a name for the phylogenetic taxon that includes the "purple bacteria and their relatives. Int J SystBacteriol *38*, 321-325.

Stackebrandt,E. and Ebers,J. (2006). Taxonomic parameters revisited: tarnished gold standards. Microbiol Today *33*, 152-155.

Stackebrandt,E., Frederiksen,W., Garrity,G.M., Grimont,P.A., Kampfer,P., Maiden,M.C., Nesme,X., Rossello-Mora,R., Swings,J., Truper,H.G., Vauterin,L., Ward,A.C., and Whitman,W.B. (2002). Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst. Evol. Microbiol. *52*, 1043-1047.

Stackebrandt,E. and Goebel,B.M. (1994). Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. International Journal of Systematic Bacteriology *44*, 846-849.

Stackebrandt,E., Pauker,O., Steiner,U., Schumann,P., Straubler,B., Heibei,S., and Lang,E. (2007). Taxonomic characterization of members of the genus Corallococcus: molecular divergence versus phenotypic coherency. Syst. Appl. Microbiol. *30*, 109-118.

Stanier,R.Y. and Van Niel,C.B. (1941). The Main Outlines of Bacterial Classification. J Bacteriol. *42*, 437-466.

Stanier,R.Y. and Van Niel,C.B. (1962). The concept of a bacterium. Arch. Mikrobiol. *42*, 17-35.

Staudt,L.M. (2003). Molecular diagnosis of the hematologic cancers. N. Engl. J Med. *348*, 1777-1785.

Suwastika,I.N., Denawa,M., Yomogihara,S., Im,C.H., Bang,W.Y., Ohniwa,R.L., Bahk,J.D., Takeyasu,K., and Shiina,T. (2014). Evidence for lateral gene transfer (LGT) in the evolution of eubacteria-derived small GTPases in plant organelles. Front Plant Sci. *5*, 678.

Swithers,K.S., Gogarten,J.P., and Fournier,G.P. (2009). Trees in the web of life. J Biol. *8*, 54.

Tatusova,T., Ciufo,S., Fedorov,B., OGÇÖNeill,K., and Tolstoy,I. (2014). RefSeq microbial genomes database: new representation and annotation strategy. Nucleic Acids Res *42*, D553-D559.

Thiergart,T., Landan,G., and Martin,W.F. (2014). Concatenated alignments and the case of the disappearing tree. BMC. Evol. Biol. *14*, 2624.

Thiergart,T., Landan,G., Schenk,M., Dagan,T., and Martin,W.F. (2012). An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin. Genome Biol. Evol. *4*, 466-485.

Thomas,C.M. and Nielsen,K.M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat. Rev. Microbiol. *3*, 711-721.

Tindall,B.J., Kampfer,P., Euzeby,J.P., and Oren,A. (2006). Valid publication of names of prokaryotes according to the rules of nomenclature: past history and current practice. Int J Syst. Evol. Microbiol. *56*, 2715-2720.

Tindall,B.J., Rossello-Mora,R., Busse,H.J., Ludwig,W., and Kampfer,P. (2010). Notes on the characterization of prokaryote strains for taxonomic purposes. Int J Syst. Evol. Microbiol. *60*, 249-266.

Treangen,T.J. and Rocha,E.P. (2011). Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS. Genet. *7*, e1001284.

Uzzell,T. and Spolsky,C. (1974). Mitochondria and plastids as endosymbionts: a revival of special creation? Am. Sci. *62*, 334-343.

van Dijk,E.L., Auger,H., Jaszczyszyn,Y., and Thermes,C. (2014). Ten years of next-generation sequencing technology. Trends Genet. *30*, 418-426.

Wayne,L.G., Brenner,D.J., Colwell,R.R., Grimont,P.A.D., Kandler,O., Krichevsky,M.I., Moore,L.H., Moore,W.E.C., Murray,R.G.E., Stackebrandt,E., Starr,M.P., and Truper,H.G. (1987). Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. International Journal of Systematic Bacteriology *37*, 463-464.

Whittaker,R.H. (1969). New concepts of kingdoms or organisms. Evolutionary relations are better represented by new classifications than by the traditional two kingdoms. Science *163*, 150-160.

Williams,D., Fournier,G.P., Lapierre,P., Swithers,K.S., Green,A.G., Andam,C.P., and Gogarten,J.P. (2011). A rooted net of life. Biol. Direct. *6*, 45.

Williams,K.P., Gillespie,J.J., Sobral,B.W., Nordberg,E.K., Snyder,E.E., Shallom,J.M., and Dickerman,A.W. (2010). Phylogeny of *Gammaproteobacteria*. J Bacteriol.

Woese,C.R. (1987). Bacterial evolution. Microbiol Rev *51*, 221-271.

Woese,C.R. (1998). The universal ancestor (vol 95, pg 6854, 1998). Proceedings of the National Academy of Sciences of the United States of America *95*, 9710.

Woese,C.R. and Fox,G.E. (1977). Phylogenetic Structure of Prokaryotic Domain - Primary Kingdoms. Proceedings of the National Academy of Sciences of the United States of America *74*, 5088-5090.

Woese,C.R., Kandler,O., and Wheelis,M.L. (1990). Towards A Natural System of Organisms - Proposal for the Domains Archaea, Bacteria, and Eucarya. Proceedings of the National Academy of Sciences of the United States of America *87*, 4576-4579.

Woese,C.R., Magrum,L.J., and Fox,G.E. (1978). Archaebacteria. Journal of Molecular Evolution *11*, 245-252.

Woese,C.R., Stackebrandt,E., and Ludwig,W. (1985a). What Are Mycoplasmas - the Relationship of Tempo and Mode in Bacterial Evolution. Journal of Molecular Evolution *21*, 305-316.

Woese,C.R., Weisburg,W.G., Hahn,C.M., Paster,B.J., Zablen,L.B., Lewis,B.J., Macke,T.J., Ludwig,W., and Stackebrandt,E. (1985b). The Phylogeny of Purple Bacteria - the Gamma-Subdivision. Systematic and Applied Microbiology *6*, 25-33.

Wong,S.Y., Paschos,A., Gupta,R.S., and Schellhorn,H.E. (2014). Insertion/Deletion-Based Approach for the Detection of *Escherichia coli* O157:H7 in Freshwater Environments. Environ. Sci. Technol. *48*, 11462-11470.

Xiong,J. (2006). Photosynthesis: what color was its origin? Genome Biol *7*, 245.

Yap,W.H., Zhang,Z., and Wang,Y. (1999). 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. *181*, 5201-5209.

Yarza,P., Yilmaz,P., Pruesse,E., Glockner,F.O., Ludwig,W., Schleifer,K.H., Whitman,W.B., Euzeby,J., Amann,R., and Rossello-Mora,R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat. Rev. Microbiol. *12*, 635-645.

Ying,J., Wang,H., Bao,B., Zhang,Y., Zhang,J., Zhang,C., Li,A., Lu,J., Li,P., Ying,J., Liu,Q., Xu,T., Yi,H., Li,J., Zhou,L., Zhou,T., Xu,Z., Ni,L., and Bao,Q. (2015). Molecular Variation and Horizontal Gene Transfer of the Homocysteine Methyltransferase Gene mmuM and its Distribution in Clinical Pathogens. Int J Biol. Sci. *11*, 11-21.

Zhang,Y.J., Tian,H.F., and Wen,J.F. (2009). The evolution of YidC/Oxa/Alb3 family in the three domains of life: a phylogenomic analysis. BMC. Evol. Biol. *9*, 137.

Zhaxybayeva,O., Gogarten,J.P., Charlebois,R.L., Doolittle,W.F., and Papke,R.T. (2006). Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. Genome Res. *16*, 1099-1108.

Zhi,X.Y., Zhao,W., Li,W.J., and Zhao,G.P. (2012). Prokaryotic systematics in the genomics era. Antonie Van Leeuwenhoek *101*, 21-34.

Zhou,X., Hou,X.X., Geng,Z., Zhao,R., Wan,K.L., and Hao,Q. (2014). Establishment of Multiple Locus Variable-number Tandem Repeat Analysis Assay for Genotyping of *Borrelia burgdorferi* sensu lato Detected in China. Biomed. Environ. Sci. *27*, 665-675.

Zotin,A.I., Ozerniuk,N.D., Zotin,A.A., and Konoplev,V.A. (1975). [Possible route for the origin of prokaryotes]. Zh. Obshch. Biol. *36*, 163-172.