APPLICATIONS OF CONSERVED INDELS FOR MICROBIAL PHYLOGENY

COMPARATIVE ANALYSES OF MICROBIAL GENOMES TO IDENTIFY MOLECULAR MARKERS FOR DIFFERENT GROUPS OF PROKARYOTES

By

VAIBHAV BHANDARI, B.Sc.

A Thesis
Submitted to the School of Graduate Studies
In Partial Fulfillment of the Requirements for the Degree
Master of Science

McMaster University

MASTER OF SCIENCE (2013)
(Biochemistry)

McMaster University
Hamilton, Ontario, Canada

TITLE: Comparative Analyses of Microbial Genomes to Identify Molecular Markers for Different groups of Prokaryotes

AUTHOR: Vaibhav Bhandari, B.Sc. (McMaster University)

SUPERVISOR: Dr. Radhey S. Gupta

NUMBER OF PAGES: [xviii], [134]

ABSTRACT

Currently centered on molecular data, bacterial and archaeal relationships are often based on their relative branching in 16S rRNA based phylogenetic trees. The availability of numerous bacterial genome sequences over the past two decades has provided new information for insights previously inaccessible to the field of taxonomy. Through utilization of comparative genomics, numerous molecular markers in the form of insertions and deletions within conserved regions of proteins, also known as Conserved Signature Indels or CSIs, have been discovered for various prokaryotic taxa. Using these techniques, we have analyzed relationships among the bacterial phyla of Thermotogae and Synergistetes and the conglomeration of bacterial organisms known as the PVC super-phylum. Through identification of large numbers of CSIs we have described the phyla Thermotogae and Synergistetes, and their sub-groups, in molecular terms for the first time. The identified molecular markers support a reconstruction of the current taxonomic divisions of these phyla. Similarly, previously only observed to group in phylogenetic trees, we have identified molecular markers for the PVC clade of bacterial phyla which are indicative of their shared ancestry. Further, in response to recent suggestions of extensive lateral gene transfer masking evolutionary relationships, an argument in favour of Darwinian mode of evolution for prokaryotic organisms is made using the identified molecular markers identified here along with markers previously identified in similar studies. Due to their taxonomic specificity, the markers that we have discovered provide useful tools for biochemical tests aiming for an understanding of the unique characteristics of the bacterial groups to which they are specific.

ACKNOWLEDGEMENTS

For his advice and guidance throughout the course of this work, I would like to express my immense gratitude to Dr. Radhey S. Gupta, my supervisor. It was an enjoyable experience getting to know him and working as a student under his tutelage. His mentorship has provided me with a great experience and invaluable insight into the scientific field. Without him, this work would not have been possible.

My appreciation is further extended to Dr. Murray Junop and Dr. Herb E. Schellhorn, members of my supervisory committee, for their insight and assistance. I must also acknowledge students at McMaster University who have provided me with new knowledge and viewpoints, enriching my experience as a graduate student; to fellow graduate student Sohail Naushad for his patient assistance as a sounding board throughout the process. Finally, I would like to thank my family for allowing me the freedom to pursue my passion.

TABLE OF CONTENTS

LIST OF FIGURES

Page Number

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

LIST OF TABLES

Page Number

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

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
BLAST	Basic Local Alignment Search Tool
Blastp	Standard protein BLAST
CheD	Glutamine deaminase chemoreceptor
CSI	Conserved Signature Indel
CSP	Conserved Signature Protein
CysE	Serine O-acetyltransferase
DAK phosphatase	Dihydroxyacetone kinase phosphatase
DGC	Diguanylate cyclase
DnaA	Chromosomal replication initiation protein DnaA
DnaK	Chaperone protein DnaK (also called Hsp 70)
DXR	1-deoxy-Dxylulose-5-phosphate reductoisomerase
E-value	Expect value
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
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, β subunit

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
IclR	Transcriptional regulator IclR
IleRS	Isoleucine-tRNA synthetase
Indel	Insert or Deletion
KorA	2-Oxoglutarate synthase
LGT	Lateral gene transfer
Mb	Mega base pair
MinD	Septum site-determining protein MinD
ML	Maximum Likelihood
MLSA	Multilocus Sequence Analysis
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
NJ	Neighbor Joining
NrdR	Transcriptional regulator NrdR
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
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
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 α-subunit
RpoB	RNA polymerase β-subunit
RpoC	RNA polymerase β’-subunit
RppK	Ribose phosphate pyrophosphokinase
SecA	Protein translocase subunit SecA
SMC	Chromosome segregation protein SMC
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

PREFACE

This following work is a sandwich thesis. Presented in chapters 2, 3, 4 and 5 are the unaltered manuscripts, published in the years 2011 and 2012, illustrating comparative genomic analysis for use in prokaryotic systematics. 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 6 reflects on the presented data. References for chapters 1 and 6 are provided at the end of this thesis. All chapters have been reproduced with consent of all co-authors.

Chapter 1

Introduction

An Introduction to Prokaryotic Systematics - A Representation of the Natural Order

Representing the earliest forms of life, prokaryotic species have been present on the planet for approximately 3.5 billion years (Forterre and Gribaldo, 2007; Schopf, 2006). Throughout this time, prokaryotic species have to a great extent influenced the history of the planet. Bacterial organisms capable of photosynthesis have been implicated in permitting the rise of organisms dependant on aerobic respiration (Hall, 1971; Horvath, 1974; Nisbet and Sleep, 2001); further, their symbiotic relationships allowed for the emergence of plants and animals, forms of complex multicellular life (Gray, 2012; Uzzell and Spolsky, 1974). They continue to play an important role in all aspects of life including in the carbon and nitrogen cycles as well as in the life cycles of plants and animals as symbionts and pathogens (Azam, 1998; Jetten, 2008; McCarren et al., 2010). Prokaryotes harbour many roles in life and are present in many extremes of life-permitting situations (Pikuta et al., 2007).

Recent evaluations suggest that ~12,000 species of bacteria and archaea have been identified, many of which remain to be cultured in a lab or accurately characterized (Federhen, 2012; Sayers et al., 2011). Additionally, it is estimated that a larger percentage of the prokaryotic diversity remains unknown, with estimates projecting presence of over 10 million species (Curtis et al., 2002; Schloss and Handelsman, 2004). With only a few thousand characterized species, and an ever-growing list of newly discovered organisms, the classification of prokaryotes into a comprehensible, structured system reflecting their evolutionary history is amongst the most challenging problems in microbiology. Apart from its utility in record-keeping, taxonomy provides a consistent system for communication among scientists concerning different species. Additionally, consistent with the goal of these attempts, ideal taxonomical divisions allow for a quick description of the common characters shared by related groups of organisms and the elucidation of evolutionary relationships.

The earliest attempts in the classification of microbes were by Otto Frederik Müller and Christian Gottfried Ehrenberg in the 18th and 19th centuries, respectively (Oren, 2010a). With limited means to observe microorganisms and few known microorganisms, species were differentiated by their cell shapes into a small number of groups termed genera. Previously considered animals, in 1866 Ernst Haeckel famously differentiated heterotrophic bacteria (placing them into the order Monera) from protists, due to the absence of a nucleus (Oren, 2010a; Sapp, 2005). Ferdinand Cohn is attributed with providing a more systematic taxonomic approach for classification with morphology defining a species placement into higher groups (distant relationships) and physiological characters used for differentiation of closely related organisms (Oren, 2010a). For much of the 20th century phenotypic cellular characteristics were the dominant determinants for species classification. Advancements in biochemical analyses brought greater depth and perception to the field. Much effort was put into improving bacterial phylogeny with debates on whether morphological or physiological criteria were to take precedence in depicting prokaryotic relationships. Cytological data led to the acceptance of the first of the major phylogenetic divisions of life forms recognized today with the distinction made between Eukaryotes and Prokaryotes (Stanier and Niel, 1962).

With progression in laboratory techniques, chemotaxonomic criteria were developed for differentiation of prokaryotic groups (Rossello-Mora and Amann, 2001). However, due to the wide variety of prokaryotes, their small size, and sharing of characteristics through convergent evolution, morphological and biochemical criteria often fall short of accurately establishing their relationships (Woese, 1987). These difficulties in classifying bacteria based on simply physical criteria were acknowledged by the 1970s (Oren, 2010a; Stanier, 1970; Woese, 1987).

DNA-DNA Hybridization and 16S rRNA Sequence – A Move towards Molecular Data

Looking for other means for identifying phylogeny, along with the knowledge that DNA was the genetic information storage molecule, gene comparison methods were sought after (Zuckerkandl and Pauling, 1965). Beginning the era of DNA sequence based phylogenetics, the DNA-DNA hybridizations method was established allowing for classification of prokaryotes lacking well-defined phenotypic characters (Brenner et al., 1969; Gevers et al., 2005; Richter and Rossello-Mora, 2009; Xu, 2010). Due to the fact that the hybridization data encompasses the entire genome of an organism, the method has sometimes been considered the gold-standard for confirmation of a species' taxonomic classification (Amann et al., 1992; Richter and Rossello-Mora, 2009; Stackebrandt and Goebel, 1994). However, the usage of this method comes with its own issues. DNA-DNA hybridization is time-consuming and expensive (Gevers et al., 2005). The hybridization analysis is only useful in differentiating among species and strains, relationships among distantly groups cannot be accurately ascertained through this methodology. Additionally, the method is variable as different experiments can produce slight differences in hybridization data (Grimont et al., 1980; Goris et al., 2007). Like morphologically and physiologically derived data, the method can only be utilized for species which can be cultured. As estimates suggest that 99% of prokaryotes cannot be cultured, much of the prokaryotic biodiversity cannot be analyzed through this method (Amann et al., 1995; Gevers et al., 2005). Further, due to the comparative nature of the method, an incremental database cannot be built (Gevers et al., 2005).

A major breakthrough in the field of evolutionary sciences was the development of gene sequencing technology and the use of 16S rRNA as a tool in identification of species relationships (Woese and Fox, 1977; Woese, 1987). 16S rRNA sequences are valued for being universally present and highly conserved among species of bacteria and archaea (Woese, 1987). For insights into their relationships, prokaryotes can be subjected to 16S rRNA sequence based phylogenetic trees or direct sequence comparisons. Among the early results of the use of this method was the introduction of the currently accepted three-domain system of classification for cellular life forms with the division of the prokaryotic species into bacteria and archaea (Woese, 1987; Woese and Fox, 1977).

The relative simplicity and ease of the 16S rRNA approach has rendered it to be widely used for the classification of species. Currently, 16S rRNA based phylogenetic trees and 16S rRNA sequence comparisons are the most commonly used tools for identification of species relationships or for determining the placement of a species in phylogenetic context (Stackebrandt, 2006). The prevalence of the 16S rRNA in prokaryotic systematics is such that the definition of a prokaryotic species has also been based upon 16S rRNA sequence similarity. In such cases, two organisms with 16S rRNA greater than 97 % similarity are defined as members of the same species group (Stackebrandt and Goebel, 1994). DNA-DNA hybridization is occasionally used to identify if two organisms are the same species based on a 70% threshold (Wayne et al., 1987; Goodfellow et al., 1997). Biochemical and morphological characters are used for additional support for claims of species placements; however, most prokaryotic taxonomy is based on phylogenetic inferences derived from ribosomal RNA sequences (Tindall et al., 2010; Zhi et al., 2012).

Though widely used, some concerns are being raised towards the use of this method for prokaryotic classification. Primarily, being a single gene within genomes that contain hundreds or thousands of other genes, it is suggested that the 16S rRNA may not accurately reflect the evolution of a genome (Ciccarelli et al., 2006). Secondly, some species are known to contain multiple copies of the gene and 16S rRNA is thought to be susceptible to lateral genetic transfer (Janda and Abbott, 2007). Another problem with the 16S rRNA based classification of species is that due to the highly conserved nature of the molecule, there exists a lack of resolution leading to the misidentification of closely related species (Janda and Abbott, 2007). Due to the conserved nature of 16S rRNA, organisms may be misclassified as members of the same taxonomic group while expressing characteristics suggesting otherwise (Janda and Abbott, 2007; Fox et al., 1992). Due to these shortcomings of the 16S rRNA based analyses; alternative means for performing taxonomic and phylogenetic tasks are required.

Introduction of the genomics era

The improvement of prokaryotic taxonomic classifications has been highly dependent on technology and its innovations. A long awaited advancement in evolutionary studies, and perhaps all of biology, has been for the availability of speedy and cost-effective genomic sequencing. As the genome contains the entire genetic data for the organism, decoding the genome was expected to allow for insight into prokaryotic life and their relationships (Boussau and Daubin, 2010). The first prokaryotic genome, *Haemophilus influenzae*, was published in 1995 (Fleischmann et al., 1995). Since then, over 5000 prokaryotic genomes have been made available and sequences are increasing at an exponential level (Markowitz et al., 2012). The availability of extensive genomic data has been useful in several aspects of microbiology and medicine (Staudt, 2003; West et al., 2006; Medini et al., 2008). It has been suggested that genomic analysis can eventually replace current phylogenetic means for species identification and classification (Coenye et al., 2005).

Using the availability of genomes, newer methods have been developed for the purpose of determining species relationships or identifying differences among their sequences. Some of these methods include Average Nucleotide Identity (ANI), Multi-Locus Sequence Analysis (MLSA), comparison of gene content and comparisons of gene order (Coenye et al., 2005; Konstantinidis and Tiedje, 2005; Snel et al., 1999). ANI compares the nucleotide sequence similarity for the conserved genes among a pair of genomes (Konstantinidis et al., 2005; Goris et al., 2007). As a genome encompassing analysis, ANI is compared favourably to DNA-DNA hybridization due to its simplicity but also because it is useful for cultured and uncultured organisms. However, like DNA-DNA hybridization, ANI is limited to pairwise comparisons between two organisms and an incremental database cannot be built based on such analyses. MLSA, the use of a large number of homologous genes to construct a phylogenetic tree or gene similarity comparisons, has been introduced as an alternative to simple phylogenetics based on single genes/proteins. It is often argued that phylogenetic studies based on single genes/proteins or even a few genes concatenated together, only use a fraction of the genomic information while discarding the majority. Such phylogenetic analyses are pejoratively referred to as trees of 1% (Doolittle and Bapteste, 2007). Phylogenetically, it is now easier to compare large numbers of genes, comprising a significant percentage of an organism's genome, and use them for thorough phylogenetic trees. A larger dataset of genes for such analyses is believed to filter the effect that LGT may have of one or a few genes and also to provide more resolution and greater robustness when compared to single gene phylogenetics.

Though gene similarity and phylogenetic methods are useful in deducing associations among prokaryotes, they fail to provide discernible characteristics for defining a related group of organisms. All of the methodologies described above depict prokaryotic relationships on degrees of relatedness rather than providing characteristics that may distinguish groups of related organisms. These systems lead to arbitrary or subjective designations. The subjectivity of the prokaryotic species classification procedures poses a problem as larger numbers of organisms are discovered. Correction of previous taxonomic mistakes is a time consuming process and requires valuable resources to ameliorate. Thus, a more robust system is highly sought after. It has been suggested that ideal characteristics used for defining taxonomic divisions must be synapomorphies, characters that are shared by a group of organisms and their most recent common ancestor (Gao, 2010; Rokas and Holland, 2000; Stackebrandt, 2006).

Use of Conserved Signature Indels as Taxonomic Tools

Over the past decade and a half, comparative genomics has been utilized by Dr. R. S. Gupta and colleagues for the identification of molecular markers indicative of relationships shared by prokaryotes (Gupta, 1998; Gupta, 2000; Gupta and Griffiths, 2002; Griffiths et al., 2005; Naushad and Gupta, 2012). One form of these molecular markers is termed Conserved Signature Indels or CSIs. CSIs are amino acid Inserts or Deletions (i.e. Indels) present within conserved regions of proteins. The conserved regions of the proteins flanking the CSIs ensure that the presence of the CSIs is not due to alignment errors or artifacts (Gupta, 1998). The CSIs represent rare genomic changes that have resulted in presence or absence of amino acids within conserved regions of a protein. When the CSIs are found in related group of organisms, they function as synapomorphies to distinguish the group from other prokaryotic organisms (Gupta, 1998). Due to the rarity of mutations affecting conserved regions within functionally important proteins, the shared presence of CSIs parsimoniously suggests towards common inheritance of rare genetic changes from an ancestor to its progeny (Gupta, 1998). CSIs have previously been utilized for identification of taxonomic divisions from genus level (viz. Clostridium) to phyla level (e.g. Aquificae, Actinobacteria) and even observe relationship shared among different phyla of the bacterial and archaeal kingdoms (Gao et al., 2006; Gao and Gupta, 2007; Griffiths and Gupta, 2006; Gupta and Gao, 2009). In the succeeding chapters of this thesis, the utilization of Conserved Signature Indels for description of the species relationships among the bacterial phyla Thermotogae and Synergistetes are described; similar comparative genomic methods for elucidation of relationships among members of the PVC group of bacteria are also presented.

Taxonomic and Phylogenetic Issues Regarding the phylum Thermotogae, the phylum Synergistetes and the PVC Group of bacteria

The Thermotogae phylum and its species

The Thermotogae is a phylum composed of a group of mostly thermophilic, heterotrophic, anaerobic gram-negative bacteria (Huber and Hannig, 2006; Reysenbach, 2001). A member of the group was first discovered with the isolation of the hyperthermophilic bacterium *Thermotoga maritima* MSB8 from geothermal vents located on the sea floor around the Azores (Huber et al., 1986). *Tt. maritima* is noted to be the first identified bacterial extremophile, harboring extreme temperature environments previously thought to only contain archaea. The species from this phylum have a characteristic balloon-like sheath or “toga” present outside the cell membrane, providing the group with the latter part of its name (Reysenbach, 2001). Most known species from the group are thermophilic. Due to the stability of their proteins at high temperatures, the Thermotogae attract great attention for potential usage in industrial processes (Conners et al., 2006; Kallnik et al., 2011; Park et al., 2010).

The hierarchical classification of the Thermotogae is fairly simplistic for the variety of its species. All cultured species from the phylum Thermotogae are currently limited to a single family, Thermotogaceae (Reysenbach, 2001). Apart from their signature sheath structure, the species of the phylum Thermotogae are primarily ascribed to this group and divided into its different sub-groups (i.e. genera), primarily based on 16S rRNA similarity and 16S rRNA trees. Therefore, no clear signature biochemical or molecular characteristics were known that could distinguish the Thermotogae from other species or differentiate the sub-groups within the Thermotogae from each other. Additionally, relationships among the genera were depicted to be different in phylogenetic trees constructed by different methodologies. Thus, characteristics were required to unambiguously define the inter-relationships among the Thermotogae. Chapter 2 presents the utilization of comparative genomics for identification of CSIs for the group. Several CSIs for the entire phylum Thermotogae are presented along with CSIs identifying interrelationships among the genera.

The Synergistetes phylum

The Synergistetes group is a recently recognized phylum of anaerobic bacteria (Hugenholtz et al., 2009). The phenotypic characteristics shared by the species from this phylum include their gram-negative cell wall structure, and rod/vibrioid cell shape (Jumas-Bilak et al., 2009). While a few species have been shown to be asaccharolytic, all known Synergistetes have the ability to ferment amino acids (Jumas-Bilak et al., 2009). The Synergistetes inhabit a majority of anaerobic environments including the soil, oil wells, wastewater treatment plants and animal gastrointestinal tracts. In humans, they can be found in healthy individuals in the umbilicus and the vaginal flora; they are also present in sites of human diseases such as cysts, abscesses, and areas of periodontal disease (de Lillo et al., 2006; Horz et al., 2006; Godon et al., 2005; Vartoukian et al., 2007; Zijnge et al., 2010; Jumas-Bilak et al., 2007; Kumar et al., 2005). Though environments data depicted these organisms to be widespread and ubiquitous in anaerobic environments, the group is relatively unknown (Godon et al., 2005). The lack of data on this group can be placed on the fact that the Synergistetes are not known to be used in industrial processes or to be pathogenic to plants and animals. However, the Synergistetes may attract more attention as the evolutionary position of the phylum is of interest due to the atypical diderm cell-wall (Gupta, 2011; Sutcliffe, 2010).

Members of the Synergistetes were first discovered in 1992 as symbionts in goat rumen and as amino acid degrading thermophiles (Allison M.J et al., 1992; Guangsheng et al., 1992). However, due to lack of distinct characteristics, the various species currently placed into this group were regularly misclassified into various other taxa, including the phylum Deferribacteres and multiple families of the phylum Firmicutes (Baena et al., 1998; Baena et al., 1999; Diaz et al., 2007; Garrity et al., 2004; Guangsheng et al., 1992; Magot et al., 1997). The members of the phylum were misclassified up till 2009, when the group was brought together as a distinct phylum-level entity using 16S rRNA sequence based phylogenetic analysis (Jumas-Bilak et al., 2009). While the Synergistetes were classified as a distinct phylum, no characteristic of these bacteria was known that could easily differentiate a Synergistetes species from other bacteria or distinguish among its different sub-groups. Like the Thermotogae, due to a lack of defining criteria, all characterized Synergistetes species are currently placed under a single family-level grouping termed the Synergistaceae (Jumas-Bilak et al., 2009). Thus, novel characteristics that may assist in defining the phylum and its sub-groups are required. CSIs that perform this function are highlighted in chapter 3.

The PVC group of Bacteria

The PVC group is an acronymic term used to describe an often observed phylogenetic grouping of gram-negative bacteria which contains the phyla Planctomycetes, Verrucomicrobia and Chlamydiae. In addition to these three phyla, the little known groups Lentisphaera, Poribacteria, candidate division OP3 and candidate division WWE2 are observed to branch in a distinct clade that separates them from other bacteria in phylogenetic trees (Schloss and Handelsman, 2004; Wagner and Horn, 2006). This branching of species in a monophyletic clade implies a shared common ancestor. No official taxonomical designation exists to describe bacterial relationships above the phylum level. Nevertheless, the relatively common association of some or all of these groups in phylogenetic trees has led to the unofficial term “superphylum” to be adopted.

Among this so called superphylum are species that play important biological and ecological roles. Chlamydiae species are renowned as animal pathogens (Peeling and Brunham, 1996; Sachse et al., 2009). Some species of the Planctomycetes, known as Anammox, are known for their anaerobic ammonium oxidization (Strous et al., 1999). Anammox species have the ability to convert ammonium to dinitrogen, an important process in the global nitrogen cycle (Devol, 2003; Kartal et al., 2010). Members of the Verrucomicrobia, Poribacteria and Lentisphaerae are also the only known prokaryotic organisms observed to have a compartmentalized cellular geography, a characteristic previously thought to be a limited to eukaryotic organisms (Fieseler et al., 2004; Fuerst and Sagulenko, 2011; Lee et al., 2009). Verrucomicrobia, though less understood, are ubiquitously found in the soil where they can make up ~10% of some microbial populations (Sangwan et al., 2005).

It is interesting to note that the species of these various phyla are not known to be phenotypically similar or share phenotypic characters that are exclusive to them; though overlapping characteristics are present among sub-groups of the superphylum (McInerney et al., 2011; Wagner and Horn, 2006). The collection of the multiple groups into a cohesive unit is solely based on observations where the different phyla branch together in phylogenetic trees (Wagner and Horn, 2006). The repeated grouping of these taxa in phylogenetic data suggests towards the likelihood of these organisms sharing a common evolutionary ancestor. However, without known shared characteristics, it is unclear whether these bacteria are evolutionary cousins or if their grouping is a phylogenetic artifact. Thus, in chapter 4, we attempt to use comparative genomics for the identification of molecular markers that may elucidate the relationships among these phyla through molecular means.

LGT, prokaryotic evolution and phylogeny

The Darwinian mode of tree-like evolution has been well-established and is entrenched as the model for prokaryotic and eukaryotic species alike. Vertical transfer of genomes from parent to progeny is deemed the major method of genetic transmission.

The term “tree of life” is used to describe the bifurcating connection linking all existing species to a common ancestor (Darwin, 1859; Gogarten and Townsend, 2005). Based on the inferences drawn through acceptance of the bifurcating tree-like model for evolution, species relationships can be identified based on similarity. Simply put, closely related species should have more shared characteristics than species more distantly related. This has been the criteria for all biological classification systems whether based on morphology, biochemistry or genetic sequences. The Linnaean taxonomy, an expression of these classification systems, divides organisms of the three domains of life into divisions from phylum level to species level so as to reflect their evolutionary relationships.

Recently, the mode of evolution for prokaryotes along with their observed relationships to each other 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; Doolittle and Bapteste, 2007; Doolittle, 2000; Nelson et al., 1999). Lateral gene transfer is the process whereby an organism integrates foreign genetic material into its genome through transformation, transduction or conjugation (Davison, 1999). LGT was first experimentally demonstrated in the 1951 with the transduction of a virulence gene into a non-virulent *Corynebacterium diptheriae* strain (Freeman, 1951). It was introduced into widespread conscience due to its role in the spread of antibiotic resistance in pathogenic bacteria (Akiba et al., 1960; Davies, 1995; Ochiai et al., 1959; Watanabe and Fukasawa, 1961).

Though the processes leading to LGT are well known, the relative abundance of such genetic events and the rate of successful incorporation of foreign genetic material into prokaryotic genomes is a contentious matter engendering much debate (Daubin et al., 2003; Doolittle and Bapteste, 2007; Gogarten et al., 2002; Kurland et al., 2003). Initially deemed important for quick adaptation to specialized environments, LGT was thought to be a limited process assisting bacteria in acquiring secondary metabolites for their survival (Cohan, 1994; Lawrence and Hendrickson, 2003). Genes involved in large networks and essential functions were thought to be minimally affected by LGT (Jain et al., 1999; Rivera et al., 1998). However, it has been suggested that LGT may be a more intrusive process such that even informational genes, including ribosomal rRNA and ribosomal proteins, previously thought to be immune to the process, may have undergone LGT in some species (Brochier et al., 2000; Yap et al., 1999; Zhaxybayeva et al., 2006).

Genomic sequencing brought about further suggestions of extensive LGT among prokaryotes. With the publishing of the *Thermotoga maritima* genome, it was estimated that about a quarter of *T. maritima* genes were closer in their relationships to Archaeal gene sequences than to any other bacterial sequences, inferring high LGT between the species and archaeal organisms which shared its extreme environment (Nelson et al., 1999). Analysis of the *Aquifex aeolicus* genome depicted similar results (Aravind et al., 1998). It has been said that species capable of freely exchanging their genetic material may eventually become indistinct (Darwin, 1859; Eisen, 2000). Thus evidence of extensive LGT among prokaryotes has led to the belief that perhaps LGT diminishes, possibly extinguishes, the ability to ascertain traditional prokaryotic relationships with any certainty (Bapteste and Boucher, 2008; Bapteste et al., 2009; Doolittle, 2000; Eisen, 2000). Therefore, scientists sharing this view of immense LGT among prokaryotes often view the visualization of species relationships in tree-form to be inaccurate. Rather, postulations of prokaryotic evolution as a tangled web-like structure rather than a simple bifurcating tree have been forwarded (Swithers et al., 2009; Williams et al., 2011).

However, though prevalent, the view of excessive LGT among bacteria is not the consensus as numerous analyses suggest a lower incidence of lateral genetic transfer (Kurland et al., 2003; Kunin et al., 2005). 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). Based on available data on comparative genomic based analysis performed over the past fifteen years, the utility of CSIs and CSPs for discerning of taxonomic groups is reviewed in chapter 5. The presence of these molecular markers is used to support the model of evolution through vertical inheritance and demonstrate the limited influence of lateral gene transfer in prokaryotic evolution.

Research Objective

The genomic database is large and growing exponentially. Taking advantage of the available genomic data, comparative genomic analysis is performed. The goal of the analyses is to identify molecular synapomorphies, present within the sequences, which define a group of related organisms. More specifically, presence of rare genomic changes in protein sequences, in the form of CSIs, is utilized for the identification of species relationships among the Thermotogae, the Synergistetes and the PVC group of bacteria.

The CSIs are also utilized to depict the limited influence of LGT on prokaryotic evolution.

Due to the similar methodology used for similar purposes, the methods, introduction and discussion sections in chapters 2, 3 and 4 contain overlapping data. The manuscript comprising chapter 5 is a review reflecting some of the conclusions referenced in prior chapters on the utility of CSIs and limits of LGT. The manuscript in chapter 5 also refers to results from chapter 2 in order to express how bacterial associations are identified despite presence of LGT.

CHAPTER 2

Phylogeny and molecular signatures for the phylum Thermotogae and its subgroups[1]

The work presented in this chapter examines, through use of comparative genomics, the relationships of the species of the phylum Thermotogae. Primarily, conserved signature indels for the phylum and its sub-groups are identified. The CSIs are also compared with phylogenetic trees to highlight the groupings of organisms within the phylum. The identified CSIs are also utilized to examine the role of lateral gene transfer in the Thermotogae and the relationships of the phylum to other bacteria are touched upon. Also, functional aspects of CSIs are reviewed.

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. I was also involved in the preparation of the manuscript, including the figures and tables provided.

[1] Due to limited space, supplementary figures (1-65) are not included in the chapter but can be accessed along with the rest of the manuscript at:

Gupta, R. S., and Bhandari, V. (2011). Phylogeny and molecular signatures for the phylum Thermotogae and its subgroups. Antonie van Leeuwenhoek *100*: 1–34. DOI 10.1007/s10482-011-9576-z

The following publication has been reproduced with kind permission from Springer Science+Business Media B.V.

Antonie van Leeuwenhoek (2011) 100:1–34
DOI 10.1007/s10482-011-9576-z

REVIEW PAPER

Phylogeny and molecular signatures for the phylum Thermotogae and its subgroups

Radhey S. Gupta · Vaibhav Bhandari

Received: 4 January 2011 / Accepted: 11 March 2011 / Published online: 19 April 2011

Abstract Thermotogae species are currently identified mainly on the basis of their unique toga and distinct branching in the rRNA and other phylogenetic trees. No biochemical or molecular markers are known that clearly distinguish the species from this phylum from all other bacteria. The taxonomic/evolutionary relationships within this phylum, which consists of a single family, are also unclear. We report detailed phylogenetic analyses on Thermotogae species based on concatenated sequences for many ribosomal as well as other conserved proteins that identify a number of distinct clades within this phylum. Additionally, comprehensive analyses of protein sequences from Thermotogae genomes have identified >60 Conserved Signature Indels (CSI) that are specific for the Thermotogae phylum or its different subgroups. Eighteen CSIs in important proteins such as PolI, RecA, TrpRS and ribosomal proteins L4, L7/L12, S8, S9, etc. are uniquely present in various Thermotogae species and provide molecular markers for the phylum. Many CSIs were specific for a number of Thermotogae subgroups. Twelve of these CSIs were specific for a clade consisting of various *Thermotoga* species except *Tt. lettingae*, which was separated from other *Thermotoga* species by a long branch in phylogenetic trees; Fourteen CSIs were specific for a clade consisting of the *Fervidobacterium* and *Thermosipho* genera and eight additional CSIs were specific for the genus *Thermosipho*. In addition, the existence of a clade consisting of the deep branching species *Petrotoga mobilis, Kosmotoga olearia* and *Thermotogales bacterium mesG1* was supported by seven CSIs. The deep branching of this clade was also supported by a number of CSIs that were present in various Thermotogae species, but absent in this clade and all other bacteria. Most of these clades were strongly supported by phylogenetic analyses based on two datasets of protein sequences and they identify potential higher taxonomic grouping (viz. families) within this phylum. We also report 16 CSIs that are shared by either some or all Thermotogae species and some species from other taxa such as Archaea, Aquificae, Firmicutes, Proteobacteria, Deinococcus, Fusobacteria, Dictyoglomus, Chloroflexi and eukaryotes. The shared presence of some of these CSIs could be due to lateral gene transfers between these groups. However, no clear preference for any particular group was observed in this regard. The molecular probes based on different genes/proteins, which contain these Thermotogae-specific CSIs, provide novel and highly specific means for identification of both known as well as previously unknown Thermotogae species in different environments.

Electronic supplementary material The online version of this article (doi:10.1007/s10482-011-9576-z) contains supplementary material, which is available to authorized users.

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

Additionally, these CSIs also provide valuable tools for genetic and biochemical studies that could lead to discovery of novel properties that are unique to these bacteria.

Keywords Conserved indels · Thermotogae taxonomy · *Fervidobacterium–Thermosipho* clade · *Kosmotoga–Petrotoga* clade · Branching order · *Thermotoga lettingae* · Protein structures · Lateral gene transfers

Introduction

Bacterial species belonging to the phylum Thermotogae were first discovered with the isolation of the hyperthermophilic bacterium *Thermotoga maritima MSB8* (*Tt. maritima*) from geothermal vents located on the sea floor around the Azores (Huber et al. 1986; Nelson et al. 1999). The latter part of this phylum name derives from the characteristic balloon-like sheath or "toga" present outside the cell membrane of these bacteria (Reysenbach 2001). The phylum is composed of a homogenous group of thermophilic, heterotrophic, anaerobic Gram-negative bacteria (Reysenbach 2001; Huber and Hannig 2006). Since the isolation of *Tt. maritima*, over 70 bacterial species belonging to this phylum have been isolated from a variety of geothermal or volcanically heated environments including oil reservoirs, hydrothermal vents, terrestrial hot springs, etc. (Patel et al. 1985; Antoine et al. 1997; Reysenbach 2001; Alain et al. 2002; Balk et al. 2002; Urios et al. 2004; Huber and Hannig 2006; Dipippo et al. 2009; Frock et al. 2010). Recently, the presence of bacteria related to the Thermotogales has also been reported in mesophilic, low temperature environments (Nesbo et al. 2010). All cultured species from the Thermotogae phylum are currently limited to a single family, *Thermotogaceae*, within the order Thermotogales (Reysenbach 2001). The NCBI taxonomy database (The NCBI Taxonomy Homepage 2010), currently lists nine genera within the family *Thermotogaceae*. These are: *Thermotoga*, *Petrotoga*, *Thermosipho*, *Fervidobacterium*, *Marinitoga*, *Kosmotoga*, *Geotoga*, *Thermopallium* and *Thermococcoides* (Reysenbach 2001; Huber and Hannig 2006; Dipippo et al. 2009; The NCBI Taxonomy Homepage 2010; Feng et al. 2010). Due to the stability of these bacteria at high temperatures and their ability to utilize diverse complex carbohydrates including xylan and cellulose for fuel (H_2) production (Eriksen et al. 2010; Kim et al. 2010), Thermotogae species are of much interest from biotechnological viewpoints (Fardeau et al. 1997; Huber and Hannig 2006; Conners et al. 2006).

Thermotogae species are also of much interest as they form one of the deepest branching lineages within the Bacteria (Woese 1987; Huber and Hannig 2006). In the 16S rRNA trees, they branch near the root of the bacterial tree in proximity of the Aquificae species, which are also hyperthermophilic bacteria (Woese 1987; Olsen and Woese 1993). The clustering of these organisms in this position has provided support for the hypothesis that life originated at high temperature and the last common ancestor of all living organisms was hyperthermophilic (Di Giulio 2003; Wong et al. 2007). Although a close relationship between Thermotogae and Aquificae is also supported by phylogenetic trees based on many individual as well as concatenated ribosomal protein sequences (Zhaxybayeva et al. 2009), phylogenetic studies based on numerous other proteins do not support a grouping of these two thermophilic phyla (Karlin et al. 1995; Eisen 1995; Klenk et al. 1999; Griffiths and Gupta 2004; Kunisawa 2005; Ciccarelli et al. 2006; Boussau et al. 2008). In a recent detailed study by Zhaxybayeva et al., the majority of genes other than those encoding for ribosomal proteins supported the grouping of Thermotogae with the Firmicutes (Clostridia) phylum (Zhaxybayeva et al. 2009). Our earlier work based on conserved indels in a number of universally distributed proteins has also provided evidence that while Thermotogae is an early diverging phylum that branches in the proximity of the Firmicutes and Actinobacteria (i.e. monoderm prokaryotes), the Aquificae, which are diderm bacteria, is a late diverging lineage that branches in between the Epsilon-proteobacteria and the Chlamydiae–Verrucomicrobiae–Planctomycetes groups (Gupta 1998, 2003; Griffiths and Gupta 2004, 2007; Kunisawa 2005).

In 1999, the genome sequence for the first Thermotogae species (*Tt. maritima*) became available (Nelson et al. 1999). The Blastp analyses of different proteins from this genome indicated that about 24% of the proteins from *Tt. maritima* had their best matches to proteins from archaeal species whereas ~21% of the proteins were most similar to the Firmicutes (Nelson et al. 1999). However, the closest blast hits are often not the nearest neighbours (Koski and Golding 2001) and recent analyses on Thermotogae genomes have led to significant changes in the numbers of closest blast hits that are observed for the *Archaea* (~10%) and the Firmicute (~45%) taxa (Zhaxybayeva et al. 2009; Nesbo et al. 2009). Nevertheless, these studies suggested that many genes in the *Tt. maritima* genome have been acquired from other taxa, particularly *Archaea* and Firmicutes, by means of lateral gene transfers (LGTs) and genetic recombination (Nelson et al. 1999; Worning et al. 2000; Nesbo et al. 2001, 2006, 2009; Zhaxybayeva et al. 2009).

The genome sequences for 12 Thermotogae species (Table 1) covering the phylogenetic diversity of this phylum are now available in the NCBI database (NCBI 2011). These sequences provide a valuable resource for different types of studies that could lead to better understanding of the taxonomy and evolution of these bacteria as well as their unique biological and biochemical characteristics. The main focus of earlier comparative genomic studies on Thermotogales had been on identifying LGT events and their effects on the evolution of these genomes (Nelson et al. 1999, 2001, 2006; Zhaxybayeva et al. 2009). However, thus far no detailed study has been carried out to identify genetic or molecular characteristics that are uniquely shared by either all Thermotogae species or by different subgroups of species within this phylum. It is important to note that Thermotogae species, except for their distinctive toga, are presently identified solely on the basis of their branching in the rRNA (or protein) trees (Reysenbach 2001; Huber and Hannig 2006). Therefore, identification of molecular markers that are unique to the Thermotogae species or their subgroups should provide novel and useful means for different types of studies. Additionally, the presence of some molecular markers that are uniquely shared by Thermotogae and certain other groups will also provide additional evidence for LGTs between them.

Using genome sequence data, our recent work has focused on identifying molecular markers that are specific for different bacterial groups. One type of molecular markers that have proven very useful for these studies consists of Conserved Signature Inserts or deletions i.e. Indels (CSIs) of defined lengths that are present at specific locations in widely distributed proteins and are specific for particular groups of organisms (Gupta 1998, 2009a; Gupta and Griffiths 2002; Griffiths and Gupta 2006b). The simplest and most parsimonious explanations for these CSIs are that the rare genetic changes responsible for them first occurred in a common ancestor of these groups (or clade) of species and they were then vertically passed on to various descendants (Gupta 1998, 2009a; Rokas and Holland 2000). However, the shared presence of CSIs in some cases can also result from independent genetic events or by means of LGTs (Boucher et al. 2003; Zhaxybayeva et al. 2006). Hence, it is useful to interpret the results of

Table 1 Sequence characteristics of Thermotogae genomes

Organism	GenBank Accession no.	Size (Mb)	No. of proteins	% GC content	Optimal temp. (°C)	Reference
Fervidobacterium nodosum Rt17-B1	CP000771.1	1.9	1,750	35.0	70	Zhaxybayeva et al. (2009)
Kosmotoga olearia T.B.F 19.5.1	CP001634.1	2.3	2,118	41.5	65	DOE-JGI[a]
Thermotogales bacterium MesG1.Ag.4.2	AEDC00000000	2.9	2,613	45.0	–	DOE-JGI[a]
Petrotoga mobilis SJ95	CP000879.1	2.2	1,898	34.1	58–60	DOE-JGI[a]
Thermosipho africanus TCF52B	CP001185.1	2.0	1,954	30.8	75	Nesbo et al. (2009)
Thermosipho melanesiensis B1429	CP000716.1	1.9	1,879	31.4	75	Zhaxybayeva et al. (2009)
Thermotoga lettingae TMO	CP000812.1	2.1	2,040	38.7	65	Zhaxybayeva et al. (2009)
Thermotoga maritima MSB8	AE000512.1	1.9	1,858	46.2	80	Nelson et al. (1999)
Thermotoga naphthophila RKU-10	CP001839.1	1.8	1,768	46.1	80	DOE-JGI[a]
Thermotoga neapolitana DSM 4359	CP000916.1	1.9	1,937	46.9	70–75	Lee et al. (2009)
Thermotoga petrophila RKU-1	CP000702.1	1.8	1,785	46.1	80	Zhaxybayeva et al. (2009)
Thermotoga sp. RQ2	CP000969.1	1.9	1,819	46.2	76–82	DOE-JGI[a]

[a] DOE-JGI—These genomes have been sequenced by the United States Department of Energy Joint Genomic Institute

these studies in conjunction with phylogenetic approaches. Additionally, depending upon the presence or absence of these CSIs in outgroup species, it is possible to infer whether the indel under consideration is an insert or a deletion and they can be used to develop rooted phylogenetic relationships (Rivera and Lake 1992; Baldauf and Palmer 1993; Gupta 1998, 2001, 2010). In addition, the shared presence of some CSIs in the Thermotogae species and another well-defined group(s) of bacteria could identify possible cases of LGTs among these taxa (Griffiths and Gupta 2006a).

In this work, we report detailed phylogenetic studies and comparative analyses of protein sequences from Thermotogae genomes to identify CSIs that are specific for these organisms at different phylogenetic depths. Our analyses have identified many CSIs that are specific for either all sequenced Thermotogae species or a number of distinct subclades within this phylum that are supported by phylogenetic analyses. Additionally, we also describe several CSIs that are shared by Thermotogae species and some other organisms (viz. Aquificae, Archaea, Deinococcus–Thermus, Firmicutes, Proteobacteria, eukaryotes, etc.) providing possible examples of LGTs between these groups. These molecular signatures provide valuable means for taxonomic, evolutionary as well genetic and biochemical studies on these bacteria.

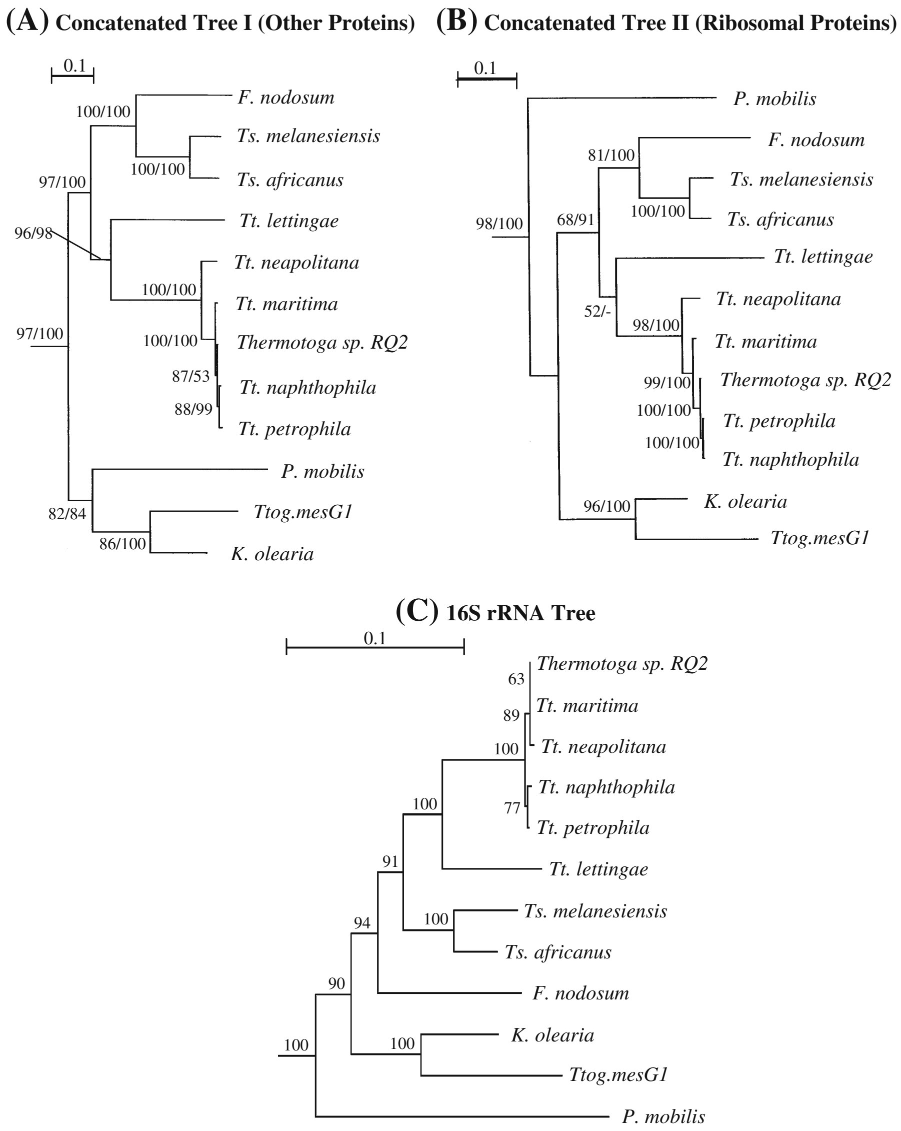
Phylogenetic analyses of Thermotogae

The complete genomes for 12 Thermotogae species are now available (Table 1). These include *Thermotoga (Tt.) maritima MSB8* (Nelson et al. 1999), *Tt. lettingae TMO*, *Tt. neapolitana DSM 4359*, *Tt. naphthophila RKU-10*, *Tt. petrophila RKU-1*, *Thermotoga sp. RQ2*, *Petrotoga (P.) mobilis* SJ95, *Kosmotoga (K.) olearia TBF 19.5.1*, *Fervidobacterium (F.) nodosum Rt17-B1*, *Thermosipho (Ts.) melanesiensis BI429*, *Ts. africanus TCF52B* and *Thermotogales bacterium mesG1.Ag.4.2 (Ttog. mesG1)*. Some characteristics of these genomes are listed in Table 1. The sequenced genomes include representatives from all known Thermotogae genera except *Geotoga*, *Marinitoga*, *Thermopallium* and the newly described *Thermococcoides* genus (Feng et al. 2010). Genome sequence for an unclassified species, *Thermotogales bacterium mesG1.Ag.4.2*, was also available. Although the genome sizes of Thermotogae species varied in a range from 1.8 to 2.9 Mb, their G+C content showed a large variation (from 30.8 to 46.9%).

The branching order of species within the Thermotogae phylum has been previously determined mainly on the basis of 16S and 23S rRNA trees (Reysenbach 2001; Mongodin et al. 2005; Huber and Hannig 2006; Dipippo et al. 2009; Feng et al. 2010). Due to the availability of genome sequences, it is now possible to determine the branching order of Thermotogae species based upon concatenated sequences for large numbers of proteins. The trees based upon large numbers of characters derived from multiple proteins are better able to resolve phylogenetic relationships 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; Gupta and Mathews 2010). Recently, phylogenetic analyses for 5 Thermotogae species based upon concatenated sequences for 29 ribosomal proteins was reported by Zhaxybayeva et al. (2009). Because, sequence information for many other Thermotogae genomes is now available (Table 1), phylogenetic trees for them were constructed based upon concatenated sequences for two different datasets of proteins. The first dataset consisted of 12 large proteins (viz. EF-Tu and EF-G, Gyrase A and Gyrase B, RNA polymerase β and β' subunits, SecA, UvrD, RecA, GroEL chaperone, DNA polymerase I and alanyl-tRNA synthetase) found in most extant bacteria (Harris et al. 2003; Ciccarelli et al. 2006; Gao et al. 2009; Gupta and Mathews 2010), which have been widely used for phylogenetic analyses (Gupta 1995; Eisen 1995; Karlin and Brocchieri 1998; Brocchieri and Karlin 2000; Bocchetta et al. 2000; Watanabe et al. 2001; Seo and Yokota 2003). Sequences for these proteins were obtained for all 12 Thermotogae spp. from the NCBI database. Sequences for these proteins from *Bacillus subtilis*, *Deinococcus radiodurans*, *Staphylococcus aureus* and *Thermus thermophilus* were also obtained to be used as outgroup in these studies. The sequences were aligned using the ClustalX 1.83 program (Jeanmougin et al. 1998). After concatenation of these sequence alignments, the poorly aligned regions were removed using the Gblocks_0.91b program (Castresana 2000), leaving a total of 8073 aligned positions for phylogenetic analyses (dataset I). The second dataset of proteins, which was created in a similar manner, was based on 15 ribosomal proteins (viz. L1, L2, L4, L5, L6, L15, S2, S5, S8, S9, S11, S13, S15, S17 and S19) and it contained 2795 aligned positions. The neighbour-joining (NJ) trees based on 100 bootstrap replicates for the two sets of protein sequences were constructed using the TREECON 1.3b program (Van de Peer and De Wachter 1997) using Kimura's distance calculation (Kimura 1983). The maximum likelihood (ML) trees based upon them were created using the TREE-PUZZLE program employing WAG+F model with gamma distribution of evolutionary rates with four categories and 10000 puzzling steps (Schmidt et al. 2002). In parallel, a NJ tree based upon 16S rRNA sequences for the same species (downloaded from the ribosome database project) (Cole et al. 2009) was also constructed using the TREECON 1.3b program based upon Kimura's two-parameter model (Kimura 1980).

The results of these analyses are presented in Fig. 1. The trees based upon the two datasets of protein sequences (Fig. 1a, b) were generally very similar in their branching pattern and most nodes were resolved with high degree of statistical support. The main difference between these trees was that whereas in the tree based on dataset I a clade consisting of *P. mobilis*, *K. olearia* and *Ttog. megG1* was supported by both NJ and ML analyses, in the trees based upon dataset II and the 16S rRNA,

Fig. 1 Phylogenetic trees for Thermotogae species based on proteins and rRNA sequences. (**A**) A ML tree based upon concatenated sequences for 12 conserved and universally distributed proteins (Dataset I). (**B**) A ML tree based upon concatenated sequences for 15 ribosomal proteins (Dataset II). For both datasets, the *numbers on the nodes* indicate the % support for various nodes in the ML and NJ analyses, respectively. (**C**) A NJ distance tree for the Thermotogae species based upon 16S rRNA sequences



P. mobilis showed deeper branching in comparison to a clade consisting of *K. olearia* and *Ttog. megG1*. However, phylogenetic trees based on both sets of protein sequences differed from the 16S rRNA tree in one important respect i.e. whereas in both these trees a clade consisting of *F. nodosum*, *Ts. africanus* and *Ts. melanesiensis* was strongly supported by both NJ and ML analyses (Fig. 1a, b), this clade was not resolved in the rRNA tree (Fig. 1c) (Reysenbach 2001; Huber and Hannig 2006). However, similar to the protein trees, these two genera were also found to group together in the phylogenetic trees based on 23S rRNA or 16S + 23S rRNA sequences (Zhaxybayeva et al. 2009; Dipippo et al. 2009).

Identification of CSIs that are specific for the Thermotogae phylum

To identify conserved indels in protein sequences, Blastp searches were performed on each protein in the genome of *Tt. petrophila*. For all proteins for whom high scoring homologs were present in most Thermotogae species as well as in some other bacteria, top 10–15 high scoring homologs from diverse Thermotogae and other bacterial groups were retrieved and their multiple sequence alignments were constructed using the ClustalX 1.83 program (Jeanmougin et al. 1998). These sequence alignments were visually inspected to identify conserved inserts or deletions that were restricted to either some or all Thermotogae species 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 considered as they do not provide useful molecular markers (Gupta 1998, 2001, 2009a). The conserved indels, which in addition to the Thermotogae species were also present in some other taxa, were also retained. The species distribution patterns of all conserved indels were further evaluated by detailed Blastp searches on short sequence segments containing the indels and their flanking conserved regions (Gupta 2009a). The sequence information for various conserved indels from different Thermotogae species as well representative species from other bacterial groups were compiled into signature files that are shown here. Unless otherwise noted, all of these CSIs are specific for the indicated groups. Some characteristics of the identified CSIs are discussed below.

Our analyses have identified 18 CSI in widely distributed proteins that are specific for various sequenced Thermotogae and provide potential molecular markers for this phylum. Some characteristics of these CSIs are listed in Table 2 and four examples are shown here. In two of the ribosomal proteins S8 and L7/L12, which are essential for protein synthesis and whose homologs are present in all bacteria (Wu et al. 1994; Gudkov 1997), 3 aa inserts in highly conserved regions are specifically present in all Thermotogales but not in any other bacteria (Fig. 2). Similarly, in the universally distributed RecA protein and in DNA polymerase I, which are essential for DNA repair and replication (Karlin et al. 1995; Eisen 1995), 5 aa and 3 aa inserts in conserved regions are uniquely present in all Thermotogae species (Fig. 3). Except for the Thermotogales, these CSIs are not found in any other bacteria (0 in >250), indicating that they provide reliable molecular markers for this phylum. The absence of these indels in all other bacterial phyla provides evidence that they constitute inserts rather than deletions in the Thermotogae species. The information for other CSIs, which are largely specific for the order Thermotogales, is provided in Table 2 and Sup. Figs. 1–14. These other proteins in which Thermotogales-specific CSIs are found include ribosomal protein S9 (Sup. Fig. 1), ribosomal protein L4 (Sup. Fig. 2), tryptophanyl-tRNA synthetase (Sup. Fig. 3), the enzymes pyruvate phosphate dikinase (Sup. Fig. 4) and ribonucleoside-diphosphate reductase (Sup. Fig. 5), MinD protein involved in septum site determination (Sup. Fig. 6), glucose inhibited division protein A (GidA) (Sup. Fig. 7), the enzyme UDP-*N*-acetylenolpyruvoylglucosamine reductase involved in the synthesis of UDP-*N*-acetylmuramic acid (Sup. Fig. 8), a DEAD/DEAH box helicase domain-containing protein (Sup. Fig. 9), the protein DnaA involved in chromosomal replication initiation (Sup. Fig. 10), the enzymes adenylosuccinate synthase (Sup. Fig. 11), phosphoribosyl-formylglycinamidine synthase II (Sup. Fig. 12) and aspartate ammonia-lyase (Sup. Fig. 13) and a MazG family protein (Sup. Fig. 14). Most of these proteins involved in important cellular functions are broadly distributed in bacteria. However, in a few cases (Sup. Figs. 11–14), their homologs were not detected in one of the Thermotogae species.

Table 2 Conserved Signature Indels that are specific for the Thermotogae phylum

Protein	50S ribosomal protein L7/L12 (RplL)	30S ribosomal protein S8 (RpsH)	DNA recombination protein (RecA)	DNA Polymerase I (PolA)	30S ribosomal protein S9 (RpsI)	50S ribosomal protein L4 (RplD)
GenBank Identifier	150020401	15644234	160901572	154249057	170289134	437924
Accession no.	YP_001305755	NP_229286	YP_001567153	YP_001409882	YP_001739372	CAA79778
Indel/size	3 aa ins	3 aa ins	5 aa ins	2–3 aa ins	3 aa ins	10–15 aa ins
Indel position[a]	82–124	48–90	292–324	34–70	11–71	166–214
Figure no.	Fig. 2a	Fig. 2b	Fig. 3a	Fig. 3b	Sup. Fig. 1	Sup. Fig. 2
Tt. petrophila	+	+	+	+	+	+
Tt. maritima	+	+	+	+	+	+
Tt. neapolitana	+	+	+	+	+	+
Tt. sp. RQ2	+	+	+	+	+	+
Tt. naphthophila	+	+	+	+	+	+
Tt. lettingae	+	+	+	+	+	+
F. nodosum	+	+	+	+	+	+[d]
Ts. melanesiensis	+	+	+	+	+	+[d]
Ts. africanus	+	+	+	+	+	+[d]
P. mobilis	+	+	+	+	+	+[d]
K. olearia	+	+	+	+[d]	+	+[d]
Ttog. mesG1	+	+	+	+[d]	+	+[d]
Other species with indel[b]	0/250	0/250	0/250	0/250	0/250	0/250

Protein	Tryptophanyl-tRNA synthetase (TrpRS)	Pyruvate phosphate dikinase (PPDK)	Ribonucleoside diphosphate reductase (RNR)	Septum site-determining protein (MinD)	Glucose inhibited division protein A (GidA)	UDP-*N*-actylenol-pyruvoyl-glucosamine reductase (MurB)
GenBank Identifier	15643258	148269788	160903303	15644613	170288483	157363403
Accession no.	NP_228302	YP_001244248	YP_001568884	NP_229666	YP_001738721	YP_001470170
Indel/size	1 aa ins	2 aa ins	27 aa ins	2 aa ins	2 aa ins	1 aa ins
Indel position[a]	13–47	222–253	79–139	151–184	244–302	200–231
Figure no.	Sup. Fig. 3	Sup. Fig. 4	Sup. Fig. 5	Sup. Fig. 6	Sup. Fig. 7	Sup. Fig. 8
Tt. petrophila	+	+	+	+	+	+
Tt. maritima	+	+	+	+	+	+
Tt. neapolitana	+	+	+	+	+	+
Tt. sp. RQ2	+	+	+	+	+	+
Tt. naphthophila	+	+	+	+	+	+
Tt. lettingae	+	+	+	+	+	+
F. nodosum	+	+	+	+	+	+
Ts. melanesiensis	+	+	+	+	+	+
Ts. africanus	+	+	+	+	+	+
P. mobilis	+	+	+	+	+	+
K. olearia	+	+	0[c]	+	+	+
Ttog. mesG1	+	+	0[c]	+	+	+
Other species with indel[b]	1/250	1/250	0/250	0/250	1/250	1/250

Table 2 continued

Protein	DEAD/DEAH box helicase domain-containing protein	Chromosomal replication initiation protein (DnaA)	Adenylo-succinate synthase (ADSS)	Phospho-ribosylformyl-glycinamidine synthase II (FGAM Synthase II)	Aspartate ammonia-lyase (AspA)	MazG family protein
GenBank Identifier	150021721	281411445	170289498	148270649	148269616	148269159
Accession no.	YP_001307075	YP_003345524	YP_001739736	YP_001245109	YP_001244076	YP_001243619
Indel/size	1–2 aa del	1 aa del	1 aa del	5–6 aa del	1 aa del	4 aa del
Indel position[a]	688–719	119–141	222–260	27–73	365–392	98–122
Figure no.	Sup. Fig. 9	Sup. Fig. 10	Sup. Fig. 11	Sup. Fig. 12	Sup. Fig. 13	Sup. Fig. 14
Tt. petrophila	+	+	+	+	+	+
Tt. maritima	+	+	+	+	+	+
Tt. neapolitana	+	+	+	+	+	+
Tt. sp. RQ2	+	+	+	+	+	+
Tt. naphthophila	+	+	+	+	+	+
Tt. lettingae	+	+	+	−	+	+
F. nodosum	+	+	+	+	+	+
Ts. melanesiensis	+	+	+	+[e]	+	+
Ts. africanus	+	+	+	+[e]	+	+
P. mobilis	+	+	+	+	0[c]	0[c]
K. olearia	+[d]	+	0[c]	0[c]	+	+
Ttog. mesG1	+[d]	+	+	0[c]	0[c]	+
Other species with indel[b]	0/250	0/250	0/250	0/250	0/250	0/250

[a] The indel position indicates the region of the protein containing the CSI

[b] The presence or absence of the CSIs in the top 250 Blast hits is indicated. The number of non-Thermotogae organisms, which were observed to contain the CSI, is specified. Species containing other indels in this region, which are likely of independent origin, were not included in the total

[c] Homologous sequences corresponding to the region containing the CSI's could not be identified in these species

[d] The CSI's in these organisms were 1–5 aa shorter than other Thermotogae species

[e] A 1 aa deletion specific for *Thermosipho* genus is present in the 6 aa insert

The CSIs in most of these proteins are highly specific for the Thermotogales species. However, in a few cases, 1–2 isolated species belonging to other bacterial groups also contained indels of similar lengths (see Table 2). For example, a 2 aa insert in pyruvate phosphate dikinase, in addition to all Thermotogae, is also present in *Korarchaeum cryptofilum* (Sup. Fig. 4). The shared presence of this CSI in *K. cryptofilum* could result from either a LGT event or due to independent occurrence of a similar genetic change in this archaeum. In a few cases, CSIs of different lengths were also present in limited numbers of species from other groups, which due to their different lengths, have likely resulted from independent genetic events.

Conserved indels that are specific for the Thermotogae subgroups

The Thermotogae phylum is presently comprised of a single order and a single family containing nine genera. Of the 12 complete genomes from this order that are currently available, six are from the *Thermotoga* genus (viz. *Tt. maritima*, *Tt. petrophila*, *Thermotoga sp. RQ2*, *Tt. lettingae*, *Tt. naphthophila* and *Tt. neapolitana*), whereas the remaining 6 are from at least 4 other genera. In the 16S rRNA tree, no specific relationship is observed among these genera (Fig. 1c) (Reysenbach 2001; Huber and Hannig 2006). However, our phylogenetic analyses and many

CSIs that we have identified strongly support the existence of a number of distinct clades within this phylum.

In the protein trees, all *Thermotoga* species except *Tt. lettingae* formed a robust clade, where *Tt. lettingae* was separated from them by a long-branch. These relationships are also supported by the identified CSIs. During our analyses, we have come across only 1 CSI, consisting of a 1 aa insert in a highly conserved region of the protein isoleucyl-tRNA synthetase, that is commonly shared by all *Thermotoga* species including *Tt. lettingae* (Fig. 4a). In contrast, 12 additional CSIs are commonly shared by all *Thermotoga* species (except *Tt. lettingae*). One example of a CSI showing this latter relationship is presented in Fig. 4b. In this case, a 7 aa insert in a highly conserved region of the universally distributed RNA polymerase β' subunit (RpoC) is commonly shared by various *Thermotoga* species (Fig. 4b), but it is not found in *Tt. lettingae* or any other Thermotogales or other bacteria. In addition to this CSI, RpoC also contains another large CSI that shows similar species distribution profile (Sup. Fig. 15). The information for other CSIs that are specific for various *Thermotoga* spp. except *Tt. lettingae* is provided in Table 3 and Sup. Figs. 16–27. The proteins in which these CSIs are found include the enzyme purine nucleoside phosphorylase (PNP) (Sup. Fig. 16); a patatin-like protein (Sup. Fig. 17); the flagellar motor switch protein FliM (Sup. Fig. 18); tRNA modification GTPase, TrmE (Sup. Fig. 19); a protein related to metalloendopeptidase glycoprotease family (Sup. Fig. 20); the enzymes aspartate aminotransferase (Sup. Fig. 21) and Dak phosphatase (Sup. Fig. 22); two different CSIs in ATP-dependent protease La (Sup. Figs. 23 and 24) and 1 aa insert in adenylate kinase (Sup. Fig. 25). In addition, a 6 aa insert in the RpoB is also present in various *Thermotoga* species (except *Tt. lettingae*) (Sup. Fig. 26). However, this insert was also present in *F. nodosum*. Lastly, a 1 aa insert in various *Thermotoga* was also identified in the protein 7-cyano-7-deazaguanine reductase, but the homologs of this protein were not detected in most other *Thermotogaceae* species (Sup. Fig. 27). All of these CSIs are present in conserved regions and most of them are exclusively present in the indicated groups of species, thereby providing strong evidence that these species are specifically related.

Fervidobacterium and *Thermosipho* are two other genera that formed a strongly supported clade in phylogenetic trees based upon both datasets of protein sequences (Fig. 1a, b). A closer relationship between these two genera as well as *Geotoga* was also observed in earlier phylogenetic trees (Reysenbach 2001; Mongodin et al. 2005; Zhaxybayeva et al. 2009; Dipippo et al. 2009). Our analyses have identified 14 CSIs that are commonly shared by species from these two genera. Three of these CSIs are shown in Fig. 5. These include a 2 aa deletion in the enzyme DNA polymerase III (PolC) (Fig. 5a), a 1 aa insert in the DNA Gyrase B subunit (Fig. 5b) and a 1 aa insert in the MreB protein (Fig. 5c), which is responsible for cell shape determination in prokaryotic organisms (Osborne et al. 2004). The MreB protein also contains a 1 aa deletion in a different region that is specific to these genera (Sup. Fig. 28). The sequence information for these proteins is mainly shown for the Thermotogae species. However, these CSIs are not found in other phyla of bacteria. Information for other proteins, which contain CSIs that are specific for the *Fervidobacterium* and *Thermosipho* genera is presented in Table 4 and Sup. Figs. 29–37. The proteins containing these CSIs include chromosome segregation protein (Sup. Fig. 29), diguanylate cyclase (Sup. Fig. 30), two different CSIs in the enzyme glucose-1-phosphate thymidyltransferase (Sup. Fig. 31), an ExsB family protein (Sup. Fig. 32), ornithine decarboxylase (Sup. Fig. 33) and the enzyme phosphomannomutase (Sup. Fig. 34). In addition to these proteins, a 5 aa insert in a basic membrane lipoprotein (Sup. Fig. 35) and 1 aa insert in the protein phosphate butyryltransferase (Sup. Fig. 36) are also specifically present in these two genera. However, for both these proteins, two homologs are present in *Fervidobacterium* and *Thermosipho* species and the insert is present in only one of them. Lastly, in the GidA protein, which contains a 2 aa insert that is specific for all Thermotogales (Sup. Fig. 7), a 1 aa deletion is also uniquely present in these two genera as well as in *Petrotoga mobilis* (Sup. Fig. 37). The presence of this CSI in *P. mobilis* could be due to LGT or its forming a deeper branching clade with these two genera, though little evidence exists for the *Thermosipho-Fervidobacterium-Petrotoga* clade.

For the *Thermosipho* genus, sequence information is available for two species i.e. *Ts. melanesiensis*

(A)

```
                                                     82                                        124
Thermotogae 12/12
  Thermosipho melanesiensis         150020401       TGLGLKEAKDLVEKAG TPD AVIKQGVNKDEAEEIKKKLEEAGA
  Fervidobacterium nodosum          154250448       ---------------- AA- -------K-E--------------
  Thermosipho africanus             217077414       ---------------- --- --V---AA-E----------A---
  Petrotoga mobilis                 160901840       -N-------------- --- ----E--P-E--------------
  Kosmotoga olearia                 239618203       ---------E------ S-- -I--E-IS-S-------Q------
  Thermotogales bac. mesG1          307297336       ---------------- --E G-V-ENLP-A----L--Q------
  Thermotoga petrophila             148269603       ---------------- S-- -I--S--P-Q---D----------
  Thermotoga naphthophila           281411679       ---------------- S-- -I--S--P-Q---D----------
  Thermotoga sp. RQ2                170288279       ---------------- S-- -I--S--P-Q---D----------
  Thermotoga lettingae              157363340       ---------------- S-- -IV-S-IP-N---D----------
  Thermotoga neapolitana            222099190       ---------------- --- ----S--S-E--------------
  Thermotoga maritima               132655          ---------------- S-- ----S--S-E--------------
Other bacteria 0/>250
  Escherichia coli                  15804576        -------------S-P     -AL-E--S--D--AL--A------
  Haemophilus influenzae            16272584        -------------S-P     -NL-E--S-E---AL--E------
  Legionella pneumophila            52840566        -------------G-P     STV-E--S----AS---E------
  Xanthomonas campestris            21230356        -----------T-AG-      IL-E--S-----K---EMT----
  Aquifex aeolicus                  15606948        ---------E--DN-P     KP--E--P-E---Q----------
  Synechococcus elongatus           56750903        -------------A-P     KP--E--S--D--AA--E------
  Magnetospirillum magneticum       83312240        -------------A-P     KSV-D--S-----K---V------
  Rhodopseudomonas palustris        115525597       -------------G-P     KPV-E--------KV-AQ--K---
  Rhodospirillum rubrum             83594028        -N-----------G-P     KPV-EA-S----AS----------
  Gloeobacter violaceus             37521171        -------------A-P     KAV-E-----D-AT----------
  Helicobacter pylori               15645813        ----------AT--TP     H-L-E----E---T-------V--
  Geobacillus kaustophilus          56418631        ------------DNTP     KP--E-IA-E------A-------
  Bacillus thuringiensis            228937402       ---------E--DNTP     KA--E--S-E----M-A----V--
  Cytophaga hutchinsonii            110639543       ------------DG-P     KPV-E-AS-----A---Q------
  Flavobacterium psychrophilum      150025249       ------------DA-P     SNV-E--S-----GL--S------
  Verrucomicrobiae bacterium        254445976       -------------G-P     KPV-E--T-E------S---A---
  Meiothermus silvanus              297567058       -----------T-QG-      ---E--S-E---K---Q--D---
  Thermus thermophilus              16974117        -----------A--G-      P--E--S-Q----------A---
  Spirochaeta thermophila           307718199       M---------F--AP-     KAV-E--S-Q----L---------
  Deferribacter desulfuricans       291280156       ---------A--DG-P     SPV-E--A-E---Q--A-------
```

(B)

```
                                                    48                                          90
Thermotogae 12/12
  Thermotoga maritima               15644234        YKYIEDGKQGILRVYLKYK GGR KNRERVIHGIVRVSHAGRRIY
  Thermotoga neapolitana            222099981       ------------------- --- -------------------V-
  Thermotoga petrophila             148270436       ------------------- --- ---------------S-----
  Thermotoga sp. RQ2                170289169       ------------------- --- ---------------S-----
  Thermotoga naphthophila           281412743       ------------------- --- ---------------S-----
  Thermosipho melanesiensis         150020859       -T---------I-IQM--- -T- R-----------I-KP-----
  Thermosipho africanus             217077297       -T---------I-IHM--- -T- R-------------KP-----
  Thermotogales bac. mesG1.         307297380       ----------V-K-F---- -D- RHKVN--S----I-KP-K-L-
  Kosmotoga olearia                 239618251       ------------K-F---- -D- R--Q---S------KP-----
  Thermotoga lettingae              157363456       -------------IHM--- -Q- RD-----K------KP---L-
  Fervidobacterium nodosum          154249803       -T-----------IQM--- -T- --------------KP-----
  Petrotoga mobilis                 160902242       --F-D-------K------ -T- RDKKPIME--I---KS---V-
Other bacteria 0/>250
  Clostridium difficile             126697654       -DV--------I-IQ---G     QEG----T-LKKI-KP-M-V-
  Halothermothrix orenii            220930979       --V--KKP-NA--I----S     --G-K--S-LK-I-KP-L-V-
  Heliobacterium modesticaldum      167629481       VE-V--N---L--I---FG     P---K--T-VK-I-KP-L-V-
  Symbiobacterium thermophilum      51894196        FEW-DN-H--VI-I----G     P-KT---S-LR-I-KP-L-V-
  Thermoanaerobacter tengcongensis  20808646        VEE-D---G---KIT---G     P-K----S-LK-I-KP-L-V-
  Bacillus subtilis                 1644197         VEFV--S----I--S---G     Q-N----T-LK-I-KP-L-V-
  Enterococcus faecalis             29374865        VE----D---VI--F---G     --E----TNLK-I-KP-L-A-
  Listeria monocytogenes            255024985       VE----DNA-TI--F---G     ATG----T-LK-I-KP-L-V-
  Staphylococcus aureus             15925226        VE-V--D---V--LF---G     Q-D----T-LK-I-KP-L-V-
  Bdellovibrio bacteriovorus        42524363        F-VAK-S----M------D     EAGGHA-NN-D---RP---V-
  Desulfurivibrio alkaliphilus      297569408       FQVS--D-----KIT-R-D     DR-QQA-T--K-C-SP-----
  Geobacter metallireducens         78221860        F-V-A-N---V-----RFI     DEK-P--RE-K---KP-S-V-
  Pelobacter propionicus            118579133       --V-S-NL----------I     DDKDG--NE-K-I-KP-G-V-
  Thermodesulfo. yellowstonii       206889491       --I-K-K----I-IN---T     SEGDS--SNLQ-I-KP---V-
  Fusobacterium ulcerans            257470840       --VVT--NKKSI------D      GKD-I-K--K-I-KP---V-
  Fusobacterium varium              253581378       --VVT--NKKSI------D      GKD-I-K--K-I-KP---V-
  Salinibacter ruber                83815774        FVN-D-E---L-------D     EYDQPA-RMLE---RP---E-
  Bacteroides capillosus            154498898       FQL-D--T--VI--T---N  N  PGK-KA-S-LR---KP-L-V-
  Acidobacterium capsulatum         225873064       --TT-EEGRA--------G     S-N-AA-RDLS---RP-C-V-
```

◀ **Fig. 2** Partial sequence alignments for (**A**) the 50S ribosomal protein L7/12 and (**B**) the 30S ribosomal protein S8, showing two conserved CSIs (*boxed*) that are uniquely present in all Thermotogae species. The *dashes* (–) in this and all other sequence alignments shown here indicate identity with the corresponding amino acid on the *top line*. The position of the indel-containing sequence for the species on the *top line* is noted above the sequence. Sequence information for only representative species is shown in this and other sequence alignments. However, no other species within the indicated numbers of blast hits contained this indel

BI429 and *Ts. africanus TCF52B* (Table 1). Our analyses have identified 8 CSIs that are mainly found in these two species thus providing molecular markers for this genus. Two examples of the CSIs that are specific for this genus are shown in Fig. 6. The proteins 1-deoxy-D-xylulose-5-phosphate reductoisomerase (Fig. 6a) as well as glycerol kinase (Fig. 6b) both contain 2 aa deletions that are only found in these species. The other proteins containing CSIs that are specific for the *Thermosipho* spp. are listed in Table 5 and information for these is presented in Sup. Figs. 38–43.

Petrotoga mobilis, *Kosmotoga olearia* and *Thermotogales bacterium mesG1* (*Ttog. mesG1*) show deep branching in phylogenetic trees based on both sets of protein sequences as well as 16S rRNA (Fig. 1). Although a grouping of these species was only observed in phylogenetic trees based on dataset I protein sequences (Fig. 1a), a specific association of them is strongly suggested by a number of CSIs that are uniquely present in them. In the protein O-antigen polymerase, the homologs of which are only found in various Thermotogales, two different CSIs consisting of a 5 aa insert (Fig. 7a) and a 1 aa insert (Sup. Fig. 44) are specifically found in *P. mobilis, K. olearia* and *Ttog. mesG1*. Likewise, in the CaCA family Na(+)/Ca(+) antiporter protein, a 1 aa deletion is uniquely present in these species (Fig. 7b). Some other proteins that contain CSIs that are specific for these genera include a 1 aa deletion in a protein of unknown function (Sup. Fig. 45), a 3 aa insert in the peptidase S15 protein (Sup. Fig. 46) and an 11 aa insert in the ATPase subunit H of the oligopeptide/dipeptide ABC transporter protein (Sup. Fig. 47). Interestingly, in this latter protein, a 4 aa insert that is specific for the *Thermosipho* and *Fervidobacterium* is also present in the same region. Whether these two CSIs have originated independently or if they are derived from each other is unclear. The information regarding species distribution of these CSIs is provided in Table 6.

We have also identified a number of CSIs that support the deeper branching of the clade consisting of *P. mobilis, K. olearia* and *Ttog. mesG1* in comparison to the other Thermotogales. In the α subunit of RNA polymerase (RpoA), a 2 aa insert is present in various Thermotogales except *P. mobilis, K. olearia* and *Ttog. mesG1* (Fig. 8a). Because, this insert is absent in all other bacterial phyla, the absence of this insert is the ancestral character state of this protein. Thus, the genetic lesion responsible for this insert was introduced in a common ancestor of the other Thermotogales after the divergence of this deep branching group of species. The other CSIs showing similar species distribution include a 5 aa deletion in the protein phosphoglucomutase/phosphomannomutase (Fig. 8b), a 6 aa deletion in the enzyme glycyl-tRNA synthetase (Sup. Fig. 48), a 2 aa deletion in the single strand DNA specific exonuclease RecJ (Sup. Fig. 49). The enzyme phosphoglucosamine mutase, involved in peptidoglycan synthesis, contains a 2 aa deletion that is uniquely present in all other Thermotogales except *P. mobilis* (Sup. Fig. 50). Additionally, in the PhoH family of proteins, which are induced in response to phosphate starvation, a 1–2 aa deletion is present in various Thermotogae except *Petrotoga*, *Kosmotoga* and *Ttog. mesG1* (Sup. Fig. 51). The *Thermosipho* and *Fervidobacterium* genera contain a 2 aa deletion, whereas *Thermotoga* species have a 1 aa indel in this position. This CSI provides evidence both for the deep branching of *Petrotoga, Kosmotoga* and *Ttog. mesG1* and it also distinguishes the *Thermosipho-Fervidobacterium* clade from the *Thermotoga* species. The information for these CSIs is provided in Table 7.

As indicated earlier, *Tt. lettingae* is separated from all other *Thermotoga* by a long branch and most of the CSIs that are commonly shared by other *Thermotoga* spp. are absent in it. The deeper branching and distinctness of *Tt. lettingae* from the other *Thermotoga* spp. is also supported by two CSIs (a 4 aa deletion in the protein phosphoribosylamine-glycine ligase (Fig. 9a) and an 8 aa deletion in the enzyme phosphoribosylformylglycinamidine synthase II (Fig. 9b)), which are shared in common by all Thermotogales species except *P. mobilis* and *Tt. lettingae*. The homologs of both these purine biosynthesis proteins were not detected in *Kosmotoga or Ttog. mesG1*.

```
(A)
                                                                       292                               324
                 Petrotoga mobilis               160901572   RKGAWFSFINEN GEEIS LGQGKTNSVSYLMENP
                 Thermosipho africanus           217077528   -R-S---YVDLK -V-H- -----N--IA--L-H-
                 Thermotogales bac. mesG1        307297691   -----YTY-S-D NN-V- -----S-GVEFMK---
                 Kosmotoga olearia               239617936   -----YTY-SQD -K--- -----S-GINF--A-S
                 Thermosipho melanesiensis       150021044   -R-S---YVDST -T-H- -----N--LD--LNH-
Thermotogae      Thermotoga lettingae            157364891   ---S--FYMADD -K-Y- -----N-V-N-FV---
12/12            Fervidobacterium nodosum        154249970   -R-S--TYEDLS -K-H- -----S-A-N--V-HS
                 Thermotoga petrophila           148270072   ---S-YYYTTLK ---V- ----SS-A-QF-KD--
                 Thermotoga naphthophila         281412047   ---S-YYYTTLK ---V- ----SS-A-QF-KD--
                 Thermotoga maritima             15644602    ---S-YYYTTLK ---V- ----SS-A-QF-KD--
                 Thermotoga sp. RQ2              170288756   ---S-YYYTTLK ---V- ----SS-A-QF-KD--
                 Thermotoga neapolitana          222099768   ---S-YYYTTLK ---V- ----GS-V-QF-K---
                 Clostridium botulinum           251779396   KS-----YGDIR       ----RE-AKQ--K---
                 Mycobacterium phlei             31540561    KS-S--TYEG-Q       -----E-ARNF-L---
                 Bacillus subtilis               296330907   KS-S-Y-YEE-R       ----RE-AKQF-K--K
                 Lactobacillus vaginalis         227529688   KS---Y-YGD-R       I---RE-AKNW-A-H-
                 Staphylococcus aureus           87126813    KS---Y-YNG-R       M----E-VKM--K---
                 Actinomyces urogenitalis        227495813   KS----TYGTDQ       -----E-ARTF-KD--
                 Mycobacterium avium             41408946    KS-S--TYEG-Q       -----E-ARTF----D
                 Streptomyces albus              291454543   KA---YTYEGDQ       -----E-ARNF-KD--
                 Edwardsiella tarda              294634757   KS---Y-YNGDK       I----A--MKF-Q---
Other bacteria   Thiomicrospira crunogena        78485933    KA---Y-YQGQK       I----D-VRQF-KD--
0/>250           Geobacter lovleyi               189426210   KS-----YNK-R       I---RE--RQF-K---
                 Erythrobacter litoralis         85374373    KS-S---YDSIR       I---RE-AKT--K---
                 Sphingomonas sp. SKA58          94496514    KS-----YDSIR       I---RE-AKT--K-H-
                 Burkholderia fungorum           30141696    KA---Y-YNG-R       I----D-AREF-R---
                 Thiomonas intermedia            296135282   KS-S-YAYNG-K       I----D-AREF-KS--
                 Leeuwen. blandensis             86141084    KS-S---YQDTK       ----RDAVKAI-KD--
                 Zunongwangia profunda           295131969   KS-S---YEDTK       ----RDAVKTI-KD--
                 Sphingomonas wittichii          148553436   KS-----YDSIR       I---RE--KV--R---
                 Sordaria macrospora             289607381   KS-----YDSVR       I---RE-AKTF-T-H-
```

```
(B)
                                                                       34                                    70
                 Fervidobacterium nodosum        154249057   ALYGLSRMLVKFLKEH VII NEDYCAFMLDVKGGSTYR
                 Thermosipho melanesiensis       150021780   ----IAK--I------ -NM EK-A---I--S----KK-
                 Thermosipho africanus           217076341   -V---TK--I------ IS- GK-A-V-V--S----KK-
                 Thermotoga sp. RQ2              170289072   -T--VA----R-I-D- I-V GK--V-VAF- -KAA-F-
                 Thermotoga maritima             15644367    -T--VA----R-I-D- I-V GK--V-VAF- -KAA-F-
Thermotogae      Thermotoga neapolitana          222099813   -V--VA------I--- I-P EK--A-VAF- -KAA-F-
12/12            Thermotoga petrophila           148270302   -T--VA----R-I-D- I-V GK--A-VAF- -RAA-F-
                 Thermotoga lettingae            157363023   -T--VL---IR---DY -K- G --T--AM- TKTR---
                 Thermotoga naphthophila         281412608   -T--VA----R-I-D- I-V GK--T-VAF- -RAA-F-
                 Petrotoga mobilis               160901568   -I--VA---L-L---Y -KK G--SII-VM- -KTT---
                 Thermotogales bac. mesG1        307299338   --F-TA---S--TR-R  MK -G--AI-AF-R- EL-K-
                 Kosmotoga olearia               239616633   -V--TA---SR-I-NY  -A EG--AL-AF- RKEA-H-
                 Bordetella pertussis            33592382    ----VVN--R-LVQD-     KAE-AVCVF- AR-K-F-
                 Burkholderia mallei             53716619    ----IVN--RRMR-DV     SAE-S-CVF- AK-K-F-
                 Geobacter sp. FRC-32            222055465   -VF-FT---LTL-Q-N     RP--V-VVF-PPRED-F-
                 Syntrophus aciditrophicus       85858429    -V--FTN--I-L-R-R     KPE-I-ITF- LK-P-F-
                 Desulfovibrio sp. FW1012B       283853836   --FMIF-L-F-L---Q     EPSHLV-F-- GR-PNF-
                 Rickettsia peacockii            238650585   ----FTS--L-L-SDF     KPKHV-VVF- S--KNF-
                 Edwardsiella ictaluri           238921711   -M--VLN--RSLIIQY     QPSHV-VVF-A-- K-F-
                 Legionella pneumophila          296105609   -I--VAN-IK-II-DY     QPEEI-VVF-A-- K-F-
                 Vibrio cholerae                 153830617   -I--VVN-IRSMMRQF     AS-RM-VIF-A-- K-F-
                 Xanthomonas oryzae              166710273   --F-VVN--RAT---R      PA-I--VV-AP- K-F-
                 Yersinia ruckeri                238755604   -M--VLN--RSL-LQY     HPSHV-VVF-A-- K-F-
Other bacteria   Veillonella parvula             282848822   -V--FLT--I-LYE-I     -P--I-VAF- --RQ-F-
0/>250           Fusobacterium ulcerans          257469719   -V--FTNT-LSII--F     SP--IGAAF---RA-LK-
                 Planctomyces maris              149177359   -IF-IT-DILNII-T-     SP--LI-AM-SS- PGT-
                 Elusimicrobium minutum          187250908   ----FV-W---LVE-K     KPH-V-VCF-SR----K-
                 Polysphondylium pallidum        281205135   -IH-YTQSILR---DF     KP--V-LCF-PR- GSF-
                 Sulfurihydro. yellowstonii      237756074   -V--FI---L-T-SVF     -TP-V-VAF-LP- K-L-
                 Aquifex aeolicus                15606735    -I--FL---FSLI-KE     RPQ-LVVVF-AP AK-K-
                 Geodermatophilus obscurus       284991623   -V--FTS--INV-RDE     QPTHV-VAF--- RK-F-
                 Holdemania filiformis           223984403   -I--FAM-IN-AVQII     QP-AMLVAF--- KH-F-
                 Chloroflexus aggregans          219850431   -VF-FAQI-LTA-A-Y     RP--V-VAF--  -R-F-
                 Thermomicrobium roseum          29569818    VVF-FAS--LEV-NDF     EP--VIVCF-T  -RSF-
```

◀ **Fig. 3** Excerpts from the sequence alignments for (**A**) RecA and (**B**) DNA polymerase I (*Pol*I) showing two additional conserved CSI (*boxed*) that are specific for the Thermotogae species. All other details are the same as in Fig. 2. Information for many other CSIs that are specific for the Thermotogae phylum is provided in Table 2

The CSIs that are commonly shared by Thermotogae and other taxa

Earlier studies on Thermotogae have indicated that many genes in their genomes have been either laterally acquired from other groups or transferred to other groups of organisms (Nelson et al. 1999; Nesbo et al. 2001, 2006, 2009; Zhaxybayeva et al. 2009). Recent comparative genomic analyses of protein sequences from 5 Thermotogae genomes (viz. *Tt. maritima*, *Tt. lettingae*, *Tt. petrophila*, *Ts. melanesiensis* and *F. nodosum*) have indicated that although these species are most closely related to the Firmicutes (Clostridia) phylum, many ORFs in these genomes showed closer affiliation to other prokaryotic taxa including Archaea, Aquificae, Proteobacteria, Deinococcus-Thermus, Bacteroidetes, etc. and in a number of cases to eukaryotic species (Zhaxybayeva et al. 2009). Our analyses have also identified many examples where a given CSI, in addition to being present in some or all Thermotogae species, is also present in other prokaryotic and eukaryotic organisms. Information for these CSIs is provided in Table 8 and two examples are shown in Figs. 10 and 11.

In the protein synthesis elongation factor EF-Tu, which is essential for protein synthesis (Rodnina et al. 1995), a 1 aa insert is commonly shared by various Thermotogae and Aquificae species (Fig. 10). These two bacterial groups also commonly share a large insert in the SecA protein (Sup. Fig. 52) (Griffiths and Gupta 2004) as well as a 1 aa deletion in a small GTP binding protein EngB (Sup. Fig. 53). In the other example shown here, a 3 aa insert in the enzyme ribonucleoside diphosphate reductase (RNR) is commonly present in various Thermotogae as well as several archaeal species belonging to the orders Thermococcales and Thermoproteales (Fig. 11). The information for other CSIs that are commonly shared between Thermotogae and some other groups of prokaryotic as well as eukaryotic organisms is provided in Table 8 and Sup. Figs. 54–65. These other prokaryotic and eukaryotic lineages which were found to contain a CSI in the same position as Thermotogae included Archaea, Aquificae, Firmicutes, Proteobacteria, Deinococcus, Fusobacteria, Dictyoglomi, Chloroflexi and certain eukaryotic algae. However, in most cases only 1–2 CSIs were shared with these taxa and it was present in only a limited number of species from these groups.

Functional significance of the Thermotogae-specific indels

Most of the discovered CSIs are present in highly conserved regions of proteins that carry out essential functions e.g. replication, transcription and translation, in various organisms. Hence, it is of much importance to understand the cellular functions of these evolutionary conserved Thermotogae-specific characteristics. For a number of proteins that contain Thermotogae-specific CSIs [viz. ribosomal protein L4 (Sup. Fig. 2), ribosomal protein L12 (Fig. 2a) and tryptophanyl-tRNA synthetase (Sup. Fig. 3)], structural information was available both from *Tt. maritima* and some other bacteria that lacked these indels. Hence, we have carried out structural comparison of these proteins to determine the location of these CSIs. The results of these studies reveal that the structures of these proteins from *Tt. maritima* and the insert-lacking bacteria are almost completely super-imposable except in the insert region (Fig. 12). Further, as observed in our earlier work (Singh and Gupta 2009; Gupta 2010), the Thermotogae-specific CSIs are present in the surface loops of these proteins.

It is commonly assumed that the surface loops in protein sequences due to their being distal from the active sites are functionally neutral or they are minimally constrained. Hence, genetic changes in these regions including addition or removal of the loop segments (i.e. indels) can occur readily and independently in different lineages without significant functional consequences (Gaget et al. 2011). Another common assumption is that smaller indels (viz. 1–2 aa) in protein sequences are of minimal functional significance. While these assumptions might be true for indels of varying lengths that are sporadically present in non-conserved regions of proteins, *they are totally incorrect for evolutionary conserved indels present in highly conserved regions of the proteins such as those that are studied in this work*. This is convincingly demonstrated by our

(A)

			320		346
Thermotoga 6/6	Thermotoga sp. RQ2	170289257	VHIAPGHGEEDYIYG	H	VQYGLPIVSPV
	Thermotoga maritima	15644113	---------------	-	-----------
	Thermotoga petrophila	148270551	---------------	-	-----------
	Thermotoga naphthophila	281412859	---------------	-	-----------
	Thermotoga neapolitana	222100204	------------V--	-	-K---------
	Thermotoga lettingae	157363246	------------EF-	-	LVN--AVI---
Other Thermotogae 0/6	Fervidobacterium nodosum	154250101	--T---------QT-		LK-N---L---
	Thermosipho africanus	217076920	--T-----A---LT-		LK-N--VL---
	Thermosipho melanesiensis	150020359	--T-----A---LT-		IK-N--VL---
	Petrotoga mobilis	160903062	--T-----M----T-		TK-N-QVI---
	Kosmotoga olearia	239616422	--T---------GT-		LR-R--VI---
	Thermotogales bac. mesG1	307298670	--M--AF--D-N-L-		RKF---LLQL-
Other species 0/>250	Staphylococcus aureus	257425247	--T------D---V-		QK-E--VI--I
	Bacillus pseudofirmus	288553142	--T---------LV-		QK---DVLC--
	Moorella thermoacetica	83589714	--T-----L---EV-		MR-H--VL--L
	Arthrospira platensis	284053011	--T-----Q----V-		QR-----L---
	Cyanothece sp. PCC 7425	220908492	--T-----Q---VV-		QH-----L---
	Microcystis aeruginosa	166363857	--T-----Q----V-		QR-----L---
	Synechococcus sp. PCC 7002	170078363	--T-----Q----T-		QK-----L---
	Anabaena variabilis	75909978	--T-----Q---VV-		LR-----LA--
	Fusobacterium varium	253581993	--T-----QD--VV-		-R----VI--I
	Fusobacterium ulcerans	257469006	--T-----QD--VV-		-R----VI--I
	Anaeromyx. dehalogenans	220915135	--T-----Q---VV-		LR---EVLN--
	Ferroglobus placidus	288931973	--------L---EL-		--H--EVFN--
	Acholeplasma laidlawii	162447562	--T------D--FV-		KK-N-DLL---
	Arabidopsis thaliana	222422851	--T-----Q---AT-		LK----L----
	Vitis vinifera	270238845	--T-----Q---VT-		MK-----L---
	Oryza sativa	218191678	--T-----Q----T-		LK---------
	Sorghum bicolor	242063240	--T-----Q----T-		LK-----I---
	Ricinus communis	255551751	--T-----Q----T-		MK----VL---
	Populus trichocarpa	224107153	--T-----R---VT-		LK-----I---
	Physcomitrella patens	168008836	--T-----Q----T-		LK----LL---

(B)

			577		618
Thermotoga 5/6	Thermotoga neapolitana	222099188	RNEKRMLQEAVDALIHNG	SDSEGKR	SRRAVLKDRNGRPLKSL
	Thermotoga sp. RQ2	170288277	------------------	-------	-----------------
	Thermotoga maritima	15643225	------------------	-------	-----------------
	Thermotoga petrophila	148269601	------------------	-------	-----------------
	Thermotoga naphthophila	281411681	------------------	-------	-----------------
	Thermotoga lettingae	157363342	---------S--S--Y--		RIGKAVA-----A----
Other Thermotogae 0/6	Thermosipho melanesiensis	150020399	K-----------S--Y--		RMGKAVT-----A----
	Thermosipho africanus	217077412	K-----------S--Y--		RMGKAVT----------
	Fervidobacterium nodosum	154250446	------------N--F--		KIGKAYV-----Q----
	Kosmotoga olearia	239618201	------------S--Y--		RVGKAVT-KS-------
	Thermotogales bac. mesG1	307297334	------------S--Y--		RVGKAVS-KS-------
	Petrotoga mobilis	160901838	-----I--Q---S-FY--		RVGKPMT---R---R--
Other species 0/>250	Lawsonia intracellularis	94987346	--------------FD--		R-GRAIAGT--------
	Desulfovibrio vulgaris	46581333	--------------FD--		R-GRAITGT--------
	Aquifex pyrophilus	7531199	---------------D--		K-GNPV- Q--------
	Sulfurihydro. yellowstonii	237755890	---------------D--		R-GRMVT Q-N------
	Clostridium botulinum	188590157	---------------D--		R-GRPVTGPGN------
	Halothermothrix orenii	220930958	---------------D--		R-GRPVTGAGN------
	Ammonifex degensii	260893379	---------------D--		R-GRPVTGPGN------
	Moorella thermoacetica	83591283	---------------D--		R-GRPVTGPGN------
	Defer. desulfuricans	291280154	--------------FD--		R-GR-IRGS-K------
	Psychroflexus torquis	91218673	---------S----LD--		R-GRAILGT-K------
	Thrmavib. acidaminovorans	269792801	---------S-----D--		R-GKPVPGAG-------
	Anaeroba. hydrogeniformans	289523376	---------------D--		RVGKAVLGAGN------
	Aminobacterium colombiense	294101621	---------S-----D--		RMGKAVLGAGN------
	Koribacter versatilis	94971701	--------------FD--		R-GR--RGA-N------
	Acidobacterium capsulatum	225874480	--------------FD--		R-GR--RGA-N------

◀ **Fig. 4** Partial sequence alignments of (**A**) Isoleucyl-tRNA synthetase and (**B**) RNA polymerase β' subunit (RpoC) showing two CSIs of different lengths (*boxed*) in highly conserved regions of these proteins that are specific for species from the genus *Thermotoga*. The sequence information for only a limited number of species is presented here. The other CSIs showing similar specificity are listed in Table 3

recent work on a number of CSIs in the GroEL and DnaK proteins. The GroEL/Hsp60 protein contains a 1 aa insert that is specifically present in many phyla of Gram-negative (diderm) bacteria (viz. Alpha-, Beta-, Gamma-, Delta- and Epsilon-proteobacteria,

Table 3 Conserved Signature Indels that are specific for the genus *Thermotoga*

Protein	Isoleucine-tRNA synthetase (IleRS)	RNA polymerase β' subunit (RpoC)	RNA polymerase β' subunit (RpoC)	Purine nucleoside phosphorylase I (PNP)	Patatin like protein
GenBank Identifier	170289257	222099188	148269601	254485330	254483880
Accession no.	YP_001739495	YP_002533756	YP_001244061	ZP_05098542	ZP_05097097
Indel/size	1 aa ins	7 aa ins	19 aa ins	3 aa ins	3 aa del
Indel position[a]	320–346	577–618	1013–1069	43–81	6–40
Figure no.	Fig. 4a	Fig. 4b	Sup. Fig. 15	Sup. Fig. 16	Sup. Fig. 17
Tt. petrophila	+	+	+	+	+
Tt. maritima	+	+	+	+	+
Tt. neapolitana	+	+	+	+	+
Tt. sp. RQ2	+	+	+	+	+
Tt. naphthophila	+	+	+	+	+
Tt. lettingae	+	–	–	–	–
F. nodosum	–	–	–	–	–
Ts.melanesiensis	–	–	–	–	–
Ts. africanus	–	–	–	–	–
P. mobilis	–	–	–	–	–
K. olearia	–	–	–	–	–
Ttog. mesG1	–	–	–	–	–
Other species with indel[b]	0/250	0/250	0/250	0/250	0/250

Protein	Flagellar motor switch protein (FliM)	tRNA modification GTPase (TrmE)	Metalloendopeptidase glycoprotease family protein	Aminotrans-ferase class I and II (AAT)	Dihydroxy-acetone kinase phosphatase (DAK Phosphatase)
GenBank Identifier	15643443	15643037	148269915	281412530	170288844
Accession no.	NP_228487	NP_228080	YP_001244375	YP_003346609	YP_001739082
Indel/size	1 aa del	1 aa ins	2 aa del	2 aa del	1 aa del
Indel position[a]	64–95	290–316	40–87	124–176	81–129
Figure no.	Sup. Fig. 18	Sup. Fig. 19	Sup. Fig. 20	Sup. Fig. 21	Sup. Fig. 22
Tt. petrophila	+	+	+	+	+
Tt. maritima	+	+	+	+	+
Tt. neapolitana	+	+	+	+	+
Tt. sp. RQ2	+	+	+	+	+
Tt. naphthophila	+	+	+	+	+
Tt. lettingae	–	–	–	–	–
F. nodosum	–	–	–	–	–
Ts.melanesiensis	–	–	–	–	–
Ts. africanus	–	–	–	–	–
P. mobilis	–	–	–	–	–
K. olearia	0[c]	–	–	–	–
Ttog. mesG1	0[c]	–	–	–	–
Other species with indel[b]	0/250	1/250	0/250	0/250	0/250

Table 3 continued

Protein	ATP-dependent protease La	ATP-dependent protease La	Adenylate kinase (adk)	RNA polymerase β subunit (RpoB)	7-Cyano-7-deazaguanine reductase (QueF)
GenBank Identifier	170289086	254484975	281412750	281411680	15643554
Accession no.	YP_001739324	ZP_05098190	YP_003346829	YP_003345759	NP_228600
Indel/size	2 aa del.	3 aa del.	1 aa ins.	6 aa ins.	1 aa ins.
Indel position[a]	519–561	561–596	28–64	912–954	45–84
Figure no.	Sup. Fig. 23	Sup. Fig. 24	Sup. Fig. 25	Sup. Fig. 26	Sup. Fig. 27
Tt. petrophila	+	+	+	+	+
Tt. maritima	+	+	+	+	+
Tt. neapolitana	+	+	+	+	+
Tt. sp. RQ2	+	+	+	+	+
Tt. naphthophila	+	+	+	+	+
Tt. lettingae	–	–[d]	–	–	0[c]
F. nodosum	–	–	–	+	0[c]
Ts. melanesiensis	–	–	–	–	+
Ts. africanus	–	–	–	–	0[c]
P. mobilis	0[c]	–	–	–	0[c]
K. olearia	–	–[d]	–	–	0[c]
Ttog. mesG1	–	–[d]	–	–	0[c]
Other species with indel[b]	0/250	0/250	0/250	0/250	1/250

[a] The indel position provided indicates the region of the protein containing the CSI

[b] The presence or absence of the CSIs in the top 250 Blast hits was examined. The number of non-Thermotogae organisms, which were observed to contain the CSI, is indicated. Species containing longer or shorter CSIs than indicated were not included in the total

[c] Homologous sequences corresponding to the region containing the CSI's could not be identified in these species

[d] A 1 aa insert was present in the indicated species rather than the 3 aa insert found in some other Thermotogae

Aquificae, Chlamydiae–Verrucomicrobiae–Planctomyctes, Bacteroidetes–Chlorobi–Fibrobacter, Spirochaetes and Cyanobacteria), but which is absent in Chloroflexi, Deinococcus–Thermus, Fusobacteria, Thermotogae, Actinobacteria and Firmicutes (Gupta 1998, 2000; Singh and Gupta 2009). When this CSI was originally identified, sequence information for GroEL was available from only a small number (<50) of bacteria (Gupta 1997, 1998; Gupta et al. 1999). However, our recent analyses show that this CSI is present in all of the >1700 sequences that are available from the indicated groups of Gram-negative bacteria, but it is lacking in virtually all of the >1400 sequences from other phyla of bacteria with <10 exceptions. These exceptions are seen mostly in cases where two homologs of the proteins are present within the same species, one containing the indel and the other lacking the indel. Thus, these exceptions could be caused by either LGTs or other non-specific causes such as incorrect information for the source species for some sequences. The species distribution profile of this CSI provides strong evidence that this indel is a highly reliable characteristic of the above noted phyla of Gram-negative bacteria and it occurred once in a common ancestor of them and since then it has not been lost from any species from these groups (Gupta 1998, 2000; Griffiths and Gupta 2004; Singh and Gupta 2009; Gupta and Shami 2011). At the same time, the absence of this CSI in all other bacterial phyla strongly indicates that the genetic change that gave rise to this CSI was a highly specific event and that similar changes, even though it involves a protein surface loop, do not occur randomly or frequently in different lineages. Similar results were observed for three other CSIs in the DnaK/Hsp70 protein that were examined (Gupta

Fig. 5 Excerpts from the sequence alignments for (**A**) DNA polymerase III, (**B**) DNA gyrase subunit B and (**C**) the cell shape determining protein MreB showing three different CSIs in conserved regions of these proteins that are uniquely present in *Thermosipho* and *Fervidobacterium* genera. The other CSIs showing similar specificity are listed in Table 4

▶

(A)			861		883
Thermosipho and Fervidobacterium 3/3	Fervidobacterium nodosum	154250388	HYRCPDCKYFEL		HEEFGSGYDLP
	Thermosipho africanus	217077387	--I--N---L-F		SK-V-------
	Thermosipho melanesiensis	150020931	--L--K---L-F		SKDM-------
Other Thermotogae 0/9	Thermotogales bac. mesG1	307298437	--L--E-GKSVF	P	ETDLE------
	Kosmotoga olearia	239617281	--V--A--HS-F	VL	DGSVE-----K
	Petrotoga mobilis	160902928	--I--S--NV-F	FD	K--I------V
	Thermotoga neapolitana	222099063	-----R-----I	VE	DDRY-A-----
	Thermotoga sp. RQ2	170288161	-----E-----V	VE	DDRY-A-----
	Thermotoga maritima	15643342	-----E-----V	VE	DDRY-A-----
	Thermotoga lettingae	157363106	--F-SK-R-V-F	IE	SN-Y-------
	Thermotoga naphthophila	281411795	-----E----KV	VE	DDRY-A-----
	Thermotoga petrophila	148269487	-----E----KV	VE	DDRY-A-----
Other species 0/>250	Bacillus licheniformis	52080261	--V----RHS-F	FN	DGSV---F---
	Enterococcus faecalis	21314372	--Y--E-Q-S-F	YE	DGSY---F-M-
	Lactococcus lactis	116513078	-----E-Q---C	YD	DGS------M-
	Streptococcus pyogenes	21911223	--V--S-QHS-F	IT	DGSV-------
	Anaerocellum thermophilum	222529790	--I--N---S-F	IT	DGSV-C----E

(B)			305		336
Thermosipho and Fervidobacterium 3/3	Thermosipho melanesiensis	150020566	GEDIREGIVAIVSVLM	S	KTPEFEGQTKSKLGN
	Thermosipho africanus	217076995	-D-V---M--VI----	-	S--------------
	Fervidobacterium nodosum	154249648	---L---MT-V-N---	M	G--Q-----------
Other Thermotogae 0/12	Thermotoga lettingae	157363918	---V---LI---T---		-E------------S
	Thermotoga petrophila	14587796	---V---LT-VI--YV		-N-------------
	Thermotoga maritima	14587800	---V---LT-VI--YV		-N-------------
	Thermotoga neapolitana	14587802	-D-V---LT-VI--YV		-N-------------
	Thermotoga naphthophila	14587798	---V---LT-V---YV		-N-------------
	Thermotoga sp. RQ2	170287901	---V---LT-VI--YV		-N-------------
	Kosmotoga olearia	239616758	---V---LT-VL--FV		-E-Q------A----
	Thermotogales bac. mesG1	307298739	---L---LS--L--FV		-E-Q------A----
	Petrotoga mobilis	160901666	---V---LL--IHIK-		PN-V------GR--S
	Thermotoga sp. KU10	19909679	---V---LT-V---YV		-N-------------
	Thermotoga sp. KU11	19909681	---V---LT-V---YV		-N-------------
	Thermotoga sp. KU1	19909675	---V---LT-VI--YV		-N-------------
Other species 0/>250	Clostridium acetobutylicum	15893304	-------LT-V---KL		TE-Q------T----
	Halothermothrix orenii	220930856	--------T--I--RL		TD-Q------T----
	Mycobacterium fortuitum	308153164	-------LA-VI--KV		SQ-Q------T----
	Nocardia acidivorans	261343282	-D-----LA-----KI		SD-Q------T----
	Rhodococcus tukisamuensis	108860569	-D-----LA-----KV		SE-Q------T----
	Streptomyces pallidus	20387202	-D-----LT--I--KL		SE-Q------T----
	Thermomicrobium roseum	221632838	---V---LT--I--ML		PE-Q------T----
	Mycoplasma feliminutum	183238622	---T---L-CVI--KL		-Q-------------
	Methanocella paludicola	282163382	-------LT--I--KL		TN-Q------T----
	Chlamydia muridarum	15835080	-------L------KV		PN-Q------Q----

(C)			47		73
Thermosipho and Fervidobacterium 3/3	Fervidobacterium nodosum	154249602	EAKEMLGKTPED	K	ILAVKPMRDGVIAD
	Thermosipho africanus	217076899	------------	T	-T--R--K------
	Thermosipho melanesiensis	150020362	-----I-----E	S	-I--R--K------
Other Thermotogae 0/9	Thermotoga naphthophila	281412684	---K-------G		LK-IR--K------
	Thermotoga neapolitana	222100101	---K-------G		LK-IR--K------
	Thermotoga maritima	15644292	---K-------G		LK-IR--K------
	Thermotoga petrophila	148270378	---K-------G		LK-IR--K------
	Thermotoga sp. RQ2	170288996	---K-------G		LK-IR--K------
	Thermotoga lettingae	157363237	---K-------Y		FK-I---K------
	Petrotoga mobilis	160902865	---K-I----AN		-I-IR-LKE-----
	Thermotogales bac. mesG1	307297756	D----I------		-K--R-I-------
	Kosmotoga olearia	239617857	Q----I----K-		-M--R-V-------
Other species 0/>250	Clostridium botulinum	148378169	--R--I-R--GN		-V-IR-------S-
	Bacillus coagulans	229542503	--RR-V-R--GN		-V-IR-LK------
	Listeria monocytogenes	226225073	--RD-V-R--G-		-T-I---K------
	Chloroflexus aggregans	219848103	---A-V----AN		-V--R-LK------
	Defer. desulfuricans	291279548	---S---R--AN		-V-IR--K-----N
	Sulfurihydro. yellowstonii	237755903	-----I-----H		-QVIR-LK------
	Brachyspira murdochii	296126110	---R----V-NS		-A-IR---------
	Fusobacterium sp. D12	257462533	----------DS		-V--R-LSE-----
	Streptomyces roseosporus	239941120	---K-I-R--GN		-V--R-LK------

Table 4 Conserved Signature Indels that are specific for *Thermosipho* and *Fervidobacterium* genera

Protein	DNA polymerase III (PolC)	DNA gyrase, subunit B (GyrB)	Cell shape determining protein (MreB)	Cell shape determining protein (MreB)	Chromosome segregation protein (SMC)
GenBank Identifier	154250388	150020566	154249602	154249602	150020781
Accession no.	YP_001411213	YP_001305920	YP_001410427	YP_001410427	YP_001306135
Indel/size	2 aa del	1 aa ins	1 aa ins	1 aa del	2 aa del
Indel position[a]	861–883	305–336	47–73	69–113	1097–1130
Figure no.	Fig. 5a	Fig. 5b	Fig. 5c	Sup. Fig. 28	Sup. Fig. 29
Tt. petrophila	–	–	–	–	–
Tt. maritima	–	–	–	–	–
Tt. neapolitana	–	–	–	–	–
Tt. sp. RQ2	–	–	–	–	–
Tt. naphthophila	–	–	–	–	–
Tt. lettingae	–	–	–	–	–
F. nodosum	+	+	+	+	+
Ts. melanesiensis	+	+	+	+	+
Ts. africanus	+	+	+	+	+
P. mobilis	–	–	–	–	–
K. olearia	–	–	–	–	–
Ttog. mesG1	–	–	–	–	–
Other species with indel[b]	0/250	0/250	0/250	0/250	0/250

Protein	Diguanylate cyclase (DGC)	Glucose-1-phosphate thymidyltrans ferase (RmlA)	Glucose-1-phosphate thymidyltransferase (RmlA)	ExsB family protein	Ornithine decarboxylase (ODC)
GenBank Identifier	154250171	154250125	154250125	150021480	150021213
Accession no.	YP_001410996	YP_001410950	YP_001410950	YP_001306834	YP_001306567
Indel/size	4 aa ins	1 aa ins	1 aa ins	1 aa del	2 aa ins
Indel position[a]	83–134	123–163	240–276	71–108	127–154
Figure no.	Sup. Fig. 30	Sup. Fig. 31a	Sup. Fig. 31b	Sup. Fig. 32	Sup. Fig. 33
Tt. petrophila	–	–	–	–	–
Tt. maritima	–	–	–	–	–
Tt. neapolitana	–	–	–	–	–
Tt. sp. RQ2	–	–	–	–	–
Tt. naphthophila	–	–	–	–	–
Tt. lettingae	–	–	–	–	–
F. nodosum	+	+	+	+	+
Ts. melanesiensis	+	+	+	+	+
Ts. africanus	+	+	+	+	+
P. mobilis	0[c]	–	–	–	–
K. olearia	0[c]	0[c]	0[c]	–	–
Ttog. mesG1	0[c]	0[c]	0[f]	–	–
Other species with indel[b]	0[d]	1/100	0/100	0[d]	0/100

Table 4 continued

Protein	Phosphomannomutase (PMM)	Basic membrane lipoprotein	Phosphate butyryltransferase (Ptb)	Glucose inhibited division protein A (GidA)
GenBank Identifier	150020471	150019927	150020016	154250326
Accession no.	YP_001305825	YP_001305281	YP_001305370	YP_001411151
Indel/size	3–4 aa del	5 aa ins	1 aa ins	1 aa del
Indel position[a]	307–334	60–113	151–179	281–311
Figure no.	Sup. Fig. 34	Sup. Fig. 35	Sup. Fig. 36	Sup. Fig. 37
Tt. petrophila	–	–	–	–
Tt. maritima	–	–	–	–
Tt. neapolitana	–	–	–	–
Tt. sp. RQ2	–	–	–	–
Tt. naphthophila	–	–	–	–
Tt. lettingae	–	–	–	–
F. nodosum	+[e]	+	+	+
Ts. melanesiensis	+	+	+	+
Ts. africanus	+	+	+	+
P. mobilis	–	–	–	+
K. olearia	–	–	–	–
Ttog. mesG1	–	–	–	–
Other species with indel[b]	0/100	0/100	0/250	0/250

[a] The indel position provided indicates the region of the protein containing the CSI

[b] The presence or absence of the CSIs in the top 250 (or 100) Blast hits was examined. The number of non-Thermotogae organisms, which were observed to contain the CSI, is indicated. Species containing larger or shorter CSI than indicated were not included in the total

[c] Homologous sequences corresponding to the region containing the CSIs could not be identified in these species

[d] BLAST searches provided only the indicated Thermotogae sequences as significant homologous matches to the query sequence

[e] A 3 aa insert is present in the indicated species rather than the 4 aa insert found in *Thermosipho*

[f] A 1 aa deletion is present in the organism rather than the 2 aa deletion found in the *Thermosipho–Fervidobacterium* clade

1998, 2000; Griffiths and Gupta 2004; Singh and Gupta 2009; Gupta and Shami 2011). These results provide convincing evidence that the conserved CSIs in protein sequences, despite their locations in the surface loops, are highly reliable and stable genetic characteristics of different lineages and they are not commonly lost or acquired, as is erroneously assumed.

The evolutionary conservation of these CSIs in all species from these taxa over eons of time indicates that there should be strong selection pressure for retention of these genetic changes and hence they likely serve important functions in these species. This inference is strongly supported by our recent work where the functional significances of a number of CSIs in the *groEL* and *dnaK* genes for cellular growth were examined by complementation studies with temperature-sensitive (*Ts*) mutants of *E. coli* (Singh and Gupta 2009). The results of these studies demonstrated that deletion as well as most changes in these CSIs (including replacement of the 1 aa insert in GroEL with other amino acids) were incompatible with the growth of *E. coli* cells (Singh and Gupta 2009). These results established that the evolutionary conserved CSIs are essential for the groups of bacteria where they are found. Recent analyses of available protein structures have shown that the surface loops in proteins are important determinants in mediating protein–protein interactions (Itzhaki et al. 2006; Akiva et al. 2008; Hormozdiari et al. 2009). Thus, these CSIs could be involved in facilitating novel protein–protein interactions that are unique and essential for these groups of bacteria. Recently, the importance of two large CSIs in the RpoC and Gyrase B proteins in mediating protein–protein interactions that were essential for the functioning of these proteins was experimentally demonstrated (Chlenov et al. 2005; Schoeffler et al.

```
(A)                                                                           150                              187
Thermosipho Genus   Thermosipho melanesiensis        150019947   ITASGGAVRN    YKNIENLTPSDILNHPVWNMGQKITVDS
2/2                 Thermosipho africanus            217076380   -------I-D    KE--DD-SVE--------S--K------
                    Fervidobacterium nodosum         154249632   ---------D LP LDK-AG---EEV-K--T----GR-----
                    Thermotoga maritima              15643651    L------L-D WK ISK-DRAR-E-V-K-------AR-----
                    Thermotoga neapolitana           222100661   L------L-D WD LEK--TA--N-V-K----S--AR-----
                    Thermotoga naphthophila          281411481   L------L-D WK ISK-DRAR-E-V-K-------AR-----
Other Thermotogae   Thermotoga sp. RQ2               170287845   L------L-D WE ISK-DRAR-E-V-R-------AR-----
                    Thermotoga petrophila            148269183   L------L-D WE ISK-DRAR-E-V-R-------AR-----
0/10                Thermotoga lettingae             157363521   --S----L-D WE LH-LH-AK-K-V-K----K--ER-----
                    Kosmotoga olearia                239617300   -------L-D YP LDR-NEV-LK-V------S--VR-----
                    Petrotoga mobilis                160903368   L------L-D YP VETL--VPVEEV------S--KR-----
                    Thermotogales bac. mesG1.        307298637   -------L-D WP LER-QDA-IR-V-R----S--NR--I--
                    Thermodesulfovibrio yellowstonii 206890733   L-----PF-G KK SYE---V--QEA----K-K--KR--I--
                    Chlamydia trachomatis            166154292   L-----PL-- KS KEELQKVSLQEV-R-------P------
                    Ehrlichia ruminantium            57239203    --G----LLY MD -DQMR-I-VQETIK----K--K--S---
                    Anaplasma marginale              56416872    L-----PFLR WT REQMQAV----A-A----K--R--S---
                    Wolbachia endosymbiont           58584439    L-----SFL- YS LEQLR-V-VGQA-S--T----K--S---
                    Francisella tularensis           56708600    L-----PF-D KQ LHELTDV--EQAC---N-Q--R--S---
                    Helicobacter bilis               237751295   -------L-D FS LEK-PYA-L--V-K--T-S------I--
                    Campylobacter jejuni             153951073   -------FYK YK I-DLNQVSVK-A-K--N----A---I--
                    Bacillus coagulans               229544428   ------SF-D YT REQL-TV-VK-A----N-S--A------
Other species       Clostridium botulinum            253681399   L-----PF-- RK KEELI-I--EEAIK--K----K--SI--
0/>100              Fusobacterium nucleatum          254302826   ------TF-G KD LEYL--V-VEQA-K--N-S--K---I--
                    Fibrobacter succinogenes         261416248   ------PF-E WP IEKF-KI-VA-A------S--K---I--
                    Bifidobacterium longum           213691674   V-----PF-G WK RADM--I--EQA-H--T----PVV-IN-
                    Micromonospora aurantiaca        302865922   V-----PF-G RR RDELTAV--EQA-A--T----PVV-IN-
                    Salinispora arenicola            159036941   V-----PF-G WR RDELTHV--EQA-A--T-D--PV--IN-
                    Selaginella moellendorffii       302816053   L------F-D WP VERLKDVK-A-A-K--N-S--K------
                    Synechococcus sp. WH 8102        33865232    L------F-D WK AEDL--A-VA-ATS--N-S--R------
                    Blastopirellula marina           87306772    L-----PF-H HS QAQL-QV-TAEA----T-K--P------
                    Alistipes shahii                 291513734   V-C----F-D FR AEELA-V-VEQA-R--Q-D--A---I--
```

```
(B)                                                                           434                       463
Thermosipho Genus   Thermosipho melanesiensis        150020512   LGAAYLAGLATGVWKTKEEI    KWNLNKRFEP
2/2                 Thermosipho africanus            217076764   ----------V-L--S----    A---------
                    Thermotoga neapolitana           222100038   ----------V-Y--DQ--- SK L-Q-DR----
                    Thermotoga lettingae             157363960   ----------V-Y--DHS-- KK Q-R-DR----
                    Thermotoga maritima              15644181    ----------V-Y--DQ--- AS L-Q-DR----
Other Thermotogae   Thermotoga petrophila            148270494   ----------V-Y--DQ--- AS L-Q-DR----
0/8                 Thermotoga naphthophila          281412800   ----------V-Y--DQ--- AS L-Q-DR----
                    Thermotoga sp. RQ2               170289111   ----------V-Y--DQ--- AS L-Q-DR----
                    Thermotogales bac. MesG1.        307297314   ----M---IC--L-DLESLK N  -RISEVV---
                    Petrotoga mobilis                160901768   ----------V-Y-NGQ--L LR --KRDAL-T-
                    Macrococcus caseolyticus         222151141   ----------V-Y--S-D-- RN R---E-Q-D-
                    Bacillus subtilis                296332350   --------I-V-F--DRS-- AN Q-K-D-----
                    Staphylococcus lugdunensis       289550906   ----------V-F-DS-D-- AH R-K-E-E---
                    Faecalibacterium prausnitzii     160944948   ---------------SR--- SH L-SC-QL---
                    Thermoanaerobacter mathranii     297545189   ---S------V-F-NSR--- EK N--VD-H---
                    Selenomonas sputigena            260886262   --------------AS--DL KK S-Q-DT--T-
                    Desulfovibrio desulfuricans      78355457    ----------V-F--SED-- TA M-Q-DR----
                    Anaeromyxobacter dehalogenans    86159876    ----F-----V-F--DR--- RR A-QVG-----
Other species       Leptotrichia hofstadii           260890399   ----------V-F--N---- KK N-R---E-T-
                    Fusobacterium periodonticum      262065913   ------------F-EN---- KQ --V-D-E-T-
0/>100              Clostridium ramosum              167755520   ------------F-AS---- KN N-K--YE-V-
                    Stenotrophomonas maltophilia     190575956   ----------V-F-SSR-Q- AA Q-G-DR----
                    Xanthomonas campestris           66766713    ------------F-SSR--- AK Q-AVDR----
                    Cupriavidus metallidurans        94311176    ------------Y-TDP--- TR Q-QVEQ----
                    Herminiimonas arsenicoxydans     134093498   ----------VAF-ES---- AT Q-QMER----
                    Thermus thermophilus             6016137     ----------V-Y-NSRDD- AA Q-Q-ER----
                    Leeuwenhoekiella blandensis      86142038    ----------V-F-ES---- QE M-QED-----
                    Sphaerobacter thermophilus       269929022   ----------V-F--DEQD- AQ N-Q--R----
```

Fig. 6 Partial sequence alignments for the proteins (**A**) 1-deoxy-D-xylulose-5-phosphate reductoisomerase and (**B**) Glycerol kinase showing two different CSIs that are specific for the genus *Thermosipho*. Other CSIs showing similar specificity are listed in Table 5

Table 5 Conserved Signature Indels that are specific for the *Thermosipho* genus

Protein	1-deoxy-D-xylulose-5-phosphate reducto-isomerase (DXR)	Glycerol kinase (GK)	Aspartate ammonia-lyase (AspA)	Glutamine deaminase chemoreceptor (CheD)	Beta lactamase domain-containing protein	Radical SAM domain-containing protein	NADH dehydrogenase (NDH)	Amido-hydrolase
GenBank Identifier	150019947	150020512	150021238	150021310	150021733	150021382	150021053	150020398
Accession no.	YP_001305301	YP_001305866	YP_001306592	YP_001306664	YP_001307087	YP_001306736	YP_001306407	YP_001305752
Indel/size	2 aa del	2 aa del	3 aa del	2 aa del	1 aa ins	4 aa del	2 aa del	1 aa ins
Indel position[a]	150–187	434–463	102–134	64–97	156–197	451–485	361–386	51–97
Figure no.	6a	6b	Sup. Fig. 38	Sup. Fig. 39	Sup. Fig. 40	Sup. Fig. 41	Sup. Fig. 42	Sup. Fig. 43
Tt. petrophila	–	–	–	–	–	–	–	–
Tt. maritima	–	–	–	–	–	–	–	–
Tt. neapolitana	–	–	–	–	–	–	–	–
Tt. sp. RQ2	–	–	–	–	–	–	–	–
Tt. naphthophila	–	–	–	–	–	–	–	–
Tt. lettingae	–	–	–	–	–	–	–	–
F. nodosum	–	0[c]	–	–	–	–	–	–
Ts. melanesiensis	+	+	+	+	+	+	+	+
Ts. africanus	+	+	+	+	+	+	+	+
P. mobilis	–	–	0[c]	0[c]	–	0[c]	–	+
K. olearia	–	0[c]	–	0[c]	0[c]	–	–	+
Ttog. mesG1	–	–	0[c]	0[c]	0[c]	–	–	–
Other species with indel[b]	0/100	0/100	0/100	2/100	0/100	0/100	2/100	0/100

[a] The indel position provided indicates the region of the protein containing the CSI

[b] The presence or absence of the CSIs in the top 100 Blast hits is indicated here. The number of non-Thermotogae organisms, which were observed to contain the CSI, is specified. Species containing larger or shorter CSI were not included in the total

[c] Homologous sequences corresponding to the region containing the CSI's could not be identified in these species

2010). Based on these results, it is expected that the Thermotogae-specific CSIs identified in the present work will also play important functional roles in these bacteria.

Discussion

The Thermotogae species are presently distinguished from other bacteria primarily on the basis of their distinct branching in phylogenetic trees. No molecular or biochemical characteristics are known that can clearly distinguish species from this phylum from all other bacteria. Further, although this phylum is comprised of at least 9 genera, due to lack of reliable information about their interrelationships, they are all placed into a single family. We report here for the first time >60 molecular signatures that are distinctive characteristics of either all sequenced Thermotogae or a number of well-defined subgroups within this phylum. These signatures provide novel means for different types of studies on these bacteria. Of the signatures described here, 18 CSIs in widely distributed proteins are largely specific for the Thermotogae species (Table 2). Due to their Thermotogae-specificity, the rare genetic changes responsible for them most likely occurred 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). Thus, these CSIs represent molecular synapomorphies that distinguish Thermotogae species from all other prokaryotic and eukaryotic organisms and provide strong evidence that this phylum is distinct from all other taxa including Firmicutes and *Archaea* (Olsen et al. 1994; Reysenbach 2001; Ludwig and Klenk 2005; Huber and Hannig 2006; Zhaxybayeva et al. 2009).

(A)

			299		334
Petrotoga-Kosmotoga Clade 3/3	Petrotoga mobilis	160901686	VSSWDERSLSWLSAVKQF	ADENH	KNHFIIGSGINTY
	Thermotogales bac. mesG1.	307297528	SG--N--FSA-YNSIF-W	LESEN	-LRIPF----G-F
	Kosmotoga olearia	239616427	SK--K--FSA-FNSIY-W	I-PS-	PSRI-F-T--G-F
Other Thermotogae 0/9	Fervidobacterium nodosum	154249210	---R---F---F-TIYIW		---K-L-Q--G--
	Thermotoga neapolitana	222100746	----H--L-----S-Y-W		-D-R-F----G--
	Thermotoga lettingae	157363637	---I---L-A--G-IY-W		-ENKL--T--G--
	Thermotoga maritima	15643569	----H--L---F-SIY-W		RT-K-L-T--G--
	Thermotoga sp. RQ2	170287927	----H--L---F-SIY-W		RT-K-L-T--G--
	Thermotoga petrophila	148269267	----H--L---F-SIY-W		RT-K-L-T--G--
	Thermotoga naphthophila	281411563	----H--L---F-SIY-W		RT-K-L-T--G--
	Thermosipho africanus	217076706	---K---F---FTSLELW		ED-K-F----A--
	Thermosipho melanesiensis	150020344	---R---F---FTSLELW		ED-K-F-T--G--

(B)

			47		78
Petrotoga-Kosmotoga Clade 3/3	Petrotoga mobilis	160901682	GTSAPELVVTISSSIKG		ASVGLGNVLGSNVAN
	Thermotogales bac. mesG1.	307299187	-------A-SVQAAF--		SDIA--------I--
	Kosmotoga olearia	239618486	-------A-S-QAAV--		SGIAI-------I--
Other Thermotogae 0/8	Thermosipho melanesiensis	150021452	---L----ASLV-V---	H	S--SIS--V---I--
	Thermosipho africanus	217077958	---L--I--SVI-TV--	E	NDILV--IV----F-
	Thermotoga neapolitana	222100834	---L----SS-V-AS--	Y	T-IAIS--V---I--
	Thermotoga maritima	15643487	---L-----S-V-TA--	E	SDILV--IV----F-
	Thermotoga petrophila	148269350	---L----SSLV-A---	Y	S-IAIS--V---I--
	Thermotoga sp. RQ2	170288008	---L----SSLV-A---	Y	S-IAIS--V---I--
	Thermotoga naphthophila	281411953	---L----SSLV-A---	Y	S-IAIS--V---I--
	Fervidobacterium nodosum	154249428	-------A-N-V-----	T	SNIS----I---IF-
Other species 0/>250	Bacteroides intestinalis	189464616	-------T-SV--AL--	S	-DIAI---V---IF-
	Porphyromonas gingivalis	34540782	-------T-SLMAAL--	S	-DIAI---I---IF-
	Dinoroseobacter shibae	159046212	---T---L-SVNAALD-	A	PEIA----V---I--
	Parvibaculum lavamentivorans	154252726	-------L-S-RAALA-	S	PGIA----V---I--
	Rhodobacter sphaeroides	126462063	---T---L-SLNAALD-	A	S-IAV-------I--
	Roseovarius sp. TM1035	149204125	---T---L-SVKAALD-	A	SEIA----V---I--
	Methanosarcina acetivorans	20090856	---L------V-AARQ-	Y	G-IA----I---IT-
	Arthrospira platensis	284051069	---L--------AAR--	D	-E-AI---V---IF-
	Lyngbya sp. PCC 8106	119487014	---L---A--F-A-RR-	D	-E-AI-------IF-
	Clostridium perfringens	168213535	-------A-S--A-L--	S	NDITM-------LF-
	Blautia hansenii	260588011	---L----TS-VAAR--	E	SDIA----V---IF-
	Saccharophagus degradans	90022815	------VM-S--A-L-D	A	GDLAI--A----I--
	Shewanella baltica	126172603	-------A-SVK-ALA-	N	SGIA----I---I--
	Truepera radiovictrix	297624647	-------AASLTAALR-	A	PEIA----I---I--
	Kuenenia stuttgarti	91202036	---S---A-S-Q-ALN-	N	D-IA----I------
	Blastopirellula marina	87308846	-------A-SL-A-LQ-	N	TDI-V---V---IF-

Fig. 7 Partial sequence alignments of the proteins (**A**) O-antigen polymerase and (**B**) CaCA family Na(+)/Ca(+) antiporter, showing a 5 aa insert and a 1 aa deletion, respectively, that are uniquely found in all three species from the *Petrotoga–Kosmotoga* clade. The presence of these CSIs in *Thermotogales bacterium mesG1* provide evidence for its placement in this clade. Some other CSIs showing similar specificity are listed in Table 6

The Thermotogae phylum presently consists of a single Class, Order and Family and no higher taxonomic groups are recognized within it. Based upon our phylogenetic analyses and the species distribution patterns of various CSIs, a number of distinct clades within this phylum can now be identified (Fig. 13). Some of these clades should be recognized as the higher taxonomic grouping (i.e. Families) within this phylum. One of these clades consists of the *Fervidobacterium* and *Thermosipho* genera. This clade is strongly supported by phylogenetic analyses based on both sets of protein sequences (Fig. 1) and also by 14 CSIs that are uniquely shared by species from these two genera (Table 4; Fig. 13). The species from these two genera also group together in the phylogenetic trees based on 23S rRNA and 16S + 23S rRNA (Zhaxybayeva et al. 2009; Dipippo et al. 2009). Several CSIs that are specific for the *Thermosipho* genus were also identified in this work. All of these observations make a strong case that the clade consisting of these two genera be recognized as a new family (viz. *Themosiphoaceae*) within the order Thermotogales.

Our analyses also suggest that *Petrotoga*, *Kosmotoga* and *Ttog. mesG1*, which show deep branching within the Thermotogae (Fig. 1), form another clade

Table 6 Conserved Signature Indels that are specific for *Petrotoga mobilis*, *Kosmotoga olearia* and *Thermotogales bacterium MesG1.Ag.4.2*

Protein	O-antigen polymerase (Wzy)	CaCA family Na(+)/Ca(+) antiporter	O-antigen polymerase (Wzy)	Hypothetical protein	Peptidase S15	Oligopeptide/dipeptide ABC transporter, ATPase subunit
GenBank Identifier	160901686	160901682	160901686	160901800	160901735	160902771
Accession no.	YP_001567267	YP_001567263	YP_001567267	YP_001567381	YP_001567316	YP_001568352
Indel/size	5 aa ins.	1 aa del.	1 aa ins.	1 aa del.	3 aa ins.	11 aa ins.
Indel position[a]	299–334	47–78	328–373	141–171	171–211	41–110
Figure no.	7a	7b	Sup. Fig. 44	Sup. Fig. 45	Sup. Fig. 46	Sup. Fig. 47
Tt. petrophila	–	–	–	–	–	–
Tt. maritima	–	–	–	–	–	–
Tt. neapolitana	–	–	–	–	–	–
Tt. sp. RQ2	–	–	–	–	–	–
Tt. naphthophila	–	–	–	–	–	–
Tt. lettingae	–	0[c]	–	–	-[e]	–
F. nodosum	–	–	–	+[d]	–	-[f]
Ts. melanesiensis	–	–	–	–	–	-[f]
Ts. africanus	–	–	–	0[c]	–	-[f]
P. mobilis	+	+	+	+	+	+
K. olearia	+	+	+	+	+	+
Ttog. mesG1	+	+	+	–	0[c]	+
No. of other species with indel[b]	0[g]	0/250	0[g]	0[g]	0[g]	0/250

[a] The indel position provided indicates the region of the protein containing the CSI

[b] The presence or absence of the CSIs in the top 250 Blast hits was examined. Of those, the number of non-Thermotogae organisms observed to contain the CSI is indicated. Species containing larger or shorter CSI than indicated were not included in the total

[c] Homologous sequences corresponding to the region containing the CSI's could not be identified in these species

[d] A 2 aa deletion is present in the indicated species rather than the 1 aa deletion found in some other Thermotogae

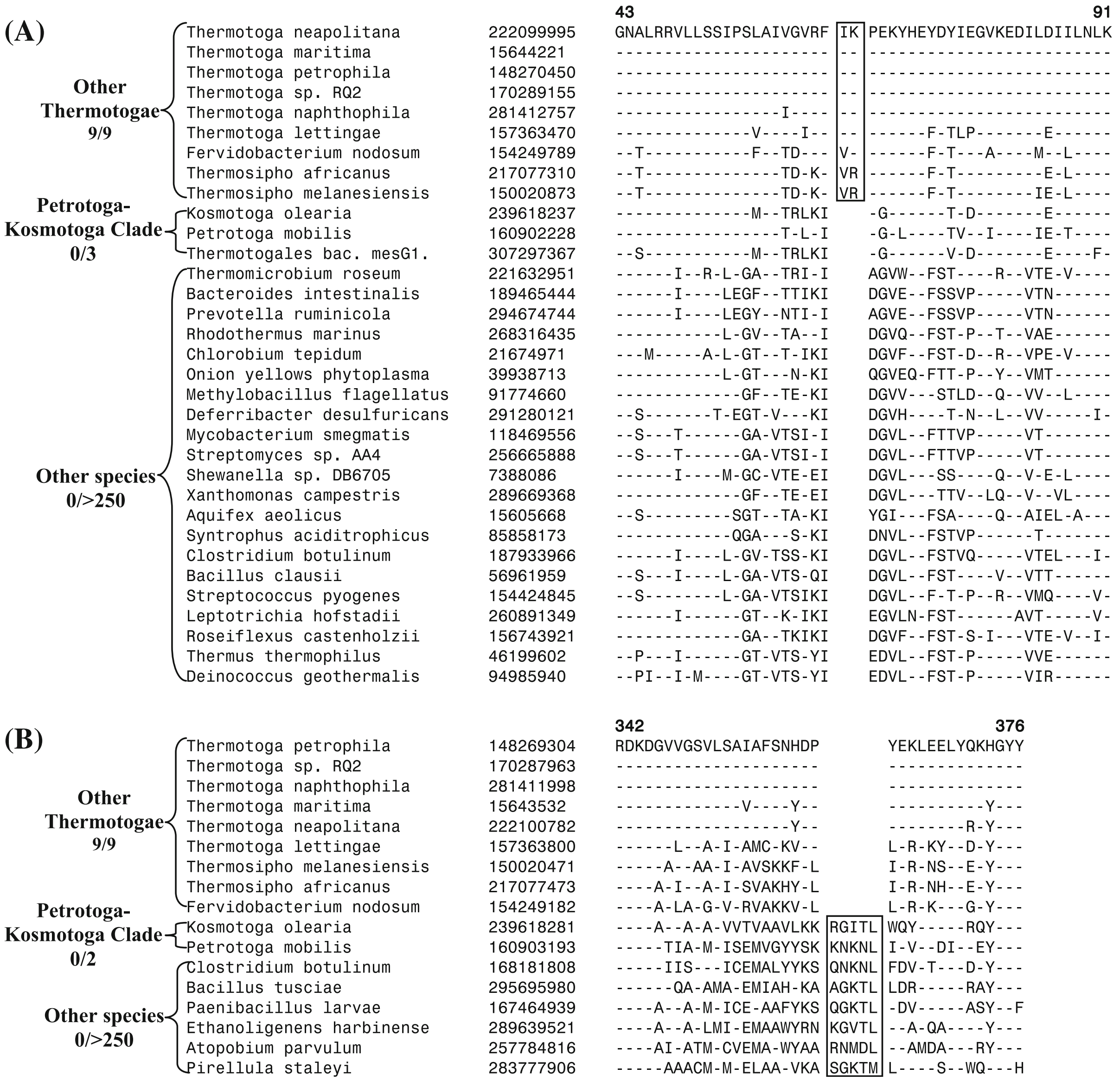
[e] A 1 aa insert is present in the indicated species rather than the 3 aa insert found in some other Thermotogae

[f] A 4 aa insert is present in the indicated species rather than the 11 aa insert found in some other Thermotogae

[g] BLAST searches provided only the indicated Thermotogae sequences as significant homologous matches to the query sequence

within this phylum. Although this clade was supported only by the phylogenetic trees based on dataset I protein sequences (Fig. 1), our identification of six CSIs that are uniquely shared by these bacteria make a strong case that these species are specifically related to each other. The conserved indels in protein sequences are known to be more effective in resolving deeper branching relationships than phylogenetic trees (Rivera and Lake 1992; Gupta and Golding 1993; Baldauf and Palmer 1993; Gupta 1998; Gupta and Mathews 2010). Hence, we believe that the clade consisting of *Petrotoga*, *Kosmotoga* and *Ttog. mesG1*, which is supported by six different CSIs, is meaningful. Within this clade, a specific grouping of the *Ttog. mesG1* with *K. olearia* was also supported by different trees. Lastly, our results provide evidence that *Tt. lettingae* is very distantly related to all other species from the genus *Thermotoga* and also to other Thermotogae species for which sequence information was available (Fig. 1). Hence, it is suggested that this species should be assigned to a distinct genus (and possibly a distinct family) within the phylum Thermotogae.

Thermotogae species are indicated to have undergone extensive LGTs with other prokaryotic taxa particularly Firmicutes, Archaea and Aquifex (Nelson et al. 1999, 2001, 2006, 2009; Mongodin et al. 2005; Zhaxybayeva et al. 2009). These inferences were



```
(A)                                                              43                                                91
                 Thermotoga neapolitana          222099995       GNALRRVLLSSIPSLAIVGVRF IK PEKYHEYDYIEGVKEDILDIILNLK
                 Thermotoga maritima             15644221        ---------------------- -- -------------------------
Other            Thermotoga petrophila           148270450       ---------------------- -- -------------------------
                 Thermotoga sp. RQ2              170289155       ---------------------- -- -------------------------
Thermotogae      Thermotoga naphthophila         281412757       -----------------I---- -- -------------------------
9/9              Thermotoga lettingae            157363470       --------------V----I-- -- ------F-TLP-------E------
                 Fervidobacterium nodosum        154249789       --T-----------F--TD--- V- ------F-T---A----M--L----
                 Thermosipho africanus           217077310       --T--------------TD-K- VR ------F-T---------E-L----
                 Thermosipho melanesiensis       150020873       --T--------------TD-K- VR ------F-T--------IE-L----
Petrotoga-       Kosmotoga olearia               239618237       --------------M--TRLKI    -G------T-D-------E------
Kosmotoga Clade  Petrotoga mobilis               160902228       -----------------T-L-I    -G-L----TV--I----IE-T----
0/3              Thermotogales bac. mesG1.       307297367       --S-----------M--TRLKI    -G------V-D-------E----F-
                 Thermomicrobium roseum          221632951       ------I--R-L-GA--TRI-I    AGVW--FST----R--VTE-V----
                 Bacteroides intestinalis        189465444       ------I----LEGF--TTIKI    DGVE--FSSVP-----VTN------
                 Prevotella ruminicola           294674744       ------I----LEGY--NTI-I    AGVE--FSSVP-----VTN------
                 Rhodothermus marinus            268316435       -----------L-GV--TA--I    DGVQ--FST-P--T--VAE------
                 Chlorobium tepidum              21674971        ---M-----A-L-GT--T-IKI    DGVF--FST-D--R--VPE-V----
                 Onion yellows phytoplasma       39938713        -----------L-GT---N-KI    QGVEQ-FTT-P--Y--VMT------
                 Methylobacillus flagellatus     91774660        -------------GF--TE-KI    DGVV---STLD--Q--VV--L----
                 Deferribacter desulfuricans     291280121       --S-------T-EGT-V---KI    DGVH----T-N--L--VV-----I-
                 Mycobacterium smegmatis         118469556       --S---T------GA-VTSI-I    DGVL--FTTVP-----VT-------
                 Streptomyces sp. AA4            256665888       --S---T------GA-VTSI-I    DGVL--FTTVP-----VT-------
Other species    Shewanella sp. DB6705           7388086         ------I----M-GC-VTE-EI    DGVL---SS----Q--V-E-L----
0/>250           Xanthomonas campestris          289669368       -------------GF--TE-EI    DGVL---TTV--LQ--V--VL----
                 Aquifex aeolicus                15605668        --S---------SGT--TA-KI    YGI---FSA----Q--AIEL-A---
                 Syntrophus aciditrophicus       85858173        ------------QGA---S-KI    DNVL--FSTVP------T-------
                 Clostridium botulinum           187933966       ------I----L-GV-TSS-KI    DGVL--FSTVQ-----VTEL---I-
                 Bacillus clausii                56961959        --S---I----L-GA-VTS-QI    DGVL--FST----V--VTT------
                 Streptococcus pyogenes          154424845       --S--------L-GA-VTSIKI    DGVL--F-T-P--R--VMQ----V-
                 Leptotrichia hofstadii          260891349       ------I------GT--K-IKI    EGVLN-FST------AVT-----V-
                 Roseiflexus castenholzii        156743921       -------------GA--TKIKI    DGVF--FST-S-I---VTE-V--I-
                 Thermus thermophilus            46199602        --P---I------GT-VTS-YI    EDVL--FST-P-----VVE------
                 Deinococcus geothermalis        94985940        --PI--I-M----GT-VTS-YI    EDVL--FST-P-----VIR------

(B)                                                              342                                     376
                 Thermotoga petrophila           148269304       RDKDGVVGSVLSAIAFSNHDP         YEKLEELYQKHGYY
                 Thermotoga sp. RQ2              170287963       ---------------------         --------------
Other            Thermotoga naphthophila         281411998       ---------------------         --------------
                 Thermotoga maritima             15643532        -------------V----Y--         ----------Y---
Thermotogae      Thermotoga neapolitana          222100782       ------------------Y--         --------R-Y---
9/9              Thermotoga lettingae            157363800       ------L--A-I-AMC-KV--         L-R-KY--D-Y---
                 Thermosipho melanesiensis       150020471       -----A--AA-I-AVSKKF-L         I-R-NS--E-Y---
                 Thermosipho africanus           217077473       ----A-I--A-I-SVAKHY-L         I-R-NH--E-Y---
Petrotoga-       Fervidobacterium nodosum        154249182       ----A-LA-G-V-RVAKKV-L         L-R-K---G-Y---
Kosmotoga Clade  Kosmotoga olearia               239618281       ----A--A-A-VVTVAAVLKK RGITL   WQY-----RQY---
0/2              Petrotoga mobilis               160903193       -----TIA-M-ISEMVGYYSK KNKNL   I-V--DI--EY---
                 Clostridium botulinum           168181808       -----IIS---ICEMALYYKS QNKNL   FDV-T---D-Y---
                 Bacillus tusciae                295695980       ------QA-AMA-EMIAH-KA AGKTL   LDR-----RAY---
Other species    Paenibacillus larvae            167464939       ----A--A-M-ICE-AAFYKS QGKTL   -DV-----ASY--F
0/>250           Ethanoligenens harbinense       289639521       ----A--A-LMI-EMAAWYRN KGVTL   --A-QA----Y---
                 Atopobium parvulum              257784816       ----AI-ATM-CVEMA-WYAA RNMDL   --AMDA---RY---
                 Pirellula staleyi               283777906       -----AAACM-M-ELAA-VKA SGKTM   L----S--WQ---H
```

Fig. 8 Partial sequence alignments showing highly conserved regions of the proteins (**A**) RNA polymerase α subunit (RpoA) and (**B**) phosphoglucomutase–phosphomannomutase α/β subunit, showing two different CSIs that are commonly present in all other Thermotogae species except those from the *Petrotoga–Kosmotoga* clade. These CSIs provide evidence for the deep branching of this clade of species in comparison to the other Thermotogae species. Some other CSIs showing similar specificity are listed in Table 7

initially based mainly on identification of closest blast hits for various ORFs from the genomes of Thermotogae spp., which is not a reliable means to infer LGTs (Koski and Golding 2001). However, recently these inferences have also been supported by other analyses (Podell and Gaasterland 2007; Zhaxybayeva et al. 2009; Nesbo et al. 2009). A recent detailed study on 5 Thermotogae genomes showed that for 42–48% of their ORFs the closet blast hits were from the Firmicutes phylum (Zhaxybayeva et al. 2009). The other prokaryotic taxa including Archaea, Aquificales, Proteobacteria, etc., in comparison had much smaller numbers of closest blast hits. Earlier studies based on species distribution profiles of a number of CSIs in universal distributed proteins (Gupta 1998, 2003; Gupta and Griffiths 2002;

Table 7 Conserved Signature Indels that are shared by most Thermotogae except the species from the *Kosmotoga* and *Petrotoga* clade

Protein	RNA polymerase, subunit A (RpoA)	Phospho-glucomutase/ phospho-mannomutase α/β/subunit (PGM/PMM)	Phospho-ribosylamine-glycine ligase (PurD)	Phosphoribo-sylformyl-glycinamidine synthase II (FGAM synthase II)	Glycyl-tRNA synthetase, β subunit (GlyS)	Single stranded, DNA specific exonuclease (RecJ)	Phospho-glucosamine mutase (GlmM)	PhoH family protein
GenBank Identifier	222099995	148269304	150020286	150021120	154249735	157363235	148269875	150020477
Accession no.	YP_002534563	YP_001243764	YP_001305640	YP_001306474	YP_001410560	YP_001470002	YP_001244335	YP_001305831
Indel/size	2 aa ins.	5 aa del.	4 aa del.	8 aa del.	6 aa del.	2 aa del.	2 aa del.	1–2 aa del.
Indel position[a]	43–91	342–376	52–82	92–126	413–450	246–288	201–235	248–283
Figure no.	8a	8b	9a	9b	Sup. Fig. 48	Sup. Fig. 49	Sup. Fig. 50	Sup. Fig. 51
Tt. petrophila	+	+	+	+	+	+	+	+[e]
Tt. maritima	+	+	+	+	+	+	+	+[e]
Tt. neapolitana	+	+	+	+	+	+	+	+[e]
Tt. sp. RQ2	+	+	+	+	+	+	+	+[e]
Tt. naphthophila	+	+	+	+	+	+	+	+[e]
Tt. lettingae	+	+	–	–	+	+	+	+[e]
F. nodosum	+	+	+	+	+	+	+	+
Ts. melanesiensis	+	+	+	+	+	+	+	+
Ts. africanus	+	+	+	+	+	+	+	+
P. mobilis	–	–	–	–	–	–	–	–
K. olearia	–	–	0[c]	0[c]	-[d]	–	+	–
Ttog. mesG1	–	0[c]	0[c]	0[c]	-[d]	–	+	–
Other species with indel[b]	0/250	0/250	0/250	0/250	0/250	3/250	0/250	0/250

[a] The indel position provided indicates the region of the protein containing the CSI

[b] BLAST searches were carried out for the top 250 hits. The number of non-Thermotogae organisms, which were observed to contain the CSI, is indicated. Species containing a larger or a shorter CSI than indicated were not included in the total

[c] Homologous sequences corresponding to the region containing the CSI's could not be identified in these species

[d] An intermediate deletion of 2–3 aa is present in the indicated species

[e] These species contain a 1 aa deletion rather than a 2 aa deletion present in *Thermosipho* and *Fervidobacterium*

Griffiths and Gupta 2004) as well phylogenetic analyses based on large datasets of protein sequences (Ciccarelli et al. 2006; Wu et al. 2009) also support the branching of Thermotogae species in the proximity of Firmicutes. All of these observations indicate that the Thermotogae and Firmicutes are neighboring phyla. Therefore, the observance of majority of the first blast hits for the Thermotogae proteins from the Firmicutes species is an expected result and it is not due to LGTs. The question arises whether these two phyla are simply neighboring or did they share a common ancestor exclusive of other bacterial phyla. If the latter possibility was true, then many CSIs should have been found that were commonly shared by species from these two groups, as was seen in our earlier work on Bacteroidetes and Chlorobi phyla (Gupta 2004) or Chlamydia and Verrucomicrobia phyla (Griffiths and Gupta 2007). However, in this work, we have come across very few CSIs that are commonly shared by the Thermotogae and Firmicutes and their numbers are similar to those seen for many other groups. Further, these shared CSIs were present in only a limited number of Firmicutes/ Clostridia species rather than in all (or most) of the species from this phylum, which would be expected if the Thermotogae and Firmicutes were sister taxa. These results strongly indicate that the phylum Thermotogae is distinct from the Firmicutes.

```
                                                            52                         82
(A)                Thermosipho melanesiensis      150020286  GSEEYLAKGIAE       KYDNVFGPTSKGAKLESSK
                   Thermosipho africanus          217076830  -------N----       --A-----D-RASR-----
Other              Fervidobacterium nodosum       154250339  ----F-V----D       GKA-----D----R--G--
Thermotogae        Thermotoga neapolitana         222100298  ------VE-V-N       WKS-----IKEV-R--G--
8/8                Thermotoga maritima            15644006   ----F-VE-VSN       WRS-----VKEV-R--G--
                   Thermotoga petrophila          148270645  ----F-VE-VSN       WRS-----VKEV-R--G--
                   Thermotoga naphthophila        281412955  ----F-VE-VSN       WRS-----VKEV-R--G--
Petrotoga          Thermotoga sp. RQ2             170289351  ----F-VE-VSN       WRS-----VKEV-R--G--
and Tt. lettingae  Thermotoga lettingae           157364759  -P-N--VD---D  KFEE RGLR----RKQA-LV-G--
                   Petrotoga mobilis              160903207  -------N--VD  KFEE SNLKII--SQNA-Q-----
0/2                Clostridium cellulovorans      242259485  -P-V--VQ-VVD  LFKK NNLKI---AQA--E--G--
                   Bacillus coagulans             229542648  -P-AP-SA--VD  EFQN AGLKI----KQA-RI----
                   Hirschia baltica               254293364  -P-VP--M-L-D  KLRM ADIP----SKAA-Q-----
                   Bordetella avium               187478699  -P-AP--A-VVD  VFRA RNLKI----KAA-Q-----
                   Burkholderia glumae            238028184  -P-AP--A--VN  HFRS RGLK-----REA-Q-----
                   Ralstonia solanacearum         83746482   -P-AP--A--VD  LFRA -GLRI----RAA-Q-----
                   Campylobacter coli RM2228      57168152   ---SF--E-VVD  IFKQ HNIPI----KAA-M--T—
                   Coxiella burnetii              212218162  -P-IP--A--VD  HFQQ ENLI-----QAA-Q--T--
                   Jonquetella anthropi           260655589  -P-AP--A-L-D  ELRQ AGLS----GAA--R---D-
                   Thrmavib. acidaminovorans      269792367  -P-AP-VA-V-D  PLLD AGIR-L--G----Q--G--
Other species      Acidobacterium capsulatum      225873624  -P-LP-SL-LVD  ELER RGIRA----QAA-M--T--
0/>250             Aquifex aeolicus               15606133   -P-AP-VE--VD  EFEK RGLKI---NKEA----G--
                   Clavispora lusitaniae          260940170  -P-QP-VD--ST  VFTK VGIP----SA-A-QM-G--
                   Candida dubliniensis           241949339  -P-QP-VD---D  IFYA VGIP----SA-A--M-G--
                   Streptomyces sp. AA4           256673079  -P-VP-VA-V-D  AVRE AGIAC---SAAA-RI-G--
                   Actinomyces urogenitalis       227496885  -P-AP-VA-V-D  ALRE AGIP----GAEA-R--G--
                   Nasonia vitripennis            156545144  -P-D-----L-D  ELKK SGIPC---QKDA-RI-AD-
                   Kuenenia stuttgarti            91203298   -P-DP-LS--VD  TFKD GNLDI---SKRASLI-G--
                   Gemmata obscuriglobus          168700183  -P-DP--A-L-D  FLRG -GLK----SKEA-RI----
                   Lentisphaera araneosa          149196425  -P-VP-CE--VN  IFRD -GLD----DKDA-Q--G--
                   Neorickettsia risticii         254797248  -Q--H--R--TD  ALMA EGIP-----KNA-------
```

```
                                                            92                                          126
(B)                Thermosipho melanesiensis      150021120  RDVLAMGARPTAILDSL             HMHKIIDGIIEGIADYGN
                   Thermosipho africanus          217077537  --I--------------             ------------------
                   Fervidobacterium nodosum       154249777  -----------------             --DR--E--V--------
Other              Thermotoga naphthophila        281412959  -------------F---             --SR--------------
Thermotogae        Thermotoga maritima            15644002   -------------F---             --SR--------------
8/8                Thermotoga neapolitana         222100302  -------------F---             --AR--------------
                   Thermotoga sp. RQ2             170289355  -----------V-F---             --SR--------------
Petrotoga          Thermotoga petrophila          148270649  -----------V-F---             --SR--------------
and Tt. lettingae  Thermotoga lettingae           157364764  --IF------I-L----   RFGDLKSA  KVKYLFS-VVS---H---
                   Petrotoga mobilis              160903388  --I-------I-L----   KFGNVFDP  KVKN-FE-VVS--S----
0/2                Aquifex aeolicus               15606878   ----S-----I-LA---   RFGEFNYH  ETKRLVK-VVS--SF---
                   Listeria monocytogenes         217964083  ---FS-----I-M-N--   RFGELDTP  -AKYLVSEVVA---G---
                   Enterococcus faecalis          257422353  --IFS-----I-L----   RFGELTDA  RTRYLFQEVVA--SG---
                   Chlorobium ferrooxidans        110598338  --IFT-----V-S-N--   RFGSPKDP  -VRYLV--VVR--G----
                   Chloroherpeton thalassium      193216237  --IFT-----I-S-N--   RFGNIQDS  KVQYLF--VVR--G----
                   Deinococcus deserti            226357170  --IF------F-V----   RFGNPDSP  RTRFLVN-VV---SH---
                   Thermus thermophilus           46199457   --IMS-----I-L----   RFGPPEEA  RSRYLLK-VVS---F---
                   Microcystis aeruginosa         166364281  --IFT-----I---N--   RFGNLQDA  RNRR-FA-VV---SH---
                   Synechococcus sp. PCC 7002     170078274  --IFT-----I-L-N--   RFGTLDNA  QTRR-FQ-VV---SH---
                   Nostoc punctiforme             186681364  --IFT-----I-L-N--   RFGSLEDA  KTQRLFQ-VVA--SH---
Other species      Thermococcus sp. AM4           254172306  --I-C-----I-L--PI   RFGPLEKE  RNRYLFE-VVK-------
0/>250             Pyrococcus abyssi              14521941   --I-C-----I-L--PI   RFGPLEKE  KNRYLFEYVVK-------
                   Thermobispora bispora          296271290  --I-S-----I-VM---   RFGGADQP  DTKRVLP-VV---SS---
                   Halorhabdus utahensis          257051721  --I-S---F-I-LT---   YFGNFDRE  -SRYLLE-VV---S----
                   Salinibacter ruber             294508687  --IFT-----ICA----   RFGSLEES  RVRYLF--VVR--G----
                   Rhodothermus marinus           268315754  --IFT-----ICA-N--   RFGRLDNL  RVRYLL--VVR-------
                   Methanosphaera stadtmanae      84490248   --IIS---K-VVL--A-   RFGHMSNQ  RSKY-F-YVVK--S----
                   Methanobrevibacter smithii     222445644  --IIS-----I-L----   RFGSLDDE  KSKYLFEHVV---S----
                   Caulobacter crescentus         16126739   ---FT-----I-L-NA-   RFGDPSHP  KTKRLV--VVA---G---
                   Roseomonas cervicalis          296532417  ---FT-----I-N-NA-   RFGAPDHP  KTRR-L--VVR--GG---
                   Sagittula stellata             126728665  ---FT-----V-AMN--   SFGVAEHP  KTRQLVH-VV---GG---
```

Fig. 9 Partial-sequence alignment of the proteins (**A**) phosphoribosylamine-glycine ligase and (**B**) phosphoribosylformylglycinamidine synthase II; showing conserved deletions in these proteins are uniquely present in various in Thermotogae homologs except *P. mobilis* and *Tt. lettingae*. The homologs of these proteins were not detected in *Kosmotoga* and *Thermotogales bacterium mesG1*. These indels support the deeper branching of *Tt. lettingae* in comparison to the other *Thermotoga* species

Table 8 Conserved indels common to Thermotogae and some other groups

Protein	Translation elongation factor Tu (EF-TU)	RNR	Preprotein translocase, SecA subunit (SecA)	Small GTP-binding protein (EngB)	ATP-dependant protease La	Translation elongation factor G (EF-G)	Bifunctional GMP synthase/glutamine amidotransferase protein (GuaA)	Cation transporting ATPase, P-type
GenBank Identifier	552037	160903303	15644326	170289148	148270288	222099964	15644564	15643086
Accession no.	AAA27415	YP_001568884	NP_229378	YP_001739386	YP_001244748	YP_002534532	NP_229617	NP_228129
Indel/size	1 aa ins.	3 aa ins.	50 aa ins.	1 aa del.	2 aa del.	1–2 aa ins.	1 aa del.	19 aa ins.
Indel position[a]	332–362	211–237	123–233	35–62	217–273	501–539	130–162	336–404
Figure no.	10	11	Sup. Fig. 52	Sup. Fig. 53	Sup. Fig. 54	Sup. Fig. 55	Sup. Fig. 56	Sup. Fig. 57
Thermotogae species containing indel	All detected Thermotogae	All detected Thermotogae	All detected Thermotogae except *P. mobilis*, *K. olearia* and *Ttog. mesG1*	All detected Thermotogae	All detected Thermotogae	All detected *Thermotoga* genus species *Ttog. mesG1* and *P. mobilis*	All detected species from the *Thermotoga* genus and *Ttog. mesG1*	All detected Thermotogae except *K. olearia* and *Ttog. mesG1*
Other species with indel	Various Aquificales, *Coxiella burnetii* and *Thermomicrobium rosetii*	Thermococci and Thremoprotei	Various Aquificales	Various Aquificales and *Caldicellulosiruptor bescii*	Acidobacteria and *Thermobaculum terrenum*	Some γ-proteobacteria and *Geobacter lovleyi*	Various Fusobacteria	*A. ehrlichii*, *A. metalliredigens*, *H. orenii* and *delta proteobacterium MLMS-1*

Protein	Threonyl tRNA synthetase (ThrS)	RNA polymerase, β' subunit (RpoC)	Binding-protein-dependent transport systems inner membrane component	Pyruvate phosphate dikinase (PPDK)	DNA gyrase, subunit A (GyrA)	Chaperone protein DnaK	Phosphoribosylformylglycinamidine synthase I (FGAM synthase I)	Metal dependant phosphohydrolase
GenBank Identifier	15643452	148269601	150020493	254484873	148270781	222099278	150021119	150020476
Accession no.	NP_228498	YP_001244061	YP_001305847	ZP_05098089	YP_001245241	YP_002533846	YP_001306473	YP_001305830
Indel/size	1 aa ins.	2–4 aa ins.	1 aa ins.	1 aa del.	1 aa ins.	1–4 aa ins.	2 aa del.	1 aa ins.
Indel position[a]	262–298	663–709	374–400	191–225	314–354	344–384	160–183	266–306
Figure no.	Sup. Fig. 58	Sup. Fig. 59	Sup. Fig. 60	Sup. Fig. 61	Sup. Fig. 62	Sup. Fig. 63	Sup. Fig. 64	Sup. Fig. 65
Thermotogae species containing indel	All detected Thermotogae	All detected Thermotogae	All detected Thermotogae	All detected Thermotogae	All Thermotogae except *P. mobilis* and *Tt. lettingae*	All detected Thermotogae except *P. mobilis*, *Tt. lettingae*, *K. olearia* and *Ttog. mesG1*	*Thermosipho* genus	*Thermosipho* and *Fervidobacterium* genera
Other species with indel	Various Eukaryotes, species from the *Leuconostoc* genus and *B. tusciae*	Many Deinococcus-Thermus species with 3 aa insert	*Dictyoglomus thermophilum* and *Dictyoglomus turgidum*	Species from Eukaryotic groups of Stramenophiles, Chlorophyta and Polymastigidae	*Thermomicrobium roseum* and *Sphaerobacter thermophilus*	Many Eukaryotes, various species in the *Burkholderia* genus, *Acidobacterium sp. MP5ACTX8*, *M. petroleiphilum* *D. radiodurans*, *N. risticii* and *P. ruminicola.*	All detected Thermoproteales	Dictyoglomi species, *P. limnophilus* and *C. morbi*

[a] The indel position provided indicates the region of the protein containing the CSI

			332		362
Thermotogae 12/12	Thermotoga maritima	552037	YKPQFYIRTADVTGEI	V	GLPEGVEMVMPGDH
	Thermotoga naphthophila	281412727	----------------	-	--------------
	Thermotoga sp. RQ2	170289185	----------------	-	--------------
	Thermotoga petrophila	148270420	----------------	-	--------------
	Thermotoga neapolitana	222099965	----------------	-	--------------
	Fervidobacterium nodosum	154249819	----------------	-	D--A---------N
	Kosmotoga olearia	239618267	-R-------------L	-	D------------N
	Thermotogales bac. mesG1.	307297396	-R---F-K--------	A	D--------I---N
	Thermosipho africanus	217077281	-----F---------L	I	DF-A---------N
	Thermosipho melanesiensis	150020843	-----F---------L	I	EF-A---------N
	Petrotoga mobilis	160902258	-R------------TL	-	EFSS-A-------N
	Thermotoga lettingae	157363440	-R---F----------	T	E-GNNA-------N
Aquificales 11/11	Hydrogenobaculum sp. SN	289631482	-R------------TV	-	K----Q-------N
	Aquifex aeolicus	15605614	-R----F-------TV	-	K------------N
	Aquifex pyrophilus	3913574	-R----F-------TV	-	K------------N
	Hydrogenobaculum sp. Y04AAS1	195952628	-R------------TV	-	K----Q-------N
	Sulfurihydrogenibium azorense	225849550	-R------------TV	-	-----Q-------N
	Sulfurihydrogenibium yellowstonii	237755894	-R---------I--TV	I	-----Q-------N
	Sulfurihydrogenibium sp. YO3AO	188996241	-R---------I--TV	-	E----Q-------N
	Hydrogenobacter thermophilus	288818166	-R----F-------TV	-	K----Q-------N
	Thermocrinis albus	289548406	-R----F-------VV	-	K----Q-------N
	Hydrogenivirga sp. 128-5-R1-1	163783908	-R------------TV	-	E----Q-------N
	Persephonella marina EX-H1	225850720	-R---------I--TV	-	E----Q-------N
Other species 2/>250	Thermomicrobium roseum	221632690	-R-------T-----V	-	-------------N
	Coxiella burnetii	29653588	-R----F--T----QL	L	S----I-------N
	Sphaerobacter thermophilus	269837639	-R-------T------		Q------------N
	Aliivibrio salmonicida	209693919	-R----F--T----D-		T------------N
	Vibrio fischeri	59710840	-R----F--T----D-		T------------N
	Chlamydia trachomatis	166154532	-R---FF--T----VV		T------------N
	Alkaliphilus metalliredigens	150392162	-R----F--T----S-		K------------N
	Ammonifex degensii	260893374	-R----F--T------		K------------N
	Clostridium botulinum	168187816	-R----F--T----S-		A-------------
	Halothermothrix orenii	220930949	-R----F--T----T-		S------------N
	Moorella thermoacetica	83591278	-R----F--T----VV		N------------N
	Natranaerobius thermophilus	188584819	-R----F--T----V-		E------------N
	Staphylococcus epidermidis	242372766	-R----F--T----VV		N----T-------N
	Syntrophomonas wolfei	114567858	-R----F--T----I-		Q------------N
	Leptotrichia hofstadii	260891597	------F--T-I---V		N------------N
	Fusobacterium varium	253581338	-R----F--T-I--AV		T------------N
	Dictyoglomus thermophilum	206901073	------F--T------		K-----Q------N
	Bacillus cereus	42779189	-R----F--T----I-		Q----T-------N
	Listeria grayi	299820941	-R----F--T----IV		T----T-------N
	Staphylococcus aureus	258422605	-R----F--T----VV		H----T-------N
	Bacteroides intestinalis	189465404	-R----L--M-C----		T----T-------N
	Prevotella buccalis	282878707	-R----L--M-C----		T------------N
	Roseiflexus sp. RS-1	148655339	-R-------T----A-		H------------N
	Chloroflexus aurantiacus	1169486	-R-------T----A-		---A-M-------N
	Wolinella succinogenes	587590	-R----V--T----S-		S------------N
	Caminibacter mediatlanticus	149195237	-R-------T----T-		Q------------N
	Dehalococcoides ethenogenes	57234266	-----FFG-T------		H--------V----
	Mycoplasma mycoides	42560712	-R----F--T-----V		T----TD------N
	Roseomonas cervicalis	296537446	-R----F--T----VV		Q------------N

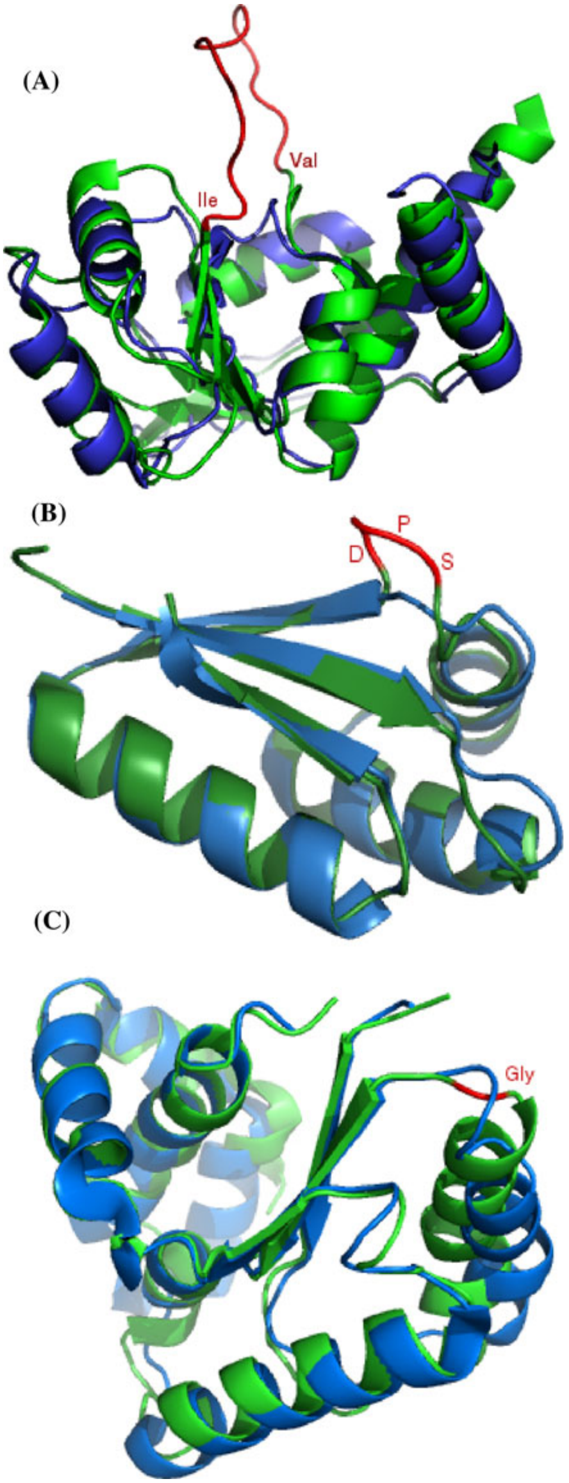
Fig. 10 Partial-sequence alignment for a highly conserved region of the protein synthesis elongation factor Tu showing a 1 aa insert that is commonly shared by various Thermotogae and Aquificae species. Additionally, an insert in this position is also present in one chloroflexus species (*T. roseum*) and a γ-proteobacterium (*C. burnetii*). The *shared* presence of this insert in these different groups could be due to LGTs

			211		237
Thermotogae 10/10	Petrotoga mobilis	160903303	HPDVFDFINSKKDN	KGN	NVLNYFNISV
	Thermotoga neapolitana	222099545	---IEE--DA-RE-	T-E	A---F--L--
	Thermotoga naphthophila	281412178	---IEE--DA--E-	T-E	A---F--L--
	Thermotoga sp. RQ2	170288626	---IEE--DA--E-	T-E	A---F--L--
	Thermotoga petrophila	148269941	---IEE--DA--E-	T-E	A---F--L--
	Thermotoga maritima	15642893	---IEE--DA--E-	T-E	A---F--L--
	Thermosipho africanus	217076775	---IEE--TA-EN-	D-E	K--K------
	Fervidobacterium nodosum	154249118	---IEE--TA-EG-	D-E	K--KF-----
	Thermotoga lettingae	157363198	---IM---SA--G-	D-E	K--RF-----
	Thermosipho melanesiensis	150020504	---IEE--TA-EN-	D-E	KI-K------
Thermococci 4/4	Pyrococcus abyssi	14521806	---IEK--HA-EK-	I-T	---SN-----
	Pyrococcus furiosus	18976812	---IEK--HA-EK-	I-T	---SN-----
	Thermococcus sp. AM4	254173913	---IEK--HA-ER-	T-T	---SN-----
	Thermococcus barophilus	254171320	---IEK--HA-EK-	I-T	---SN-----
Thermoprotei 4/4	Caldivirga maquilingensis	159041217	---IEK--T--TGR	LKD	VQ-QN-----
	Pyrobaculum calidifontis	126460402	---IRK--KA-TGE	LKD	VH-QN-----
	Pyrobaculum islandicum	119872039	----RK--KA-TGE	LKD	AH-QN-----
	Thermoproteus neutrophilus	171186058	----RK--KA-TGE	LKD	AH-QN-----
Other species 0/>250	Halothermothrix orenii	220933052	---IME--TA-A-E		-R--N-----
	Selenomonas sputigena	260888488	---IYE-LHM-QE-		K-ASM-L-I
	Clostridium difficile	296451257	--EIEE--DI-T-L		EKVTKA----
	Jonquetella anthropi	260654687	---IER-MDA-TED		GR-P------
	Geobacter metallireducens	78222511	---IM---MC-D-Q		KH--N-----
	Syntrophus aciditrophicus	85723395	---IMN--MC-A-K		HQ--N-----
	Prevotella melaninogenica	252119097	---SEA--DA-MEE		GKVTGA-V--
	Bacteroides plebeius	198275618	---SES--DA-MTE		GKVTGA-V--
	Deferribacter desulfuricans	291280756	---IEE--KI-R--		KT-TN-----
	Leptospirillum rubarum	124516151	---ILE--DC-M-L		TQVTN-----
	Dictyoglomus turgidum	217967765	---IWE-VTC-DQE		G--SN-----
	Ferroglobus placidus	288931372	---IEE--TA-WEE		G--RN-----
	Archaeoglobus fulgidus	11499254	---IEE--KA-WEE		G--RN-----
	Roseobacter litoralis	163732870	----E---AA-S-P		AR-RM--M--
	Fusobacterium ulcerans	257468300	----SE--SI-S-L		-KIQKA----
	Geobacillus sp. Y412MC10	261408340	----E---TV-QTM		GQVTNA-L--
	Picrophilus torridus	48478313	---IM---T--DSE	-	K--SN-----
	Ferroplasma acidarmanus	257076070	--NIM------DAE	-	KI-SN-----
	Thermoplasma volcanium	13540923	---IME-VL--DSE	-	K--SN-----
	Aciduliprofundum boonei	254168812	---ILE--T--DSE	-	K--SN-----
	Mycobacterium sp. MCS	108798058	---IY---A--SGS	T	AK-PH--L-I

Fig. 11 Partial sequence alignment of the protein RNR showing a 3 aa insert that is commonly shared by various Thermotogae species as well as by various Thermococci and Thermoproteales. The *shared* presence of this insert in these groups could again be due to LGTs. A 1 aa insert is also present in this position in some other species, which is likely of independent origin. Other examples of CSIs that are commonly present in the Thermotogae species and other groups are presented in Table 8

Although these two phyla are indicated to be phylogenetic neighbors, no evidence was obtained that they shared a common ancestor exclusive of other bacterial groups.

In the present work, we have also discovered several examples where a given CSI, in addition to being shared by all or most Thermotogae species, was also present in certain other groups of organisms. These other groups included Archaea, Aquificae, Firmicutes, Proteobacteria, Deinococcus, Fusobacteria, Dictyoglomi, Chloroflexi and Eukaryotes. The numbers of CSIs that are commonly shared between Thermotogae and any of these other groups are generally very few (between 1 and 3) and based on them no clear pattern or relationship can be inferred. Further, in most cases, these CSIs were present in



◀ **Fig. 12** The locations of the Thermotogae-specific inserts in the structures of some of the proteins. (**A**) Structural comparison of the ribosomal protein L4 from *Tt. maritima* (PDB number 1DMG; shown in *green*) (Worbs et al. 2000) and *Escherichia coli* (depicted in *blue*, PDB number 3OFC) (Dunkle et al. 2010) showing the 15 aa insert that is present in *Tt. maritima* (Sup. Fig. 2); (**B**) Structural comparison of the C-terminal fragment (residues 54–128) of ribosomal protein L12 from *Tt. maritima* (in *green*, PDB number 1DD4) (Wahl et al. 2000) with the homologous protein from *E. coli* (depicted in *blue*, PDB number 1CTF) (Leijonmarck and Liljas 1987) showing the location of the 3 aa insert that is specific for the Thermotogae phylum. (i.e. The locations of the Thermotogae-specific inserts in the structures of some of the proteins.) The amino acid corresponding to the insert are depicted in *red*. (**C**) The structural comparison of the N-terminal fragments (residues 1–151) of tryptophanyl-tRNA synthetase from *Tt. maritima* (depicted in *green*; PDB number 2G36) with the homologous protein from *Yersinia pestis* is (shown in *blue*, PDB number 3N9I) indicating the location of the 1 aa insert (shown in *red*) that is found in Thermotogae species. The structures of all of the above proteins were obtained from the Protein Data Bank and they were aligned using the PyMol program (Delano 2002). (Color figure online)

only a limited number of species from these other taxa. The shared presence of these CSI in Thermotogae and these other groups could result from a number of possibilities including LGT of the indel-containing gene from Thermotogae to these other groups or vice versa. However, it is also possible that similar genetic changes in some of these lineages have occurred independently. Although we are unable to distinguish between these possibilities at the present time, the shared presence of many CSIs by Thermotogae and other prokaryotic/eukaryotic phyla support the inference from other studies that genes for a number of proteins have been laterally transferred between these groups (Zhaxybayeva et al. 2009; Nesbo et al. 2009).

The molecular signatures for the phylum Thermotogae and a number of its clades described here also provide novel means for the identification of these bacteria. Based upon our work on many other CSIs (Gupta and Griffiths 2006; Gao et al. 2009; Singh and Gupta 2009; Gupta 2009a), most of the CSIs identified in the present work are expected to retain their specificity for the indicated clades and have high predictive values. Thus, based upon their presence or absence, it should be possible to identify known or even previously unknown species belonging to these groups in different environments. Because these CSIs are present in highly conserved proteins, degenerate PCR primers for these genes/proteins (that flank these CSIs) can be readily designed (Galley et al. 1992; Griffiths and Gupta 2002; Gao and Gupta 2005) and

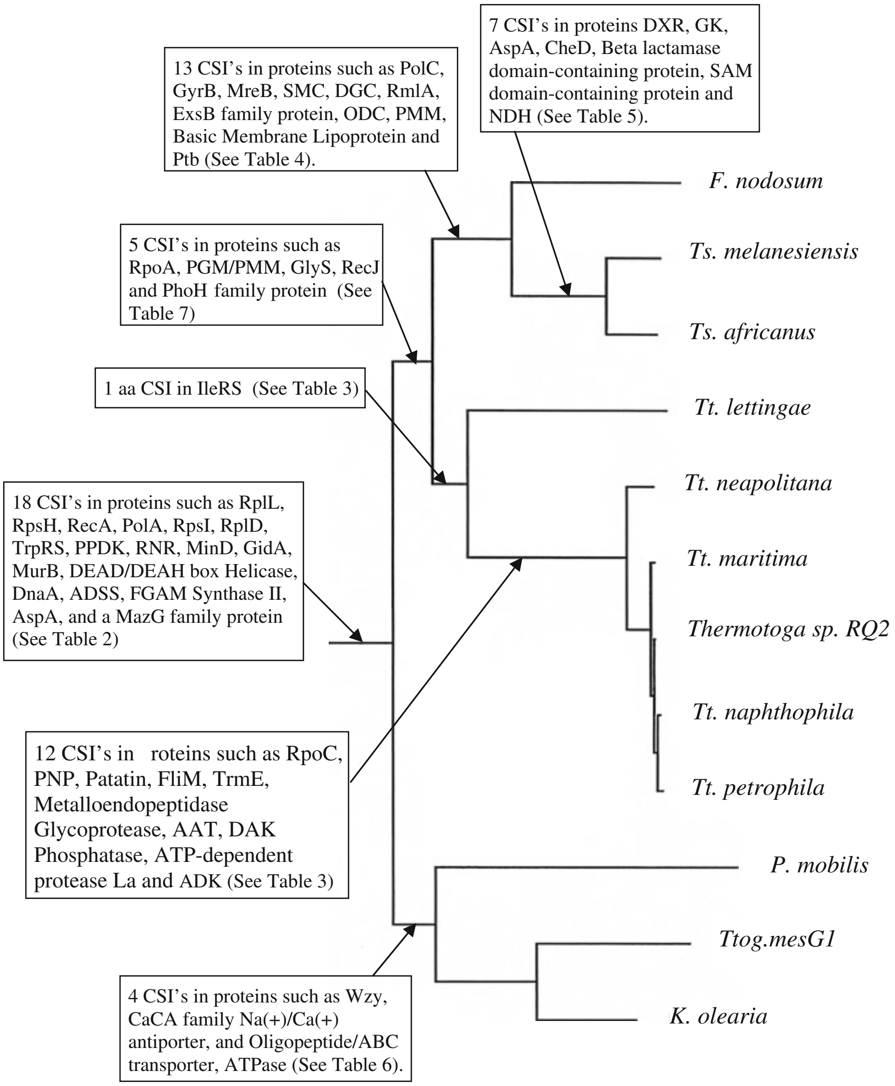

13 CSI's in proteins such as PolC, GyrB, MreB, SMC, DGC, RmlA, ExsB family protein, ODC, PMM, Basic Membrane Lipoprotein and Ptb (See Table 4).
7 CSI's in proteins DXR, GK, AspA, CheD, Beta lactamase domain-containing protein, SAM domain-containing protein and NDH (See Table 5).
5 CSI's in proteins such as RpoA, PGM/PMM, GlyS, RecJ and PhoH family protein (See Table 7)
1 aa CSI in IleRS (See Table 3)
18 CSI's in proteins such as RplL, RpsH, RecA, PolA, RpsI, RplD, TrpRS, PPDK, RNR, MinD, GidA, MurB, DEAD/DEAH box Helicase, DnaA, ADSS, FGAM Synthase II, AspA, and a MazG family protein (See Table 2)
12 CSI's in roteins such as RpoC, PNP, Patatin, FliM, TrmE, Metalloendopeptidase Glycoprotease, AAT, DAK Phosphatase, ATP-dependent protease La and ADK (See Table 3)
4 CSI's in proteins such as Wzy, CaCA family Na(+)/Ca(+) antiporter, and Oligopeptide/ABC transporter, ATPase (See Table 6).
F. nodosum
Ts. melanesiensis
Ts. africanus
Tt. lettingae
Tt. neapolitana
Tt. maritima
Thermotoga sp. RQ2
Tt. naphthophila
Tt. petrophila
P. mobilis
Ttog.mesG1
K. olearia

Fig. 13 A summary diagram based upon phylogenetic analyses and the species distribution patterns of various Thermotogae-specific CSIs indicating the evolutionary relationships among Thermotogae species and different clades within this phylum that are supported by large numbers of molecular signatures

they should provide novel means for identification of new as well as existing Thermotogae species (or isolates) in different environments or sequence databases. Further, the presence or absence of various clade-specific CSIs in other Thermotogales species should enable their placement into one of the identified clades, or possibly new clades if all of these signatures are absent in them. Lastly, the identified CSIs, due to their specificity for the Thermotogae species, also provide novel tools for genetic and biochemical studies on these bacteria. Studies on understanding the cellular functions of these CSIs could lead to discovery of novel genetic and biochemical properties that are unique to these bacteria and which could provide insights into their unique morphology and physiological characteristics.

Acknowledgments This work was supported by a research grant from the Natural Science and Engineering Research Council of Canada. We thank Sanjan George and Zainab Motala for assistance in carrying out initial Blast searches and creation of the signature files. We thank the DOE Joint Genome Research Institute for making genome sequences for a

number of Thermotogae species publicly available that enabled some of these analyses.

References

Akiva E, Itzhaki Z, Margalit H (2008) Built-in loops allow versatility in domain–domain interactions: lessons from self-interacting domains. Proc Natl Acad Sci USA 105:13292–13297

Alain K, Marteinsson VT, Miroshnichenko ML, Bonch-Osmolovskaya EA, Prieur D, Birrien JL (2002) *Marinitoga piezophila* sp. nov., a rod-shaped, thermo-piezophilic bacterium isolated under high hydrostatic pressure from a deep-sea hydrothermal vent. Int J Syst Evol Microbiol 52:1331–1339

Antoine E, Cilia V, Meunier JR, Guezennec J, Lesongeur F, Barbier G (1997) *Thermosipho melanesiensis* sp. nov., a new thermophilic anaerobic bacterium belonging to the order Thermotogales, isolated from deep-sea hydrothermal vents in the southwestern Pacific Ocean. Int J Syst Bacteriol 47:1118–1123

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

Balk M, Weijma J, Stams AJ (2002) *Thermotoga lettingae* sp. nov., a novel thermophilic, methanol-degrading bacterium isolated from a thermophilic anaerobic reactor. Int J Syst Evol Microbiol 52:1361–1368

Bocchetta M, Gribaldo S, Sanangelantoni A, Cammarano P (2000) Phylogenetic depth of the bacterial genera Aquifex and Thermotoga inferred from analysis of ribosomal protein, elongation factor, and RNA polymerase subunit sequences. J Mol Evol 50:366–380

Boucher Y, Douady CJ, Papke RT et al (2003) Lateral gene transfer and the origins of prokaryotic groups. Annu Rev Genet 37:283–328

Boussau B, Gueguen L, Gouy M (2008) Accounting for horizontal gene transfers explains conflicting hypotheses regarding the position of aquificales in the phylogeny of Bacteria. BMC Evol Biol 8:272

Brocchieri L, Karlin S (2000) Conservation among HSP60 sequences in relation to structure, function, and evolution. Protein Sci 9:476–486

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

Chlenov M, Masuda S, Murakami KS, Nikiforov V, Darst SA, Mustaev A (2005) Structure and function of lineage-specific sequence insertions in the bacterial RNA polymerase beta′ subunit. J Mol Biol 353:138–154

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

Cole JR, Wang Q, Cardenas E et al (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 37:D141–D145

Conners SB, Mongodin EF, Johnson MR, Montero CI, Nelson KE, Kelly RM (2006) Microbial biochemistry, physiology, and biotechnology of hyperthermophilic Thermotoga species. FEMS Microbiol Rev 30:872–905

Delano WL (2002) The Pymol user's manual. Delano Scientific, Palo Alto

Di Giulio M (2003) The universal ancestor was a thermophile or a hyperthermophile: tests and further evidence. J Theor Biol 221:425–436

Dipippo JL, Nesbo CL, Dahle H, Doolittle WF, Birkland NK, Noll KM (2009) *Kosmotoga olearia* gen. nov., sp. nov., a thermophilic, anaerobic heterotroph isolated from an oil production fluid. Int J Syst Evol Microbiol 59:2991–3000

Dunkle JA, Xiong L, Mankin AS, Cate JH (2010) Structures of the *Escherichia coli* ribosome with antibiotics bound near the peptidyl transferase center explain spectra of drug action. Proc Natl Acad Sci USA 107:17152–17157

Eisen JA (1995) The RecA protein as a model molecule for molecular systematic studies of bacteria: comparison of trees of RecAs and 16S rRNAs from the same species. J Mol Evol 41:1105–1123

Eriksen NT, Riis ML, Holm NK, Iversen N (2010) H(2) synthesis from pentoses and biomass in *Thermotoga* spp. Biotechnol Lett 33:293–300

Fardeau ML, Ollivier B, Patel BK et al (1997) *Thermotoga hypogea* sp. nov., a xylanolytic, thermophilic bacterium from an oil-producing well. Int J Syst Bacteriol 47:1013–1019

Feng Y, Cheng L, Zhang X, Li X, Deng Y, Zhang H (2010) *Thermococcoides shengliensis* gen. nov., sp. nov., a new member of the order Thermotogales isolated from oil-production fluid. Int J Syst Evol Microbiol 60:932–937

Frock AD, Notey JS, Kelly RM (2010) The genus Thermotoga: recent developments. Environ Technol 31:1169–1181

Gaget V, Gribaldo S, Tandeau dM (2011) An rpoB signature sequence provides unique resolution for the molecular typing of Cyanobacteria. Int J Syst Evol Microbiol 61:170–183

Galley KA, Singh B, Gupta RS (1992) Cloning of HSP70 (dnaK) gene from *Clostridium perfringens* using a general polymerase chain reaction based approach. Biochem Biophys Acta 1130:203–208

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

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 (2004) Signature sequences in diverse proteins provide evidence for the late divergence of the order *Aquificales*. Int Microbiol 7:41–52

Griffiths E, Gupta RS (2006a) 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

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

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

Gudkov AT (1997) The L7/L12 ribosomal domain of the ribosome: structural and functional studies. FEBS Lett 407:253–256

Gupta RS (1995) Evolution of the chaperonin families (Hsp60, Hsp10 and Tcp-1) of proteins and the origin of eukaryotic cells. Mol Microbiol 15:1–11

Gupta RS (1997) Protein phylogenies and signature sequences: evolutionary relationships within prokaryotes and between prokaryotes and eukaryotes. Antonie van Leeuwenhoek 72:49–61

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 (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 RS (2003) Evolutionary relationships among photosynthetic bacteria. Photosynth Res 76:173–183

Gupta RS (2004) The phylogeny and signature sequences characteristics of *Fibrobacters*, *Chlorobi* and *Bacteroidetes*. Crit Rev Microbiol 30:123–143

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, Golding GB (1993) Evolution of HSP70 gene and its implications regarding relationships between archaebacteria, eubacteria, and eukaryotes. J Mol Evol 37: 573–582

Gupta RS, Griffiths E (2002) Critical issues in bacterial phylogenies. Theor Popul Biol 61:423–434

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, Shami A (2011) Molecular signatures for the Crenarchaeota and the Thaumarchaeota. Antonie van Leeuwenhoek 99:133–157

Gupta RS, Mukhtar T, Singh B (1999) Evolutionary relationships among photosynthetic prokaryotes (*Heliobacterium chlorum*, *Chloroflexus aurantiacus*, cyanobacteria, *Chlorobium tepidum* and proteobacteria): implications regarding the origin of photosynthesis. Mol Microbiol 32:893–906

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

Hormozdiari F, Salari R, Hsing M et al (2009) The effect of insertions and deletions on wirings in protein–protein interaction networks: a large-scale study. J Comput Biol 16:159–167

Huber R, Hannig M (2006) Thermotogales. Prokaryotes 7:899–922

Huber R, Langworthy TA, Konig H et al (1986) *Thermotoga maritima* sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90°C. Arch Microbiol 144:324–333

Itzhaki Z, Akiva E, Altuvia Y, Margalit H (2006) Evolutionary conservation of domain–domain interactions. Genome Biol 7:R125

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

Karlin S, Brocchieri L (1998) Heat shock protein 70 family: multiple sequence comparisons, function, and evolution. J Mol Evol 47:565–577

Karlin S, Weinstock GM, Brendel V (1995) Bacterial classifications derived from recA protein sequence comparisons. J Bacteriol 177:6881–6893

Kim JY, Kavas M, Fouad WM, Nong G, Preston JF, Altpeter F (2010) Production of hyperthermostable GH10 xylanase Xyl10B from *Thermotoga maritima* in transplastomic plants enables complete hydrolysis of methylglucuronoxylan to fermentable sugars for biofuel production. Plant Mol Biol

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Klenk HP, Meier TD, Durovic P et al (1999) RNA polymerase of *Aquifex pyrophilus*: Implications for the evolution of the bacterial *rpoBC* operon and extremely thermophilic bacteria. J Mol Evol 48:528–541

Koski LB, Golding GB (2001) The closest BLAST hit is often not the nearest neighbor. J Mol Evol 52:540–542

Kunisawa T (2005) Dichotomy of major bacterial phyla inferred from gene arrangement comparisons. J Theor Biol 234:1–5

Lee D, Seo H, Park, C, Park K (2009) WeGAS: a web-based microbial genome annotation system. Biosci Biotechnol Biochem 73:213–216

Leijonmarck M, Liljas A (1987) Structure of the C-terminal domain of the ribosomal protein L7/L12 from Escherichia coli at 1.7 A. J Mol Biol 195:555–579

Ludwig W, Klenk H-P (2005) Overview: a phylogenetic backbone and taxonomic framework for prokaryotic systematics. In: Brenner DJ, Krieg NR, Staley JT, Garrity GM (eds) Bergey's manual of systematic bacteriology. Springer, Berlin, pp 49–65

Mongodin EF, Hance IR, DeBoy RT et al (2005) Gene transfer and genome plasticity in *Thermotoga maritima*, a model hyperthermophilic species. J Bacteriol 187:4935–4944

NCBI Completed microbial genomes (2011) http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html

Nelson KE, Clayton R, Gill S et al (1999) Evidence for lateral gene transfer between *Archaea* and *Bacteria* from genome sequence of *Thermotoga maritima*. Nature 399:323–329

Nesbo CL, L'Haridon S, Stetter KO, Doolittle WF (2001) Phylogenetic analyses of two "Archaeal" genes in *Thermotoga maritima* reveal multiple transfers between *Archaea* and *Bacteria*. Mol Biol Evol 18:362–375

Nesbo CL, Dlutek M, Doolittle WF (2006) Recombination in Thermotoga: implications for species concepts and biogeography. Genetics 172:759–769

Nesbo CL, Bapteste E, Curtis B et al (2009) The genome of *Thermosipho africanus* TCF52B: lateral genetic connections to the Firmicutes and Archaea. J Bacteriol 191:1974–1978

Nesbo CL, Kumaraswamy R, Dlutek M, Doolittle WF, Foght J (2010) Searching for mesophilic Thermotogales bacteria: "mesotogas" in the wild. Appl Environ Microbiol 76: 4896–4900

Olsen GJ, Woese CR (1993) Ribosomal RNA: a key to phylogeny. FASEB J 7:113–123

Olsen GJ, Woese CR, Overbeek R (1994) The winds of (evolutionary) change: breathing new life into microbiology. J Bacteriol 176:1–6

Osborne AR, Clemons WM Jr, Rapoport TA (2004) A large conformational change of the translocation ATPase SecA. Proc Natl Acad Sci USA 101:10937–10942

Patel BKC, Morgan HW, Daniel RM (1985) *Fervidobacterium nodosum* gen. nov. and spec. nov., a novel chemoorganotrophic, caldoactive, anaerobic bacterium. Arch Microbiol 141:63–69

Podell S, Gaasterland T (2007) DarkHorse: a method for genome-wide prediction of horizontal gene transfer. Genome Biol 8:R16

Reysenbach A-L (2001) Phylum BII. Thermotogae phy. nov. In: Boone DR, Castenholz RW (eds) Bergey's manual of systematic bacteriology. Springer, Berlin, pp 369–387

Rivera MC, Lake JA (1992) Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science 257:74–76

Rodnina MV, Pape T, Fricke R, Wintermeyer W (1995) Elongation factor Tu, a GTPase triggered by codon recognition on the ribosome: mechanism and GTP consumption. Biochem Cell Biol 73:1221–1227

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

Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504

Schoeffler AJ, May AP, Berger JM (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

Seo PS, Yokota A (2003) The phylogenetic relationships of cyanobacteria inferred from 16S rRNA, gyrB, rpoC1 and rpoD1 gene sequences. J Gen Appl Microbiol 49:191–203

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

The NCBI Taxonomy Homepage (2010) http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/

Urios L, Cueff-Gauchard V, Pignet P et al (2004) *Thermosipho atlanticus* sp. nov., a novel member of the Thermotogales isolated from a Mid-Atlantic Ridge hydrothermal vent. Int J Syst Evol Microbiol 54:1953–1957

Van de Peer Y, De Wachter R (1997) Construction of evolutionary distance trees with TREECON for Windows: accounting for variation in nucleotide substitution rate among sites. Comput Appl Biosci 13:227–230

Wahl MC, Bourenkov GP, Bartunik HD, Huber R (2000) Flexibility, conformational diversity and two dimerization modes in complexes of ribosomal protein L12. EMBO J 19:174–186

Watanabe K, Nelson J, Harayama S, Kasai H (2001) ICB database: the gyrB database for identification and classification of bacteria. Nucleic Acids Res 29:344–345

Woese CR (1987) Bacterial evolution. Microbiol Rev 51: 221–271

Wong JT, Chen J, Mat WK, Ng SK, Xue H (2007) Polyphasic evidence delineating the root of life and roots of biological domains. Gene 403:39–52

Worbs M, Huber R, Wahl MC (2000) Crystal structure of ribosomal protein L4 shows RNA-binding sites for ribosome incorporation and feedback control of the S10 operon. EMBO J 19:807–818

Worning P, Jensen LJ, Nelson KE, Brunak S, Ussery DW (2000) Structural analysis of DNA sequence: evidence for lateral gene transfer in *Thermotoga maritima*. Nucleic Acids Res 28:706–709

Wu H, Jiang L, Zimmermann RA (1994) The binding site for ribosomal protein S8 in 16S rRNA and spc mRNA from *Escherichia coli*: minimum structural requirements and the effects of single bulged bases on S8–RNA interaction. Nucleic Acids Res 22:1687–1695

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

Zhaxybayeva O, Gogarten JP, Charlebois RL, Doolittle WF, Papke RT (2006) Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. Genome Res 16:1099–1108

Zhaxybayeva O, Swithers KS, Lapierre P et al (2009) On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales. Proc Natl Acad Sci USA 106:5865–5870

CHAPTER 3

Molecular signatures for the phylum Synergistetes and some of its subclades[1]

This chapter follows the species of the phylum Synergistetes. Utilizing comparative genomics, conserved signature indels for the phylum and its sub-groups are identified. The CSIs are also compared with phylogenetic trees to highlight the groupings of organisms within the phylum. 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.

[1] Due to limited space, supplementary figures (1-74) for this manuscript are not included in the chapter but can be accessed along with the rest of the manuscript at:

Bhandari, V., and Gupta, R. S. (2012). Molecular signatures for the phylum Synergistetes and some of its subclades. Antonie van Leeuwenhoek *102*: 517–540. DOI 10.1007/s10482-012-9759-2

The following publication has been reproduced with kind permission from Springer Science+Business Media B.V.

Antonie van Leeuwenhoek
DOI 10.1007/s10482-012-9759-2

REVIEW PAPER

Molecular signatures for the phylum Synergistetes and some of its subclades

Vaibhav Bhandari · Radhey S. Gupta

Received: 4 April 2012 / Accepted: 30 May 2012

Abstract Species belonging to the phylum Synergistetes are poorly characterized. Though the known species display Gram-negative characteristics and the ability to ferment amino acids, no single characteristic is known which can define this group. For eight Synergistetes species, complete genome sequences or draft genomes have become available. We have used these genomes to construct detailed phylogenetic trees for the Synergistetes species and carried out comprehensive analysis to identify molecular markers consisting of conserved signature indels (CSIs) in protein sequences that are specific for either all Synergistetes or some of their sub-groups. We report here identification of 32 CSIs in widely distributed proteins such as RpoB, RpoC, UvrD, GyrA, PolA, PolC, MraW, NadD, PyrE, RpsA, RpsH, FtsA, RadA, etc., including a large >300 aa insert within the RpoC protein, that are present in various Synergistetes species, but except for isolated bacteria, these CSIs are not found in the protein homologues from any other organisms. These CSIs provide novel molecular markers that distinguish the species of the phylum Synergistetes from all other bacteria. The large numbers of other CSIs discovered in this work provide valuable information that supports and consolidates evolutionary relationships amongst the sequenced Synergistetes species. Of these CSIs, seven are specifically present in *Jonquetella*, *Pyramidobacter* and *Dethiosulfovibrio* species indicating a cladal relationship among them, which is also strongly supported by phylogenetic trees. A further 15 CSIs that are only present in *Jonquetella* and *Pyramidobacter* indicate a close association between these two species. Additionally, a previously described phylogenetic relationship between the *Aminomonas* and *Thermanaerovibrio* species was also supported by 9 CSIs. The strong relationships indicated by the indel analysis provide incentives for the grouping of species from these clades into higher taxonomic groups such as families or orders. The identified molecular markers, due to their specificity for Synergistetes and presence in highly conserved regions of important proteins suggest novel targets for evolutionary, genetic and biochemical studies on these bacteria as well as for the identification of additional species belonging to this phylum in different environments.

Keywords Conserved indels · Molecular signatures · Synergistetes phylogeny and taxonomy · *Jonquetella-Pyramidobacter-Dethiosulfovibrio* clade · *Aminomonas-Thermanaerovibrio* clade · Lateral gene transfers

Electronic supplementary material The online version of this article (doi:10.1007/s10482-012-9759-2) contains supplementary material, which is available to authorized users.

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

Published online: 19 June 2012

Introduction

The Synergistetes group of bacteria is a recently recognized phylum for which 40 organisms have been isolated and over three hundred 16S rRNA sequences are available (Hugenholtz et al. 2009; NCBI Taxonomy 2012). The phenotypic characteristics shared by the species from this phylum include their gram-negative cell wall, anaerobic existence, and rod/vibrioid cell shape (Jumas-Bilak et al. 2009). Although the presence of lipopolysaccharides, which is an important characteristic of diderm cell envelopes, in Synergistetes species has not yet been reported, they do contain genes for various proteins that are involved in lipopolysaccharide biosynthesis (Sutcliffe 2010). While a few species have been shown to be asaccharolytic, all Synergistetes have the ability to ferment amino acids (Magot et al. 1997; Baena et al. 1998, 1999a; Surkov et al. 2001; Hongoh et al. 2007; Downes et al. 2009; Jumas-Bilak et al. 2009). The Synergistetes inhabit primarily anaerobic environments including animal gastrointestinal tracts, soil, oil wells and wastewater treatment plants and they are also present in sites of human diseases such as cysts, abscesses and areas of periodontal disease (Godon et al. 2005; Kumar et al. 2005; de Lillo et al. 2006; Horz et al. 2006; Jumas-Bilak et al. 2007; Vartoukian et al. 2007; Zijnge et al. 2010). Due to their presence at illness related sites, the Synergistetes are suggested to be opportunistic pathogens but they can also be found in healthy individuals in the microbiome of the umbilicus and in normal vaginal flora (Vartoukian et al. 2007; Marchandin et al. 2010). Other species from this phylum have been identified as significant contributors in the degradation of sludge for production of biogas in anaerobic digesters and are potential candidates for use in renewable energy production through their production of hydrogen gas (McSweeney et al. 1993; Maune and Tanner 2012; Delbes et al. 2001; Riviere et al. 2009; Ziganshin et al. 2011).

Synergistetes species were first identified with the isolation of *Synergistes jonesii* from which the phylum name "Synergistetes" is derived. *S. jonesii* was isolated from the rumen of a goat in 1992 and described as a gram-negative staining, anaerobic, rod-shaped, commensal bacteria with the ability to degrade the toxic compound pyridinediol 3-hydroxy-4-1(*H*)-pyridone (Allison et al. 1992; McSweeney et al. 1993). *S. jonesii* 16S rRNA was not closely related to that of any bacteria characterized at the time but the species was later misclassified as a member of the Deferribacteres group (Garrity et al. 2004). Around the same time, the species *Thermoanaerovibrio acidaminovorans* was isolated from methanogenic sludge and indicated to be a member of the *Selenomonas* genus within the Firmicutes phylum (Guangsheng et al. 1992; Baena et al. 1999b). These misclassifications of Synergistetes continued, as species from the genera *Aminobacterium* and *Dethiosulfovibrio* were described as forming a deep-branching clade beside cluster V within Clostridia, consisting of a group composed of *Thermoanaerobacter* species (Magot et al. 1997; Baena et al. 1998). Several other organisms now considered Synergistetes were also placed among the Syntrophomonadaceae family of the Firmicutes (Garrity et al. 2004; Dahle and Birkeland 2006; Diaz et al. 2007). Eventually, efforts based on 16S rRNA sequences by Jumas-Bilak et al. (2009), identified the monophyletic nature of the Synergistetes within the bacterial domain and proposed that these "*Synergistia jonesii*-like" species form a distinct phylum, now named the Synergistetes (Jumas-Bilak et al. 2009). All characterized Synergistetes species are currently placed under the class Synergistia, the order Synergistiales and the family Synergistaceae (Jumas-Bilak et al. 2009). This family until recently was comprised of 11 genera, namely: *Aminiphilus*, *Aminobacterium*, *Aminomonas*, *Anaerobaculum*, *Cloacibacillus*, *Dethiosulfovibrio*, *Jonquetella*, *Pyramidobacter*, *Synergistes*, *Thermoanaerovibrio* and *Thermovirga* (Hugenholtz et al. 2009; Jumas-Bilak et al. 2009; NCBI Taxonomy 2012). Recently, a new genus *Fretibacterium* has also been described, which contains a single species *Fretibacterium fastidiosum* that was previously known as *Synergistes bacterium* SGP1. A candidate genus *Tammella*, composed of a group of related and uncultured species found within termite guts, has also been suggested to belong to the phylum Synergistetes (Hongoh et al. 2007; Hugenholtz et al. 2009).

While the Synergistetes are currently classified as belonging to a separate phylum based on their 16S rRNA sequences, no characteristic of these bacteria is known that can easily differentiate a Synergistetes species from other bacteria. Though all cultured Synergistetes can ferment amino acids, various species from other taxa also share this ability (Hou et al. 2004; Fonknechten et al. 2010). The availability of genome sequences has allowed for the employment of comparative genome approaches for the identification of molecular markers that are specific for different bacterial groups at various taxonomic levels (Gupta 1998; Griffiths et al. 2005; Gupta and Bhandari 2011; Gupta and Shami 2011). Using genomic sequences, our lab has pioneered the discovery of conserved signature insertions/deletions (i.e. indels, CSIs) present in protein sequences that are specific for particular groups of organisms (Gupta 1998, 2009; Gao and Gupta 2005; Griffiths and Gupta 2006; Gupta and Bhandari 2011). The group specific presence of CSIs can be parsimoniously explained through rare genetic changes occurring in a common ancestor to the particular groups of species and then being passed down through vertical descent (Gupta 1998, 2000, 2009). Such CSIs, which are present in a related group of species and absent in other organisms, are useful as molecular markers for the identification of species belonging to a taxonomic group and the demarcation of the group's boundaries. Additionally, through comparison of sequences and based on the presence or absence of the indicated CSIs in outgroup species, a rooted phylogenetic relationship can be inferred among the species (Rivera and Lake 1992; Baldauf and Palmer 1993; Gupta 1998, 2001).

From the species identified as Synergistetes, complete or annotated draft genomes are now available for nine species (described below). In the present work, we have carried out detailed comparative analyses on protein sequences from these genomes to identify molecular markers (CSIs) that are specific for the phylum Synergistetes and some of its subgroups, as well as those that provide information regarding its relationship to other bacterial phyla. Our work has identified numerous CSIs that provide highly specific markers for all sequenced members of the Synergistetes phylum as well as a number of its sub-groups. Additionally, several CSIs that are commonly shared by Synergistetes and some species from other bacterial phyla suggest potential cases of lateral gene transfers. These CSIs provide novel and powerful means for the identification/circumscription of species from the phylum Synergistetes and for different types of studies on them.

Phylogenetic analysis of the genome sequenced Synergistetes

The complete genomes for *Aminobacterium (Amb.) colombiense* (Chertkov et al. 2010), *T. acidaminovorans* (Chovatia et al. 2009) and *Thermovirga (Tv.) lienii* (Dahle and Birkeland 2006) have been published while annotated draft genomes were accessible for *Dethiosulfovibrio peptidovorans* (Labutti et al. 2010), *Aminomonas (Amm.) paucivorans* (Pitluck et al. 2010), *Anaerobaculum (An.) hydrogeniformans*, *Jonquetella anthropi* and *Pyramidobacter piscolens* (NCBI genomic database 2012). Limited sequence data for *F. fastidiosum*, which is currently referred to as *Synergistetes bacterium SGP1* in the NCBI database, was also available (Vartoukian et al. 2012). These species represent nine of twelve characterized genera from the phylum. Some characteristics of these organisms and their genomes are provided in Table 1.

The relationships of the species in the Synergistetes phylum have thus far been primarily analyzed through 16S rRNA sequence data. However, it is now recognized that trees based on a larger dataset of genes or proteins representing diverse functional categories are more reliable in resolving phylogenetic relationships than a single gene such as the 16S rRNA or a single protein (Rokas et al. 2003; Ciccarelli et al. 2006; Wu and Eisen 2008). Therefore, in order to visualize the relationship among the sequenced Synergistetes species, phylogenetic trees based upon concatenated sequences of ten housekeeping proteins were constructed. The 10 proteins that were used for phylogenetic analysis (viz. ArgRS, GyrB, Hsp70, ribosomal proteins L1 and L5, RpoB, RpoC, TrxB, UvrD and ValRS) are found in most bacteria and they have been extensively used for other phylogenetic studies (Bocchetta et al. 2000; Ciccarelli et al. 2006; Soria-Carrasco et al. 2007; Zhaxybayeva et al. 2009; Naushad and Gupta 2012). In addition to the Synergistetes species, the dataset that was employed for phylogenetic analyses also contained information for species from several other bacterial phyla including those in whose proximity the Synergistetes species were observed to branch in earlier studies (Guangsheng et al. 1992; Magot et al. 1997; Baena et al. 1998; Garrity et al. 2004; Diaz et al. 2007; Herlemann et al. 2009). The results for the multi-protein concatenated phylogenetic analysis are presented in Fig. 1a. In parallel, a 16S rRNA tree was also created to investigate the congruence with the protein tree (Fig. 1b).

In both the protein tree and the 16S rRNA tree, the Synergistetes species formed a monophyletic clade that was distinct from all other bacterial groups,

Table 1 Characteristics of the Synergistetes species with sequenced genomes

Synergistetes species	GC (%)	Isolation source	Asaccharolytic	Optimum temperature (°C)	Size (Mb)	Reference
Aminobacterium colombiense	45.3	Anaerobic lagoon of dairy wastewater treatment plant	Yes	37	1.98	DOE-JGI[a]
Aminomonas paucivorans	43	Anaerobic lagoon of dairy wastewater treatment plant	Yes	35	2.6	DOE-JGI[a]
Anaerobaculum hydrogeniformans	46.6	Oil-well production water	No	55	2.3	GSC-WashU[b]
Dethiosulfovibrio peptidovorans	54.4	Oil well	Yes	42	2.6	DOE-JGI[a]
Jonquetella anthropi	59.4	Peritoneal fluid, Breast and pelvic abscess, sebaceous cyst and wounds	Yes	37	1.7	DOE-JGI[a]
Pyramidobacter piscolens	59	Human oral cavity	Yes	37	2.6	JCV[c]
Thermanaerovibrio acidaminovorans	63.8	Granular methanogenic sludge	No	55	1.85	DOE-JGI[a]
Thermovirga lienii	46.6	Oil-well production water	Yes	58	2.06	DOE-JGI[a]
Fretibacterium fastidiosum	63	Subgingival plaque	Yes	37	Unknown	Vartoukian et al. (2010)

[a] DOE-JGI—these genomes have been sequenced by the United States Department of Energy Joint Genomic Institute

[b] GSC-WashU—Genome Sequencing Center (GSC) at Washington University (WashU) School of Medicine

[c] JCV—J. Craig Venter Institute

supporting their assignment into a separate phylum. The species such as *Syntrophomonas wolfei* or *Selenomonas sputigena* in whose proximity some of the Synergistetes species were indicated to branch in earlier studies, branched distinctly from them. Although the relationships among other bacterial species/phyla differed within the two trees and they were mostly unresolved, within the Synergistetes clade both the concatenated protein tree and the rRNA tree displayed a similar branching order. In both trees, the Synergistetes species showed a split into two clades at the highest level. One clade is comprised of *Thermanaerovibrio* and *Aminomonas* sharing a distant relationship with the *Thermovirga* and *Anaerobaculum* species while the other clade is comprised of the five species from the genera *Aminobacterium*, *Jonquetella*, *Pyramidobacter, Dethiosulfovibrio* and *Fretibacterium*. The concatenated protein tree shows, with high statistical support, that the *J. anthropi* and *P. piscolens* species branch together and that *D. peptidovorans* is the closest relative of these two species. The trees also show a well-supported, close relationship between *T. acidaminovorans* and *Amm. paucivorans.* The species *Amb. colombiense* and *F. fastidiosum* are observed to robustly branch together, though their relationship to the *Dethiosulfovibrio–Pyramidobacter–Jonquetella* clade was strongly supported only by the NJ concatenated protein tree. The ML analysis and the rRNA tree weakly supported this relationship. The position of *An. hydrogeniformans* and *Tv. lienii* species within the phylum was poorly resolved in both the concatenated protein and the rRNA trees. In both the trees, short branches connect these species to the *Thermoanaerovibrio–Aminomonas* clade and their grouping in this clade was weakly supported by the ML tree and the rRNA tree.

Identification of CSIs that are specific for the Synergistetes species

For the identification of CSIs, BlastP searches against the non-redundant protein sequence (nr) database were carried out on all proteins from the genome of the species *Amb. colombiense DSM 12261* and *T. acidaminovorans DSM 6589* from the Synergistetes phylum (Altschul et al. 1997, 2005). Using the ClustalX program, multiple sequence alignments were

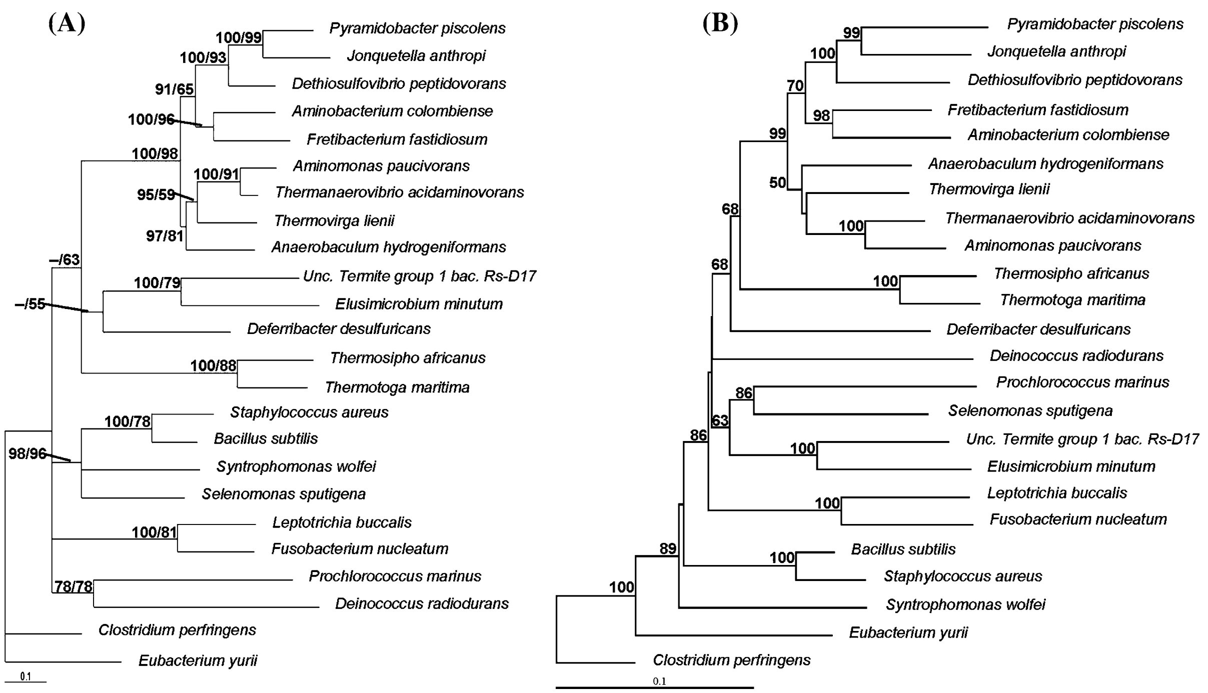


Fig. 1 Phylogenetic tree for sequenced Synergistetes species and representative species from some closely related bacterial phyla. **a** A neighbour joining (NJ) distance tree based upon concatenated sequences for 10 highly conserved and widely distributed proteins (ArgRS, GyrB, Hsp70, ribosomal proteins L1 and L5, RpoB, RpoC, TrxB, UvrD and ValRS). The numbers on the node indicate the % bootstrap score (or statistical support) for each node in the NJ and maximum-likelihood analyses, respectively. The *dashes* (–) at nodes indicate that the statistical support for this particular branching relationship was <50 % in the NJ or ML analysis. **b** A NJ tree for the same species as shown in (**a**) based upon 16S rRNA sequences. The trees were constructed as described in our earlier work (Gupta and Bhandari 2011)

created for all proteins for which high scoring homologues were available from most Synergistetes species as well as several other groups of organisms. The aligned proteins were visually inspected to identify insertions or deletions that were flanked by conserved amino acids on both sides. Insertions and deletions that were not flanked by at least 4–5 conserved residues within the neighbouring 30–40 residues were not further considered as they do not provide useful molecular markers (Gupta 1998, 2001; Gupta and Bhandari 2011). More detailed BlastP searches (searching for 250 of the closest sequence matches) were then carried out on 50–80 aa long segments (longer in some cases) containing the indels and its flanking conserved regions to determine the species distribution for the identified indels. Indels predominantly found in Synergistetes species or those that were found in Synergistetes along with some other taxonomic group of organisms were retained and compiled into signature files. The signature files shown here contain sequence alignments of various detected indels along with the flanking conserved regions for all Synergistetes and representative species from other taxonomic groups for which information was detected in the Blast searches. However, due to spatial considerations sequence information for only limited numbers of species from other groups are shown here. In a few cases, where more than one homologue of a protein was detected for the same species, sequence information for different homologues was included only if they showed differing characteristics (viz. one homologue contained the indel while the other(s) did not). All of the indels reported in this work are independent of each other and they are not part of indels for other larger clades.

CSIs that are specific for the Synergistetes phylum

CSIs in proteins brought about by rare genomic changes that are restricted to phylogenetically well-defined groups are useful as molecular markers that provide means for evaluating evolutionary relationships (Gupta 1998; Rokas and Holland 2000). Our analyses of genome sequences from Synergistetes species have identified 32 CSIs that help define and demarcate the species of this phylum. Some characteristics for these Synergistetes specific CSIs are listed in Table 2 and two examples are provided in Fig. 2. Figure 2a depicts two inserts that are present in close proximity within the β' subunit of the RNA polymerase enzyme (RpoC), an essential enzyme responsible for transcription of genes in all organisms. The region of the protein shown is highly conserved among all organisms and it contains a 2 aa insert that is specifically present in all homologues of the RpoC enzyme from species of the phylum Synergistetes. Another example of a conserved indel that is specific for all Synergistetes species is shown in Fig. 2b. In this case, the α-subunit of DNA polymerase III contains a 1 aa insert that is specifically present in all Synergistetes species (Fig. 2b) but not in any other bacteria. The absence of amino acid residue both the CSIs shown in Fig. 2 in all other organisms except Synergistetes indicates that these CSIs constitute inserts rather than deletions. Within the RpoC protein, in the neighborhood of the conserved insert that is shown in Fig. 2a, another very large insert consisting of between 311 and 316 aa is also uniquely present in all sequenced Synergistetes species. The sequence region corresponding to this large insert is shown in Fig. 3. BlastP searches with this insert show no significant hits for any proteins from organisms outside of other Synergistetes, indicating that this insert is a distinctive characteristic of the species of this phylum. Because of its large size, this large insert likely forms a unique domain of the RpoC protein that is only found in the Synergistetes species.

Other indels present in all genome sequenced Synergistetes, and absent in species from other taxonomic groups, are depicted in Supplementary Figs. 1–12 and some characteristics of them are summarized in Table 2 . These indels are present in proteins involved in important cellular processes such as DNA replication (e.g. DNA polymerase I), protein translation (30S ribosomal protein S1) and cell metabolism (2-oxoglutarate synthase). For some of these Synergistetes specific CSIs, protein homologues for one or more Synergistetes species were not detected (Supplementary Figs. 7–13). A 3 aa insert in 2-oxoglutarate synthase (Supplementary Fig. 7) is an example of such an indel. The insert is present in all detected Synergistetes but in this case the homologue for *F. fastidiosum* was not found in BlastP searches. It is possible that the gene coding for this protein has been lost from this species due to genetic, environmental or physiological factors. However, as fully published genome sequences among the Synergistetes species are available for only *T. acidaminovorans*, *Tv. lienii* and *Amb. colombiense*, the lack of a protein homologue for some of these species could also be due to the fact that their entire genomes have not yet been sequenced and/or annotated. Nevertheless, since these CSIs are only found in the Synergistetes species and not in any other bacteria (0/250; top 250 blast hits), they also provide reliable molecular markers for this group.

The indels identified above are completely specific for the Synergistetes species. However, for a small number of other CSIs discovered in this work, along with their presence in all Synergistetes, these CSIs were also found in a small number (usually 1–2) of species belonging to other taxonomic groups. Two such examples are shown in Fig. 4. The first of these is another 2 aa long insert in the RpoC protein (Fig. 4a). This CSI is found in all Synergistetes in a highly conserved region of the protein, however it is also present in the species *Eubacterium yurii* from the Clostridia class of the phylum Firmicutes. The CSI is not present in any other organisms, including other Firmicutes species. Different possibilities exist for the presence of the CSI in a single species outside of the phylum. The shared presence of the CSI in *E. yurii*, a species not considered to be directly related to the Synergistetes (see Fig. 1), might be the result of a lateral gene transfer event wherein the Synergistetes gene containing the indel might have been introduced into *E. yurii*. Alternatively, it is possible that two separate genetic events led to the presence of similar CSIs in Synergistetes and *E. yurii*. A second example of such an indel is shown in Fig. 4b. Here a 2 aa insert is present within the 30S ribosomal protein S8 of Synergistetes species and an *uncultured Termite group 1 phylotype RS-D17* considered to belong to the phylum Elusimicrobia. As shown, *Elusimicrobium minutum* itself contains a 1 aa insert in a similar position in the protein. It is possible that the Elusimicrobia are a sister taxon of the Synergistetes and the indel has been passed on to both phyla through a

Table 2 Characteristics of the CSIs that are specific for the Synergistetes phylum

Protein name	Gene name	GenBank identifier	Fig. no.	Indel size	Indel position[a]	Species distribution of indel	
						Synergistetes[b]	Other organisms[c]
DNA-directed RNA polymerase, β' subunit	rpoC	294101621	Fig. 2a	2 aa ins	1197–1241	9/9	–
DNA polymerase III, α subunit	polC	288574785	Fig. 2b	1 aa ins	762–807	9/9	–
DNA-directed RNA polymerase, β' subunit	rpoC	294101621	Fig. 3	313 aa ins	1197–1601	9/9	–
30S ribosomal protein S1	rpsA	282856368	Suppl. Fig. 1	1 aa del	341–386	9/9	–
S-adenosylmethyltransferase MraW	mraW	294101810	Suppl. Fig. 2	1 aa ins	182–220	9/9	–
Ribose phosphate pyrophosphokinase	prsA	294101745	Suppl. Fig. 3	5 aa ins	137–186	9/9	–
UvrD/REP helicase	uvrD	295112231	Suppl. Fig. 4	2–3 aa ins	61–125	9/9	–
GTP-binding protein TypA	typA	288574323	Suppl. Fig. 5	11–13 aa ins	219–287	9/9	–
2-Oxoglutarate synthase	korA	288574831	Suppl. Fig. 6	3 aa ins	126–167	8/8	–
MazG family protein	–	288574801	Suppl. Fig. 7	2 aa ins	154–183	8/8	–
Integrase family protein	–	294102004	Suppl. Fig. 8	1 aa del	264–294	8/8	–
FAD dependent oxidoreductase	–	288574196	Suppl. Fig. 9	2 aa ins	263–290	8/8	–
DNA gyrase, A subunit	gyrA	288573325	Suppl. Fig. 10	1 aa del	562–594	8/8	–
DEAD/DEAH box helicase domain protein	–	294101046	Suppl. Fig. 11	2–3 aa del	278–324	7/7	–
Serine *O*-acetyltransferase	cysE	294101111	Suppl. Fig. 12	4 aa ins	18–53	6/6	–
DNA directed RNA polymerase, β' subunit	rpoC	288574655	Fig. 4a	2 aa ins	44–93	9/9	*Eubacterium yurii*
Ribosomal protein S8	rpsH	294101642	Fig. 4b	2 aa ins	39–86	9/9	*Termite group 1 Rs-D17*
RecA protein	recA	294101403	Suppl. Fig. 13	2 aa ins	179–228	8/8	*Eubacterium yurii*
UvrD/REP helicase	uvrD	295112231	Suppl. Fig. 14	3–4 aa ins	88–122	9/9	*Ktedonobacter racemifer*
DNA polymerase I	polA	260654849	Suppl. Fig. 15	3–4 aa ins	520–552	9/9	*Bacteriovorax marinus*
DNA-directed RNA polymerase, β subunit	rpoB	282857671	Suppl. Fig. 16	6–9 aa ins	977–1020	9/9	*Rubrobacter xylanophilus*
Polyribonucleotide nucleotidyltransferase	pnp	289523245	Suppl. Fig. 17	2 aa ins[d]	342–390	8/8	*Gemmatimonas aurantiaca*

Table 2 continued

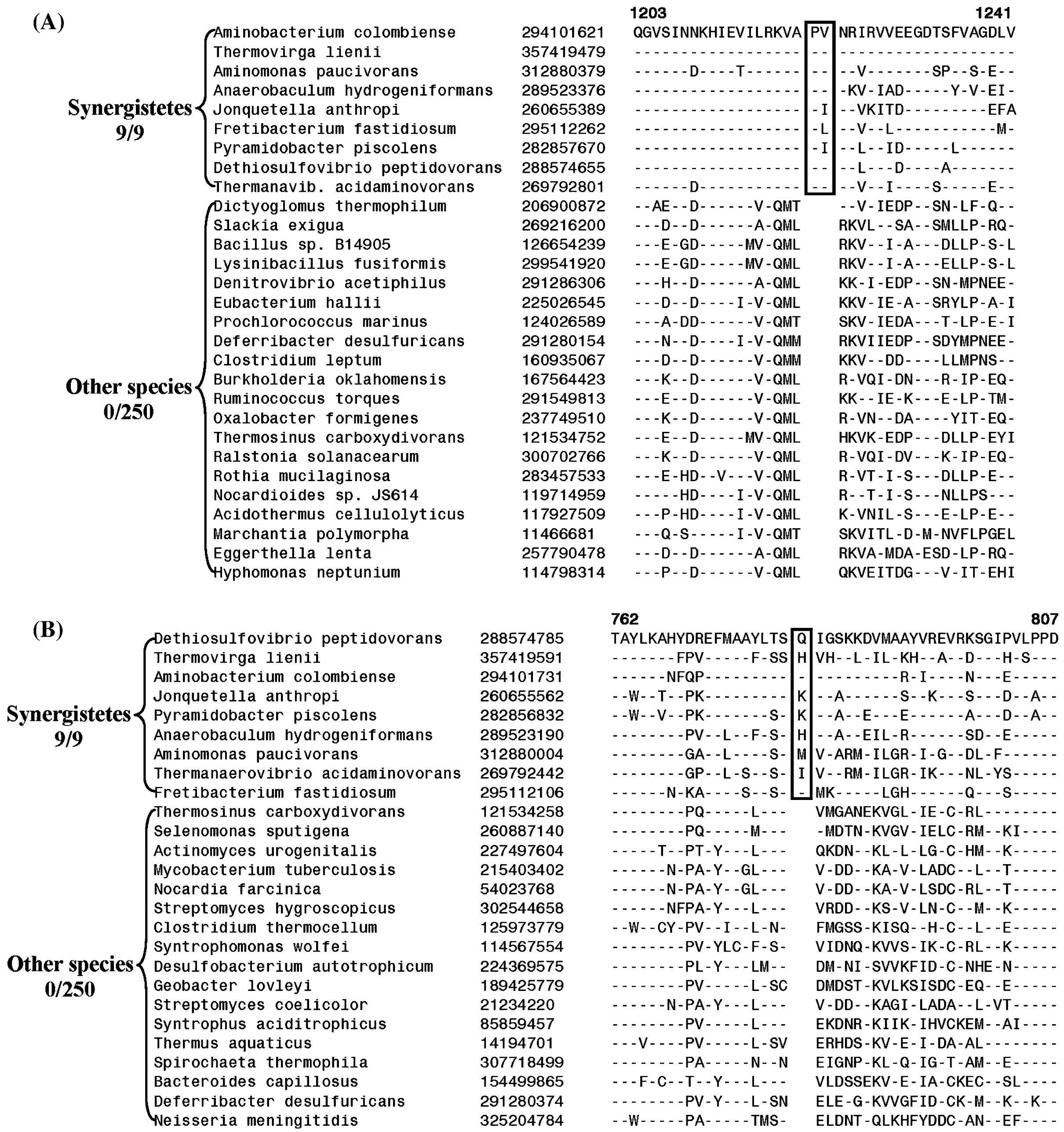
Protein name	Gene name	GenBank identifier	Fig. no.	Indel size	Indel position[a]	Species distribution of indel	
						Synergistetes[b]	Other organisms[c]
Protein of unknown function DUF1385	–	288574776	Suppl. Fig. 18	2 aa del	249–289	7/7	*Elusimicrobium minutum*, *Thermotoga lettingae*, *Fervidobacterium nodosum*
Helicase domain protein	–	288574739	Suppl. Fig. 19	3 aa del	550–604	5/5	*Syntrophus aciditrophicus*
ABC transporter, periplasmic substrate binding protein	–	289523650	Suppl. Fig. 20	2 aa del	121–164	4/4	*Desulfobacterium autotrophicum*
3,4-Dihydroxy-2-butanone 4-phosphate synthase	ribB	294101989	Suppl. Fig. 21	3 aa ins	221–270	7/8	–
DEAD/DEAH box helicase domain protein	–	282856966	Suppl. Fig. 22	6–9 aa ins	28–74	7/8	–
DNA repair protein RadA	radA	289523258	Suppl. Fig. 23	2 aa del	129–184	7/8	–
Nicotinate nucleotide adenylyltransferase	nadD	294101840	Suppl. Fig. 24	2 aa ins	94–121	7/8	–
Orotate phosphoribosyltransferase	pyrE	269792697	Suppl. Fig. 25	1 aa del	82–128	8/9	–
Cell division protein FtsA	ftsA	260655410	Suppl. Fig. 26	1–2 aa ins	10–45	7/8	*Sutterella wadsworthensis*
Putative metal dependent phosphohydrolase	–	260654227	Suppl. Fig. 27	3–4 aa ins	208–250	8/9	*Syntrophothermus lipocalidus*, *Thermotogales bacterium mesG1*

[a] The indel position provided indicates the region of the protein containing the CSI

[b] Homologous sequences corresponding to the region containing the CSI's could not be detected for some of the eight Synergistetes species and this is indicated accordingly by the second of the two numbers

[c] BLAST searches were carried out for the top 250 hits. The number of non-Synergistetes organisms, which were observed to contain the CSI, is indicated. Species containing a larger or a shorter CSI than indicated were not included in the total

[d] The 2 aa insert indicated here consists of two separate 1 aa inserts in close proximity of each other



```
(A)
                                                                   1203                                   1241
                    Aminobacterium colombiense         294101621   QGVSINNKHIEVILRKVA PV NRIRVVEEGDTSFVAGDLV
                    Thermovirga lienii                 357419479   ------------------ -- -------------------
                    Aminomonas paucivorans             312880379   ------D----T------ -- --V-------SP--S-E--
                    Anaerobaculum hydrogeniformans     289523376   ------------------ -- -KV-IAD-----Y-V-EI-
Synergistetes       Jonquetella anthropi               260655389   ------------------ -I --VKITD---------EFA
    9/9             Fretibacterium fastidiosum         295112262   ------------------ -L --V--L-----------M-
                    Pyramidobacter piscolens           282857670   ------------------ -I --L--ID-----L------
                    Dethiosulfovibrio peptidovorans    288574655   ------------------ -- --L---D----A-------
                    Thermanavib. acidaminovorans       269792801   ------D----------- -- --V--I----S-----E--
                    Dictyoglomus thermophilum          206900872   --AE--D------V-QMT    --V-IEDP--SN-LF-Q--
                    Slackia exigua                     269216200   ---D--D------A-QML    RKVL--SA--SMLLP-RQ-
                    Bacillus sp. B14905                126654239   ---E-GD-----MV-QML    RKV--I-A---DLLP-S-L
                    Lysinibacillus fusiformis          299541920   ---E-GD-----MV-QML    RKV--I-A---ELLP-S-L
                    Denitrovibrio acetiphilus          291286306   ---H--D------A-QML    KK-I-EDP--SN-MPNEE-
                    Eubacterium hallii                 225026545   ---D--D----I-V-QML    KKV-IE-A--SRYLP-A-I
                    Prochlorococcus marinus            124026589   ---A-DD------V-QMT    SKV-IEDA---T-LP-E-I
                    Deferribacter desulfuricans        291280154   ---N--D----I-V-QMM    RKVIIEDP--SDYMPNEE-
                    Clostridium leptum                 160935067   ---D--D------V-QMM    KKV--DD----LLMPNS--
Other species       Burkholderia oklahomensis          167564423   ---K--D------V-QML    R-VQI-DN---R-IP-EQ-
   0/250            Ruminococcus torques               291549813   ---E--D------V-QML    KK--IE-K---E-LP-TM-
                    Oxalobacter formigenes             237749510   ---K--D------V-QML    R-VN--DA----YIT-EQ-
                    Thermosinus carboxydivorans        121534752   ---E--D-----MV-QML    HKVK-EDP---DLLP-EYI
                    Ralstonia solanacearum             300702766   ---K--D------V-QML    R-VQI-DV---K-IP-EQ-
                    Rothia mucilaginosa                283457533   ---E-HD--V---V-QML    R-VT-I-S---DLLP-E--
                    Nocardioides sp. JS614             119714959   -----HD----I-V-QML    R--T-I-S---NLLPS---
                    Acidothermus cellulolyticus        117927509   ---P-HD----I-V-QML    K-VNIL-S---E-LP-E--
                    Marchantia polymorpha              11466681    ---Q-S-----I-V-QMT    SKVITL-D-M-NVFLPGEL
                    Eggerthella lenta                  257790478   ---D--D------A-QML    RKVA-MDA-ESD-LP-RQ-
                    Hyphomonas neptunium               114798314   ---P--D------V-QML    QKVEITDG---V-IT-EHI

(B)
                                                                    762                                            807
                    Dethiosulfovibrio peptidovorans     288574785   TAYLKAHYDREFMAAYLTS Q IGSKKDVMAAYVREVRKSGIPVLPPD
                    Thermovirga lienii                  357419591   -------FPV-----F-SS H VH--L-IL-KH--A--D---H-S---
                    Aminobacterium colombiense          294101731   ------NFQP--------- - ---------R-I----N---E-----
                    Jonquetella anthropi                260655562   --W--T--PK--------- K --A------S--K---S---D--A--
Synergistetes       Pyramidobacter piscolens            282856832   --W--V--PK-------S- K --A--E---E------A---D--A--
    9/9             Anaerobaculum hydrogeniformans      289523190   --------PV--L--F-S- H --A--EIL-R------SD--E-----
                    Aminomonas paucivorans              312880004   --------GA--L----S- M V-ARM-ILGR-I-G--DL-F------
                    Thermanaerovibrio acidaminovorans   269792442   --------GP--L-S--S- I V--RM-ILGR-IK---NL-YS-----
                    Fretibacterium fastidiosum          295112106   ------N-KA----S--S- - MK-----LGH------Q---S-----
                    Thermosinus carboxydivorans         121534258   --------PQ-----L---   VMGANEKVGL-IE-C-RL--------
                    Selenomonas sputigena               260887140   --------PQ-----M---   -MDTN-KVGV-IELC-RM--KI----
                    Actinomyces urogenitalis            227497604   -----T--PT-Y---L---   QKDN--KL-L-LG-C-HM--K-----
                    Mycobacterium tuberculosis          215403402   ------N-PA-Y--GL---   V-DD--KA-V-LADC--L--T-----
                    Nocardia farcinica                  54023768    ------N-PA-Y--GL---   V-DD--KA-V-LSDC-RL--T-----
                    Streptomyces hygroscopicus          302544658   ------NFPA-Y---L---   VRDD--KS-V-LN-C--M--K-----
                    Clostridium thermocellum            125973779   --W--CY-PV--I--L-N-   FMGSS-KISQ--H-C--L--E-----
                    Syntrophomonas wolfei               114567554   --------PV-YLC-F-S-   VIDNQ-KVVS-IK-C-RL--K-----
Other species       Desulfobacterium autotrophicum      224369575   --------PL-Y---LM--   DM-NI-SVVKFID-C-NHE-N-----
   0/250            Geobacter lovleyi                   189425779   --------PV-----L-SC   DMDST-KVLKSISDC-EQ--E-----
                    Streptomyces coelicolor             21234220    ------N-PA-Y---L---   V-DD--KAGI-LADA--L-VT-----
                    Syntrophus aciditrophicus           85859457    --------PV-----L---   EKDNR-KIIK-IHVCKEM--AI----
                    Thermus aquaticus                   14194701    ---V----PV-----L-SV   ERHDS-KV-E-I-DA-AL--------
                    Spirochaeta thermophila             307718499   --------PA-----N--N   EIGNP-KL-Q-IG-T-AM--E-----
                    Bacteroides capillosus              154499865   ---F-C--T--Y---L---   VLDSSEKV-E-IA-CKEC--SL----
                    Deferribacter desulfuricans         291280374   --------PV-Y---L-SN   ELE-G-KVVGFID-CK-M--K--K--
                    Neisseria meningitidis              325204784   --W-----PA-----TMS-   ELDNT-QLKHFYDDC-AN--EF----
```

Fig. 2 Partial sequence alignments of conserved region within the **a** RNA polymerase β' subunit (RpoC) and **b** DNA polymerase III α subunit showing two CSIs (*boxed*) that are uniquely present in species from the Synergistetes phylum, but not in any other bacteria. The *dashes* (–) in this and all other alignments indicate identity with the corresponding amino acid on the top line. The numbers in the second column are the GenBank identifier numbers of the particular proteins. The numbers below the taxon identifiers indicate the number of species detected with the indel and the total species of the respective taxon which were detected. Only representative species are shown in the alignments, however, no other species in the indicated number of blast hits contained the indel (0/250). Information for 12 other CSIs in widely distributed proteins that are specifically present in all sequenced species from the Synergistetes phylum is provided in Supplemental Figs. 1–11 and summarized in Table 2

```
                              1197
Aminobac. colombiense         QGVYRSQGVSINNKHIEVILRKVAPVNRIRVVEEGDTSFVAGDLVWTDEIEDENEAIRRENEKNIMEATRIFS
Jonquetella anthropi          QEVYESQGVSINNKHIEVILRKVAPINRVKITDEGDTSFVAGEFAWTSDVEKEIEEIKASNEKNLTEAVESLK
Therm. acidaminovorans        QDVYRSQGVSINDKHIEVILRKVAPVNRVRVIEEGDSSFVAGELVWKEDLEAESRRISEQN-ARFLEEASFLC
Bacillus sp. B14905           QKVYRMQGVEIGDKHIEVMVRQML--RKVRVIEAGDTDLLPGSLL----------------------------
Prochlorococcus marinus       QNVYKSQGVAIDDKHIEVIVRQMT--SKVRIEDAGDTTFLPGELI----------------------------
Pseudomonas fluorescens       QDVYRLQGVKINDKHIETILRQML--RKVEIAESGDSSFIKGDQM----------------------------
                              * **  *** * :**** ::*::    :: : : **: :: *

Aminobac. colombiense         GRVVKSIAASRLTDTILQYQNMPLTEEAIRTLLRPGYLISQMVLEGDEGSDLILVIGEAAFRKRMEGLELIET
Jonquetella anthropi          GKKLSDRGGALKGLLPDGAFDKPVTEDVLRKVLAPGGAVTELYFTDEEGP-LRVVVGEAAFRKEMRGMELVEA
Therm. acidaminovorans        GAVVKDAVG--SLDGTGITRDEELTMDKLGRLLSPGVGASELICQDREGL-LRVIVGEASFKRELEGLELLSD
Bacillus sp. B14905           -------------------------------------------------------------------------
Prochlorococcus marinus       -------------------------------------------------------------------------
Pseudomonas fluorescens       -------------------------------------------------------------------------

Aminobac. colombiense         FKSEDGKE-IAAGTILTPGQLGIVTSGDPVSICVRDHETIEKLVDSSYLAEDIVVDNEVMAEKDHIFTQAMAA
Jonquetella anthropi          VTKGD-KVVPADEP-LTVAQLSVITQGAPVPHMFRSVEKLRKQQDEGYVAEAVIAESGVLAKADQLLTEDLIE
Therm. acidaminovorans        LPVADGCVISAGSR-LTASDASRIVAMGPQPLLVKDLKAMEDMVGEVYLAEDVTVDGRVLSLKDRLFDLEVFQ
Bacillus sp. B14905           -------------------------------------------------------------------------
Prochlorococcus marinus       -------------------------------------------------------------------------
Pseudomonas fluorescens       -------------------------------------------------------------------------

Aminobac. colombiense         ICFEHNIQAVKIWHSVERISVLDALQERLMNNIWGRHLTQAVDSEGNGLTDVSQMVDARIIRGLVDDEISAVD
Jonquetella anthropi          QIRRSSAPSVKIWKTVDTISVRDLLQERLINRIWGRQLKLAIGADGQALSGSVHMVDGSVVKGIVEGQIQGLV
Therm. acidaminovorans        ELRSLPVESVRIWRNPERLDLCKDVYEYLIGNYLSQRVLRVITRDGAVTEPLDNRISMEIAEGIRSGEVEAIE
Bacillus sp. B14905           -------------------------------------------------------------------------
Prochlorococcus marinus       -------------------------------------------------------------------------
Pseudomonas fluorescens       -------------------------------------------------------------------------

Aminobac. colombiense         LEG-EILTREKALIELLNTLIYGKVLLEPVVDEKGQILVDSGQEINRAMIDLLVRSQAGEFVVRPLSARHDEK
Jonquetella anthropi          FADDSNTSRETELTEALSSVVSGKVLLEDVKNSDGTVALQAGQEIGKKQLAKIVAADPTILTVRPVLDQTETV
Therm. acidaminovorans        LDGNNVVSRERVLKALLTEKVYGKVLLEPVRDVDGNVVVPSGREVSHQVMDQIVAACPGEMVVRPILAQGEHR
Bacillus sp. B14905           ---------------------------------------------------------------------DIHQ
Prochlorococcus marinus       ---------------------------------------------------------------------ELRQ
Pseudomonas fluorescens       ---------------------------------------------------------------------ELTH

                                                                                 1617
Aminobac. colombiense         TLIWDVTFVRKLREGPKCRPFVHGITKAALATESFLSAASFQQTAQVLAGAAVKGEMD
Jonquetella anthropi          QLISRVSFVRRLRLGPQWRPFIHGVTKAALATDSFLSAASFQQTAQILAGAAVRNQVD
Therm. acidaminovorans        RLIQRISFVRRLRELPTWKPVLHGITKAALATDSFLSAASFQQTAQVLASAAVRGEVD
Bacillus sp. B14905           FAEANADAVMNGKNPATCRPVILGITKASLETESFLSAASFQETTRVLTDAAIKGKRD
Prochlorococcus marinus       VEDTNQAISITGGAPSEFTPVLLGITKASLNTDSFISAASFQETTRVLTEAAIEGKSD
Pseudomonas fluorescens       VLVENERLSTEDKFVAKFTRVLLGITKASLSTESFISAASFQETTRVLTEAAVTGKRD
                                                   : *:***:* *:**:******:*:::*: **:  : *
```

Fig. 3 Partial amino acid sequence alignment of the RpoC protein showing a large insert that is specifically present in all sequenced Synergistetes species. Partial sequence for the neighbouring regions is also shown in the alignment. The *dashes* in this particular alignment represent sequence gaps. The identical and conserved residues in this alignment are indicated by * and semicolons (:), respectively. Blastp searches with the insert sequence (without the flanking region) show no significant hit for any protein except for the RpoC homologs from the Synergistetes species. Sequence information is shown for only a few Synergistetes, but this insert is present in all sequenced species

common ancestor. However, this postulation is not supported by the phylogenetic trees (Fig. 1) and it is possible that the CSI in these two taxa occurred independently or by means of LGT. The information for other CSIs where indels found in Synergistetes are also present in one or two species from other taxa is summarized in Table 2 and the sequence alignments for these are presented in Supplementary Figs. 13–20.

A further seven CSIs, specific for species of the Synergistetes phylum, were discovered where one species from the phylum was detected to lack the indel. A 3 aa insert in the 3-4-dihydroxy-2-butanone 4-phosphate synthase (Supplementary Fig. 21) is an

```
(A)                                                                44                                                  93
                  Dethiosulfovibrio peptidovorans      288574655  GLFCERIFGPTKSFECACGKYKKS GP KFKGVICDRCGVEVTDNRVRRERM
                  Thermovirga lienii                   357419479  -------------Y---------- -- -----V-------I--S------L
                  Pyramidobacter piscolens             282857670  -------------Y---------- -- -----V------------------
                  Aminobacterium colombiense           294101621  -------------Y--------R- -- ----I-------------------
Synergistetes     Aminomonas paucivorans               312880379  -----------R---------R- -- --R-I-------------------
    9/9           Thermanaerovibrio acidaminovorans    269792801  -----------R---------R- -- -----V------------------
                  Anaerobaculum hydrogeniformans       289523376  ----------VR-Y--------RN -- ----IV----------SK------
                  Jonquetella anthropi                 260655389  -------------Y---------- -- --Q--V--H-------KS------
                  Fretibacterium fastidiosum           295112262  -----------R-Y--T-----R- -- --R---------------------
                  Eubacterium yurii                    306821196  -----K------DY--N----R-D KL -HR-----K------SSK------
                  Anaerococcus lactolyticus            227486351  -----K------DY--S-----RM    RY---V-EK------KSK------
                  Cryptobacterium curtum               256826829  ------------DW--------RI    R---IV-E-------RAK------
                  Slackia exigua                       269216200  -----K------DW--------RI    R---IV-E-------RSK------
                  Corynebacterium diphtheriae          38233061   -----------RDW--------RV    RY--I--E-------KSK------
                  Micrococcus luteus                   289705072  -----K-----RDW--------RV    R---I--E-------RSK---D--
                  Brevibacterium linens                260905736  -----K-----RDW--------RV    R---I--E-------RAK------
                  Slackia heliotrinireducens           257064681  -----K------DW--S-----RV    R---IV-E-------RAK------
                  Olsenella uli                        302336390  -----K----V-DW--------GI    R---IV-E-------SAK---D--
Other species     Acidothermus cellulolyticus          117927509  -----------RDW--Y-----RV    R---I--E-------RSK------
   1/250          Collinsella intestinalis             229815794  -----K----A-DW--S-----GI    R---IV-E-------TAK------
                  Brochothrix thermosphacta            1495299    ------------DW--S-----RV    RY---V---------KSK------
                  Staphylococcus epidermidis           242372761  ------------DW--S-----RV    RY--MV---------KSK------
                  Leuconostoc citreum                  170017896  ---D--------DY--------RI    RY--IV---------SSK------
                  Bacillus subtilis                    221307920  ------------DW--H-----RV    RY---V---------RAK------
                  Enterococcus faecalis                29377681   ------------DW--------RI    RY--IV---------RSK------
                  Pediococcus pentosaceus              116493169  ---D--------DY--------RI    RY--IV---------KSK------
                  Lactobacillus brevis                 116334276  ---D--------DW--------RI    RY---V---------RSK------
                  Catenibacterium mitsuokai            224542339  ------------DW---------V    RY---V---------RSS------

(B)                                                                39                                              86
                  Aminobacterium colombiense           294101642  ILKEEGYIRNYKVINDPKK PY AVVRVFLNYGPNKERVIQGLRRISKPG
                  Thermovirga lienii                   357419500  --------K--R------L -- G-LKLY----------------------
                  Fretibacterium fastidiosum           295111409  ---D----K-V-T-T---- -- -SI-I--S---ER--------------
                  Dethiosulfovibrio peptidovorans      288574676  ---G----K---------- -- GIL----S----R--------------
Synergistetes     Pyramidobacter piscolens             282857349  ---S---------L----- -A GTIK--M-----R---V-----M----
    9/9           Jonquetella anthropi                 260655369  ---S--F------------ N- GTLKI--S----R---T---T-V----
                  Thermanaerovibrio acidaminovorans    269792780  ------------TVT-GEG SF P-L-I-------R------I--V----
                  Anaerobaculum hydrogeniformans       289523355  V--D----K--------G- -- QML-IY-H--------------V----
                  Aminomonas paucivorans               312880358  ---D--------T-T-A-- -V P-L-L-M-----R------I-------
                  Termite group 1 Rs-D17               189485099  --------S--EIKGIENQ TQ S-L-IH-R-TSKGKV-L--IK----SS
                  Elusimicrobium minutum               187251808  V------IA---AVHNET-  G G------K-T-ENDVI-N--K-V-R--
                  Ammonifex degensii                   260893358  ------F-KD-E--E-G-Q    GII-IY-K----R----T--K------
                  Caldicellulosiruptor bescii          222529716  --L---F-KD-EI-D-G-N    GII-IR-K-------A-T--K------
                  Clostridium difficile                126697654  --L---F--G-D--E-G-Q    GII-IQ-K--QEG----T--KK-----
                  Desulfotomaculum acetoxidans         258513654  ------FV-DVEY-E-G-Q    GI---Y-K--TT-----T--K------
                  Halothermothrix orenii               220930979  --------KD----EKKPQ    NAL-IY-K-SK-G-K--S--K------
                  Moorella thermoacetica               83591262   ---N----K--EY-E-N-Q    GIL-LY-K------K--T--K---C--
                  Ruminococcaceae bacterium            307694138  --LD----K-FQL-D-GTQ    G-I-IT-K--AG--K--S----V----
                  Syntrophomonas wolfei                114567827  -MQ-----KD-EFVE-G-Q    GII-IY-K---D-KK--T-IK------
Other species     Thermosediminibacter oceani          302388726  T-----F-QD-E--E-G-Q    GIIKIH-K------K--T-IK------
   1/250          Bacillus halodurans                  15612711   ---R--F--D-EY-E-S-Q    G-I-I--K--SSN----T--K------
                  Brevibacillus brevis                 226309822  ------F--DAEFVE-N-Q    GII----K--AGN----T--K------
                  Enterococcus faecalis                29374865   ---R--F--DVEY-E-D-Q    G-I----K--K-E----TN-K------
                  Geobacillus kaustophilus             56418655   ---R--F--D-EY-E-N-Q    GIL-I--K----N----T--K------
                  Lactobacillus brevis                 227509945  ---R--FV-DVEY-E-D-Q    G-I----K--KD-Q---T--K------
                  Streptococcus equi                   195977222  ---R--F-K-VE--E-G-Q    GII----K--Q-G----TN-K-V----
                  Weissella paramesenteroides           241895174  ---S--FV-DVEY-E-D-Q    G-I----K-SAD-T---T--K------
                  Thermosinus carboxydivorans          121534648  ------F-KD-E--D-G-Q    GIL--S-K--A-R-K--T-IK------
                  Ilyobacter polytropus                310779582  --------S-F---T-GN-    KNI--Y-R-S G---I-K-IK------
                  Fusobacterium ulcerans               257470840  --------A----VT-GN-    KSI--Y-K-D G-D-I-K-IK------
```

Fig. 4 Partial sequence alignments of RpoC and RpsH proteins showing two CSIs, which in addition to the Synergistetes species are also present in isolated other bacterial species. **a** Excerpt from RpoC sequence alignment depicting a 2 aa conserved insert which in addition to the Synergistetes is also present in *Eubacterium yurii*. **b** Sequence alignment of the ribosomal protein S8 (RpsH) showing a 2 aa insert which in addition to the Synergistetes is also found in an *Uncultured Termite group 1 bacterium phylotype Rs-D17*. A 1 aa insert in this position is also present in *Elusimicrobia minutum*. Sequence information for 8 other CSIs in different protein containing an isolated exception is provided in Supplementary Figs. 13–20 and Table 2

example of such a CSI. The insert is present in all detected species of the Synergistetes except for *P. piscolens*. No species outside of the phylum contain the insert. The information for other such CSIs is summarized in Table 2 and sequence alignments for them are presented in Supplementary Figs. 21–27. It is possible that these CSIs were also originally introduced in a common ancestor of the Synergistetes phylum but they were lost in some species over time due to ecological/physiological pressures or by mechanisms such as LGT followed by gene loss. In some of the CSIs described above, in addition to the CSIs that were specific for the Synergistetes, indels of different lengths were also present in species from other taxonomic groups. Due to their different lengths, these CSIs have likely originated from independent genetic events.

CSIs that are specific for subgroups of the Synergistetes phylum

All Synergistetes species are currently classified as part of a single class (Synergistia), order (Synergistales) and family (Synergistaceae) (Jumas-Bilak et al. 2009; www.bacterio.cict.fr). The relationships among the species/genera of this phylum are not well understood. In the phylogenetic trees based upon concatenated protein sequences and the 16S rRNA a number of strongly supported relationships among the species within this phylum are observed (Fig. 1). Importantly, in the present work, our analyses of protein sequences from Synergistetes have led to discovery of several CSIs that are commonly shared only by species from this phylum and that are absent in all others. These CSIs independently support a specific evolutionary relationship among these species and they, in conjunction with the results from phylogenetic analyses, can be used for determination of the relationships among the members of the phylum Synergistetes.

In the phylogenetic trees shown in Fig. 1, a clade consisting of *D. peptidovorans*, *J. anthropi* and *P. piscolens* is supported with high statistical support in both the concatenated protein tree and the rRNA tree. In our analysis, we have identified seven indels (Table 3) that are uniquely present in these three species supporting independently that these three species are closely related and form a distinct clade within the Synergistetes phylum. The first of these is a 4 aa deletion in the penicillin-binding protein 1A family protein which is involved in cell wall construction (Fig. 5). This deletion is found only in homologues of the protein from *D. peptidovorans*, *P. piscolens* and *J. anthropi* and all other Synergistetes, as well as non-Synergistetes species, lack this deletion. An additional 6 CSIs specific to these three organisms were discovered in the proteins tRNA modification enzyme TrmE, ribonucleoside diphosphate reductase, putative DEAD/DEAH box helicase, RpoB and the PlsC proteins. Information for these CSIs is summarized in Table 3 and their sequence alignments are presented in Supplementary Figs. 28–33. Among the three organisms which are part of this clade, *J. anthropi* and *P. piscolens* were observed as being more closely related to each other than either is to *D. peptidovorans*. This close association is underscored by a total of 15 CSIs, including an example that is shown

Table 3 Characteristics of the CSIs that are Specific for a Clade Consisting of *J. anthropi*, *P. piscolens* and *D. peptidovorans*

Protein name	Gene name	GenBank identifier	Figure no.	Indel size	Indel position[a]	Other species containing indel[b]
Penicillin binding protein, 1A family	–	288574813	Fig. 5	4 aa del	102–140	–
Ribonucleoside diphosphate reductase	nrdA	260654687	Suppl. Fig. 28	1 aa ins	193–225	–
Putative DEAD/DEAH box helicase	–	260655128	Suppl. Fig. 29	1 aa del	398–457	–
Putative DEAD/DEAH box helicase	–	260655128	Suppl. Fig. 30	6-8 aa ins	437–496	–
DNA directed RNA polymerase, β subunit	rpoB	282857671	Suppl. Fig. 31	13 aa ins	358–407	–
1-Acyl-sn-glycerol-3-phosphate acyltransferase	plsC	282855432	Suppl. Fig. 32	1 aa ins	57–84	*P. staleyi*, *T. mathranii*
tRNA modification GTPase TrmE	trmE	260655716	Suppl. Fig. 33	1 aa ins	263–299	*E. minutum*, *Termite group 1 Rs-D17*

[a] The indel position provided indicates the region of the protein containing the CSI

[b] BLAST searches were carried out for the top 250 hits. Organisms, other than *J. anthropi*, *P. piscolens* and *D. peptidovorans*, which were observed to contain the CSI are indicated. Species containing a larger or a shorter CSI than indicated were not included in the total

			102		140
Synergistetes 3/8	Dethiosulfovibrio peptidovorans	288574813	EDSSFYSHHGIRPLAILRSIFS		GDGGHGASTITQQLARN
	Pyramidobacter piscolens	282855808	---Q----K----------L--		-E--Q------------
	Jonquetella anthropi	260655943	---D----K-------F--LAT		-EK-Q------------
	Thermovirga lienii	357419625	---E--K-K-VDFS--I-AFWK	NVTS	-RVEQ------------
	Aminomonas paucivorans	312880037	---E--E-A-L--A----ALWI	DLSH	QRARQ-G----------
	Thermanavib. acidaminovorans	269792666	---D--E-Q-VS-T----AVLV	DLIH	RGARQ-G-------S--
	Aminobacterium colombiense	294101764	------Q-G---VT--G-ALMV	DILH	RGARQ-G----------
	Anaerobaculum hydrogeniformans	289523163	--DN--N-R--DIKG-I-AAWV	NLTK	KGTFQ-G----------
Other species 0/250	Veillonella dispar	238019116	---R-------D-VG---AVWV	NIVH	SGVSE-G----------
	Alteromonas macleodii	239996582	---R--E----D-IG-M-AAV-	LVLT	-EKRQ----L-M----G
	Enterobacter cloacae	311277669	---R--E---VD-VG-F-AASV	ALFS	-HASQ------------
	Escherichia coli	209756604	---R--E---VD-AG-F-AASV	ALFS	-HASQ------------
	Haemophilus influenzae	16272388	---R--D---LD-IG-A-AL-V	AVSN	-GASQ------------
	Klebsiella pneumoniae	206577457	---R--E---VD-VG-F-AASV	AMFS	-HASQ------------
	Pasteurella dagmatis	260912704	--AR--H---VD-IG-A-A-KV	AISK	-GASQ------------
	Salmonella enterica	160867288	---R--E---VD-VG-F-AASV	ALFS	-HASQ------------
	Shewanella amazonensis	119776506	--AR--E-Q--D-IG-I-AA-V	LAAT	-EKKQ--------V---
	Sodalis glossinidius	229258577	---R--E---VD-VG---AASI	ALLS	-NASQ------------
	Vibrio cholerae	121727599	---RY-E-Y-FD-IG-T-AA-A	VLAS	-SASQ------------
	Xenorhabdus bovienii	290473153	---R--E---VD-IGVI-AVSV	MMTS	-HASQ------------
	Yersinia pestis	262364265	---R--D---VD-VG---AVSI	AMLS	-RASQ------------
	Hydrogenobacter thermophilus	288817433	--RN----F--D-I-V--ALIA	NIRE	REITQ------------
	Veillonella atypica	303228365	---R-------D-IG---A-WV	NVVH	SGVSE-G----------
	Cylindrospermopsis raciborskii	282899343	---RY-W-F-VD--G---AV-I	NTQS	--VQQ----V------S
	Raphidiopsis brookii	282896024	---RY-W-F-VD--GV--AV-I	NTQS	--VQQ----V------S
	Bacillus subtilis	321311701	--AR--E----D-VR-GGALVA	NFKD	-F-AE-G------VVK-
	Listeria monocytogenes	254852796	--AR--E-D--D-IRLGGAVIA	NLTD	-F-AE----LS--IIKM
	Fusobacterium gonidiaformans	257466898	--KR--E----D-RGL--AV-V	NLRS	-HARQ---S------K-

Fig. 5 Partial sequence alignment of a family 1A penicillin-binding protein containing a 4 aa deletion that is specific for *D. peptidovorans*, *P. piscolens* and *J. anthropi*. Sequence information for five other CSIs that are specific for this clade of species is presented in Table 3 and Supplementary Figs. 28–33

			169		208
Jonquetella-Pyramidobacter 2/2	Pyramidobacter piscolens	282857671	AKIIPERGAWVEFV		MAGDSLSVNIDSRRKLPATLLLKAFG
	Jonquetella anthropi	260655390	----------L---		-T--K-M--V---K------------
Other Synergistetes 0/7	Aminobacterium colombiense	294101620	-----D----L--S	L	SP--V------N-K-----I------
	Dethiosulfovibrio peptidovorans	288574654	-----D----L--A	M	GS--A-Y----N-K---G--------
	Anaerobaculum hydrogeniformans	289523377	-----D----L--D	L	TP--V------N---V---A------
	Thermovirga lienii	357419478	--V--D--V-L--D	L	TP-EV----V-NKK---L-M----Y-
	Aminomonas paucivorans	312880380	--V--D----L--D	L	TP--L---K--N-K-I-V-M------
	Thermanaerovibrio acidaminovorans	269792802	--L--D----L--D	L	TPSEF---K--N-K-I-V-MF---L-
	Fretibacterium fastidiosum	295112261	--L--D----LD-S	M	AS--IV-----N-----V--------
Other species 0/250	Actinomyces coleocanis	227494537	-----S----L--E	I	DKR--VG-R--RK--Q-V-HF---L-
	Corynebacterium diphtheriae	38233060	V-V--S----L--D	V	DKR-TVG-R--RK-RQ-V-V----L-
	Cryptobacterium curtum	256826828	--V--S----L--E	T	DKR-I---R--RK--Q------R-L-
	Gordonia neofelifaecis	326382974	V-V--G----L--D	V	DKR-TVG-R--RK-RQ-V-I----L-
	Mycobacterium tuberculosis	167967486	V-V--S----L--D	V	DKR-TVG-R--RK-RQ-V-V----L-
	Nocardia farcinica	54026619	-RV--S----L--D	V	DKR-T-G-R--RK-RQ-V-V----L-
	Alkalilimnicola ehrlichii	114319612	-RV--Y--S-LD-E	F	DPK-NIY-R--R-------I--R-L-
	Coxiella burnetii	224493307	-RV--Y--S-LD-E	F	DPK-C-F-R--R-------VI-R-L-
	Nitrococcus mobilis	88812764	-RV--Y--S-LD-E	F	DPK-AIF-R--R-------I--R-L-
	Pseudomonas mendocina	330505214	-R---Y--S-LD-E	F	DPK-AVF-R--R------SV--R-L-
	Vibrio vulnificus	302347128	-R---Y--S-LD-E	F	DPK-N-Y-R--R-----S-II-R-L-
	Desulfovibrio aespoeensis	317153964	SR---M--S-LD-D	F	DHK-I-Y-R--R---M---I----M-
	Geobacter sulfurreducens	39997954	-RV--Y--S-LD-E	F	DHK-I-F-R--R---M---V----L-
	Myxococcus xanthus	89142894	-R---Y--S-ID-E	F	DHK-L-Y-R--R-------V-IR-L-
	Pelobacter carbinolicus	77918308	-RV--Y--S-LD-D	F	DHK-L-Y-R--R-------V----L-
	Thermomicrobium roseum	221632985	--L--T----L--E	T	SNR-V---K--RK--I-V-V--R-V-
	Natranaerobius thermophilus	188584828	GT---N----L--E	M	N-QNVI--RV-RT----V-V-IR-L-
	Burkholderiales bacterium	303256441	-R---Y--S-LD-E	F	D-K-I-HFR--R---M-G-IM---L-
	Flexistipes sinusarabici	336322745	-R---Y--S-ID-E	F	D-KNVMY-R--KK--I-V-V----L-
	Neorickettsia risticii	41017780	-R---Y--S-LD-E	F	D-K-V-YFR--KK----V----R-L-

Fig. 6 Excerpts from sequence alignment for RNA polymerase β subunit (RpoB) showing a 1 aa deletion that is specifically present in *J. anthropi* and *P. piscolens*. The region contains a 1 aa deletion specific for the three species. Sequence information for 14 other CSIs that are specific for these two species is presented in Table 4 and Supplementary Figs. 34–47

Table 4 Characteristics of CSIs that are specific for *J. anthropi* and *P. piscolens*

Protein name	Gene name	GenBank identifier	Figure no.	Indel size	Indel position[a]	Other species containing indel[b]
DNA-directed RNA polymerase, β subunit	rpoB	282857671	Fig. 6	1 aa del	169–190	–
Chlorohydrolase family protein	–	260654152	Suppl. Fig. 34	1 aa ins	241–272	–
Phospho-*N*-acetylmuramoyl-pentapeptide-transferase	mraY	260655416	Suppl. Fig. 35	1 aa ins	218–266	–
Recombination protein RecR	recR	260654559	Suppl. Fig. 36	1 aa ins	63–122	–
Transcriptional regulator NrdR	nrdR	260655521	Suppl. Fig. 37	1 aa del	84–133	–
Lipid A biosynthesis acyltransferase	–	282855413	Suppl. Fig. 38	1 aa ins	262–284	–
Glutamate synthase	gltB	282858093	Suppl. Fig. 39	6 aa ins	334–383	–
Acetate kinase	ackA	282856654	Suppl. Fig. 40	2 aa del	238–279	–
AcrB/D/F family transporter	–	260654486	Suppl. Fig. 41	1 aa del	111–154	–
Transcriptional regulator IclR	iclR	282857025	Suppl. Fig. 42	10–13 aa ins	139–202	–
Ribose phosphate pyrophosphokinase	kprS	282858157	Suppl. Fig. 43	1 aa ins	204–246	*Dichelobacter nodosus*
Phosphoribosyl-formylglycinamidine synthase II	purL	260655584	Suppl. Fig. 44	4 aa del	102–150	*Hydrogenobaculum* sp. Y04AAS1
MazG family protein	–	282858161	Suppl. Fig. 45	3–4 aa ins	33–75	*Clostridium phytofermentans*
Glutamate synthase	gltB	282858093	Suppl. Fig. 46	8 aa ins	84–130	*Desulfovibrio africanus*
S-adenosylmethyltransferase MraW[c]	mraW	289522914	Suppl. Fig. 47	1 aa ins	153–196	–

[a] The indel position provided indicates the region of the protein containing the CSI

[b] BLAST searches were carried out for the top 250 hits. Organisms, other than *J. anthropi* or *P. piscolens*, which were observed to contain the CSI are indicated. Species containing a larger or a shorter CSI than indicated were not included in the total

[c] All Synergistetes species were observed to contain the indel except *J. anthropi* and *P. piscolens*, thus, differentiating these two species from the rest of the phylum

in Fig. 6, a 1 aa deletion in a conserved region of the enzyme RNA polymerase β subunit. Other indels that provide similar molecular evidence for the observed close relationships between these two genera are presented in Supplementary Figs. 34–47 and information for them is summarized in Table 4. The fidelity of these molecular markers can be tested on cultured but unsequenced members of the phylum Synergistetes and as more species belonging to these genera are sequenced, the identified CSIs should provide molecular markers for their induction into the clade formed by this sub-group of the phylum.

The phylogenetic trees also support a cladal relationship among two other species, *Amm. paucivorans* and *T. acidaminovorans*, which branch as sister organisms with high statistical support (Fig. 1). The clade harbouring these genera has been proposed to form a higher-level taxon within the phylum (Jumas-Bilak et al. 2009). In the present work, we have identified 7 CSIs that differentiate the species representing these two genera from all other species and support a specific grouping of the genera *Thermanaerovibrio* and *Aminomonas* (Table 5). Among these CSIs is a 2 aa insert in enzyme *S*-adenosyl-methionine isomerase (Fig. 7). The information for 6 other CSIs supporting a specific relationship among these two species is provided in Table 5 and their sequence alignments are depicted in Supplementary Figs. 48–53. Two other CSIs identified in the present work, which include a 1–2 aa deletion in the ribosomal protein L13 (Supplementary Fig. 54) and 2 aa deletion in DNA gyrase B (Supplementary Fig. 55), are present in all detected Synergistetes species except *Thermanaerovibrio* and *Aminomonas*. The absence of these CSIs in the two species suggests that this clade may have diverged from the common Synergistetes ancestor before the other species of the phylum and the two indels may have been introduced after the divergence of this clade from the common Synergistetes ancestor. A loss of this signature from this clade after its divergence from other Synergistetes can also explain the

Table 5 Characteristics of the CSIs that are specific for a clade consisting of *T. acidaminovorans* and *Amm. paucivorans*

Protein name	Gene name	GenBank identifier	Figure no.	Indel size	Indel position[a]	Other species containing the indel[b]
S-adenosylmethionine/tRNA-ribosyltransferase-isomerase	queA	269792529	Fig. 7	2 aa ins	156–194	–
RecA protein	recA	269793250	Suppl. Fig. 48	1 aa ins	143–184	–
Glu/Leu/Phe/Val dehydrogenase	–	269791934	Suppl. Fig. 49	1 aa del	318–361	–
Uracil phosphoribosyltransferase[c]	upp	312880140	Suppl. Fig. 50	4–5 aa ins	133–184	–
Methyltransferase GidB	gidB	269791772	Suppl. Fig. 51	2 aa ins	75–123	*Sorghum bicolor*
Xanthine/uracil/vitamin C permease	–	269792033	Suppl. Fig. 52	5 aa ins	390–443	*Mesembryanthemum crystallinum*
Electron transport complex, RnfABCDGE type, C subunit	–	312880739	Suppl. Fig. 53	1 aa ins	166–213	*Saccharophagus degradans*, *marine gamma proteobacterium*, *Eubacterium cellulosolvens*
Ribosomal protein L13[d]	rplM	294101309	Suppl. Fig. 54	1–2 aa del	108–138	–
DNA gyrase subunit B[d]	gyrB	294102629	Suppl. Fig. 55	2 aa del	191–234	*Acidaminococcus fermentans*, *Acetonema longum*, *Seinonella peptonophila*
Hypothetical protein Taci_0455[e]	–	269792069	Suppl. Fig. 56	2 aa ins	205–256	–

[a] The indel position provided indicates the region of the protein containing the CSI

[b] BLAST searches were carried out for the top 250 hits. The number of non-Synergistetes organisms, which were observed to contain the CSI, is indicated. Species containing a larger or a shorter CSI than indicated were not included in the total

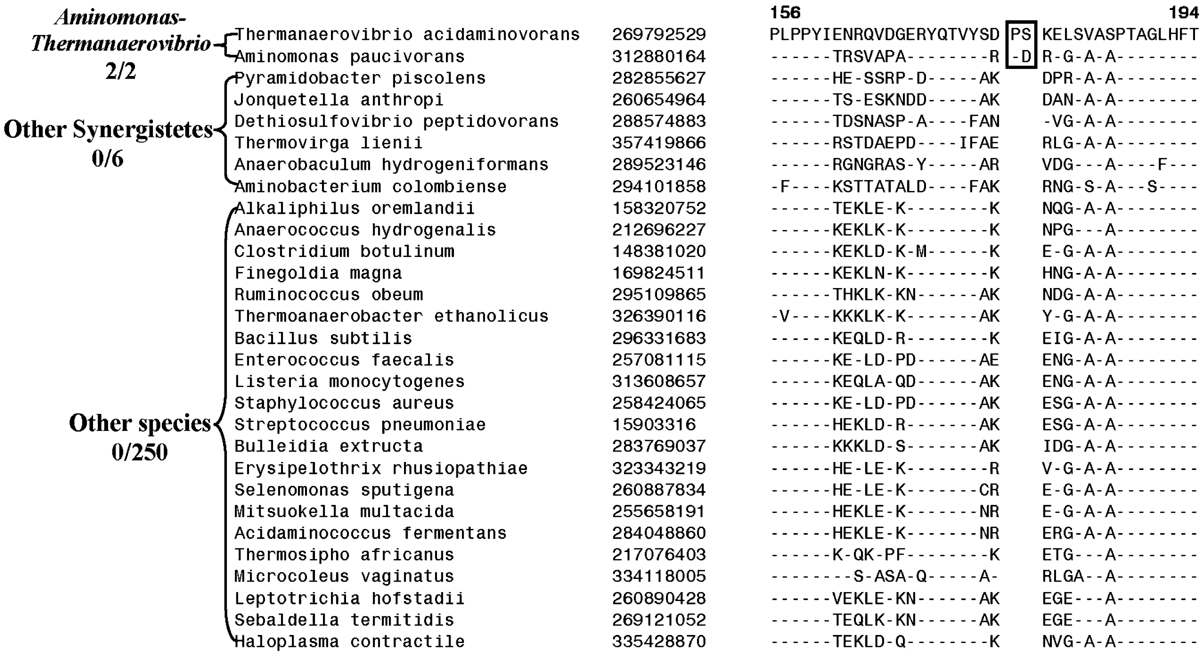
[c] BLAST searches were carried out for the top 250 hits. However, in the indicated case, no species outside of the Synergistetes phylum contained the protein homolog or the conserved region corresponding to the sequences flanking the indel

[d] All Synergistetes species were observed to contain the indel except *Amm. paucivorans* and *T. acidaminovorans*, thus, differentiating these two species from the rest of the phylum

[e] The CSI is also present in *Thermovirga lienii* species from the Synergistetes phylum

observation. These indels also support a close relationship among the genera *Thermanaerovibrio–Aminomonas* and information for them is also summarized in Table 5. These two species were observed to branch in a weakly supported clade with the *Tv. lienii* and *An. hydrogeniformans* (Fig. 1). However, only 1 CSI supporting the three-species-clade with *Tv. lienii* was identified in a protein of unknown function (Supplementary Fig. 56) and no CSI specific for all four organisms was discovered.

In the phylogenetic trees (Fig. 1), the species *Amb. colombiense* and *F. fastidiosum* were observed to branch with *J. anthropi*, *P. piscolens* and *D. peptidovorans*. A specific relationship among these species is also supported by two of the identified CSIs. The first of these CSIs consists of a 1 aa del in GyrB that is uniquely present in all five of these species (Fig. 8a). Another CSI in orotidine 5′-phosphate decarboxylase, also consisting of a 1 aa deletion, is commonly shared by *Amb. Colombiense*, *J. anthropi*, *P. piscolens* and *D. peptidovorans* (Fig. 8b). A homologue for this protein was not detected for *F. fastidiosum*, whose genome has not been fully sequenced. Thus, it is likely that this CSI will also be present in this species and could provide an additional molecular marker for this clade.



			156		194
Aminomonas-Thermanaerovibrio 2/2	Thermanaerovibrio acidaminovorans	269792529	PLPPYIENRQVDGERYQTVYSD	PS	KELSVASPTAGLHFT
	Aminomonas paucivorans	312880164	------TRSVAPA--------R	-D	R-G-A-A--------
Other Synergistetes 0/6	Pyramidobacter piscolens	282855627	------HE-SSRP-D-----AK		DPR-A-A--------
	Jonquetella anthropi	260654964	------TS-ESKNDD-----AK		DAN-A-A--------
	Dethiosulfovibrio peptidovorans	288574883	------TDSNASP-A----FAN		-VG-A-A--------
	Thermovirga lienii	357419866	------RSTDAEPD----IFAE		RLG-A-A--------
	Anaerobaculum hydrogeniformans	289523146	------RGNGRAS-Y-----AR		VDG---A----F---
	Aminobacterium colombiense	294101858	-F----KSTTATALD----FAK		RNG-S-A---S----
Other species 0/250	Alkaliphilus oremlandii	158320752	------TEKLE-K--------K		NQG-A-A--------
	Anaerococcus hydrogenalis	212696227	------KEKLK-K--------K		NPG---A--------
	Clostridium botulinum	148381020	------KEKLD-K-M------K		E-G-A-A--------
	Finegoldia magna	169824511	------KEKLN-K--------K		HNG-A-A--------
	Ruminococcus obeum	295109865	------THKLK-KN------AK		NDG-A-A--------
	Thermoanaerobacter ethanolicus	326390116	-V----KKKLK-K-------AK		Y-G-A-A--------
	Bacillus subtilis	296331683	------KEQLD-R--------K		EIG-A-A--------
	Enterococcus faecalis	257081115	------KE-LD-PD------AE		ENG-A-A--------
	Listeria monocytogenes	313608657	------KEQLA-QD------AK		ENG-A-A--------
	Staphylococcus aureus	258424065	------KE-LD-PD------AK		ESG-A-A--------
	Streptococcus pneumoniae	15903316	------HEKLD-R-------AK		ESG-A-A--------
	Bulleidia extructa	283769037	------KKKLD-S-------AK		IDG-A-A--------
	Erysipelothrix rhusiopathiae	323343219	------HE-LE-K--------R		V-G-A-A--------
	Selenomonas sputigena	260887834	------HE-LE-K-------CR		E-G-A-A--------
	Mitsuokella multacida	255658191	------HEKLE-K-------NR		E-G-A-A--------
	Acidaminococcus fermentans	284048860	------HEKLE-K-------NR		ERG-A-A--------
	Thermosipho africanus	217076403	------K-QK-PF--------K		ETG---A--------
	Microcoleus vaginatus	334118005	--------S-ASA-Q-----A-		RLGA--A--------
	Leptotrichia hofstadii	260890428	------VEKLE-KN------AK		EGE---A--------
	Sebaldella termitidis	269121052	------TEQLK-KN------AK		EGE---A--------
	Haloplasma contractile	335428870	------TEKLD-Q--------K		NVG-A-A--------

Fig. 7 Partial sequence alignment of *S*-adenosylmethionine/tRNA-ribosyltransferase-isomerase protein showing a 2 aa insert in a conserved region that is specific for *T. acidaminovorans* and *Amm. paucivorans* species. Nine other CSIs that are specific for these two species have also been identified (Table 4 and Supplementary Figs. 34–47)

CSIs that are commonly shared by species of the Synergistetes phylum with other taxonomic groups

The Synergistetes is a taxonomic group that has only recently been identified as a separate phylum within the bacterial domain. Though it branches distinctly in the 16S rRNA trees with long branches separating it from other bacterial groups (Fig. 1; Jumas-Bilak et al. 2009), species from the phylum had previously been classified as part of *Syntrophomonadaceae* family in the Firmicutes (Baena et al. 1998, 1999a; Diaz et al. 2007), grouped with *Deferribacteres* (Garrity et al. 2004) and misclassified as *Selenomonas* (Guangsheng et al. 1992). The presence or absence of CSIs that associate these groups with the Synergistetes should prove helpful in determining whether any link exists between the Synergistetes and these other groups of bacteria.

In our analysis we have identified some CSIs that, along with being present in some or all the Synergistetes species, were present in other groups of organisms. Two examples of such indels are presented in Fig. 9. Figure 9a shows a 1 aa insert in the MiaB-family RNA modification enzyme that is uniquely present in all detected Synergistetes as well as various species from the phylum *Chloroflexi*. All other bacteria lack this insert. Similarly, in the DNA polymerase III α-subunit, a 1 aa insert is present in all detected Synergistetes and also in various *Fusobacteria*, an *Opitutaceae* species as well as in *Thermomicrobium* (Fig. 9b). In phylogenetic trees constructed from these protein sequences, the Synergistetes species do not branch with species from these taxa (unpublished results) indicating that the shared presence of these CSIs is not due to their being sister taxa of Synergistetes or LGTs. The CSIs in these groups have thus likely originated independently. Other CSIs that the Synergistetes share with species from other taxonomic groups are listed in Table 6 and sequence information for them is provided in Supplementary Figs. 57–74. These other taxa include the *Fusobacteria* (Supplementary Figs. 57–61), *Elusimicrobia* (Supplementary Figs. 61, 62), class *Negativicutes* (Supplementary Figs. 63–66), *Acidobacteria* (Supplementary Fig. 67), *Proteobacteria* (Supplementary Figs. 68–70), *Aquificae* (Supplementary Fig. 71), *Erysipelotrichi* (Supplementary Fig. 72), *Actinobacteria* (Supplementary Fig. 73) and order *Lactobacillales* (Supplementary Fig. 74). The Synergistetes share the greatest number

(A)			384		413
Synergistetes 5/9	Dethiosulfovibrio peptidovorans	288574018	AREAAKKARELVR		KTAMTGLNLPGKLADCS
	Jonquetella anthropi	260655480	-------------		-S--S--S---------
	Aminobacterium colombiense	294102629	--D----------		------M----------
	Pyramidobacter piscolens	282855958	-------------		-S-LS--S---------
	Fretibacterium fastidiosum	295111747	--------K----		-G---M-----------
	Thermovirga lienii	357419158	-------------	R	---LA--D---------
	Thermanaerovibrio acidaminovorans	269791748	---------D---	R	-S--S-MD---------
	Aminomonas paucivorans	312878944	---------D---	R	-S--S--D---------
	Anaerobaculum hydrogeniformans	289523899	--D----------	R	-S-FG--D---------
Other species 0/250	Clostridium difficile	109675347	-----------T-	R	-SVLESTS--------A
	Syntrophomonas wolfei	114565581	-----------T-	R	-N-LESTA---------
	Thermocrinis albus	289549166	--------K----	R	RSPLEDTT---------
	Aquifex aeolicus	15606321	--------K----	R	-SPLEEGV---------
	Bacteroides intestinalis	189464548	--V--R----S-Q	R	-SP-S-GGM--------
	Flexibacter litoralis	9971369	--M--R----M-Q	R	---LS-TG---------
	Rhodothermus marinus	268318257	--V--R------Q	R	-N-LN-SS--------A
	Staphylococcus aureus	293497972	--V-------VT-	R	-S-LDVAS---------
	Leptospira biflexa	183219432	-----RR--D-T-	R	--VLE-GG---------
	Ilyobacter polytropus	310777811	------------M	R	-S-LEVGS---------
	Fusobacterium ulcerans	257470425	-----Q------L	R	-SVLEVGS---------
	Erysipelothrix rhusiopathiae	323342250	--V--R-----T-	R	-G-LEVSS---------
	Chlamydophila pneumoniae	15618195	-----------TL	R	-S-LDSAR-----I--L
	Leptospirillum ferriphilum	209863973	-----R--KD-AK	R	-NVLE-S---------Q
	Chlorobium phaeobacteroides	189499015	S-D--R--KD-T-	R	-S-LESSG---------
	Brachyspira murdochii	296127762	-----R---D-A-	R	-N-LESDS---------
	Microcystis aeruginosa	159028965	-A---RR--D---	R	-SVLESSP---------
	Trichodesmium erythraeum	113477398	-A---RR--D---	R	-SVLESSP---------
	Acidobacterium sp. MP5ACTX8	299136177	-----R---D-T-	R	-G-LD-GG---------

(B)			180		206
Synergistetes 4/8	Aminobacterium colombiense	294102579	PGIRPSATG		DDQARTATPKGAIIAGAD
	Jonquetella anthropi	260654639	--V-LA-A-		---T-V----D-------
	Dethiosulfovibrio peptidovorans	288575022	--V-LTSL-		---T-I---CQ--KN---
	Pyramidobacter piscolens	282856738	--V-LV-G-		---S-V---AD-FRN---
	Anaerobaculum hydrogeniformans	289523224	--V--EG-S	K	-----IM--GQ-KKK---
	Aminomonas paucivorans	312879980	----LPGD-	T	Q----VD--AA-MGR---
	Thermanaerovib. acidaminovorans	269792698	----FQGGE	V	H----VMG-RE-VAS--S
	Thermovirga lienii	357419930	-----KDFV	NK	---K-V---AE-ARS---
Other species 0/250	Pseudomonas brassicacearum	330811055	-----AGSA	Q	---R-IL--RQ-LD----
	Haemophilus parasuis	167855101	-----EGSD	F	G--R-VM---Q--EI-S-
	Actinobacillus minor	240949272	-----EGSD	F	G--R-VM---Q--ET-S-
	Dickeya dadantii	242239141	-----AGSE	A	G--R-IM--EQ-RQ--V-
	Pseudoalteromonas haloplanktis	332534575	-----EGSD	A	G--K-IM---Q--DS-S-
	Shewanella amazonensis	119774875	-----EGSE	A	G--H-IM--AQ-LQ--S-
	Lactobacillus gasseri	238853382	-----AGNA	K	---S-V---AQ-KEW-ST
	Staphylococcus epidermidis	330685500	-----EGSA	Q	N--K-IT--EQ-KQL-ST
	Streptococcus pneumoniae	149005807	-----AGAA	V	G--K-VM--AD-YQI-S-
	Salmonella enterica	161503187	-----AGSE	A	---R-IM--EQ-LS--V-
	Desulfovibrio desulfuricans	220904675	-----AGSV	A	---R-VM--AQ-VA----
	Geobacter lovleyi	189424925	--V---FAA	V	---K-IM--AE-VK----
	Brachyspira murdochii	296127421	--V--KWAS	T	---E-IM---E--EN-C-
	Acetobacter pomorum	329114068	-----AGAA	K	G--K-VM--AE-RA----
	Magnetospirillum gryphiswalden	144899231	------WAE	A	G--K-VM---E-RER---
	Nostoc punctiforme	186683999	--V--TWAD	N	A--K-SL--AQ-MK----
	Acaryochloris marina	158337488	--V--PGSV	T	G--K-AM--TA-MQ---N
	Fusobacterium gonidiaformans	257466503	--V--KWSA	T	N--E-IM---E-VQH-C-
	Erysipelothrix rhusiopathiae	323342222	-----N-SE	K	G--K-VT--AQ-KAN-SS
	Syntrophomonas wolfei	114566804	-----AWSE	K	N--K-IT--GQ-LQM---
	Eubacterium yurii	306820700	----FTDEK	T	---T-IT--SD-KRI-S-

Fig. 8 Partial sequence alignments of **a** DNA Gyrase subunit B and **b** orotidine 5′-phosphate decarboxylase showing two CSIs in conserved regions that are specific for the species *D. peptidovorans*, *J. anthropi*, *Amb. colombiense*, *P. piscolens*, *F. fastidiosum*, which form or define a higher clade within the Synergistetes group of species

Table 6 Conserved indels common to Synergistetes and shared with other groups

Protein	Beta-ketoacyl-acyl-carrier protein synthase II	Xaa-Pro dipeptidase	ATP-dependent protease La	Phosphoribosyl-formylglycin-amidine cycloligase	Translation elongation factor 1A	GTP-binding protein Era
GenBank identifier	260654633	260654796	260654937	294102156	312880374	294101756
Accession no.	ZP_05860123	ZP_05860284	ZP_05860425	YP_003554014	ZP_07740174	YP_003553614
Indel/size	2 aa del	1 aa ins	1 aa ins	1 aa del	1 aa del	1 aa ins
Indel position[a]	273–312	220–260	625–672	253–299	231–269	75–121
Figure no.	Suppl. Fig. 57	Suppl. Fig. 58	Suppl. Fig. 59	Suppl. Fig. 60	Suppl. Fig. 61	Suppl. Fig. 62
Synergistetes species containing indel	*J. anthropi* and *P. piscolens*	*D. peptidovorans*, *P. piscolens* and *J. anthropi*	*J. anthropi* and *P. piscolens*	All detected Synergistetes	*Amm. paucivorans*, *T. acidaminovorans*, *D. peptidovorans*	All detected Synergistetes
Other species with indel	Epsilon-proteobacteria	Some Fusobacteria	Fusobacteria and *Treponema pallidum*	Two δ-proteobac. and Fusobacteria	Fusobacteria, Proteobacteria, two Elusimicrobia	Some Clostridia species, *Termite group 1 Rs-D17*
Protein	NADH dehydrogenase	Translation elongation factor Tu	Penicillin-binding protein 2	Ribonuclease, Rne/Rng family	DNA-directed RNA polymerase, subunit B	Ribosomal RNA large subunit methyltransferase J
GenBank identifier	312880263	294101321	288574639	269792810	288574654	289522985
Accession no.	ZP_07740063	YP_003553179	ZP_06392996	YP_003317714	ZP_06393011	ZP_06439839
Indel/size	1 aa del	1 aa ins	3-4 aa ins	1 aa ins	1 aa ins	1 aa ins
Indel position[a]	31-86	228-270	171-205	14-47	412-457	174-233
Figure no.	Suppl. Fig. 63	Suppl. Fig. 64	Suppl. Fig. 65	Suppl. Fig. 66	Suppl. Fig. 67	Suppl. Fig. 68
Synergistetes species containing indel	All detected Synergistetes	*Amb. colombiense*, *Tv. lienii* and *An. hydrogeniformans*	*P. piscolens*, *D. peptidovorans*	*A. paucivorans*, *T. acidamino-vorans*	All detected Synergistetes	All detected Synergistetes
Other species with indel	Negativicutes	Negativicutes	*Veillonella* species	*Veillonella* species	Acidobacteria and a few clostridia	Many alpha proteobacteria
Protein	GTP-binding protein Era	Orotidine 5′-phosphate decarboxylase	Ribosomal protein S5	Homoserine kinase	1-Hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate synthase	3-Oxoacid CoA-transferase, subunit B
GenBank identifier	294101756	289523224	260655366	294101575	294101868	288573369
Accession no.	YP_003553614	ZP_06440078	ZP_05860854	YP_003553433	YP_003553726	ZP_06391726
Indel/size	4 aa ins	1 aa del	1 aa ins	1 aa ins	1 aa ins	2 aa del
Indel position[a]	213–259	149–177	105–144	222–258	171–218	43–83
Figure no.	Suppl. Fig. 69	Suppl. Fig. 70	Suppl. Fig. 71	Suppl. Fig. 72	Suppl. Fig. 73	Suppl. Fig. 74
Synergistetes species containing indel	All detected Synergistetes	Some Synergistetes	All detected Synergistetes	All detected Synergistetes	Most Synergistetes	All detected Synergistetes
Other species with indel	*Anaeromyxobacter*, *Prochlorococcus marinus*	*Pseudomonas* species and *A. caldus*	*Hydrogenothermaceae* species	*Erysipelotrichi* and *A. viridans*	Most detected Actinobacteria	Few *Streptococci*

[a] The indel position provided indicates the region of the protein containing the CSI

(five) of these CSIs with the Fusobacteria and they share only 1–2 indels with most other taxonomic groups. In many cases where the Synergistetes share CSIs with other taxa, only some species from the Synergistetes or the other taxa contain the indel. The CSIs in these other groups may have arisen independently through separate genetic events or it is also plausible that their shared presence in some of these cases is due to LGTs.

Discussion and concluding remarks

The Synergistetes are a relatively unknown group of species living ubiquitously in anaerobic environments. Though characteristics for the isolated Synergistetes are known, such as their gram-negative morphology and their ability to ferment amino acids, no single molecular, morphological or physiological characteristic is known that distinguishes them as a group from other bacterial organisms. Utilizing the available genomic data for this group of organisms, we report here identification of over 60 novel CSIs specific for the species of the Synergistetes phylum. Of the various discovered CSIs, 32 were identified to be specific for all or most Synergistetes species (maximum of three exceptions unrelated to each other). These CSIs are present in widely distributed proteins with important cellular functions and they are rarely present in protein homologues of species outside of the phylum. As they are present in most or all Synergistetes and absent in bacteria from all other taxonomic groups, they provide strong evidence that species of the Synergistetes phylum constitute a monophyletic group that is distinct from all other prokaryotic taxa. These CSIs also provide novel molecular means for identification and circumscription of species from this phylum.

The bacteria belonging to the Synergistetes have been classified into 12 different genera (and a candidate genus) within the phylum. Despite the recognition of numerous species and genera, due to the lack of reliable biological characteristics that can identify the interrelationships among these bacteria, all genera are presently grouped into a single class, order and family. Numerous CSIs were discovered during the course of the study that were present in only certain clades of species within the Synergistetes phylum and absent from others. The group specificities of these CSIs are summarized in Fig. 10. Explicitly, 7 CSI were detected to be specifically found in only the *J. anthropi*, *P. piscolens* and *D. peptidovorans* species; 15 CSIs were identified that are specific for the *J. anthropi* and *P. piscolens* species (or differentiate them from other Synergistetes) and 9 other CSIs differentiated the *T. acidaminovorans* and *Amm. paucivorans* from other members of the phyla. In addition, two of the discovered CSIs also supported a grouping together of the *J. anthropi*, *P. piscolens*, *D. peptidovorans*, *Amb. colombiense* and *F. fastidiosum* species. These relationships are also consistently observed in phylogenetic trees created for the Synergistetes group and the identified CSIs provide valuable markers that consolidate these relationships. Furthermore, it should be noted that in contrast to the CSIs supporting these relationships, very few, if any, CSIs that supported alternative relationships among these species were detected. Thus, the identified CSIs provide independent evidence for the existence of these clades and provide molecular means to demarcate and circumscribe these clades. The evidence based upon identified CSIs supports the division of the phylum Synergistetes into a number of distinct families (or other higher taxonomic groupings) and a formal proposal in this regard will be made in future work. Though the branching and interrelationships of several species within the phylum is well supported by multiple CSIs, the relationships of *Tv. lienii* and *An. hydrogeniformans*, and also to some extent *Amb. colombiense* and *F. fastidiosum* to other Synergistetes species were not resolved by the identified CSIs. This problem may be addressed as genome sequences for additional Synergistetes species become available.

As previously mentioned, the Synergistetes have often been misclassified as a lower ranked taxonomic group with bacteria belonging to other phylogenetic divisions. In our analysis, some CSIs were also discovered that were shared by Synergistetes species along with species from other taxonomic groups. Some of the organisms sharing such indels included species from the Fusobacteria, Chloroflexi, Proteobacteria, Acidobacteria, Aquificae and Firmicutes phyla. Most of these groups shared no more than 1–3 CSIs and in many cases only a few species within the groups contained the indels. Geissinger et al. (2009) presented a study suggesting a shared common ancestor for *Elusimicrobium* and Synergistetes and a recent study by Gupta (2011) also suggested that the *Negativicutes*, *Fusobacteria*, *Elusimicrobia* and

(A)

			189		224
Synergistetes 7/7	Jonquetella anthropi	260654222	YGADLYKKRSLPKLLTELEK	T	LPQSVWLRLFYLHPS
	Pyramidobacter piscolens	282855982	--S--FGRP---R--D----	E	--RG-----------
	Dethiosulfovibrio peptidovorans	288573284	--S--S-RG--SG--DAM-A	E	--EDI----------
	Aminobacterium colombiense	294101399	--R-V-G-P--IA--DS-TA	S	---E-----L----A
	Anaerobaculum hydrogeniformans	289524260	--M-WDGSSH-VE--DL---	H	V-DGM-I-PL----
	Aminomonas paucivorans	312879348	--L-RGGRE--TD--DA--P	S	--GD-F---L----
	Thermanavib. acidaminovorans	269793254	--E- LGTD -ME--DQM-A	-	VRGHGV---L----T
Chloroflexi 6/6	Sphaerobacter thermophilus	269837471	-----GI-NG--G--RMIAE	E	V-DLP---ML-IY--
	Roseiflexus castenholzii	156741270	--R--GL-DG-AI--D-ICA	V	--EN--V--M-AY-
	Roseiflexus sp. RS-1	148654700	--R--GL-DG-AL--D--CA	V	--KDR-V--M-AY-
	Chloroflexus aggregans	219848510	--R--GLRDG-AI--D--CQ	V	T-SDI-I--M-AY-
	Chloroflexus aurantiacus	163847415	--R--GL-DG-AT--A--CQ	V	T-PET-I--M-AY-
	Oscillochloris trichoides	309792044	--R--GLQDG-AT--E-MCQ	I	V-H-G-I--M-AY-
Other species 0/250	Thermotoga maritima	15644605	--I---R-QA--D--RR-NS		-NGEF-I-VM----
	Clostridium botulinum	187933390	--S-I-G-KN-HV--K--S-		IEGIK-I-VL-CY-
	Eubacterium yurii	306819660	--L-I-GEKK--N--R--S-		IEGIR-I-FL-TY-
	Roseburia intestinalis	240144123	--V---GEK--HR--D--N-		IKDLF-I-IM-CY-
	Syntrophomonas wolfei	114566788	--H-ISPQSA--T--R--S-		-DGLE-I--M----
	Thermoanaerobacter italicus	289578370	--I-I---FM--Q--K--SL		I-NLK-I--L-AY-
	Fusobacterium varium	253583638	--I----EK--AR-MK--V-		IDGLK---TY-MF-
	Ilyobacter polytropus	310779080	--I-----KA--DVMKA-S-		VEGIE-I-TY-MF-N
	Bacteroides finegoldii	255693887	--V-----QM--E-IERISD		I-GVE-I--H-AY-A
	Psychroflexus torquis	91215250	--L------N-AE--R--V-		VEGIE-I--H-AF-T
	Bacillus tusciae	295696216	--L---GR-R--D--KA-ND		VDELR-I--H-AY-
	Brachyspira murdochii	296125911	--H-I--RLA--D--K--S-		IEGIE-I-VL-QN-
	Aquifex aeolicus	15606200	--K----EYK-VE--EG---		VEGIK-I--L--Y-T
	Thermocrinis albus	289547853	--K---HRKA-VQ--KK--E		-EGIE-I--L--Y-T
	Prochlorococcus marinus	123967655	--Q-I-G-P--A---N--S-		VSIP-I-IH-AY-T
	Leptospirillum rubarum	124514462	--S--EDGEG--R--E-IDR		IGRIP-V--L-AY-T
	Dialister microaerophilus	313891928	--Q--RDGT--IL--KQ-V-		I-EVK-I-----Y-T

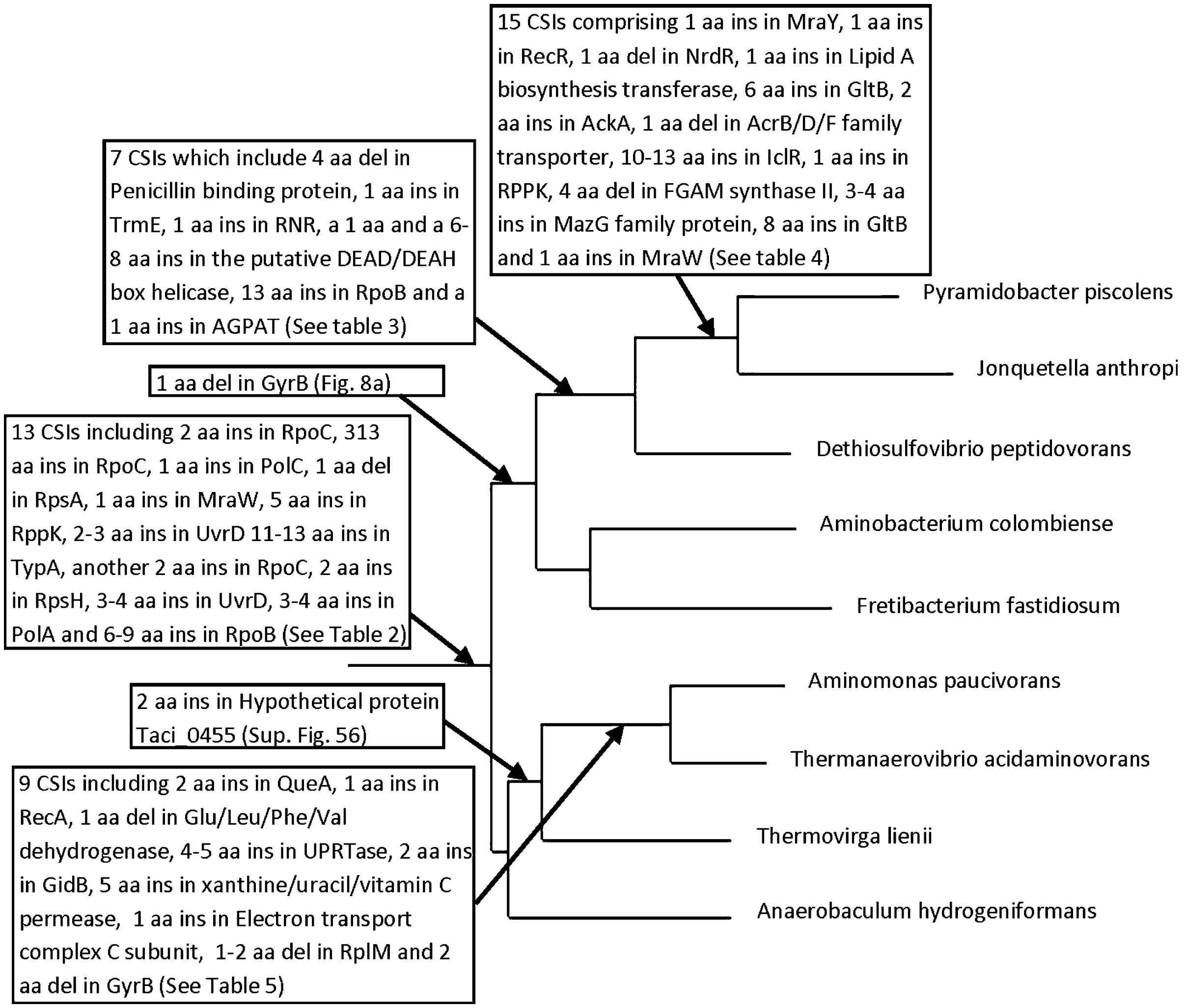
(B)

			244		274
Synergistetes 9/9	Dethiosulfovibrio peptidovorans	288574785	FYLRSAEEMWQIFG	D	DVPEALENTLKIAERC
	Thermovirga lienii	357419591	-----P----A---	K	EL----T--VN-----
	Jonquetella anthropi	260655562	-----P----NF--	A	E--D--T--VE-----
	Aminobacterium colombiense	294101731	-----P---DSL--	A	EL-D--D---A-----
	Pyramidobacter piscolens	282856832	--F---Q---GY--	S	EA-DS-----R-----
	Thermanaerovibrio acidaminovorans	269792442	----TP----SLL-	G	E-----R---LV----
	Aminomonas paucivorans	312880004	--F--P----SL--	A	EL-D--R--QE--D--
	Anaerobaculum hydrogeniformans	289523190	-----PV---K---	-	ELSD--H--VD-S---
	Fretibacterium fastidiosum	295112106	-----P---DA---	A	E-----D--VA-----
Fusobacteria 4/4	Fusobacterium sp. 3_1_5R	257452286	L--K-Y---YAVL-	E	QYQ---Q-SVE--K--
	Fusobacterium periodonticum	262066954	L--K-K---QRSLD	E	KFHK-I---NY--SL-
	Fusobacterium gonidiaformans	257465914	L--K-Y---YAVL-	E	QYQ---Q-SVE--K--
	Fusobacterium nucleatum	34763539	L--K-KD--KRFL-	E	KFEK-I--ANH--DL-
Other species 2/250	Opitutaceae bacterium	225163584	---K-RD--FK---	R	EL--S-T---AV--M-
	Thermomicrobium roseum	221632747	L-FK------RL--	A	E-----L--VR-----
	Sphaerobacter thermophilus	269837330	L-FK-P----RL--		E---------IR-----
	Bifidobacterium bifidum	224282597	Y-IK-----REL-K		-H-D-CD---E-----
	Elusimicrobium minutum	187250629	L-FK--Q--AEL-S		YI---VS---EV--K-
	Kribbella flavida	284030835	--VKT-A--REVWK		-L---CD---L---Q-
	Eubacterium limosum	310828648	---K-PD--SRL-I		-----I---N------
	Clostridium perfringens	110803674	---K-P----DM-D		YI-----------E-
	Peptostreptococcus stomatis	307243299	--IK-R---EDL-A		FS---V----------
	Desulfotomaculum acetoxidans	258514118	---K-P---AEL-P		EL-Q--Q--VS--DK-
	Symbiobacterium thermophilum	51891797	L-VK-----YWR--		H--G----S-A-----
	Bacteroides capillosus	154499865	--IK-E---A-L-P		-W---I---VRV--K-
	Geobacter sulfurreducens	39996502	--VKTP---AAA-H		YA---IS--V------
	Desulfovibrio desulfuricans	220904502	L-YK-I---EKS--		H-----A--AR---A-
	Spirochaeta thermophila	307718499	---K-P---ARL-E		-I-------VR-----
	Borrelia burgdorferi	15594924	--IK-Q---CEL-N		-L-------VR---K-
	Dictyoglomus turgidum	217967773	--FK-PD--INL-K		-----I---F------
	Selenomonas noxia	292670795	Y--K-EQ--AEL-A		-Y-G-----V------
	Leptospirillum rubarum	124514840	---KTPR--VRS-E		EL---I----R--D-I

◀ **Fig. 9** Examples of CSIs that are commonly shared by Synergistetes species and other groups of bacteria. **a** A CSI consisting of 1 aa insert in the MiaB-family of RNA modification enzyme that is commonly shared by different Synergistetes and Chloroflexi. **b** A 1 aa insert in a conserved region of DNA polymerase III, α subunit shared by all Synergistetes and Fusobacteria as well as two other bacteria belonging to the Chloroflexi and Verrucomicrobia phyla

Synergistetes phyla might be closely related to each other based on their cell membrane structure and shared indels in their DnaK and GroEL proteins (Geissinger et al. 2009; Gupta 2011). Though the *Elusimicrobia* and *Fusobacteria* share some CSIs with the Synergistetes species, no CSI was found that was specifically shared by all detected Synergistetes and species from these taxa. Furthermore, the branching of these phyla in the protein trees (Fig. 1) does not support their close relationship with the Synergistetes. Hence, based upon these results, at present no clear relationship of the Synergistetes species to other bacteria phyla can be inferred. These results provide further evidence supporting the placement of Synergistetes species into a distinct phylum.

Due to their specificity, Synergistetes-specific CSIs provide interesting prospects for future research. Since these CSIs are present in conserved regions of various proteins, degenerate primers utilizing the conserved regions can be designed for use as a means for



15 CSIs comprising 1 aa ins in MraY, 1 aa ins in RecR, 1 aa del in NrdR, 1 aa ins in Lipid A biosynthesis transferase, 6 aa ins in GltB, 2 aa ins in AckA, 1 aa del in AcrB/D/F family transporter, 10-13 aa ins in IclR, 1 aa ins in RPPK, 4 aa del in FGAM synthase II, 3-4 aa ins in MazG family protein, 8 aa ins in GltB and 1 aa ins in MraW (See table 4)

7 CSIs which include 4 aa del in Penicillin binding protein, 1 aa ins in TrmE, 1 aa ins in RNR, a 1 aa and a 6-8 aa ins in the putative DEAD/DEAH box helicase, 13 aa ins in RpoB and a 1 aa ins in AGPAT (See table 3)

1 aa del in GyrB (Fig. 8a)

13 CSIs including 2 aa ins in RpoC, 313 aa ins in RpoC, 1 aa ins in PolC, 1 aa del in RpsA, 1 aa ins in MraW, 5 aa ins in RppK, 2-3 aa ins in UvrD 11-13 aa ins in TypA, another 2 aa ins in RpoC, 2 aa ins in RpsH, 3-4 aa ins in UvrD, 3-4 aa ins in PolA and 6-9 aa ins in RpoB (See Table 2)

2 aa ins in Hypothetical protein Taci_0455 (Sup. Fig. 56)

9 CSIs including 2 aa ins in QueA, 1 aa ins in RecA, 1 aa del in Glu/Leu/Phe/Val dehydrogenase, 4-5 aa ins in UPRTase, 2 aa ins in GidB, 5 aa ins in xanthine/uracil/vitamin C permease, 1 aa ins in Electron transport complex C subunit, 1-2 aa del in RplM and 2 aa del in GyrB (See Table 5)

Pyramidobacter piscolens
Jonquetella anthropi
Dethiosulfovibrio peptidovorans
Aminobacterium colombiense
Fretibacterium fastidiosum
Aminomonas paucivorans
Thermanaerovibrio acidaminovorans
Thermovirga lienii
Anaerobaculum hydrogeniformans

Fig. 10 A summary diagram portraying the species distribution of various identified Synergistetes-specific CSIs and the evolutionary stages where the genetic changes responsible for them have occurred

identification of various species of the phylum in different environments (Gao and Gupta 2005). This might prove to be especially useful, as it is surmised that universal primers utilized in detection of organisms is metagenomic studies may not efficiently identify some Synergistetes species (Hamady and Knight 2009). As molecular markers, the phylum-specific CSIs can be useful as identification tool for detection of known and unknown species in metagenomics experiments. These CSI can also assist in the classification of newly discovered bacteria into the phylum Synergistetes and its sub-groups.

Finally, some species of Synergistetes have also been notoriously difficult to culture/isolate (Vartoukian et al. 2010) and, for others, their biological nuances have just begun to be understood. It has been suggested that Synergistetes act in concert with other oral bacteria to degrade proteinaceous compounds in periodontitis lesions (Homer and Beighton 1992; Wei et al. 1999; Vartoukian et al. 2007). Prior functional studies on taxa-specific CSIs have shown that such indels are usually present in peripheral regions of proteins and they tend to be essential for the function of the proteins in the organisms where the CSIs occur (Itzhaki et al. 2006; Akiva et al. 2008; Hormozdiari et al. 2009; Singh and Gupta 2009). Hence, agents that bind to these CSIs and inhibit their cellular functions could provide novel therapeutics, which are specifically directed against this group of bacteria. Lastly, the molecular markers discovered in this study, due to their specificity for Synergistetes species provide novel and valuable means for understanding the contribution of this group of bacteria to the environment and to the microbial communities that they inhabit. Thus, analyses devoted to the understanding of the function of these CSIs should provide important insights into the biochemical and physiological properties that define the Synergistetes and their roles in different environments.

Acknowledgments This work was supported by a research grant from the Natural Science and Engineering Research Council of Canada.

References

Akiva E, Itzhaki Z, Margalit H (2008) Built-in loops allow versatility in domain–domain interactions: lessons from self-interacting domains. Proc Natl Acad Sci USA 105:13292–13297

Allison MJ, Maynerry WR, McSweeney CS, Stahl DA (1992) *Synergistes jonesii*, gen.nov., sp.nov.: a rumen bacterium that degrades toxic pyridinediols. Syst Appl Microbiol 15:522–529

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Altschul SF, Wootton JC, Gertz EM, Agarwala R, Morgulis A, Schaffer AA, Yu YK (2005) Protein database searches using compositionally adjusted substitution matrices. FEBS J 272:5101–5109

Baena S, Fardeau ML, Labat M, Ollivier B, Thomas P, Garcia JL, Patel BK (1998) *Aminobacterium colombiensegen*. nov. sp. nov., an amino acid-degrading anaerobe isolated from anaerobic sludge. Anaerobe 4:241–250

Baena S, Fardeau ML, Ollivier B, Labat M, Thomas P, Garcia JL, Patel BK (1999a) *Aminomonas paucivorans* gen. nov., sp. nov., a mesophilic, anaerobic, amino-acid-utilizing bacterium. Int J Syst Bacteriol 49(Pt 3):975–982

Baena S, Fardeau ML, Woo TH, Ollivier B, Labat M, Patel BK (1999b) Phylogenetic relationships of three amino-acid-utilizing anaerobes, Selenomonas acidaminovorans, '*Selenomonas acidaminophila*' and *Eubacterium acidaminophilum*, as inferred from partial 16S rDNA nucleotide sequences and proposal of *Thermanaerovibrio acidaminovorans* gen. nov., comb. nov. and *Anaeromusa acidaminophila* gen. nov., comb. nov. Int J Syst Bacteriol 49(Pt 3):969–974

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

Bocchetta M, Gribaldo S, Sanangelantoni A, Cammarano P (2000) Phylogenetic depth of the bacterial genera Aquifex and Thermotoga inferred from analysis of ribosomal protein, elongation factor, and RNA polymerase subunit sequences. J Mol Evol 50:366–380

Chertkov O, Sikorski J, Brambilla E, Lapidus A, Copeland A, Glavina DR, Nolan M, Lucas S, Tice H, Cheng JF, Han C, Detter JC, Bruce D, Tapia R, Goodwin L, Pitluck S, Liolios K, Ivanova N, Mavromatis K, Ovchinnikova G, Pati A, Chen A, Palaniappan K, Land M, Hauser L, Chang YJ, Jeffries CD, Spring S, Rohde M, Goker M, Bristow J, Eisen JA, Markowitz V, Hugenholtz P, Kyrpides NC, Klenk HP (2010) Complete genome sequence of *Aminobacterium colombiense* type strain (ALA-1). Stand Genomic Sci 2:280–289

Chovatia M, Sikorski J, Schroder M, Lapidus A, Nolan M, Tice H, Glavina DR, Copeland A, Cheng JF, Lucas S, Chen F, Bruce D, Goodwin L, Pitluck S, Ivanova N, Mavromatis K, Ovchinnikova G, Pati A, Chen A, Palaniappan K, Land M, Hauser L, Chang YJ, Jeffries CD, Chain P, Saunders E, Detter JC, Brettin T, Rohde M, Goker M, Spring S, Bristow J, Markowitz V, Hugenholtz P, Kyrpides NC, Klenk HP, Eisen JA (2009) Complete genome sequence of *Thermanaerovibrio acidaminovorans* type strain (Su883). Stand Genomic Sci 1:254–261

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

Dahle H, Birkeland NK (2006) *Thermovirga lienii* gen. nov., sp. nov., a novel moderately thermophilic, anaerobic, amino-acid-degrading bacterium isolated from a North Sea oil well. Int J Syst Evol Microbiol 56:1539–1545

de Lillo A, Ashley FP, Palmer RM, Munson MA, Kyriacou L, Weightman AJ, Wade WG (2006) Novel subgingival bacterial phylotypes detected using multiple universal polymerase chain reaction primer sets. Oral Microbiol Immunol 21:61–68

Delbes C, Moletta R, Godon J (2001) Bacterial and archaeal 16S rDNA and 16S rRNA dynamics during an acetate crisis in an anaerobic digestor ecosystem. FEMS Microbiol Ecol 35:19–26

Diaz C, Baena S, Fardeau ML, Patel BK (2007) *Aminiphilus circumscriptus* gen. nov., sp. nov., an anaerobic amino-acid-degrading bacterium from an upflow anaerobic sludge reactor. Int J Syst Evol Microbiol 57:1914–1918

Downes J, Vartoukian SR, Dewhirst FE, Izard J, Chen T, Yu WH, Sutcliffe IC, Wade WG (2009) *Pyramidobacter piscolens* gen. nov., sp. nov., a member of the phylum 'Synergistetes' isolated from the human oral cavity. Int J Syst Evol Microbiol 59:972–980

Fonknechten N, Chaussonnerie S, Tricot S, Lajus A, Andreesen JR, Perchat N, Pelletier E, Gouyvenoux M, Barbe V, Salanoubat M, Le Paslier D, Weissenbach J, Cohen GN, Kreimeyer A (2010) Clostridium sticklandii, a specialist in amino acid degradation: revisiting its metabolism through its genome sequence. BMC Genomics 11:555

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

Garrity GM, Bell JA, Lilburn TG (2004) Taxonomic outline of the Prokaryotes. Bergey's manual of systematic bacteriology. Release 5.0, 2nd edn. Springer, New York

Geissinger O, Herlemann DP, Morschel E, Maier UG, Brune A (2009) The ultramicrobacterium "*Elusimicrobium minutum*" gen. nov., sp. nov., the first cultivated representative of the termite group 1 phylum. Appl Environ Microbiol 75:2831–2840

Godon JJ, Moriniere J, Moletta M, Gaillac M, Bru V, Delgenes JP (2005) Rarity associated with specific ecological niches in the bacterial world: the 'Synergistes' example. Environ Microbiol 7:213–224

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

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

Guangsheng C, Plugge CM, Roelofsen W, Houwen FP, Stams AJM (1992) *Selenomonas acidaminovorans* sp. nov., a versatile thermophilic proton-reducing anaerobe able to grow by decarboxylation of succinate to propionate. Arch Microblol 157:169–175

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 natural evolutionary relationships among prokaryotes. Crit Rev Microbiol 26:111–131

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

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

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

Hamady M, Knight R (2009) Microbial community profiling for human microbiome projects: tools, techniques, and challenges. Genome Res 19:1141–1152

Herlemann DP, Geissinger O, Ikeda-Ohtsubo W, Kunin V, Sun H, Lapidus A, Hugenholtz P, Brune A (2009) Genomic analysis of "*Elusimicrobium minutum*," the first cultivated representative of the phylum "Elusimicrobia" (formerly termite group 1). Appl Environ Microbiol 75:2841–2849

Homer KA, Beighton D (1992) Synergistic degradation of bovine serum albumin by mutans streptococci and other dental plaque bacteria. FEMS Microbiol Lett 69:259–262

Hongoh Y, Sato T, Dolan MF, Noda S, Ui S, Kudo T, Ohkuma M (2007) The motility symbiont of the termite gut flagellate *Caduceia versatilis* is a member of the "Synergistes" group. Appl Environ Microbiol 73:6270–6276

Hormozdiari F, Salari R, Hsing M, Schonhuth A, Chan SK, Sahinalp SC, Cherkasov A (2009) The effect of insertions and deletions on wirings in protein–protein interaction networks: a large-scale study. J Comput Biol 16:159–167

Horz HP, Citron DM, Warren YA, Goldstein EJ, Conrads G (2006) Synergistes group organisms of human origin. J Clin Microbiol 44:2914–2920

Hou S, Saw JH, Lee KS, Freitas TA, Belisle C, Kawarabayasi Y, Donachie SP, Pikina A, Galperin MY, Koonin EV, Makarova KS, Omelchenko MV, Sorokin A, Wolf YI, Li QX, Keum YS, Campbell S, Denery J, Aizawa S, Shibata S, Malahoff A, Alam M (2004) Genome sequence of the deep-sea gamma-proteobacterium *Idiomarina loihiensis* reveals amino acid fermentation as a source of carbon and energy. Proc Natl Acad Sci USA 101:18036–18041

Hugenholtz P, Hooper SD, Kyrpides NC (2009) Focus: Synergistetes. Environ Microbiol 11:1327–1329

Itzhaki Z, Akiva E, Altuvia Y, Margalit H (2006) Evolutionary conservation of domain–domain interactions. Genome Biol 7:R125

Jumas-Bilak E, Carlier JP, Jean-Pierre H, Citron D, Bernard K, Damay A, Gay B, Teyssier C, Campos J, Marchandin H (2007) *Jonquetella anthropi* gen. nov., sp. nov., the first member of the candidate phylum 'Synergistetes' isolated from man. Int J Syst Evol Microbiol 57:2743–2748

Jumas-Bilak E, Roudiere L, Marchandin H (2009) Description of 'Synergistetes' phyl. nov. and emended description of the phylum 'Deferribacteres' and of the family

Syntrophomonadaceae, phylum ‘Firmicutes’. Int J Syst Evol Microbiol 59:1028–1035

Kumar PS, Griffen AL, Moeschberger ML, Leys EJ (2005) Identification of candidate periodontal pathogens and beneficial species by quantitative 16S clonal analysis. J Clin Microbiol 43:3944–3955

Labutti K, Mayilraj S, Clum A, Lucas S, Glavina DR, Nolan M, Tice H, Cheng JF, Pitluck S, Liolios K, Ivanova N, Mavromatis K, Mikhailova N, Pati A, Goodwin L, Chen A, Palaniappan K, Land M, Hauser L, Chang YJ, Jeffries CD, Rohde M, Spring S, Goker M, Woyke T, Bristow J, Eisen JA, Markowitz V, Hugenholtz P, Kyrpides NC, Klenk HP, Lapidus A (2010) Permanent draft genome sequence of *Dethiosulfovibrio peptidovorans* type strain (SEBR 4207). Stand Genomic Sci 3:85–92

Magot M, Ravot G, Campaignolle X, Ollivier B, Patel BK, Fardeau ML, Thomas P, Crolet JL, Garcia JL (1997) *Dethiosulfovibrio peptidovorans* gen. nov., sp. nov., a new anaerobic, slightly halophilic, thiosulfate-reducing bacterium from corroding offshore oil wells. Int J Syst Bacteriol 47:818–824

Marchandin H, Damay A, Roudiere L, Teyssier C, Zorgniotti I, Dechaud H, Jean-Pierre H, Jumas-Bilak E (2010) Phylogeny, diversity and host specialization in the phylum Synergistetes with emphasis on strains and clones of human origin. Res Microbiol 161:91–100

Maune MW, Tanner RS (2012) Description of *Anaerobaculum hydrogeniformans* sp. nov., an anaerobe that produces hydrogen from glucose, and emended description of the genus *Anaerobaculum*. Int J Syst Evol Microbiol 62: 832–838

McSweeney CS, Mackie RI, Odenyo AA, Stahl DA (1993) Development of an oligonucleotide probe targeting 16S rRNA and its application for detection and quantitation of the ruminal bacterium *Synergistes jonesii* in a mixed-population chemostat. Appl Environ Microbiol 59:1607–1612

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

NCBI genomic database (2012) http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi. Accessed 5 May 2012

NCBI Taxonomy (2012) http://www.ncbi.nlm.nih.gov/taxonomy. Accessed 5 May 2012

Pitluck S, Yasawong M, Held B, Lapidus A, Nolan M, Copeland A, Lucas S, Del Rio TG, Tice H, Cheng JF, Chertkov O, Goodwin L, Tapia R, Han C, Liolios K, Ivanova N, Mavromatis K, Ovchinnikova G, Pati A, Chen A, Palaniappan K, Land M, Hauser L, Chang YJ, Jeffries CD, Pukall R, Spring S, Rohde M, Sikorski J, Goker M, Woyke T, Bristow J, Eisen JA, Markowitz V, Hugenholtz P, Kyrpides NC, Klenk HP (2010) Non-contiguous finished genome sequence of *Aminomonas paucivorans* type strain (GLU-3). Stand Genomic Sci 3:285–293

Rivera MC, Lake JA (1992) Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science 257:74–76

Riviere D, Desvignes V, Pelletier E, Chaussonnerie S, Guermazi S, Weissenbach J, Li T, Camacho P, Sghir A (2009) Towards the definition of a core of microorganisms involved in anaerobic digestion of sludge. ISME J 3:700–714

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

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

Soria-Carrasco V, Valens-Vadell M, Pena A, Anton J, Amann R, Castresana J, Rossello-Mora R (2007) Phylogenetic position of *Salinibacter ruber* based on concatenated protein alignments. Syst Appl Microbiol 30:171–179

Surkov AV, Dubinina GA, Lysenko AM, Glockner FO, Kuever J (2001) *Dethiosulfovibrio russensis* sp. nov., *Dethosulfovibrio marinus* sp. nov. and *Dethosulfovibrio acidaminovorans* sp. nov., novel anaerobic, thiosulfate- and sulfur-reducing bacteria isolated from ‘Thiodendron’ sulfur mats in different saline environments. Int J Syst Evol Microbiol 51:327–337

Sutcliffe IC (2010) A phylum level perspective on bacterial cell envelope architecture. Trends Microbiol 18:464–470

Vartoukian SR, Palmer RM, Wade WG (2007) The division “Synergistes”. Anaerobe 13:99–106

Vartoukian SR, Palmer RM, Wade WG (2010) Cultivation of a Synergistetes strain representing a previously uncultivated lineage. Environ Microbiol 12:916–928

Vartoukian S, Downes J, Palmer RM, Wade WG (2012) *Fretibacterium fastidiosum* gen. nov., sp. nov., isolated from the human oral cavity. Int J Syst Evol Microbiol. doi:10.1099/ijs.0.041038-0

Wei GX, van der Hoeven JS, Smalley JW, Mikx FH, Fan MW (1999) Proteolysis and utilization of albumin by enrichment cultures of subgingival microbiota. Oral Microbiol Immunol 14:348–351

Wu M, Eisen JA (2008) A simple, fast, and accurate method of phylogenomic inference. Genome Biol 9:R151

Zhaxybayeva O, Swithers KS, Lapierre P, Fournier GP, Bickhart DM, DeBoy RT, Nelson KE, Nesbo CL, Doolittle WF, Gogarten JP, Noll KM (2009) On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales. Proc Natl Acad Sci USA 106:5865–5870

Ziganshin AM, Schmidt T, Scholwin F, Il’inskaya ON, Harms H, Kleinsteuber S (2011) Bacteria and archaea involved in anaerobic digestion of distillers grains with solubles. Appl Microbiol Biotechnol 89:2039–2052

Zijnge V, van Leeuwen MB, Degener JE, Abbas F, Thurnheer T, Gmur R, Harmsen HJ (2010) Oral biofilm architecture on natural teeth. PLoS ONE 5:e9321

CHAPTER 4

Molecular signatures for the PVC clade (Planctomycetes, Verrucomicrobia, Chlamydiae, and Lentisphaerae) of bacteria provide insights into their evolutionary relationships[1]

Work presented in the following chapter examines the relationship among bacterial species of the PVC group. Phylogenetic trees along with CSIs and CSPs are used to identify the linkages among the multiple phyla indicated to belong to this superphylum. CSIs for some clades among the Verrucomicrobiae and Planctomycetes are also described. 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 data analysis, the preparation of the manuscript and construction of the figures and tables.

[1] The citation for the manuscript is:

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.

frontiers in
MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 17 September 2012
doi: 10.3389/fmicb.2012.00327

Molecular signatures for the PVC clade (Planctomycetes, Verrucomicrobia, Chlamydiae, and Lentisphaerae) of bacteria provide insights into their evolutionary relationships

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

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

Edited by:
Naomi L. Ward, University of Wyoming, USA

Reviewed by:
Garry Myers, University of Maryland, USA
John A. Fuerst, The University of Queensland, Australia

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

The PVC superphylum is an amalgamation of species from the phyla Planctomycetes, Verrucomicrobia, and Chlamydiae, along with the Lentisphaerae, Poribacteria, and two other candidate divisions. The diverse species of this superphylum lack any significant marker that differentiates them from other bacteria. Recently, genome sequences for 37 species covering all of the main PVC groups of bacteria have become available. We have used these sequences to construct a phylogenetic tree based upon concatenated sequences for 16 proteins and identify molecular signatures in protein sequences that are specific for the species from these phyla or those providing molecular links among them. Of the useful molecular markers identified in the present work, six conserved signature indels (CSIs) in the proteins Cyt c oxidase, UvrD helicase, urease, and a helicase-domain containing protein are specific for the species from the Verrucomicrobia phylum; three other CSIs in an ABC transporter protein, cobyrinic acid ac-diamide synthase, and SpoVG protein are specific for the Planctomycetes species. Additionally, a 3 aa insert in the RpoB protein is uniquely present in all sequenced Chlamydiae, Verrucomicrobia, and Lentisphaerae species, providing evidence for the shared ancestry of the species from these three phyla. Lastly, we have also identified a conserved protein of unknown function that is exclusively found in all sequenced species from the phyla Chlamydiae, Verrucomicrobia, Lentisphaerae, and Planctomycetes suggesting a specific linkage among them. The absence of this protein in Poribacteria, which branches separately from other members of the PVC clade, indicates that it is not specifically related to the PVC clade of bacteria. The molecular markers described here in addition to clarifying the evolutionary relationships among the PVC clade of bacteria also provide novel tools for their identification and for genetic and biochemical studies on these organisms.

Keywords: conserved signature indels, signature proteins, Verrucomicrobia, Planctomycetes, Chlamydia, Lentisphaerae, PVC superphylum, phylogenetic trees

INTRODUCTION

The bacteria of the Planctomycetes, Verrucomicrobia, Chlamydiae, and Lentisphaerae phyla along with the Candidate Poribacteria, Candidate phylum OP3 and Candidate division WWE2 are collectively grouped and referred to as the PVC superphylum or the PVC clade (Wagner and Horn, 2006). The PVC group is comprised of species that are of much importance due to their characteristics and the roles they play in many areas of life. Species of the Chlamydiae phylum are one of the most widely studied microorganisms due to their pathogenic capacities in humans and in animals. They are responsible for many human illnesses including sexually transmitted urinary tract infections, trachoma, and pneumonia (Sachse et al., 2009). Species of the phylum Planctomycetes are renowned for their unusual cellular features such as internal compartmentalization, sterol biosynthesis, and endocytosis-analogous pathways that are generally associated with the eukaryotes (Fuerst and Webb, 1991; Lindsay et al., 1997; Pearson et al., 2003; Ward et al., 2006; Lonhienne et al., 2010; Fuerst and Sagulenko, 2011; McInerney et al., 2011). This phylum also harbors a group of anaerobic chemoautotrophic "anammox" (anaerobic ammonium oxidation) organisms (van de Graaf et al., 1995; Strous et al., 1999). These anammox species can oxidize ammonium to dinitrogen and are therefore quite useful in decontamination of wastewater rich in ammonia (Dalsgaard et al., 2003). Their importance is underscored by estimates which suggest that anammox bacteria may contribute up to 50% of the atmospheric nitrogen (Devol, 2003). The species from the phylum Verrucomicrobia are abundant in soil based environments with estimates proposing that up to 10% of all bacteria in the soil belong to this phylum (Sangwan et al., 2005). These bacteria are also found in aquatic environments (Martiny et al., 2005; Haukka et al., 2006) and known to associate with eukaryotic species as indicated by their presence in termite guts, human intestines, nematodes, and some ciliate protozoa (Petroni et al., 2000; Vandekerckhove et al., 2002; Shinzato et al., 2005;

Wang et al., 2005). Some members of the Verrucomicrobiae are known to exist in ultramicrobial sizes, others to possess extensions of the cellular membrane termed the prosthecae and some also exist in acidophilic environments (Hedlund et al., 1997; Janssen et al., 1997; Pol et al., 2007). Thus, the species of the PVC phylum are important in our quest to better understand prokaryotic evolution, microbial ecology, and physiology.

Though much diversity exists among the bacteria of different phyla that comprises this superphylum, a close relationship among them has been suggested by the 16S rRNA trees and number of other phylogenetic studies employing single gene and multigene analyses of protein sequences (Cho et al., 2004; Wagner and Horn, 2006; Hou et al., 2008; Pilhofer et al., 2008; Glockner et al., 2010; Siegl et al., 2011). Among the members of this clade, the Planctomycetes and Chlamydiae were observed to be phylogenetically related as early as 1986 based on 16S rRNA secondary structures and phylogenetic trees (Weisburg et al., 1986; Woese, 1987; Fuerst, 1995). A close relationship of the Verrucomicrobia to the Chlamydiae and Planctomycetes was first observed by Hedlund et al. (1996) and the "sister-taxon" grouping of the Lentisphaerae to the Verrucomicrobia was recognized with the isolation of the first Lentisphaerae organism *Victivallis vadensis* (Zoetendal et al., 2003; Cho et al., 2004). The taxonomic entity labeled as the PVC superphylum was proposed in 2006, based on 16S ribosomal data, by Wagner and Horn (2006) to encompass the monophyletic group comprised of the above four phyla along with the recently discovered Candidate Poribacteria, Candidate phylum OP3 and Candidate phylum WWE2 (Hugenholtz et al., 1998; Fieseler et al., 2004; Chouari et al., 2005; Wagner and Horn, 2006). However, a monophyletic grouping of the different bacteria belonging to these phyla has also been disputed by other phylogenetic studies based upon 16S rRNA as well as several single gene and concatenated protein phylogenies (Ward et al., 2000; Jenkins and Fuerst, 2001; Ciccarelli et al., 2006; Griffiths and Gupta, 2007; Santarella-Mellwig et al., 2010).

Apart from their linkages in phylogenetic trees, little evidence exists to group the different phyla that are part of the PVC clade into a single large group. Nevertheless, some uncommon features are seen to be shared by multiple phyla of the group. The Verrucomicrobia along with the Poribacteria and Lentisphaerae share a similar intracellular structural plan with the Planctomycetes in having membranous borders dividing the cell into compartments (Fieseler et al., 2004; Lee et al., 2009; Fuerst and Sagulenko, 2011). Planctomycetes and Chlamydiae lack peptidoglycan in their cell walls (Konig et al., 1984; Liesack et al., 1986; Fox et al., 1990; Staley et al., 1992; Ward et al., 2006; Fuerst and Sagulenko, 2011). Also common among the Chlamydiae and Planctomycetes is the lack of FtsZ-based cell division (Bernander and Ettema, 2010; Fuerst and Sagulenko, 2011). However, as these features are not exclusive to the members of the PVC group and not found in all species of the phyla comprising the PVC group, they do not provide much clarity in the debate concerning the grouping of these phyla into a superphylum.

Due to the advent of rapid genomic sequencing techniques and availability of genomic sequences, comparative genomics provide powerful means for answering a variety of questions related to bacterial evolution. Using genome sequences, many approaches are being used to understand the evolutionary relationships among bacteria. While some approaches using whole genome alignments have been most used (or are mainly applicable) for studying closely related organisms (Angiuoli and Salzberg, 2011; Agren et al., 2012; Sahl et al., 2012), other comparative genomic approaches involving identification of molecular markers in the forms of either conserved signature inserts or deletions (CSIs) or conserved signature proteins (CSPs) have been extensively used to define taxonomic clades of different phylogenetic ranks in molecular terms (Gupta, 1998, 2010; Gupta and Griffiths, 2002; Dutilh et al., 2008; Gao and Gupta, 2012). The applications of these approaches previously to the Chlamydiae species have led to identification of numerous CSIs and CSPs that are specific for the species from this phylum or a number of its subclades (Griffiths et al., 2005, 2006; Gupta and Griffiths, 2006). Some interesting cases of lateral gene transfers (LGTs) between Actinobacteria and Chlamydiae were also identified by these studies (Griffiths and Gupta, 2006). Additionally, our work using these approaches also indicated that the phyla Chlamydiae and Verrucomicrobia are specifically related and they shared a common ancestor exclusive of the Planctomycetes (Griffiths and Gupta, 2007). However, thus far no molecular markers have been identified that are specific for the Planctomycetes and/or Verrucomicrobia phyla or those linking all members of the PVC group. In the present work, we describe the results of comparative genomic analysis aimed at identifying molecular markers that are uniquely shared by either the Planctomycetes or Verrucomicrobia phyla or those that are commonly shared by different main groups of the PVC superphylum. Additionally, we also report phylogenetic studies based upon concatenated protein sequences to evaluate the relationships among the PVC clade of bacteria.

MATERIALS AND METHODS

Complete or partial genomic sequences are now available for 37 species/strains belonging to the PVC group (see **Table 1**). For phylogenetic analyses, sequences for 16 housekeeping and ribosomal proteins (ArgRS, EF-G, EF-Tu, GyrA, GyrB, DnaK, IleRS, RecA, RpoB, RpoC, TrpRS, UvrD, ValRS along with ribosomal proteins L1, L5, and S12) were utilized. The protein sequences for various species of the PVC group and for species from some other bacterial phyla were retrieved from the NCBI protein database and their alignments were constructed using the ClustalX 1.83 program (Jeanmougin et al., 1998; NCBI protein database, 2012). After concatenation of all of these sequence alignments into a single file, the poorly aligned regions were removed using the Gblocks_0.91b program (Castresana, 2000). The remaining 7016 aligned and homologous characters were employed for construction of phylogenetic trees using the neighbor-joining (NJ) and maximum likelihood (ML) algorithms as described in our earlier work (Gupta and Mok, 2007; Gupta and Bhandari, 2011; Naushad and Gupta, 2012).

Identification of CSIs that are specific for the PVC group of species was carried out using similar procedures as described in our earlier work (Griffiths et al., 2005; Gupta and Bhandari, 2011; Naushad and Gupta, 2012). Briefly, BlastP searches were initially conducted on various proteins from the genomes of *Opitutus terrae* (van Passel et al., 2011a) and *Pirellula staleyi* (Clum et al.,

Table 1 | Some characteristics for sequenced species of the PVC group of bacteria.

Organism	GC%	Size (Mb)	Ref seq identity	Genome status	No. of proteins	Reference
PLANCTOMYCETES						
Candidatus Kuenenia stuttgartiensis	41.0	4.2	–	Draft	4663	Strous et al. (2006)
Phycisphaera mikurensis	73.0	3.9	NC_017080.1	Complete	3287	NCBI genome project
Gemmata obscuriglobus	67.2	9.2	NZ_ABGO00000000	Draft	7989	JCVI
Isosphaera pallida	62.4	5.5	NC_014962.1	Complete	3722	Goker et al. (2011)
Singulisphaera acidiphila	59.9	9.7	NZ_AGRX00000000	Draft	7630	DOE-JGI*
Rhodopirellula baltica	55.4	7.1	NC_005027.1	Complete	7325	Glockner et al. (2003)
Pirellula staleyi	57.5	6.2	NC_013720.1	Complete	4717	Clum et al. (2009)
Blastopirellula marina	57.0	6.6	NZ_AANZ00000000	Draft	6025	Glockner et al. (2003)
Planctomyces limnophilus	53.7	5.5	NC_014148.1	Complete	4258	Labutti et al. (2010)
Planctomyces brasiliensis	56.4	6.0	NC_015174.1	Complete	4750	DOE-JGI*
Planctomyces maris	50.5	7.8	NZ_ABCE00000000	Draft	6480	JCVI
VERRUCOMICROBIA						
Opitutaceae bacterium Tav5	61.0	7.4	NZ_AGJF00000000	Draft	6006	DOE-JGI*
Opitutaceae bacterium Tav1	63.2	7.1	NZ_AHKS00000000	Draft	5984	DOE-JGI*
Diplosphaera colitermitum	60.7	5.2	NZ_ABEA00000000	Draft	4826	DOE-JGI*
Opitutus terrae	55.3	6.0	NC_010571.1	Complete	4612	van Passel et al. (2011a)
Coraliomargarita akajimensis	53.6	3.7	NC_014008.1	Complete	3120	Mavromatis et al. (2010)
Verrucomicrobiae bacterium DG1235	54.3	5.8	NZ_ABSI00000000	Draft	4909	JCVI
Methylacidiphilum infernorum	45.5	2.3	NC_010794.1	Complete	2472	Hou et al. (2008)
Pedosphaera parvula	52.6	7.4	NZ_ABOX00000000	Draft	6510	Kant et al. (2011b)
Akkermansia muciniphila	55.8	2.7	NC_010655.1	Complete	2138	DOE-JGI*
Verrucomicrobium spinosum	60.3	8.2	NZ_ABIZ00000000.1	Complete	6509	TIGR[#]
Chthoniobacter flavus	61.1	7.8	NZ_ABVL00000000	Draft	6716	Kant et al. (2011a)
CHLAMYDIAE						
Chlamydophila abortus	39.9	1.1	NC_004552.2	Complete	932	Thomson et al. (2005)
Chlamydophila psittaci	39.1	1.2	NC_017289.1	Complete	975	Schofl et al. (2011)
Chlamydophila caviae	39.1	1.2	NC_003361.3	Complete	1005	Read et al. (2003)
Chlamydophila felis	39.3	1.2	NC_007899.1	Complete	1054	Azuma et al. (2006)
Chlamydophila pecorum	41.1	1.1	NC_015408.1	Complete	988	Mojica et al. (2011)
Chlamydophila pneumoniae	40.6	1.2	NC_002179.2	Complete	1119	Read et al. (2000)
Chlamydia trachomatis	41.3	1.0	NC_010287.1	Complete	874	Thomson et al. (2008)
Chlamydia muridarum	40.3	1.1	NC_002620.2	Complete	910	Read et al. (2000)
Simkania negevensis	41.6	2.6	NC_015713.1	Complete	2518	Collingro et al. (2011)
Waddlia chondrophila	43.8	2.1	NC_014225.1	Complete	1956	Bertelli et al. (2010)
Parachlamydia acanthamoebae	39.0	3.1	NC_015702.1	Complete	2789	Collingro et al. (2011)
Protochlamydia amoebophila	34.7	2.4	NC_005861.1	Complete	2031	Horn et al. (2004)
LENTISPHAERAE AND PORIBACTERIA						
Victivallis vadensis Lentisphaerae	59.4	5.3	NZ_ABDE00000000	Draft	4065	van Passel et al. (2011b)
Lentisphaera araneosa	41.0	6.0	NZ_ABCK00000000	Draft	5104	Thrash et al. (2010)
Candidatus Poribacteria WGA-A3	53.4	1.9	NZ_ADFK00000000	Draft	1585	Siegl et al. (2011)

**DOE-JGI – U.S. Department of Energy Joint Genomic Institute.*

[#] TIGR – The Institute for Genomic Research.

JCVI – J. Craig Venter Institute.

2009) and sequences for 10–12 species that included assorted species from the PVC group and some from other phyla were retrieved. Sequence alignments for these proteins were created and manually examined for inserts or deletions that were flanked on both sides by conserved regions (Gupta and Griffiths, 2002; Gupta and Bhandari, 2011; Naushad and Gupta, 2012). A second, more detailed BlastP search was then carried out on the identified sequence consisting of the indel and the conserved flanking region. The indels that were specific for the members of the PVC group were formatted into signature files showing the sequence alignments and GenBank identifier (GI) numbers of various proteins.

RESULTS

PHYLOGENETIC ANALYSES OF THE PVC GROUP OF BACTERIA BASED UPON CONCATENATED PROTEIN SEQUENCES

The proposal to amalgamate different bacterial groups that are part of the PVC clade is mainly based upon their branching in the 16S rRNA trees (Wagner and Horn, 2006). As indicated earlier, although close branching of species from some of these groups has been observed in a number of studies (Cho et al., 2004; Wagner and Horn, 2006; Hou et al., 2008; Pilhofer et al., 2008; Glockner et al., 2010; Siegl et al., 2011) most of these studies did not contain representatives from all bacterial phyla that are part of the PVC clade and their results have been contradicted by other analyses (Ward et al., 2000; Ciccarelli et al., 2006; Griffiths and Gupta, 2007). It is now widely accepted that in contrast to phylogenetic inferences based upon any single gene or protein, including 16S rRNA, those based upon large numbers of characters derived from multiple conserved genes/proteins are more reliable in accurately depicting the evolutionary relationships among distantly related phyla (Rokas et al., 2003; Ciccarelli et al., 2006; Wu and Eisen, 2008). Although some earlier studies are based upon concatenated protein sequences, they contained only limited numbers of Chlamydiae or Planctomycetes species (generally 4–5 Chlamydiaceae and 1–2 Planctomycetes) and no representative from the Verrucomicrobia or Lentisphaerae phyla (Ciccarelli et al., 2006; Strous et al., 2006; Hou et al., 2008). Our earlier work based upon concatenated protein sequences also included only one Verrucomicrobiae and three Planctomycetes species (Griffiths and Gupta, 2007). However, complete or partial genomic sequences are now available for 37 species belonging to the PVC clade of bacteria, including 11 species each from the Planctomycetes and Verrucomicrobia phyla, 12 from the Chlamydiae, two from the Lentisphaerae and a Poribacteria (**Table 1**). Hence, to examine the evolutionary relationship among these species, phylogenetic trees were constructed based upon a large concatenated dataset of protein sequences derived from 16 important proteins (see Methods). Most of these proteins are universally distributed and have been extensively used for phylogenetic analyses (Ciccarelli et al., 2006; Strous et al., 2006; Gupta and Mok, 2007; Hou et al., 2008). The trees were constructed using both ML and NJ methods and the results of these studies are summarized in **Figure 1**. The numbers at the nodes in this tree show the statistical significance of the node by the ML and NJ methods, respectively.

In the tree based upon concatenated protein sequences (**Figure 1**), species of the Planctomycetes, Verrucomicrobia, Chlamydiae, and Lentisphaerae phyla branched together with other members of their phylum. The monophyly and distinctness of these clades was well supported by both ML and NJ analyses with at least 75% bootstrap support by each of these methods. In this tree, Lentisphaerae and Verrucomicrobia were observed to branch together. Although a clade consisting of these two phyla has a bootstrap score of 95% by the NJ method, it was very weakly supported (supported only 54% of the time) by the ML method. Similarly, a clade consisting of the Lentisphaerae, Verrucomicrobia and Chlamydiae phyla was also strongly supported by the NJ method but not by the ML analysis. Additionally, although in this tree the four phyla that form the PVC clade were observed to branch together, a clade consisting of all four of them was poorly supported by both ML and NJ methods. Lastly, the single Poribacteria species in our dataset did not branch with the PVC group of bacteria. In addition to these observations, this tree also provides some insights into the relationships within the Verrucomicrobia and Planctomycetes phyla, which are discussed below together with the results of signature sequences for these groups of bacteria.

PHYLOGENY AND MOLECULAR SIGNATURES FOR THE PHYLUM VERRUCOMICROBIA

The sequenced Verrucomicrobia species formed a distinct clade in our phylogenetic tree (**Figure 1**), which was strongly supported by the NJ method and also had significant support by the ML analysis. Within this clade, the different Verrucomicrobia species split into two main clades, both of which were significantly supported by the NJ and ML analyses. One of these clades (marked **O1**), which we will refer to as the Opitutae clade, was comprised of the species *O. terrae*, *Diplosphaera colitermitum, Coraliomargarita akajimensis, Opitutaceae bacterium TAV5,* and *TAV1* and also *Verrucomicrobiae bacterium DG1235.* The first five of these species/strains belong to the class Opitutae, whereas *V. bacterium DG1235* is currently a part of the class Verrucomicrobiae (NCBI Taxonomy, 2012). The other members of the class Verrucomicrobiae (viz. *Verrucomicrobium spinosum*, *Akkermansia muciniphila* and *Pedosphaera parvulaparvula*) were part of the second major clade where they branched with *Chthoniobacter flavus*, a member of the class Spartobacteria and *Methylacidiphilum infernorum*, an unclassified species belonging to this phylum (Yoon et al., 2008; NCBI Taxonomy, 2012).

Currently, no molecular or biochemical marker of any kind is known that is specific for the species from the phylum Verrucomicrobia. However, of the signatures that we have identified, one consisting of a 2 aa insert in the Cytochrome c oxidase protein (**Figure 2A**) provides a potential molecular marker for this phylum. This indel is present in all members of the Verrucomicrobia phylum where the homologs of this protein could be detected, but it was not found in the homologs of this protein from any other bacteria including those from the Lentisphaerae, Chlamydiae, and Planctomycetes phyla. As this insert (CSI) is of fixed length, and it is present within a conserved region of the protein, it provides a useful and reliable molecular marker. Due to the highly specific nature of the genetic change which gave rise to this CSI and its specific presence only in this group of species, the genetic event responsible for this most likely occurred in a common ancestor of this phylum followed by vertical transmission of the gene containing this CSI to various descendant species (Gupta, 1998; Gupta and Griffiths, 2002; Gupta and Bhandari, 2011). Although a homolog for this protein was not detected in all sequenced verrucomicrobiae species, the noted genetic characteristic is specific for the species from this phylum and it provides a molecular means to distinguish species possessing the homolog from other bacteria.

Another identified CSI, shown in **Figure 2B**, consists of a 1 aa deletion in a conserved region of the UvrD helicase enzyme that is specific for the Opitutae clade (01) of Verrucomicrobia species (**Figure 1**). The species distribution of this CSI is consistent with the phylogenetic tree and it supports the grouping/placement of *V. bacterium* DG1235 within the Opitutae class rather than with

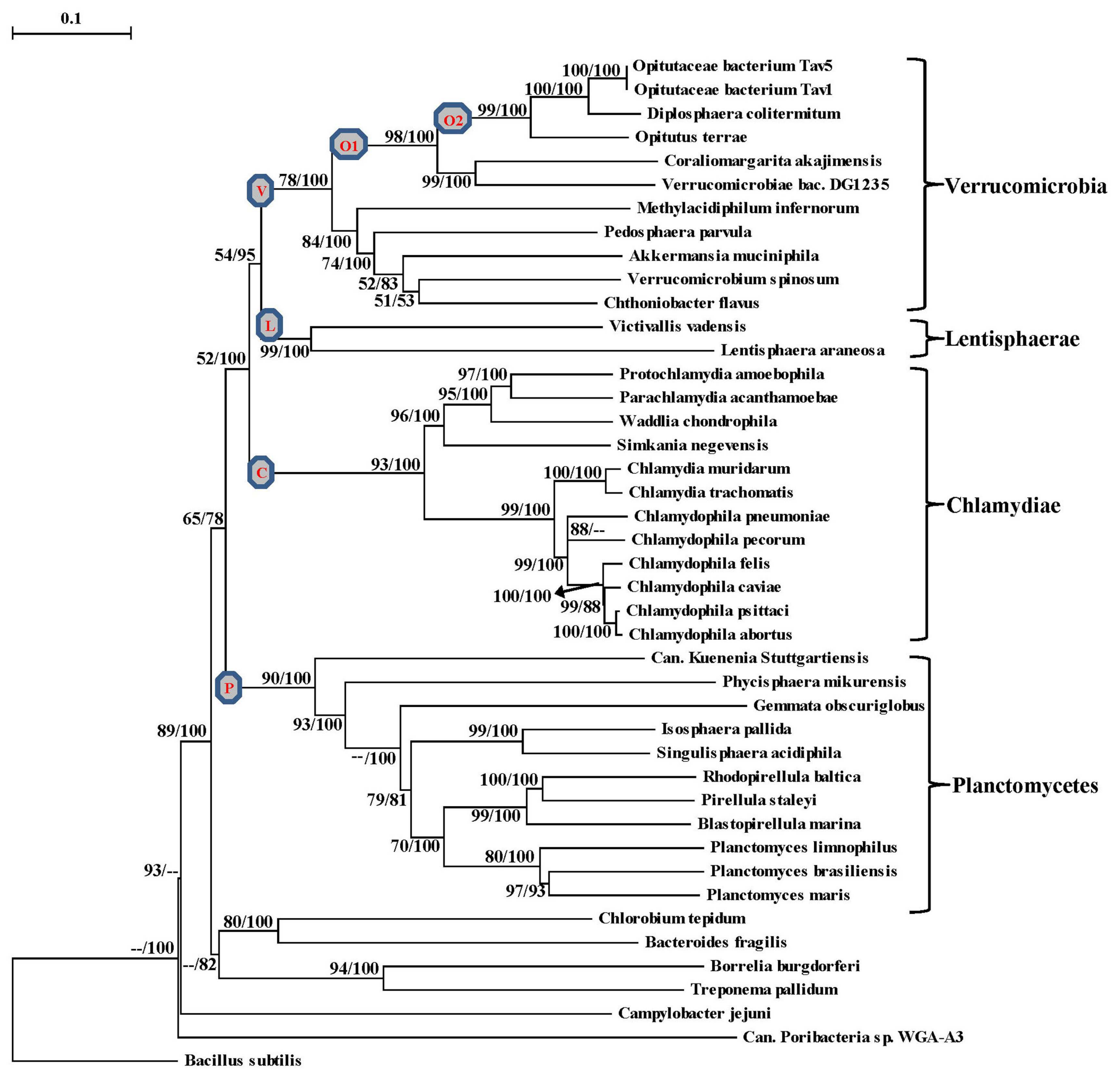


FIGURE 1 | A Neighbor-joining distance tree for the sequenced species belonging to the PVC group of bacteria based upon concatenated sequences for 16 conserved proteins. The numbers on the node indicate% statistical support for different nodes in the ML and NJ analyses, respectively. The scores that were less than 50% are not shown and represented by (–). The letters in the circle mark separate clades for the Verrucomicrobia phylum (V), Planctomycetes phylum (P), Chlamydiae (C), Lentisphaerae (L), Opitutae class (O1), Opitutaceae family (O2).

other members of the class Verrucomicrobiae. The branching of *V. bacterium DG1235* with the Opitutae class of bacteria has also been observed in earlier studies (Pilhofer et al., 2008; Wertz et al., 2012). This CSI provides a potentially useful molecular marker for the Opitutae class. Within the Opitutae class, a subclade consisting of *O. terrae*, *D. colitermitum*, and *O. bacterium TAV5* and *TAV1*, which represent the Opitutaceae family of species, was also strongly supported. During our analyses, two CSIs that are specific for this subclade were identified. The sequence information for one of these CSIs consisting of an 11 aa insert in the Urease enzyme, is shown in **Figure 2C**. Another CSI consisting of a 2 aa insert showing similar specificity is present in a helicase domain-containing protein and sequence information for this is presented in **Figure A1** in Appendix. Within the Opitutaceae family, the two unclassified species *O. bacterium TAV5* and *TAV1* exhibit closer relationship in the phylogenetic tree to *D. colitermitum* than to

A

```
                                                                                 416                                      457
                   Opitutaceae bacterium TAV5          373853807    SGITQGLMLSATTDNGTILA HP NFVETLNTIRPMMLFRLVGG
                   Opitutaceae bacterium TAV1          374592581    -------------------- -- --------------------
                   Diplosphaera colitermitum           225155868    ---------N---------- -S ------LA--W------I--
Verrucomicrobia    Opitutus terrae                     182415885    ---------N---EG--V-- Y- Y-ID-I----L---M-VI--
                   Coraliomargarita akajimensis        294054994    ---------NG--EG--L-Q Y- --LD--QS--------AI--
                   Verrucomicrobiae bacterium DG1235   254444747    ---------NNADEQ- --- Y- --L---QS---L--T-V---
                   Chthoniobacter flavus               196231228    ------------RES--V-- Y- --LDAVTSTKSL-HI-AL--
                   Verrucomicrobium spinosum           171910295    --VM-----N----G--V-- Y- -------A---L--L-VL--
                   Arcobacter butzleri                 315636574    A-----M-WR-YDEY-SLVY    S-ID-V-VLH-YYTI-A---
                   Campylobacter jejuni                2415537      A-----M-WR--DEY-NL-Y    S-ID-VVA-V-YYWI-AI--
                   Acidovorax radicis                  351729576    A-VM----WR-INPD--LTY    T---SVKATY-FYVI-VA--
                   Aromatoleum aromaticum              56478350     A-VM----WR--NPD--LTY    S---SVKASY-FWSI--A--
                   Neisseria gonorrhoeae               194099150    A-VM----W-SLN-D--LTY    S---SVKRTM-YYMI-FA--
Other Species      Aeromonas caviae                    334704993    --VM----WR-VN-D--LTY    S---A-QASY-FYFV-FL--
                   Azotobacter vinelandii              226944110    N-------WR-VNQD--LTY    S---A-EASH-GFIV-MI--
                   Pseudomonas aeruginosa              334835436    N-------WR-VNED--LTY    S---S-VASH-GFIV-----
                   Azospirillum brasilense             165874791    ---M----WR-YDTL-FLQY    S----VTATH-FHVI-AA--
                   Rhodopseudomonas palustris          115522048    ---L----WR-Y-SL-FLEY    S-I--VEAMH-FYII-AA--
                   Salinisphaera shabanensis           335424949    --VM----WR-SNPD--LTY    S-IQA-QATY-YYAV--L--
```

B

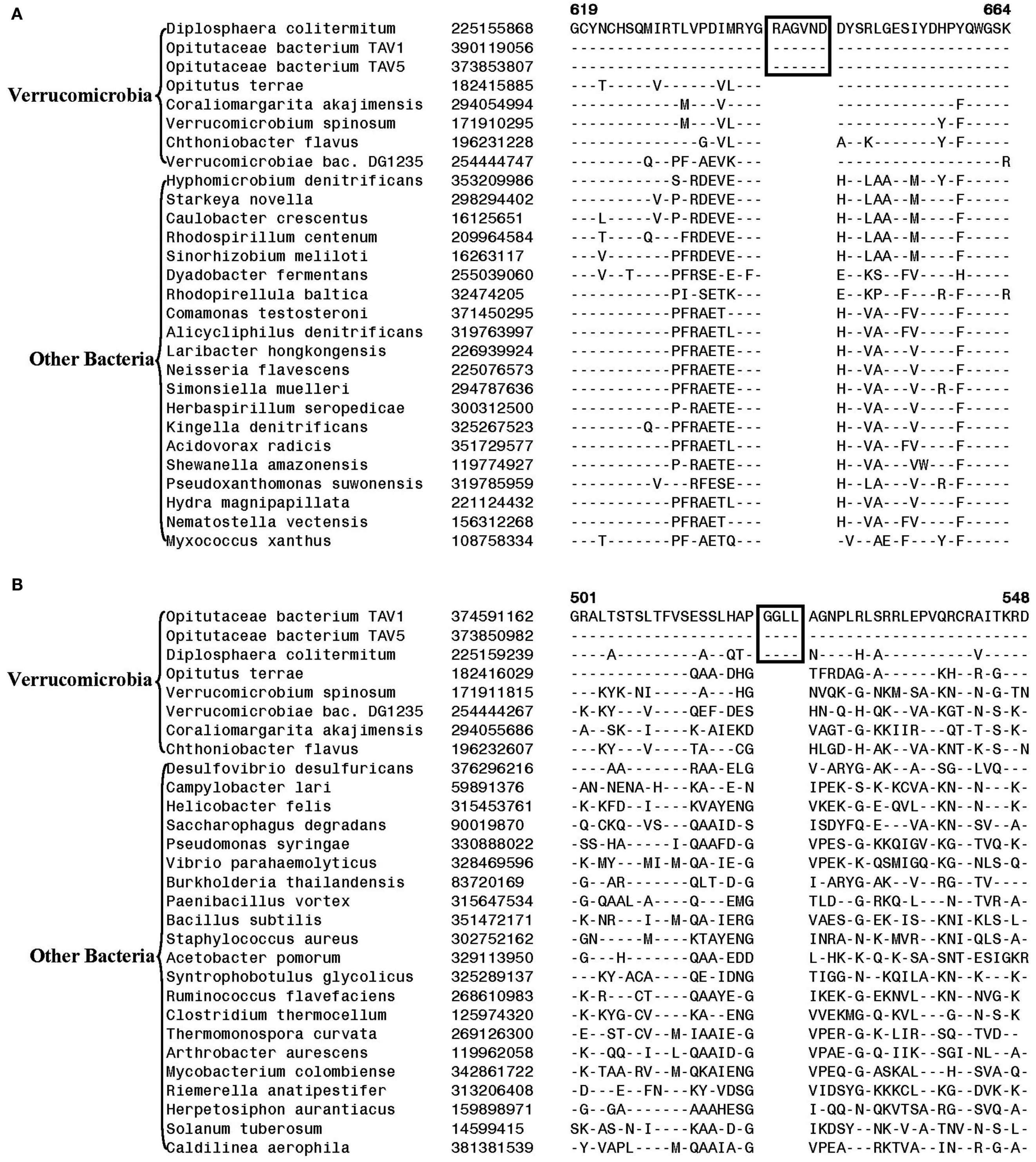
```
                                                                                 271                                   309
                   Opitutaceae bacterium TAV1          374589760    VDEYQDTNTLQSRIIDLMAA   HHRIMAVGDDAQCIYSWRG
                   Opitutaceae bacterium TAV5          373852186    --------------------   -------------------
Opitutae           Diplosphaera colitermitum           225157628    ------------Q-----G-   -------------------
                   Opitutus terrae                     182415966    -----------AQ-V-RL--   --CV---------------
                   Verrucomicrobiae bacterium DG1235   254442638    --------V--AA-V-K--P   -------------------
                   Coraliomargarita akajimensis        294054220    ---F----R---E-V--IGV   N-Q-----------T---
                   Pedosphaera parvula                 223939230    --------A---EL---L-- R --NV-V------S--A---
Other              Akkermansia muciniphila             187735189    --------Y----L--ML-R E -GQL-V------S------
Verrucomicrobia    Chthoniobacter flavus               196231880    --------RI-ADF--IL-K L QRNV-V------S------
                   Methylacidiphilum infernorum        189218921    -------SRI-ADF---LGQ S SQ-V-V------S------
                   Verrucomicrobium spinosum           171912708    --------L---EL-E-L-G P EGNL-V------S------
                   Campylobacter jejuni                380633789    -----------YK-LKNLCC M -EN-TV----D-S------
                   Myxococcus xanthus                  154933919    --------R--GDLV--F-G E RKDLTV----C-S---F--
                   Stigmatella aurantiaca              115378130    --------R--GDLV--LVG E RKNLTV----C-S---F--
                   Leptonema illini                    374587763    --------HA-Y-LVL-L-G G -RNVVV----D-S------
                   Slackia heliotrinireducens          257064330    --------HA-YA-TK-L-- K -QN--V----D-S------
Other Species      Cryptobacterium curtum              256827426    --------HA-YA-TT-L-- R -KN--V----D-S------
                   Leptospira biflexa                  183221122    --------RI-AH-AC-L-S K -QN-LV---------GF--
                   Chlorobium tepidum                  21673182     I-------RV-YLVAKMI-E K -RN-FV------S------
                   Pirellula staleyi                   283782302    --------G--Y---RAL-D E -RN-CV----D-S--G---
                   Methanosphaerula palustris          219852269    ------I----EEL-R--VG S TGNLSV----D----H---
                   Fusobacterium ulcerans              317064713    --------SV-REFLKML-G T NGN-------Y-S--GF--
                   Haloplasma contractile              335429814    --------KS-F-LVR-L-K K --N-FV---TD-S------
```

C

```
                                                                                 99                                             141
                   Opitutaceae bacterium TAV5          373850982    IVGIGHAGNPGIQSGIG SAFPDPVTGKT NPMIVGACTEVIAGE
Opitutaceae        Opitutaceae bacterium TAV1          374591162    ----------------- ----------- ---------------
                   Diplosphaera colitermitum           225159239    -A--------------- --Y---I---P ---------------
                   Opitutus terrae                     182416029    ------G--------L- -T-V--R---K -------A-------
Other              Chthoniobacter flavus               196232607    ---L---------Q--T             HGLVI--A-------
Verrucomicrobia    Verrucomicrobium spinosum           171911815    ----------LL-D--D               -VI--G-------
                   Verrucomicrobiae bacterium DG1235   254444267    ----------L--D--H             PS-T---S-------
                   Coraliomargarita akajimensis        294055686    ----------L--T-VT             DG-VI--G-----A-
                   Puccinia graminis                   331219541    -S---K----D--A-VS             DQ-V--VN-------
                   Trichoderma atroviride              358390863    -----K----DVMD-VT             PG-V--S--D-----
                   Cryptococcus neoformans             308026859    -----K----DMMD-VT             DG----SS----S--
                   Arthrobotrys oligospora             345565448    ---V-K----DVMD-VD             AGLV--SN-D-----
                   Schizophyllum commune               302679944    -----K----DVMANVH             PSL-I-SS-------
                   Helicobacter hepaticus              3859481      -----K----DT-D-VN             EA-V---A-------
Other Species      Campylobacter lari                  48958385     -----K----D--D-VD             SSL-I-TS-DI-GA-
                   Helicobacter bilis                  51102351     -F---K---KDT-D-VC             DKL---TN-------
                   Thermincola potens                  296132134    -----K----NLMD-VD             PG-V---A-------
                   Mycobacterium sp. JDM601            333988912    -----Q----YV-D-VD             PAL---TG-Q-----
                   Gordonia effusa                     359773677    ---L-R----D-SD-VD             PALVI-PS-D-----
                   Albugo laibachii                    325180658    -----KG---DVMD-V-             ANL---VS-------
                   Prochlorococcus marinus             124026631    -S---K----DT-E-VN               I----S--A----
```

FIGURE 2 | Partial sequence alignments of three different proteins showing CSIs that are specific for the Verrucomicrobia species. (A) A 2 aa CSI in a conserved region of Cytochrome c oxidase (cbb3-type) subunit 1 that is specific for all sequenced Verrucomicrobia species where homologs of this protein were identified; **(B)** A CSI consisting of 1 aa deletion in the UvrD helicase that is specific for the Opitutae class; and **(C)** An 11 aa insert in the Urease alpha subunit that is specific for the Opitutaceae family. The CSIs are boxed and the dashes (–) in this and all other alignments indicate identity with the amino acid that is present on the top line. The position of these sequence regions for the species on the top line is noted above the sequence. Except for the indicated groups of Verrucomicrobia, these CSIs are not present in any other species in the top 250 Blastp hits. Sequence information for only limited number of species from other phyla of bacteria are shown in the alignments. The GenBank identifier (GI) numbers for different proteins are shown in the second columns.

O. terrae (Yoon et al., 2008). A close relationship between these species was supported by three CSIs that were identified in the present work. The sequence information for two of these CSIs, which are present in the Cyt c oxidase and the Urease proteins are shown in **Figure 3**. The sequence information for another CSI (a 1 aa deletion) in the Cyt c oxidase protein that is also specific for



```
A                                                          619                                    664
                  Diplosphaera colitermitum      225155868 GCYNCHSQMIRTLVPDIMRYG RAGVND DYSRLGESIYDHPYQWGSK
                  Opitutaceae bacterium TAV1     390119056 --------------------- ------ -------------------
                  Opitutaceae bacterium TAV5     373853807 --------------------- ------ -------------------
Verrucomicrobia   Opitutus terrae                182415885 ---T-----V------VL---        -------------------
                  Coraliomargarita akajimensis   294054994 ------------M---V----        -------------F-----
                  Verrucomicrobium spinosum      171910295 ------------M---VL---        -----------Y-F-----
                  Chthoniobacter flavus          196231228 --------------G-VL---        A--K-------Y-F-----
                  Verrucomicrobiae bac. DG1235   254444747 --------Q--PF-AEVK---        ------------------R
                  Hyphomicrobium denitrificans   353209986 -----------S-RDEVE---        H--LAA--M--Y-F-----
                  Starkeya novella               298294402 ---------V-P-RDEVE---        H--LAA--M----F-----
                  Caulobacter crescentus         16125651  ---L-----V-P-RDEVE---        H--LAA--M----F-----
                  Rhodospirillum centenum        209964584 ---T----Q---FRDEVE---        H--LAA--M----F-----
                  Sinorhizobium meliloti         16263117  ---V-------PFRDEVE---        H--LAA--M----F-----
                  Dyadobacter fermentans         255039060 ---V--T----PFRSE-E-F-        E--KS--FV----H-----
                  Rhodopirellula baltica         32474205  -----------PI-SETK---        E--KP--F---R-F----R
                  Comamonas testosteroni         371450295 -----------PFRAET----        H--VA--FV----F-----
                  Alicycliphilus denitrificans   319763997 -----------PFRAETL---        H--VA--FV----F-----
Other Bacteria    Laribacter hongkongensis       226939924 -----------PFRAETE---        H--VA---V----F-----
                  Neisseria flavescens           225076573 -----------PFRAETE---        H--VA---V----F-----
                  Simonsiella muelleri           294787636 -----------PFRAETE---        H--VA---V--R-F-----
                  Herbaspirillum seropedicae     300312500 -----------P-RAETE---        H--VA---V----F-----
                  Kingella denitrificans         325267523 --------Q--PFRAETE---        H--VA---V----F-----
                  Acidovorax radicis             351729577 -----------PFRAETL---        H--VA--FV----F-----
                  Shewanella amazonensis         119774927 -----------P-RAETE---        H--VA---VW---F-----
                  Pseudoxanthomonas suwonensis   319785959 ---------V---RFESE---        H--LA---V--R-F-----
                  Hydra magnipapillata           221124432 -----------PFRAETL---        H--VA---V----F-----
                  Nematostella vectensis         156312268 -----------PFRAET----        H--VA--FV----F-----
                  Myxococcus xanthus             108758334 ---T-------PF-AETQ---        -V--AE-F---Y-F-----

B                                                          501                                           548
                  Opitutaceae bacterium TAV1     374591162 GRALTSTSLTFVSESSLHAP GGLL AGNPLRLSRRLEPVQRCRAITKRD
                  Opitutaceae bacterium TAV5     373850982 -------------------- ---- ------------------------
                  Diplosphaera colitermitum      225159239 ----A---------A--QT- ---- N----H-A----------V-----
Verrucomicrobia   Opitutus terrae                182416029 -------------QAA-DHG      TFRDAG-A------KH--R-G---
                  Verrucomicrobium spinosum      171911815 ---KYK-NI-----A---HG      NVQK-G-NKM-SA-KN--N-G-TN
                  Verrucomicrobiae bac. DG1235   254444267 -K-KY---V----QEF-DES      HN-Q-H-QK--VA-KGT-N-S-K-
                  Coraliomargarita akajimensis   294055686 -A--SK--I----K-AIEKD      VAGT-G-KKIIR---QT-T-S-K-
                  Chthoniobacter flavus          196232607 ---KY---V----TA---CG      HLGD-H-AK--VA-KNT-K-S--N
                  Desulfovibrio desulfuricans    376296216 ----AA-------RAA-ELG      V-ARYG-AK--A--SG--LVQ---
                  Campylobacter lari             59891376  -AN-NENA-H---KA--E-N      IPEK-S-K-KCVA-KN--N---K-
                  Helicobacter felis             315453761 -K-KFD--I----KVAYENG      VKEK-G-E-QVL--KN--N---K-
                  Saccharophagus degradans       90019870  -Q-CKQ--VS---QAAID-S      ISDYFQ-E---VA-KN--SV--A-
                  Pseudomonas syringae           330888022 -SS-HA-----I-QAAFD-G      VPES-G-KKQIGV-KG--TVQ-K-
                  Vibrio parahaemolyticus        328469596 -K-MY---MI-M-QA-IE-G      VPEK-K-QSMIGQ-KG--NLS-Q-
                  Burkholderia thailandensis     83720169  -G--AR-------QLT-D-G      I-ARYG-AK--V--RG--TV----
                  Paenibacillus vortex           315647534 -G-QAAL-A----Q---EMG      TLD--G-RKQ-L---N--TVR-A-
                  Bacillus subtilis              351472171 -K-NR---I--M-QA-IERG      VAES-G-EK-IS--KNI-KLS-L-
                  Staphylococcus aureus          302752162 -GN-----M----KTAYENG      INRA-N-K-MVR--KNI-QLS-A-
Other Bacteria    Acetobacter pomorum            329113950 -G---H-------QAA-EDD      L-HK-K-Q-K-SA-SNT-ESIGKR
                  Syntrophobotulus glycolicus    325289137 ---KY-ACA----QE-IDNG      TIGG-N--KQILA-KN--K---K-
                  Ruminococcus flavefaciens      268610983 -K-R---CT----QAAYE-G      IKEK-G-EKNVL--KN--NVG-K
                  Clostridium thermocellum       125974320 -K-KYG-CV----KA--ENG      VVEKMG-Q-KVL---G--N-S-K
                  Thermomonospora curvata        269126300 -E--ST-CV--M-IAAIE-G      VPER-G-K-LIR--SQ--TVD--
                  Arthrobacter aurescens         119962058 -K--QQ--I--L-QAAID-G      VPAE-G-Q-IIK--SGI-NL--A-
                  Mycobacterium colombiense      342861722 -K-TAA-RV--M-QKAIENG      VPEQ-G-ASKAL---H--SVA-Q-
                  Riemerella anatipestifer       313206408 -D---E--FN---KY-VDSG      VIDSYG-KKKCL--KG--DVK-K-
                  Herpetosiphon aurantiacus      159898971 -G--GA-------AAAHESG      I-QQ-N-QKVTSA-RG--SVQ-A-
                  Solanum tuberosum              14599415  SK-AS-N-I----KAA-D-G      IKDSY--NK-V-A-TNV-N-S-L-
                  Caldilinea aerophila           381381539 -Y-VAPL----M-QAAIA-G      VPEA---RKTVA--IN--R-G-A-
```

FIGURE 3 | Partial sequence alignment of (A) Cytochrome c oxidase and (B) alpha subunit of urease, showing two CSIs (boxed) that are specifically present in *D. colitermitum, Opitutaceae bacterium TAV1,* and *Opitutaceae bacterium TAV5* species.

these species is presented in **Figure A2** in Appendix. It is noteworthy that these two proteins (viz. Cyt c oxidase and Urease) also contain other CSIs in different positions that are specific for the phylum Verrucomicrobia or the class Opitutae (**Figures 2A,B**), indicating that distinct genetic changes within these genes have occurred at different evolutionary stages.

PHYLOGENY AND MOLECULAR SIGNATURES FOR THE PLANCTOMYCETES SPECIES

The 11 Planctomycetes species for which sequences are available also formed a well-supported clade in our phylogenetic tree (**Figure 1**). The Planctomycetes species have been divided into two separate classes: the Phycisphaerae and the Planctomycetia (NCBI Taxonomy, 2012). *Phycisphaera mikurensis* is the sole recognized and sequenced species for the class Phycisphaerae. The Planctomycetia class is further divided into the orders Planctomycetales and Candidatus Brocadiales (Ward, 2011). The Candidatus Brocadiales consists of several candidate species including *K. stuttgartiensis*. Complete genomes for nine organisms from the order Planctomycetales are available: *Blastopirellula marina*, *Gemmata obscuriglobus*, *Isosphaera pallida*, *P. staleyi*, *Planctomyces* (*Pl.*) *brasiliensis*, *Pl. limnophilus*, *Pl. maris*, *Rhodopirellula baltica* and *Singulisphaera acidiphila*. The nine species of the Planctomycetales order, as expected, branched together in the tree. However, in conflict with the established placement of *K. stuttgartiensis* within the class Planctomycetia, this species was observed as the deepest branching member of the phylum with *Ph. mikurensis* sharing a closer relationship to the species of the Planctomycetales order. The deeper branching of the anammox species (viz. *K. stuttgartiensis*) in comparison to Phycisphaera has also been observed in earlier studies (Fukunaga et al., 2009; Fuchsman et al., 2012). Similar to the Verrucomicrobiae, no molecular or biochemical marker is known that is specific for the Planctomycetes species. However, two of the CSIs identified in this work were specific for all of the sequenced species from this phylum. The sequence information for one of these CSIs, consisting of a 6 aa insert in a conserved region of an ABC transporter protein is shown in **Figure 4A**. This CSI is uniquely present in all of the sequenced Planctomycetes species, but it is not found in any other bacteria. Similarly, in the SpoVG protein, which is involved in methicillin and glycopeptide resistance and production of extracellular polysaccharides in virulent *Staphylococcus aureus* (Matsuno and Sonenshein, 1999; Schulthess et al., 2009), a 36 aa insert in a conserved region is present in all of the sequenced Planctomycetes species (**Figure A3** in Appendix). In view of the observed specificities of these CSIs for the species from the phylum Planctomycetes, they provide molecular markers for this phylum.

Another CSI identified in the present work supports the view that *K. stuttgartiensis* represents a deep-branching group of organisms within the phylum Planctomycetes. In this case, a 10–11 aa insert in a conserved region of the protein cobyrinic acid ac-diamide synthase is present in all of the sequenced Planctomycetes species except *K. stuttgartiensis* (**Figure 4B**). The simplest and most likely explanation for the species distribution pattern of this CSI is that the genetic change leading to this insert was introduced into a common ancestor of other sequenced Planctomycetes species after the divergence of *K. stuttgartiensis*. Hence, the absence of this CSI from *K. stuttgartiensis* supports its position as the deepest branching sequenced species from this phylum, which is in agreement with its branching position in the phylogenetic trees (**Figure 1**; Fuchsman et al., 2012).

MOLECULAR MARKERS FOR THE LARGER CLADES WITHIN THE PVC PHYLA OF BACTERIA

Although the species of the phyla Planctomycetes, Verrucomicrobia, Lentisphaerae, and Chlamydiae formed distinct clades and branched in the proximity of each other in the phylogenetic tree based upon concatenated protein sequences (**Figure 1**), the grouping of these phyla into a single clade or other multi-phyla clades was very poorly supported by ML analysis, highlighting the concerns from earlier studies regarding amalgamation of these phyla into a single "superphylum" (Cho et al., 2004; Wagner and Horn, 2006; Griffiths and Gupta, 2007). Hence, molecular markers that could provide independent support for the grouping of these phyla are of much importance. Our analysis has identified a few molecular markers that are helpful in these regards.

In our earlier work on Chlamydiae, a 3 aa insert in the β subunit of RNA polymerase (RpoB) was identified that in addition to the sequenced Chlamydiae species was also exclusively present in one Verrucomicrobia species (*V. spinosum*) whose sequence was available at that time (Griffiths and Gupta, 2007). An updating of the sequence information for this CSI (**Figure 5**) indicates that this CSI is specifically present in all members of the Chlamydiae and Verrucomicrobia phylum along with the two species of the phylum Lentisphaerae for which sequences are available. However, this CSI is not present in any other bacteria including different Planctomycetes and the Poribacteria. The unique shared presence of this conserved insert in this essential protein by all sequenced Chlamydiae, Verrucomicrobia, and Lentisphaerae species strongly indicates that the species from these three phyla shared a common ancestor exclusive of all other bacteria. Thus, the species distribution pattern of this CSI strongly supports the grouping together of these three phyla into a single large clade, consistent with their branching in the phylogenetic tree. The absence of this CSI in the Planctomycetes species is also consistent with its deeper branching in comparison to the other three phyla (**Figure 1**; Ward et al., 2000; Jenkins and Fuerst, 2001; Wagner and Horn, 2006; Griffiths and Gupta, 2007; Hou et al., 2008; Pilhofer et al., 2008).

Our detailed analysis identified no CSI that was specifically shared by all or most of species from the PVC phyla of bacteria. However, we have identified one signature protein, whose specific presence in various species belonging to the PVC clade suggests that the species from the four main phyla might be specifically related. The protein of interest is a hypothetical protein (the protein CT421.2 from *C. trachomatis*; accession number NP_219933) whose length varies from ∼53 aa in the Chlamydiaceae to more than 80 aa in the Planctomycetes. In BlastP searches with the *C. trachomatis* homolog all of the observed hits for this protein are for the PVC group of species and no hit outside of this group is observed. The 53 aa long region of this chlamydial protein is well conserved in all sequenced species belonging to the PVC clade and a sequence alignment for this region is presented in **Figure 6**. The specific presence of this protein in the PVC group of bacteria (all except Poribacteria) suggests that the gene for this protein initially

```
A
                                                                 132                                                      187
                 Pirellula staleyi                  283778435    QFYHLLPELSTLENVLTPIMISQ GVFQYR LYRRAHQQRAKELLELVGLGHRLTHKP
                 Planctomyces maris                 149174058    ---------NL-----S-L--RY STWE-K WRKKQFR-D-L-II-K---S--IK-R-
                 Phycisphaera mikurensis            383766318    -S---IG---A-Q---LAARVAT PLTRWS PT--EAKA--AD--GR----D--H-R-
                 Blastopirellula marina             87307061     ----------------------H -MWR-S WA-KQFRK--E----M---S--T----
                 Rhodopirellula baltica             327541711    ----------AI----A-A--RR S-LG-- -D-KSLRL--EAM-DR---LT-SH-Q-
Planctomycetes   Planctomyces limnophilus           296122764    -----M---TLV---VL-L--RH SIWS-K WN--S-ID---Q-IDR---S-------
                 Planctomyces brasiliensis          325108692    ---------TLI---IA-L--RY STWE-S WR--ELRD----MIDQ---S--IK-R-
                 Gemmata obscuriglobus              168700215    ---------TA-D---M-AY--N STLGW- WN-ARLRH--E----KM------K-R-
                 Singulisphaera acidiphila          373480795    ----------A----MM-QL-RH -LWS-G MQ-GRIRRE-TA---R-------N-L-
                 Isosphaera pallida                 320104203    ---------TA-----L-H--RH S-LG-Y YR-A-LTRE-AA-M-R----D--H-R-
                 Can. Kuenenia stuttgartiensis      91202042     -------DFTA-----L-CF-GK QFSKNP ASKENFHN--VS---R----G------
                 Alistipes indistinctus             354604541    --H-----FTAM---CI-GY-AR        -D-TDVERQ-R----ML-ME--TE---
                 Cellulophaga algicola              319954972    --HQ----FTA----CI-AF-KN        TTKN-AEK------DFL--S--YH---
                 Flavobacterium branchiophilum      347535880    --HQ----FTA----CI-GF-A-        KNKKETEEA--K--DYL--S--IH---
                 Prevotella histicola               357043907    --HQ----FTA---IMI-AY-AG        KNTKEAR---E----FM--SD-SS---
                 Solitalea canadensis               379653321    --H-----FTA----CI-AF-AK        QS-KEAEKY-H-----L--S---Q---
                 Escherichia coli                   170679953    --H----DFTA----AM-LL-GK        KKPAEINS--L-M-KA------AN-R-
                 Methylomonas methanica             333983933    --H---G-FTI----AM-LL-GK        -PI-EA-AE-A---KR------VM---
                 Pseudomonas aeruginosa             254235961    --H-----FTA----CM-LL-GR        TPIVEAR---A----R-------S---
                 Vibrio fischeri                    197337826    --H---ADF-A----AM-LL-GG        VNKKEAAK--T----K---A--IE-R-
Other Species    Xylella fastidiosa                 15837679     --H-----FTA----MM-VLL-G        QSVCSG-M--QM---A-E-----NY--
                 Fusobacterium gonidiaformans       315917243    -SF----K--A----EL-LVYAG        ISKKEREE---KM--I----T--H-R-
                 Pelobacter propionicus             118581421    --H-----F-A----MM-AL-AR        VP--DAISL-SA--AD---------R-
                 Geobacter uraniireducens           148265266    --H-----F-A---AMM-LL-GG        MK-GEAEAI-E---RD---S-------
                 Syntrophus aciditrophicus          85860096     --HN----F-S---TMM-AL--G        MP--NALE--ET--HD----D-M----
                 Leptospirillum ferrodiazotroph     251770985    --H-----F-A-----M-TW--G         KD-EDTSW-RT---E----D--D---
                 Ktedonobacter racemifer            298245390    -A-N-I-T-TAQ---EV-LYVGK         HKGSPAA------D----S---G-R-
                 Aromatoleum aromaticum             56478740     --H-----F-A----AM-LY-RR        ME-E-ANE--VAM-KE-------D-A-
                 Chromobacterium violaceum          34498418     --H-----FTA----MM-LL-RR        MDKGEAAR--A-M--R----K--E---
                 Thermotoga maritima                15643120     -S-N---R-TA----EL-MIYAG        VPAKERKR------------D--H-R-
                 Mycobacterium gilvum               145224707    -S-N-I-F--AV---EL-L-FEP        YD-K-LRK--I-----------IN-Q-

B
                                                                 178                                                           235
                 Pirellula staleyi                  283778187    GVLLCMYDSGTRLAAEVSSDVTEYF TRERTPECVWS EARTFQTRIRRNIRLAEAPS
                 Rhodopirellula baltica             32473812     --V-----AN--------T-ID-F- AASKDGREFF- G-KF-D--------------
                 Blastopirellula marina             87308313     --I--LFE-S----G--AG--DQF- -NGAGANTA-A D-KV-R--------------
                 Singulisphaera acidiphila          373477668    -----L--A----GG--IE-LDRF- DGR-S-PMP-- D-QVYR-----------C--
                 Isosphaera pallida                 320103913    --V----E-----SS--ID-LETF- QAK-Q-NSA-A D-KL-------------S--
Planctomycetes   Planctomyces maris                 149176663    --V--L-ET------D-TD-LSAFL  NNSD--AP-- S-KV--S------K------
                 Planctomyces limnophilus           296121620    -IV--L-ET------DITD-L-RFL  E-SD-QSP-A Q--I-SS------K------
                 Planctomyces brasiliensis          325108937    -L---L-ET------D-TD-LMRFL  D-AD-MAP-A N--I-NS-------------
                 Gemmata obscuriglobus              168705300    --VV-L--AA-K--Q--VT-L-TFL AQS-GANVP-A N-KV-G-------K---C--
                 uncultured planctomycete           374849352    -II--L--AS----N--IR-LET-L QKS-ALNVP-A Q-QI---------K---C-G
                 Phycisphaera mikurensis            383766330    --V---HEAN-L--G-IRAELDGF-  GQAGEGDA-H Q-EV-DPPV----K------
                 Can. Kuenenia stuttgartiensis      91203316     --I----S-RA--CN--VEKIR---             DEKV-D-IV-K--K-S-S--
                 Acetobacterium woodii              379013358    ---MT-F--R-N-SNQ-VDE-VD--             KDKVYE-M-P----------
                 Alkaliphilus oremlandii            158321886    --V-S-F-GR-N-SIQ-VDE-KN--             RGKVYT-I-P--V-------
                 Anaerococcus prevotii              257067211    ----T-F-KR-N-SY--VEE-KS--             KNKV-K-M-P--V-------
                 Halothermothrix orenii             220933184    ----T---AR-N-SQQ-IDE-KN--             KNKVYE-I-P--V--S----
                 Johnsonella ignava                 358068677    -I-FT---PRNN-SSQ-IEN-K-SL             NENIYN-V-P----------
                 Aerococcus urinae                  326804287    ---MT-F--R-N--N--VEE-RK--             GDKVYN-L-P--V--S----
                 Bacillus megaterium                294501983    ----T-L-AR-N-GLQ-TAE-KK--             QD-VY--I-P--V--S----
                 Enterococcus faecalis              307286502    ----T---AR-N-G---VEE-RK--             REKVYD-I-P-----S----
Other Species    Lactobacillus brevis               227509065    ----T-F-AR-N-GVQ-NQE-RK--             KNEVYE-V-P--V--S----
                 Paenibacillus larvae               167462787    ----T-F-AR-N-GIQ-IEE-KK--             QQKVY--I-P--V--S----
                 Ktedonobacter racemifer            298243815    --V-T-F-PR----GDIVRE-RNH-             PKEA---I-N--V--S----
                 Selenomonas ruminantium            383755638    --VMT----R-K--EQ-VAE-RNS-             D-VVYK-M-P--V--S----
                 Dialister invisus                  258646392    -I--T-F-GR-N-SLQ-ADE-KK--             GNKV-R-V-P-SVK-S----
                 Treponema primitia                 374813699    -LFFT----R----Q--V-Q-SA--             KQKV-T-IVP--V--S----
                 Leptospira noguchii                359725024    ----T-F-KR-N--NQ-AE--KS--             KDKVYT-I-P--VK-S----
                 Turicibacter sanguinis             293376419    ----T-L-RR---GLD-INE-KL--             KEKV-N-I-P-LV--S----
                 Desulfovibrio africanus            374299440    --V-T---KRN--SGQ-KNE-RRS-             PDKM-E-IVP--V--S----
                 Rickettsia massiliae               157964124    -I-FT---KRN--TEQ-ED--RKCL             G-LV-K-V-P---K-S----
```

FIGURE 4 | Partial sequence alignments of (A) a conserved region within an ABC transporter protein depicting a 4 aa insert that is specifically present in in all sequenced Planctomycetes species; (B) An 11 aa insert in the cobyrinic acid ac-diamide synthase that is specific for all sequenced Planctomycetes except *Candidatus Kuenenia stuttgartiensis*.

originated in a common ancestor of these organisms, followed by its vertical transmission to various descendants. Although the function of this protein is not known, its specific presence in the PVC group of bacteria provides suggestive evidence that the species from these groups shared a common ancestor exclusive of other bacteria.

			163		203
Chlamydiae	Chlamydia muridarum	301336775	IIPYRGSWLEASFDINDLIYIHID	RKK	RRRKILAMTFIRAL
	Chlamydia trachomatis	376008076	-----------V------------	---	--------------
	Chlamydophila felis	89898127	-----------I------------	---	--------------
	Chlamydophila caviae	29840449	-----------I------------	---	--------------
	Chlamydophila pecorum	330444699	-----------I--V---------	---	--------------
	Chlamydophila pneumoniae	15835616	-----------I------------	---	-------I------
	Chlamydophila psittaci	329943036	-----------I------------	---	-------I------
	Chlamydophila abortus	333410414	-----------I------------	---	-------I------
	Simkania negevensis	338733407	----------S---------VYV-	---	----V--TS---T-
	Criblamydia sequanensis	343183572	----------GA--M----H-Y--	---	-------T------
	Candidatus Protochlamydia	46446238	----------GA--T----H-Y--	---	-------T------
	Parachlamydia acanthamoebae	282889742	----------GA--S----H-Y--	---	-------T------
	Waddlia chondrophila	297620829	----------GA--T--M-H-Y--	---	-------T----S-
	Estrella lausannensis	343183585	----------GA--TG---H-Y--	---	-------S------
Verrucomicrobia	Chthoniobacter flavus	196233588	---D----V-VQ--T---L-VYL-	-R-	----F--T--L---
	Pedosphaera parvula	223936435	---D----Y--Q--TS--L-VYL-	---	----F-TT--F---
	Methylacidiphilum infernorum	189218816	---D------VA--T---L-VYL-	-R-	K---F-IT-LL---
	Akkermansia muciniphila	187735536	---D------VQ--T---L-VYL-	-RR	----F--T--M-Y-
	Opitutaceae bacterium TAV1	374590103	---D--T---VQ------L-VY--	-RR	----F-IT-LF---
	Methylacidiphilum fumariolicum	384915709	---D------VA--S---L-VYL-	-R-	K---F-IT-LL---
	Diplosphaera colitermitum	225164279	---D--T---VQ------L-VYL-	-RR	----F-IT-LF---
	Opitutaceae bacterium TAV5	373854229	---D--T---VQ------L-VYL-	-RR	----F-IT-LF---
	Verrucomicrobium spinosum	171914821	---D--T---VQ--T---L-VYL-	-RR	----F--T-LL-VI
	Coraliomargarita akajimensis	294056237	---D--T---VQ--Q---L-VYL-	-RR	----F-LT-LL--M
	Opitutus terrae	182412057	---D--T---VQ--N---L-VYL-	-RR	----F-IT-LL--I
	Verrucomicrobiae bac. DG1235	254442756	---D--T---VQ--N---L-VYL-	-RR	----F-IT-LL--V
Lentisphaerae	Victivallis vadensis	281358737	---D----MDVQ-----F---YL-	-RR	----FYIT--L--I
	Lentisphaera araneosa	149198915	---D----M-VQY-NH-----FM-	-RR	----F-IS--L--V
Planctomycetes	Phycisphaera mikurensis	383767519	V--E----I-LEVSKK-VLQMR--		QST--P-T--L---
	Planctomyces limnophilus	296120714	V--E----I-LNIGKR-TLNVR--		QSG-FS----L--M
	Isosphaera pallida	320101660	---E----I-LQVNKK-ALEVR--		QSG-FS---LL--M
	planctomycete KSU-1	386812691	---E----I-LEVGKK-ILTVR--		QSG-LP-TC-L---
	Can. Kuenenia stuttgartiensis	91200660	---E----I-LEVGKK-VLTVR--		QSG-LP-TC-L---
	Planctomyces maris	149177090	V--E----I-LVVGKK-TLGVR--		QSG-FS---LL--M
	Pirellula staleyi	283780325	---E----I-VNVTKREALS-R--		QSG-FS-L-LL--M
	Singulisphaera acidiphila	373477164	---E----I-LQVTKKETLGVR--		QSG-FS---LL--M
	Planctomyces brasiliensis	325108564	---E----I-LLISKKETLGVR--		QSG-FS---LL--M
	Blastopirellula marina	87306545	---E----I-INITKK-SFTVR--		QSG-FA-T-LL--M
	Gemmata obscuriglobus	168700810	---E----I-INATKK-TLGVR--		QSG-FS-V-LL--M
	Rhodopirellula baltica	32473688	V--E----I-VNVTKK-ALTVR--		QSG-FA-TMLL--M
Poribacteria	Can. Poribacteria sp. WGA-A3	284106476	---------DFE--AR-IL-VR--		----MP-TILLK-F
Other Species	Pseudomonas syringae	330969829	---------DFE--PK-CVFVR--		----LP-SVLL---
	Azotobacter vinelandii	226942770	---------DFE--PK-AVFVR--		----LP-SVLL---
	Vibrio cholerae	121728867	---------DFE--PK-NL-VR--		----LP-SIIL---
	Escherichia coli	378211764	---------DFE--PK-NLFVR--		----LP-TIIL---
	Rickettsia canadensis	157803327	V--------DLE--AK-I--FR--		-K--LY-T-LL--I
	Rhizobium etli	190891347	V--------DIE--AK-IV-AR--		-----PVTSLLM--
	Sorangium cellulosum	162448680	V--------DFE--PK-I--VR--		----MH-TVLL---
	Kingella kingae	333376362	---------DFE--PK--L-FR--		-----PVTILL---
	Laribacter hongkongensis	226939179	---------DFE--PK--L-FR--		----MPVT-LLK--
	Simonsiella muelleri	294789168	---------DFE--PK--L-FR--		----MPVTILL---
	Eikenella corrodens	225024705	---------DLE--PK--L-FR--		----MPVTILLK--
	Nitrosomonas europaea	30249986	---------DFE--PK-YV-FR--		----MPVT-LLK-M
	Sutterella parvirubra	378821788	V--------DFE--AK-IL-FRV-		----MPGTILLK--
	Candidatus Nitrospira	302036657	---------DFE--AR-IL-VR--		----MP-TILLK-F
	Trypanosoma congolense	343473637	---------DFE--PK-CVFVR--		----LP-SVLL---
	Holophaga foetida	373489184	--------I-FEL-TKG-F-AR--		-K--F-GS--M---
	Candidatus Koribacter	94971702	--------V-FEY-QKNIL-VR--		-K--F-GTI-L---
	Terriglobus saanensis	320105627	--------V-FEY-QKNTL-VR--		-K--F-GTI-L---
	Deferribacter desulfuricans	291280155	--------IDFE--NK-VMHVR--		KK----VT-LLK--
	Eubacterium siraeum	167749850	V--N--A---YEM-S--IF-VR--		KN---P-T------
	Ruminococcus flavefaciens	268610263	V--N--A---YEM-S--VV-VR--		KN---PIT------
	Prevotella denticola	325853552	---FK---I-FAT---NVM-AY--		-KK-LPVT-ML--I

FIGURE 5 | A 3 aa insert in a conserved region of the RNA Polymerase β subunit (RpoB) that is specifically present in all sequenced Chlamydiae, Verrucomicrobia, and Lentisphaera species, but not found in Planctomycetes or any other phyla of bacteria.

```
                 Chlamydia trachomatis            15605148     MSRHRSYGKSV KGETKRNVLKRFERIEVLRKLGRWDDATAKKATGLLKTPAIK
                 Chlamydia muridarum              15835319     ----------I ---------------------------------------V--
                 Chlamydia trachomatis            380250941    ----------- ---------------K-----------------------V--
                 Chlamydophila felis              89898577     ----------I ------------------------N----------P---VM-
                 Chlamydophila pecorum            330444256    ----------- --V--------------------------------P---IL-
                 Chlamydophila psittaci           329942557    ----------I ---------------DL-------N----------P---V--
Chlamydiae       Chlamydophila caviae             29840003     ----------I ---------------D--------N-V--------P---VM-
                 Chlamydophila abortus            62184874     ----------I ---------------DL----S--N-T--------P---V--
                 Chlamydophila pneumoniae         15618419     ----------- --V------------V--------N-S----V---P---IL-
                 Simkania negevensis              338732803    ----P-----N --GK-Q------Q--DI-------K-GEN--V---P---V
                 Can. Protochlamydia amoebophila  46446687     ----P-F--AG -TA------------D--K-I---K-SEN-RI---P---VL
                 Waddlia chondrophila             297621267    ----P-F---N --GQ----------LD--K-----EKDKD--I---P---VM
                 Parachlamydia acanthamoebae      282891329    ----P-F-S-S -AA------------D--------K-GEN-RV---P--
                 Verrucomicrobiae bac. DG1235     254444458    --Q-P--KSNA STGA----------V-L-K-R-E-K-  GDRVS--R--KPEA
                 Methylacidiphilum infernorum     189475224    --Q----RTGS LLVA-------Y---NI-K-Q-K-KE  GD-VL--P--KP-
                 Opitutus terrae                  182414531    --Q-K-LQG-S GIVI----------VDI-K-R-Q-KA  GDRVQ--R--KPDV
                 Pedosphaera parvula              223940523    --Q---LRAVA TMGG----------VGL-K-R-Q-KE  GDRI---R--KPEA
                 Coraliomargarita akajimensis     294053808    --Q-N-FKA-GGG-KKN-T-------V-L---R-E-KE  GDRVI--K--QPEA
Verrucomicrobia  Chthoniobacter flavus            196231188    --Q---LKGQS TIAA----------V-L-QQR-QYKQ  GD-VV--R--KPNE
                 Opitutaceae bacterium TAV5       373851114    --Q-N-LQGPR GIVI----------VAL-K-R-Q-K-  GERVM--R--KPDV
                 Opitutaceae bacterium TAV1       374592647    --Q-N-LQGPR GIVI----------VAL-K-R-Q-K-  GERVM--R--KPDV
                 Verrucomicrobium spinosum        171911200    --Q---LKS-GGSVG---S----G---KLMKAR-Q-NE  GRSLYN-P--KPEA
                 Diplosphaera colitermitum        225158855    --Q-N-LQGPR GIVI----------VAL-K-RSQ-KE  GDRVM--R--KPDV
                 Akkermansia muciniphila          187734755    --K-S-LKATG TVGG--S-------VKL-KER-E-KK  GQSPL--P--KHEA
Lentisphaerae    Lentisphaera araneosa            149200569    --I---LKVKG NTAG-K--------VDQ-IQE--LKP  GDQVL--P--KVNI
                 Victivallis vadensis             281357842    --Q-P-LRVGG -IRKI---M--Y---D--LAD--IKE   D-VL--P--KPTE
                 Planctomyces maris               149178348    V-LDK-LKSKS TLVRA------A----K-KFED--VE  GQG-L--P-VRVE-
                 Planctomyces brasiliensis        325110909    --LNS-LKTGS -MKRP------H----Q-Q-QD--IE  GQAPI--P-VRVF-
                 Planctomyces limnophilus         296121219    V-IDK-LKRKG RLARS----S-S---LQMIGEDKLGT  QDSPF--P-LRIV-
                 Blastopirellula marina           87306489     -TIDK-LKVKR G-ISS-S--T-V--L-QM--A-KFNP ETDSPI-IP--RV--
                 planctomycete KSU-1              386404875    --IDK-LKTKG -LVRP----T-I---KL--EEKK-EP  TISVF-IP-VKVL-
                 Singulisphaera acidiphila        373477821    --IDK-LK-GG GLSRT----T-P--LAL-QEDE--KP  -QGVFN-P--KFRR
Planctomycetes   Isosphaera pallida               320104842    --IDK-LR-GG GLGGQ------A--MAL-QED---TE  NHGPYN-P--KYR-
                 Gemmata obscuriglobus            168701653    --IDK-LKRKA GMSRQ-C--T-A---TKMLEN-KFGA  ESSPY--P--RVQ
                 Can. Kuenenia stuttgartiensis    91204170     --IDK-LKPKG -LSRP---YRKV---AI-KAEN---Q  DTSVF--P-VKVE-
                 Phycisphaera mikurensis          383766696    --LDS-LKAGG GLKRH----T-P----L-A-KDAF-K KQGDPL--P-VGNR-
                 Pirellula staleyi                283779428    -PIDN-LKVKA GAISN----T-A--L-K-KEAD--KE  GD-VL-MP-VR
                 Rhodopirellula baltica           327541007    -D-SLKVQAGA   IKS----T-A--V-RMKD-DKFNE  ESSII-MP--R
```

FIGURE 6 | Sequence alignment of a protein of unknown function that is uniquely found in various species from the PVC phylum of bacteria except Poribacteria. In Blastp searches, no homolog of this protein is detected in any other bacteria outside of the PVC clade of bacteria.

DISCUSSION AND CONCLUSION

The PVC superphylum is proposed to be composed of numerous species that are part of four phyla and three candidate phyla. With several cellular features unique to members of this group of bacteria as well as the important pathogenic organisms present within this group, the relationships that these bacteria share with other prokaryotes and with each other is of great evolutionary interest (Devol, 2003; Sachse et al., 2009; Fuerst and Sagulenko, 2011; McInerney et al., 2011). However, elucidation of the relationships among the PVC group of bacteria has thus far proven difficult and led to contradictory results by phylogenetic means. In this work, we report for the first time identification of molecular markers in the form of CSIs and CSPs that are unique and distinctive characteristics of species from the phyla Verrucomicrobia and Planctomycetes and others that provide independent support for the grouping of species from the phyla Planctomycetes, Verrucomicrobia, Chlamydiae, and Lentisphaerae into larger clades. Large numbers of CSIs and CSPs for the Chlamydiae species were identified in our earlier work (Griffiths et al., 2005, 2006; Gupta and Griffiths, 2006). Based upon the species distribution patterns of these markers, the evolutionary stages where the genetic changes responsible for them have likely occurred are depicted in **Figure 7**.

Grounded upon the identified markers, it is now possible to clearly distinguish species from each of the three main phyla (viz. Planctomycetes, Verrucomicrobia, and Chlamydiae) that comprise the PVC clade of bacteria in molecular terms. The specificities of these markers for the species from these clades provide independent evidence for the monophyly of these clades. Additionally, based upon these molecular markers a number of relationships within these bacterial phyla can also be consolidated. Within Verrucomicrobia, newly identified CSIs allow the species from the class Opitutae and family Opitutaceae to be distinguished in molecular terms. The species distribution of these CSIs strongly indicate that the species *V. bacterium DG1235*, which is currently a part of the class Verrucomicrobiae, should in fact be transferred to the class Opitutae. A number of CSIs also provide evidence that the two unclassified species belonging to the family Opitutaceae viz. *O. bacterium TAV5* and *TAV1* are closely related to *D. colitermitum* and they should perhaps be assigned to the genus *Diplosphaera.* Within Planctomycetes, the species distribution pattern of the identified CSIs strongly indicates that the anammox species *K. stuttgartiensis* constitutes the deepest branching lineage of this phylum, which is consistent with its branching in the phylogenetic tree. However, this inference is at variance with the current assignment of *K. stuttgartiensis* to the class Planctomycetia, whereas the species *Ph. mikurensis* which branches less deeply than *K. stuttgartiensis* is part of a separate class (Phycisphaerae). The anammox organisms such as *K. stuttgartiensis*

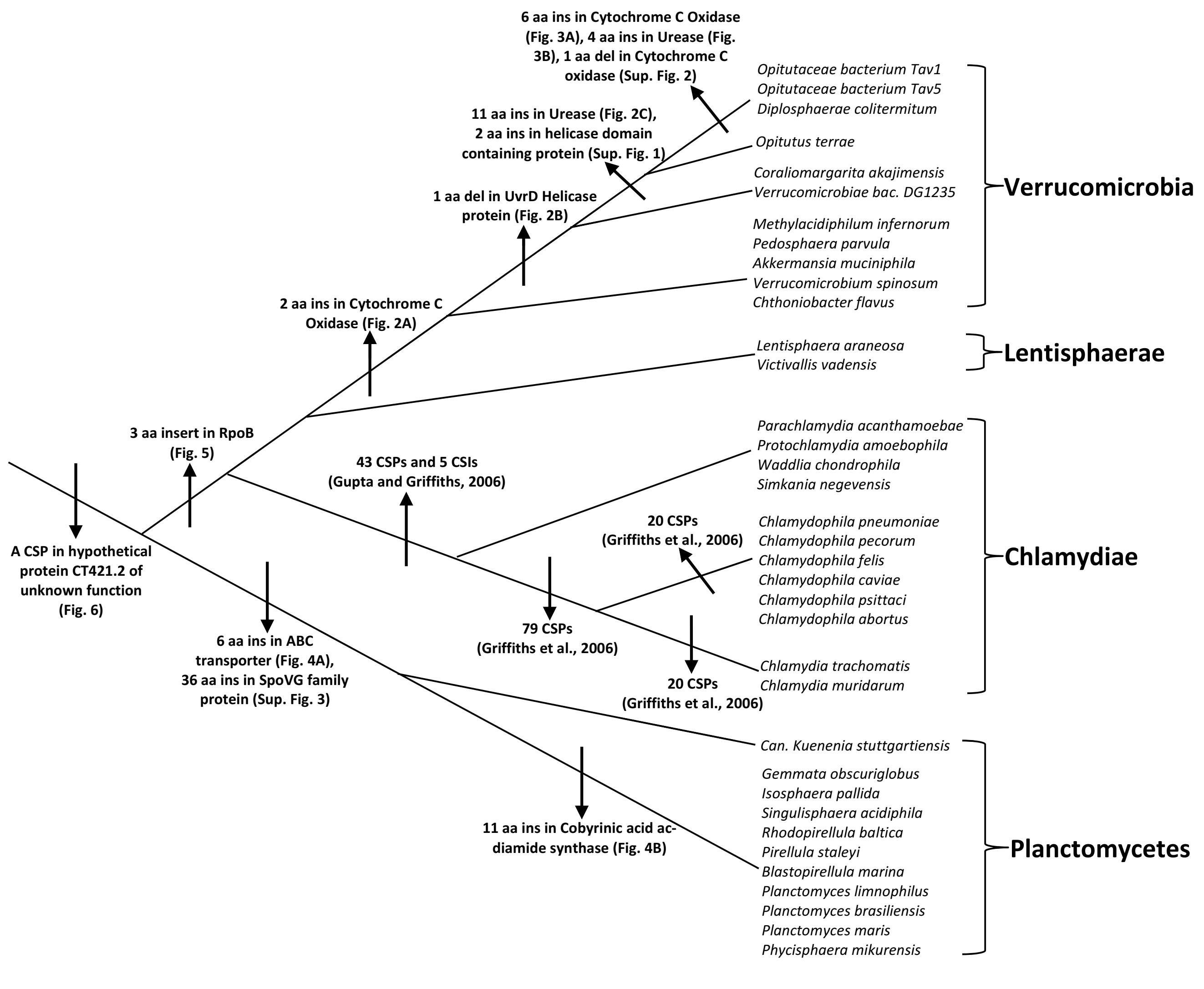

6 aa ins in Cytochrome C Oxidase (Fig. 3A), 4 aa ins in Urease (Fig. 3B), 1 aa del in Cytochrome C oxidase (Sup. Fig. 2)
11 aa ins in Urease (Fig. 2C), 2 aa ins in helicase domain containing protein (Sup. Fig. 1)
1 aa del in UvrD Helicase protein (Fig. 2B)
2 aa ins in Cytochrome C Oxidase (Fig. 2A)
3 aa insert in RpoB (Fig. 5)
43 CSPs and 5 CSIs (Gupta and Griffiths, 2006)
20 CSPs (Griffiths et al., 2006)
A CSP in hypothetical protein CT421.2 of unknown function (Fig. 6)
6 aa ins in ABC transporter (Fig. 4A), 36 aa ins in SpoVG family protein (Sup. Fig. 3)
79 CSPs (Griffiths et al., 2006)
20 CSPs (Griffiths et al., 2006)
11 aa ins in Cobyrinic acid ac-diamide synthase (Fig. 4B)
Opitutaceae bacterium Tav1
Opitutaceae bacterium Tav5
Diplosphaerae colitermitum
Opitutus terrae
Coraliomargarita akajimensis
Verrucomicrobiae bac. DG1235
Methylacidiphilum infernorum
Pedosphaera parvula
Akkermansia muciniphila
Verrucomicrobium spinosum
Chthoniobacter flavus
Verrucomicrobia
Lentisphaera araneosa
Victivallis vadensis
Lentisphaerae
Parachlamydia acanthamoebae
Protochlamydia amoebophila
Waddlia chondrophila
Simkania negevensis
Chlamydophila pneumoniae
Chlamydophila pecorum
Chlamydophila felis
Chlamydophila caviae
Chlamydophila psittaci
Chlamydophila abortus
Chlamydia trachomatis
Chlamydia muridarum
Chlamydiae
Can. Kuenenia stuttgartiensis
Gemmata obscuriglobus
Isosphaera pallida
Singulisphaera acidiphila
Rhodopirellula baltica
Pirellula staleyi
Blastopirellula marina
Planctomyces limnophilus
Planctomyces brasiliensis
Planctomyces maris
Phycisphaera mikurensis
Planctomycetes

FIGURE 7 | A summary diagram depicting the different CSIs and CSPs that have been identified for the PVC clade of bacteria and the predicted evolutionary stages where the genetic changes leading for these molecular signatures likely originated. Information for various CSIs and CSPs for the Chlamydiae is from our earlier work (Griffiths et al., 2005; Gupta and Griffiths, 2006; Griffiths and Gupta, 2007).

possess a number of distinctive features such as the presence of an ammonium oxidizing organelle called the anammoxosome and cell division by constrictive binary fission, which differentiate them from other members of the class Planctomycetia (van Niftrik et al., 2009).

More importantly, in the present work, we have also identified some signatures that are helpful in clarifying how the species from the PVC phyla of bacteria are related and providing some evidence supporting their amalgamation into larger clades. However, only a couple of signatures that are helpful in this regard were identified. The most significant of these signatures is a 3 aa long insert in the RpoB protein that is commonly and uniquely shared by all of the sequenced Chlamydiae, Verrucomicrobia, and Lentisphaerae species but not found in any other bacteria. The observed species specificity of this signature, in this important protein, strongly indicates that the species from these three phyla shared a common ancestor exclusive of all other bacteria. The RpoB protein also contains a number of other CSIs in other regions of the protein that are specific for other groups/phyla of bacteria (Griffiths and Gupta, 2007; Gupta and Mok, 2007; Gao et al., 2009; Gupta and Bhandari, 2011). The high degree of specificity of these CSIs for different groups/phyla of bacteria provides evidence that the gene for RpoB has not been laterally transferred among different bacterial groups. An other signature that is informative in this regard consists of a small protein of unknown function that is specifically found in all of the species from the above three phyla of bacteria and also in the Planctomycetes. The observed species specificity of this protein suggests that the gene for this protein very likely originated in a common ancestor of the PVC clade of bacteria. However, in this case other possibilities to account for the species distribution of this protein cannot be entirely excluded. Nonetheless, the unique shared presence of this protein by various species that are part of the PVC clade provide evidence supporting their grouping into a large clade.

The molecular markers described in the present work, in addition to their usefulness for evolutionary and taxonomic studies, also provide novel and valuable tools for the identification of these organisms in different environments. In view of the presence of the identified CSIs in conserved regions of various proteins, degenerate primers based upon conserved regions in them can be designed for selective amplification (detection) of sequences from various species from these groups. Additionally, blast searches with the sequence queries based upon these proteins also provide useful identification tools for detection of both known and unknown species from these phyla in metagenomic sequences. Finally, the identified CSIs and CSP provide novel tools for genetic and biochemical studies and functional studies on them could lead to discovery of novel biochemical and/or physiochemical properties that are commonly shared by these phyla or the PVC clade of bacteria.

ACKNOWLEDGMENTS

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

REFERENCES

Agren, J., Sundstrom, A., Hafstrom, T., and Segerman, B. (2012). Gegenees: fragmented alignment of multiple genomes for determining phylogenomic distances and genetic signatures unique for specified target groups. *PLoS ONE* 7, e39107. doi:10.1371/journal.pone.0039107

Angiuoli, S. V., and Salzberg, S. L. (2011). Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* 27, 334–342.

Azuma, Y., Hirakawa, H., Yamashita, A., Cai, Y., Rahman, M. A., Suzuki, H., Mitaku, S., Toh, H., Goto, S., Murakami, T., Sugi, K., Hayashi, H., Fukushi, H., Hattori, M., Kuhara, S., and Shirai, M. (2006). Genome sequence of the cat pathogen, *Chlamydophila felis*. *DNA Res.* 13, 15–23.

Bernander, R., and Ettema, T. J. (2010). FtsZ-less cell division in archaea and bacteria. *Curr. Opin. Microbiol.* 13, 747–752.

Bertelli, C., Collyn, F., Croxatto, A., Ruckert, C., Polkinghorne, A., Kebbi-Beghdadi, C., Goesmann, A., Vaughan, L., and Greub, G. (2010). The Waddlia genome: a window into chlamydial biology. *PLoS ONE* 5, e10890. doi:10.1371/journal.pone.0010890

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

Cho, J. C., Vergin, K. L., Morris, R. M., and Giovannoni, S. J. (2004). *Lentisphaera araneosa* gen. nov., sp. nov, a transparent exopolymer producing marine bacterium, and the description of a novel bacterial phylum, Lentisphaerae. *Environ. Microbiol.* 6, 611–621.

Chouari, R., Le Paslier, D., Dauga, C., Daegelen, P., Weissenbach, J., and Sghir, A. (2005). Novel major bacterial candidate division within a municipal anaerobic sludge digester. *Appl. Environ. Microbiol.* 71, 2145–2153.

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.

Clum, A., Tindall, B. J., Sikorski, J., Ivanova, N., Mavrommatis, K., Lucas, S., Glavina, T., Del, R., Nolan, M., Chen, F., Tice, H., Pitluck, S., Cheng, J. F., Chertkov, O., Brettin, T., Han, C., Detter, J. C., Kuske, C., Bruce, D., Goodwin, L., Ovchinikova, G., Pati, A., Mikhailova, N., Chen, A., Palaniappan, K., Land, M., Hauser, L., Chang, Y. J., Jeffries, C. D., Chain, P., Rohde, M., Goker, M., Bristow, J., Eisen, J. A., Markowitz, V., Hugenholtz, P., Kyrpides, N. C., Klenk, H. P., and Lapidus, A. (2009). Complete genome sequence of *Pirellula staleyi* type strain (ATCC 27377). *Stand. Genomic. Sci.* 1, 308–316.

Collingro, A., Tischler, P., Weinmaier, T., Penz, T., Heinz, E., Brunham, R. C., Read, T. D., Bavoil, P. M., Sachse, K., Kahane, S., Friedman, M. G., Rattei, T., Myers, G. S., and Horn, M. (2011). Unity in variety–the pan-genome of the Chlamydiae. *Mol. Biol. Evol.* 28, 3253–3270.

Dalsgaard, T., Canfield, D. E., Petersen, J., Thamdrup, B., and Acuna-Gonzalez, J. (2003). N2 production by the anammox reaction in the anoxic water column of Golfo Dulce, Costa Rica. *Nature* 422, 606–608.

Devol, A. H. (2003). Nitrogen cycle: solution to a marine mystery. *Nature* 422, 575–576.

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.

Fieseler, L., Horn, M., Wagner, M., and Hentschel, U. (2004). Discovery of the novel candidate phylum "Poribacteria" in marine sponges. *Appl. Environ. Microbiol.* 70, 3724–3732.

Fox, A., Rogers, J. C., Gilbart, J., Morgan, S., Davis, C. H., Knight, S., and Wyrick, P. B. (1990). Muramic acid is not detectable in Chlamydia psittaci or Chlamydia trachomatis by gas chromatography-mass spectrometry. *Infect. Immun.* 58, 835–837.

Fuchsman, C. A., Staley, J. T., Oakley, B. B., Kirkpatrick, J. B., and Murray, J. W. (2012). Free-living and aggregate-associated planctomycetes in the Black Sea. *FEMS Microbiol. Ecol.* 80, 402–416.

Fuerst, J. A. (1995). The planctomycetes: emerging models for microbial ecology, evolution and cell biology. *Microbiology* 141(Pt 7), 1493–1506.

Fuerst, J. A., and Sagulenko, E. (2011). Beyond the bacterium: planctomycetes challenge our concepts of microbial structure and function. *Nat. Rev. Microbiol.* 9, 403–413.

Fuerst, J. A., and Webb, R. I. (1991). Membrane-bounded nucleoid in the eubacterium *Gemmata obscuriglobus*. *Proc. Natl. Acad. Sci. U.S.A.* 88, 8184–8188.

Fukunaga, Y., Kurahashi, M., Sakiyama, Y., Ohuchi, M., Yokota, A., and Harayama, S. (2009). *Phycisphaera mikurensis* gen. nov., sp. nov., isolated from a marine alga, and proposal of Phycisphaeraceae fam. nov., Phycisphaerales ord. nov. and Phycisphaerae classis nov. in the phylum Planctomycetes. *J. Gen. Appl. Microbiol.* 55, 267–275.

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.

Glockner, F. O., Kube, M., Bauer, M., Teeling, H., Lombardot, T., Ludwig, W., Gade, D., Beck, A., Borzym, K., Heitmann, K., Rabus, R., Schlesner, H., Amann, R., and Reinhardt, R. (2003). Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1. *Proc. Natl. Acad. Sci. U.S.A.* 100, 8298–8303.

Glockner, J., Kube, M., Shrestha, P. M., Weber, M., Glockner, F. O., Reinhardt, R., and Liesack, W. (2010). Phylogenetic diversity and metagenomics of candidate division OP3. *Environ. Microbiol.* 12, 1218–1229.

Goker, M., Cleland, D., Saunders, E., Lapidus, A., Nolan, M., Lucas, S., Hammon, N., Deshpande, S., Cheng, J. F., Tapia, R., Han, C., Goodwin, L., Pitluck, S., Liolios, K., Pagani, I., Ivanova, N., Mavromatis, K., Pati, A., Chen, A., Palaniappan, K., Land, M., Hauser, L., Chang, Y. J., Jeffries, C. D., Detter, J. C., Beck, B., Woyke, T., Bristow, J., Eisen, J. A., Markowitz, V., Hugenholtz, P., Kyrpides, N. C., and Klenk, H. P. (2011). Complete genome sequence of *Isosphaera pallida* type strain (IS1B). *Stand. Genomic. Sci.* 4, 63–71.

Griffiths, E., and 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.

Griffiths, E., and Gupta, R. S. (2007). 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.

Griffiths, E., Ventresca, M. S., and Gupta, R. S. (2006). BLAST screening of chlamydial genomes to identify signature proteins that are unique for the chlamydiales, chlamydiaceae, *Chlamydophila* and *Chlamydia* groups of species. *BMC Genomics* 7, 14. doi:10.1186/1471-2164-7-14

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. (2010). "Applications of conserved indels for understanding microbial phylogeny," in *Molecular Phylogeny of Microorganisms*, ed. A. Oren and R. T. Papke (Norfolk: Caister Academic Press), 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., 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 Mok, A. (2007). Phylogenomics and signature proteins for the alpha proteobacteria and its main groups. *BMC Microbiol.* 7, 106. doi:10.1186/1471-2180-7-106

Haukka, K., Kolmonen, E., Hyder, R., Hietala, J., Vakkilainen, K., Kairesalo, T., Haario, H., and Sivonen, K. (2006). Effect of nutrient loading on bacterioplankton community composition in lake mesocosms. *Microb. Ecol.* 51, 137–146.

Hedlund, B. P., Gosink, J. J., and Staley, J. T. (1996). Phylogeny of *Prosthecobacter*, the fusiform caulobacters: members of a recently discovered division of the bacteria. *Int. J. Syst. Bacteriol.* 46, 960–966.

Hedlund, B. P., Gosink, J. J., and Staley, J. T. (1997). Verrucomicrobia div. nov., a new division of the bacteria containing three new species of *Prosthecobacter. Antonie Van Leeuwenhoek* 72, 29–38.

Horn, M., Collingro, A., Schmitz-Esser, S., Beier, C. L., Purkhold, U., Fartmann, B., Brandt, P., Nyakatura, G. J., Droege, M., Frishman, D., Rattei, T., Mewes, H. W., and Wagner, M. (2004). Illuminating the evolutionary history of chlamydiae. *Science* 304, 728–730.

Hou, S., Makarova, K. S., Saw, J. H., Senin, P., Ly, B. V., Zhou, Z., Ren, Y., Wang, J., Galperin, M. Y., Omelchenko, M. V., Wolf, Y. I., Yutin, N., Koonin, E. V., Stott, M. B., Mountain, B. W., Crowe, M. A., Smirnova, A. V., Dunfield, P. F., Feng, L., Wang, L., and Alam, M. (2008). Complete genome sequence of the extremely acidophilic methanotroph isolate V4, *Methylacidiphilum infernorum*, a representative of the bacterial phylum verrucomicrobia. *Biol. Direct* 3, 26.

Hugenholtz, P., Pitulle, C., Hershberger, K. L., and Pace, N. R. (1998). Novel division level bacterial diversity in a Yellowstone hot spring. *J. Bacteriol.* 180, 366–376.

Janssen, P. H., Schuhmann, A., Morschel, E., and Rainey, F. A. (1997). Novel anaerobic ultramicrobacteria belonging to the Verrucomicrobiales lineage of bacterial descent isolated by dilution culture from anoxic rice paddy soil. *Appl. Environ. Microbiol.* 63, 1382–1388.

Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G., and Gibson, T. J. (1998). Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.* 23, 403–405.

Jenkins, C., and Fuerst, J. A. (2001). Phylogenetic analysis of evolutionary relationships of the planctomycete division of the domain bacteria based on amino acid sequences of elongation factor Tu. *J. Mol. Evol.* 52, 405–418.

Kant, R., van Passel, M. W., Palva, A., Lucas, S., Lapidus, A., Glavina, D. R., Dalin, E., Tice, H., Bruce, D., Goodwin, L., Pitluck, S., Larimer, F. W., Land, M. L., Hauser, L., Sangwan, P., de Vos, W. M., Janssen, P. H., and Smidt, H. (2011a). Genome sequence of *Chthoniobacter flavus* Ellin428, an aerobic heterotrophic soil bacterium. *J. Bacteriol.* 193, 2902–2903.

Kant, R., van Passel, M. W., Sangwan, P., Palva, A., Lucas, S., Copeland, A., Lapidus, A., Glavina, D. R., Dalin, E., Tice, H., Bruce, D., Goodwin, L., Pitluck, S., Chertkov, O., Larimer, F. W., Land, M. L., Hauser, L., Brettin, T. S., Detter, J. C., Han, S., de Vos, W. M., Janssen, P. H., and Smidt, H. (2011b). Genome sequence of "*Pedosphaera parvula*" Ellin514, an aerobic Verrucomicrobial isolate from pasture soil. *J. Bacteriol.* 193, 2900–2901.

Konig, E., Schlesner, H., and Hirsch, P. (1984). Cell wall studies on budding bacteria of the planctomycetes/pasteuria group and on a *Prosthecomicrobium* sp. *Arch. Microbiol.* 138, 200–205.

Labutti, K., Sikorski, J., Schneider, S., Nolan, M., Lucas, S., Glavina, D. R., Tice, H., Cheng, J. F., Goodwin, L., Pitluck, S., Liolios, K., Ivanova, N., Mavromatis, K., Mikhailova, N., Pati, A., Chen, A., Palaniappan, K., Land, M., Hauser, L., Chang, Y. J., Jeffries, C. D., Tindall, B. J., Rohde, M., Goker, M., Woyke, T., Bristow, J., Eisen, J. A., Markowitz, V., Hugenholtz, P., Kyrpides, N. C., Klenk, H. P., and Lapidus, A. (2010). Complete genome sequence of *Planctomyces limnophilus* type strain (Mu 290). *Stand. Genomic. Sci.* 3, 47–56.

Lee, K. C., Webb, R. I., Janssen, P. H., Sangwan, P., Romeo, T., Staley, J. T., and Fuerst, J. A. (2009). Phylum Verrucomicrobia representatives share a compartmentalized cell plan with members of bacterial phylum Planctomycetes. *BMC Microbiol.* 9, 5. doi:10.1186/1471-2180-9-5

Liesack, W., Konig, H., Schlesner, H., and Hirsch, P. (1986). Chemical compostion of the peptidoglycan-free cell envelopes of budding bacteria of the *Pirella/Planctomycetes* group. *Arch. Microbiol.* 145, 361–366.

Lindsay, M. R., Webb, R. I., and Fuerst, J. A. (1997). Pirellulosomes: a new type of membrane-bounded cell compartment in planctomycete bacteria of the genus *Pirellula. Microbiology* 143, 739–748.

Lonhienne, T. G., Sagulenko, E., Webb, R. I., Lee, K. C., Franke, J., Devos, D. P., Nouwens, A., Carroll, B. J., and Fuerst, J. A. (2010). Endocytosis-like protein uptake in the bacterium *Gemmata obscuriglobus. Proc. Natl. Acad. Sci. U.S.A.* 107, 12883–12888.

Martiny, A. C., Albrechtsen, H. J., Arvin, E., and Molin, S. (2005). Identification of bacteria in biofilm and bulk water samples from a nonchlorinated model drinking water distribution system: detection of a large nitrite-oxidizing population associated with Nitrospira spp. *Appl. Environ. Microbiol.* 71, 8611–8617.

Matsuno, K., and Sonenshein, A. L. (1999). Role of SpoVG in asymmetric septation in *Bacillus subtilis. J. Bacteriol.* 181, 3392–3401.

Mavromatis, K., Abt, B., Brambilla, E., Lapidus, A., Copeland, A., Deshpande, S., Nolan, M., Lucas, S., Tice, H., Cheng, J. F., Han, C., Detter, J. C., Woyke, T., Goodwin, L., Pitluck, S., Held, B., Brettin, T., Tapia, R., Ivanova, N., Mikhailova, N., Pati, A., Liolios, K., Chen, A., Palaniappan, K., Land, M., Hauser, L., Chang, Y. J., Jeffries, C. D., Rohde, M., Goker, M., Bristow, J., Eisen, J. A., Markowitz, V., Hugenholtz, P., Klenk, H. P., and Kyrpides, N. C. (2010). Complete genome sequence of *Coraliomargarita akajimensis* type strain (04OKA010-24). *Stand. Genomic. Sci.* 2, 290–299.

McInerney, J. O., Martin, W. F., Koonin, E. V., Allen, J. F., Galperin, M. Y., Lane, N., Archibald, J. M., and Embley, T. M. (2011). Planctomycetes and eukaryotes: a case of analogy not homology. *Bioessays* 33, 810–817.

Mojica, S., Huot, C. H., Daugherty, S., Read, T. D., Kim, T., Kaltenboeck, B., Bavoil, P., and Myers, G. S. (2011). Genome sequence of the obligate intracellular animal pathogen *Chlamydia pecorum* E58. *J. Bacteriol.* 193, 3690.

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 protein database. (2012). Available at: http://www.ncbi.nlm.nih.gov/protein. [accessed May 18, 2012].

NCBI Taxonomy. (2012). Available at: http://www.ncbi.nlm.nih.gov/taxonomy. [accessed May 18, 2012].

Pearson, A., Budin, M., and Brocks, J. J. (2003). Phylogenetic and biochemical evidence for sterol synthesis in the bacterium *Gemmata obscuriglobus. Proc. Natl. Acad. Sci. U.S.A.* 100, 15352–15357.

Petroni, G., Spring, S., Schleifer, K. H., Verni, F., and Rosati, G. (2000). Defensive extrusive ectosymbionts of Euplotidium (Ciliophora) that contain microtubule-like structures are bacteria related to Verrucomicrobia. *Proc. Natl. Acad. Sci. U.S.A.* 97, 1813–1817.

Pilhofer, M., Rappl, K., Eckl, C., Bauer, A. P., Ludwig, W., Schleifer, K. H., and Petroni, G. (2008). Characterization and evolution of cell division and cell wall synthesis genes in the bacterial phyla Verrucomicrobia, Lentisphaerae, Chlamydiae, and Planctomycetes and phylogenetic comparison with rRNA genes. *J. Bacteriol.* 190, 3192–3202.

Pol, A., Heijmans, K., Harhangi, H. R., Tedesco, D., Jetten, M. S., and Op den Camp, H. J. (2007). Methanotrophy below pH 1 by a new Verrucomicrobia species. *Nature* 450, 874–878.

Read, T. D., Brunham, R. C., Shen, C., Gill, S. R., Heidelberg, J. F., White, O., Hickey, E. K., Peterson, J., Utterback, T., Berry, K., Bass, S., Linher, K., Weidman, J., Khouri, H., Craven, B., Bowman, C., Dodson, R., Gwinn, M., Nelson, W., DeBoy, R., Kolonay, J., McClarty, G., Salzberg, S. L., Eisen, J., and Fraser, C. M. (2000). Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Res.* 28, 1397–1406.

Read, T. D., Myers, G. S., Brunham, R. C., Nelson, W. C., Paulsen, I.

T., Heidelberg, J., Holtzapple, E., Khouri, H., Federova, N. B., Carty, H. A., Umayam, L. A., Haft, D. H., Peterson, J., Beanan, M. J., White, O., Salzberg, S. L., Hsia, R. C., McClarty, G., Rank, R. G., Bavoil, P. M., and Fraser, C. M. (2003). Genome sequence of *Chlamydophila caviae* (*Chlamydia psittaci* GPIC): examining the role of niche-specific genes in the evolution of the chlamydiaceae. *Nucleic Acids Res.* 31, 2134–2147.

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

Sachse, K., Vretou, E., Livingstone, M., Borel, N., Pospischil, A., and Longbottom, D. (2009). Recent developments in the laboratory diagnosis of chlamydial infections. *Vet. Microbiol.* 135, 2–21.

Sahl, J. W., Matalka, M. N., and Rasko, D. A. (2012). Phylomark, a tool to identify conserved phylogenetic markers from whole-genome alignments. *Appl. Environ. Microbiol.* 78, 4884–4892.

Sangwan, P., Kovac, S., Davis, K. E., Sait, M., and Janssen, P. H. (2005). Detection and cultivation of soil verrucomicrobia. *Appl. Environ. Microbiol* 71, 8402–8410.

Santarella-Mellwig, R., Franke, J., Jaedicke, A., Gorjanacz, M., Bauer, U., Budd, A., Mattaj, I. W., and Devos, D. P. (2010). The compartmentalized bacteria of the planctomycetes-verrucomicrobia-chlamydiae superphylum have membrane coat-like proteins. *PLoS Biol.* 8, e1000281. doi:10.1371/journal.pbio.1000281

Schofl, G., Voigt, A., Litsche, K., Sachse, K., and Saluz, H. P. (2011). Complete genome sequences of four mammalian isolates of *Chlamydophila psittaci*. *J. Bacteriol.* 193, 4258.

Schulthess, B., Meier, S., Homerova, D., Goerke, C., Wolz, C., Kormanec, J., Berger-Bachi, B., and Bischoff, M. (2009). Functional characterization of the sigmaB-dependent yabJ-spoVG operon in Staphylococcus aureus: role in methicillin and glycopeptide resistance. *Antimicrob. Agents Chemother.* 53, 1832–1839.

Shinzato, N., Muramatsu, M., Matsui, T., and Watanabe, Y. (2005). Molecular phylogenetic diversity of the bacterial community in the gut of the termite *Coptotermes formosanus*. *Biosci. Biotechnol. Biochem.* 69, 1145–1155.

Siegl, A., Kamke, J., Hochmuth, T., Piel, J., Richter, M., Liang, C., Dandekar, T., and Hentschel, U. (2011). Single-cell genomics reveals the lifestyle of Poribacteria, a candidate phylum symbiotically associated with marine sponges. *ISME. J.* 5, 61–70.

Staley, J. T., Fuerst, J. A., Giovannoni, S., and Schlesner, H. (1992). "The order Planctomycetales and the genera *Planctomyces*, *Pirellula*, *Gemmata* and *Isosphaera*," in *The Prokaryotes*, eds A. Balows, H. G. Truper, M. Dworkin, W. Harder, and K. H. Schleifer (New York: Springer-Verlag), 3710–3731.

Strous, M., Fuerst, J. A., Kramer, E. H., Logemann, S., Muyzer, G., van de Pas-Schoonen, K. T., Webb, R., Kuenen, J. G., and Jetten, M. S. (1999). Missing lithotroph identified as new planctomycete. *Nature* 400, 446–449.

Strous, M., Pelletier, E., Mangenot, S., Rattei, T., Lehner, A., Taylor, M. W., Horn, M., Daims, H., Bartol-Mavel, D., Wincker, P., Barbe, V., Fonknechten, N., Vallenet, D., Segurens, B., Schenowitz-Truong, C., Medigue, C., Collingro, A., Snel, B., Dutilh, B. E., Op den Camp, H. J., van der, D. C., Cirpus, I., van de Pas-Schoonen, K. T., Harhangi, H. R., van Niftrik, L., Schmid, M., Keltjens, J., vand, V, Kartal, B., Meier, H., Frishman, D., Huynen, M. A., Mewes, H. W., Weissenbach, J., Jetten, M. S., Wagner, M., and Le Paslier, D. (2006). Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature* 440, 790–794.

Thomson, N. R., Holden, M. T., Carder, C., Lennard, N., Lockey, S. J., Marsh, P., Skipp, P., O'Connor, C. D., Goodhead, I., Norbertzcak, H., Harris, B., Ormond, D., Rance, R., Quail, M. A., Parkhill, J., Stephens, R. S., and Clarke, I. N. (2008). *Chlamydia trachomatis*: genome sequence analysis of lymphogranuloma venereum isolates. *Genome Res.* 18, 161–171.

Thomson, N. R., Yeats, C., Bell, K., Holden, M. T., Bentley, S. D., Livingstone, M., Cerdeno-Tarraga, A. M., Harris, B., Doggett, J., Ormond, D., Mungall, K., Clarke, K., Feltwell, T., Hance, Z., Sanders, M., Quail, M. A., Price, C., Barrell, B. G., Parkhill, J., and Longbottom, D. (2005). The *Chlamydophila abortus* genome sequence reveals an array of variable proteins that contribute to interspecies variation. *Genome Res.* 15, 629–640.

Thrash, J. C., Cho, J. C., Vergin, K. L., Morris, R. M., and Giovannoni, S. J. (2010). Genome sequence of *Lentisphaera araneosa* HTCC2155T, the type species of the order Lentisphaerales in the phylum Lentisphaerae. *J. Bacteriol.* 192, 2938–2939.

van de Graaf, A. A., Mulder, A., de Bruijn, P., Jetten, M. S., Robertson, L. A., and Kuenen, J. G. (1995). Anaerobic oxidation of ammonium is a biologically mediated process. *Appl. Environ. Microbiol* 61, 1246–1251.

van Niftrik, L., Geerts, W. J., van Donselaar, E. G., Humbel, B. M., Webb, R. I., Harhangi, H. R., Camp, H. J., Fuerst, J. A., Verkleij, A. J., Jetten, M. S., and Strous, M. (2009). Cell division ring, a new cell division protein and vertical inheritance of a bacterial organelle in anammox planctomycetes. *Mol. Microbiol* 73, 1009–1019.

van Passel, M. W., Kant, R., Palva, A., Copeland, A., Lucas, S., Lapidus, A., Glavina, D. R., Pitluck, S., Goltsman, E., Clum, A., Sun, H., Schmutz, J., Larimer, F. W., Land, M. L., Hauser, L., Kyrpides, N., Mikhailova, N., Richardson, P. P., Janssen, P. H., de Vos, W. M., and Smidt, H. (2011a). Genome sequence of the verrucomicrobium *Opitutus terrae* PB90-1, an abundant inhabitant of rice paddy soil ecosystems. *J. Bacteriol.* 193, 2367–2368.

van Passel, M. W., Kant, R., Palva, A., Lucas, S., Copeland, A., Lapidus, A., Glavina, D. R., Dalin, E., Tice, H., Bruce, D., Goodwin, L., Pitluck, S., Davenport, K. W., Sims, D., Brettin, T. S., Detter, J. C., Han, S., Larimer, F. W., Land, M. L., Hauser, L., Kyrpides, N., Ovchinnikova, G., Richardson, P. P., de Vos, W. M., Smidt, H., and Zoetendal, E. G. (2011b). Genome sequence of *Victivallis vadensis* ATCC BAA-548, an anaerobic bacterium from the phylum Lentisphaerae, isolated from the human gastrointestinal tract. *J. Bacteriol.* 193, 2373–2374.

Vandekerckhove, T. T., Coomans, A., Cornelis, K., Baert, P., and Gillis, M. (2002). Use of the Verrucomicrobia-specific probe EUB338-III and fluorescent in situ hybridization for detection of "*Candidatus Xiphinematobacter*" cells in nematode hosts. *Appl. Environ. Microbiol.* 68, 3121–3125.

Wagner, M., and Horn, M. (2006). The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. *Curr. Opin. Biotechnol.* 17, 241–249.

Wang, M., Ahrne, S., Jeppsson, B., and Molin, G. (2005). Comparison of bacterial diversity along the human intestinal tract by direct cloning and sequencing of 16S rRNA genes. *FEMS Microbiol. Ecol.* 54, 219–231.

Ward, N., Staley, J. T., Fuerst, J. A., Giovannoni, S., Schlesner, H., and Stackebrandt, E. (2006). The Order Plancktomycetales, including the genera *Planctomycetes, Pirellula, Gemmata, Isosphaera* and the Candidatus genera *Brocadia, Kuenenia* and *Scalindua*," in *The Prokaryotes*, eds M. Dworkin, S. Falkow, K. H. Schleifer, and E. Stackebrandt (New York: Springer), 757–793.

Ward, N. L. (2011). "Class I. Planctomycetia class.nov," in *Bergey's Manual of Systematic Bacteriology*, eds N. R. Krieg, J. T. Staley, D. R. Brown, B. P. Hedlund, B. J. Paster, N. L. Ward, W. Ludwig, and W. B. Whitman (New York: Springer-Verlag), 879.

Ward, N. L., Rainey, F. A., Hedlund, B. P., Staley, J. T., Ludwig, W., and Stackebrandt, E. (2000). Comparative phylogenetic analyses of members of the order Planctomycetales and the division Verrucomicrobia: 23S rRNA gene sequence analysis supports the 16S rRNA gene sequence-derived phylogeny. *Int. J. Syst. Evol. Microbiol.* 50(Pt 6), 1965–1972.

Weisburg, W. G., Hatch, T. P., and Woese, C. R. (1986). Eubacterial origin of chlamydiae. *J. Bacteriol.* 167, 570–574.

Wertz, J. T., Kim, E., Breznak, J. A., Schmidt, T. M., and Rodrigues, J. L. (2012). Genomic and physiological characterization of the Verrucomicrobia isolate *Diplosphaera colitermitum* gen. nov., sp. nov., reveals microaerophily and nitrogen fixation genes. *Appl. Environ. Microbiol.* 78, 1544–1555.

Woese, C. R. (1987). Bacterial evolution.M*icrobiol. Rev.* 51, 221–271.

Wu, M., and Eisen, J. A. (2008). A simple, fast, and accurate method of phylogenomic inference. *Genome Biol.* 9, R151.

Yoon, J., Matsuo, Y., Adachi, K., Nozawa, M., Matsuda, S., Kasai, H., and Yokota, A. (2008). Description of *Persicirhabdus sediminis* gen. nov., sp. nov., *Roseibacillus ishigakijimensis* gen. nov., sp. nov., *Roseibacillus ponti* sp. nov., *Roseibacillus persicicus* sp. nov., *Luteolibacter pohnpeiensis* gen. nov., sp. nov. and *Luteolibacter algae* sp. nov., six marine members of the phylum "Verrucomicrobia," and emended descriptions of the class Verrucomicrobiae, the order Verrucomicrobiales and the family Verrucomicrobiaceae. *Int. J. Syst. Evol. Microbiol.* 58, 998–1007.

Zoetendal, E. G., Plugge, C. M., Akkermans, A. D., and de Vos, W. M.

(2003). Victivallis vadensis gen. nov., sp. nov., a sugar-fermenting anaerobe from human faeces. *Int. J. Syst. Evol. Microbiol.* 53, 211–215.

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: 25 June 2012; accepted: 23 August 2012; published online: 17 September 2012.

*Citation: Gupta RS, Bhandari V and Naushad HS (2012) Molecular signatures for the PVC clade (Planctomycetes, Verrucomicrobia, Chlamydiae, and Lentisphaerae) of bacteria provide insights into their evolutionary relationships. Front. Microbio. **3**:327. doi: 10.3389/fmicb.2012.00327*

This article was submitted to Frontiers in Evolutionary and Genomic Microbiology, a specialty of Frontiers in Microbiology.

APPENDIX

			11		47
Opitutaceae	Opitutus terrae	182416332	LSEREPELGLGIVASVD	RA	ARRLGIDFPATGEKRLYA
	Opitutaceae bacterium TAV1	374591695	M------------V---	--	QS-I--G------R----
	Opitutaceae bacterium TAV5	373853298	M------------V---	--	QS-I--G------R----
	Diplosphaera colitermitum	225165936	M----------L-T---	--	QG-I-VGY-----S----
Other Verrucomicrobia	Coraliomargarita akajimensis	294055537	I--T--------LE---		RLQ-RLI---SS-A-I—
	Pedosphaera parvula	223935786	M--S------AT-VHTG		EG-VQVL-S----T-T--
	Verrucomicrobium spinosum	171911735	V-DT--------ILKAS		FG--EVL---AA-T-Q--
	Chthoniobacter flavus	196233912	V-DT--------ILKAE		YG-VEVF---AM-Q-Q--
	Coraliomargarita akajimensis	294055537	I--T--------LE---		RLQ-RLI---SS-A-I--
Other Species	Candidatus Hamiltonella	238899318	V-DT-T-----TITEI-		L-MVTLA--------R--
	Legionella pneumophila	296107886	I-NT-SQ-----ITE--		G-QVS-S---A--D-I--
	marine gamma proteobacterium	119503297	V-HADV-------VEL-		G--ITLH---V--E-T--
	Legionella drancourtii	374261199	I-NT-SQ-----ITELS		G-QVH-S---A--E-I--
	Allochromatium vinosum	288941218	---TQT-----L-EA--		G-HVR-AY-----T----
	marine gamma proteobacterium	119476990	V-NT-S-------LE-I		G--VE-S---AA-R-T--
	Haemophilus pittmaniae	343518823	---T-NN------T-IE		G--VQL----AQ-N-I--
	Idiomarina baltica	85711786	---T-TD----V-TA--		G-QVSVM--S---T-M--
	Cellvibrio japonicus	192361560	I--A-VD------LD-A		N---TLS---A--R-T--
	Haemophilus parasuis	167855864	I--S-NN------TA--		N-TVTLS---SD-Q-V--
	Thiocystis violascens	350555366	---TQG-----L-EA-E		G-HVRMAY-----T----
	Salmonella enterica	213581710	I-DT-S-----T-VAM-		--TVTLL--S---N----
	Hahella chejuensis	83647726	I-NT--------ITEN-		N--IT-S---AT-Q-T-S
	Methylococcus capsulatus	53803369	I-DT-------T-VG-E		HD-VTVS-I-S--R-I--
	Enterobacter hormaechei	334123487	I-DT-S-----T-VA--		--MVTLL------N----
	Moritella sp. PE36	149907517	I-DT-S-----TIVAIE		G-M-TLL---S--N----
	Providencia rettgeri	268590572	I-DT-S-----A-VAI-		--MVTLL---S--N---S
	Mannheimia haemolytica	261493688	---S-NN------T---		H-TVT-N---AE-E-V--
	Xenorhabdus nematophila	300724845	I-DT-S-----TIVAL-		--MVTLL---S--N----
	Serratia sp. M24T3	383298216	I-DT-S-----T-VA--		T-MVTLL------N----
	Pectobacterium wasabiae	261823055	I-DT-S-----T-VA--		T-MITLL---S--N---S
	Escherichia coli	323975774	I-DT-S-----T-VA--		--TVTLL--S---N----
	Citrobacter koseri	157147523	I-DT-S-----T-VAM-		--TVTLL------N----
	Salmonella enterica	194446131	I-DT-S-----T-VAM-		--TVTLL--S---N----
	Legionella longbeachae	270158478	I-NT-SK-----IID-S		G-QVSVS---AE-E-I-S
	Salmonella enterica	16763486	I-DT-S-----T-VAM-		--TVTLL--S---N----
	Actinobacillus succinogenes	152979645	I--S-NN-----ITE-N		S-AVT-F---AD-T-I--
	Yersinia mollaretii	238797482	I-DT-S-----T-VAI-		V-MITLL------N----

FIGURE A1 | Partial sequence alignment of the helicase domain-containing protein showing 2 aa insert that is specific for the family Opitutaceae. The insert in not present in other Verrucomicrobia or in any other group of bacteria.

```
                                                            674                           703
                Opitutaceae bacterium TAV5     373853807    GGKYTDDWHYNHMRDPR   MSPGSNMPAYPWL
                Opitutaceae bacterium TAV1     390119056    -----------------   -------------
                Diplosphaera colitermitum      225155868    ---------F-------   -------------
Verrucomicrobia Verrucomicrobiae bac. DG1235   254444747    ----S-S---D--L--- Q T-------V----
                Verrucomicrobium spinosum      171910295    ----PNV------K--- A V-------N-A-M
                Chthoniobacter flavus          196231228    ----PSI--FH------ Q I----I--N----
                Coraliomargarita akajimensis   294054994    --LRS-----Y--LN-- D V-------N----
                Opitutus terrae                182415885    ----PNI--IR--A--- S I-A--I--N----
                Marivirga tractuosa            313675717    ----P-S--F---Y--- S -----T--P----
                Leadbetterella byssophila      312131435    ----P-S------F--T S -----I--S----
                Niabella soli                  374374046    ----P-S--F---L--- S -A---V--S----
                Chitinophaga pinensis          256419727    ----PHS------L--T S -----I--Q----
                Solitalea canadensis           379653421    ----P-S------M--S S -A---I--Q--S-
                Runella slithyformis           338214615    -A--P-S------E--T S -----I--K----
                Halomonas elongata             307546127    --R-S-N--RA-LYN-- D VV---V-------
                Alishewanella jeotgali         375108644    --R-S-----A-LM--- S VV-Q---------
                Idiomarina baltica             85712907     --R-S-----V-LMN-- N VV-E-----F---
                Marinobacter aquaeolei         120554703    --R-S-A-QRQ-LY--- S VV-E-----F---
                Thiorhodospira sibirica        350553069    --R-S-E--RL-LI--- S VV-E---------
Other Bacteria  Photobacterium damselae        269102672    --R-S-E---V-LM--- A VV-E-----F---
                Vibrio parahaemolyticus        28898317     --R-S-E--RV-LL--- E LV-E----GF---
                Saccharophagus degradans       90022066     -QR-S-T--KA-LYN-- N VV-E-----F---
                Pseudomonas fulva              333901048    --R-S-E--RA-LYN-- N VV-E-K--S----
                Alteromonas macleodii          332141481    --R-S-E--RV-LLN-- N VV-E----GF---
                Colwellia psychrerythraea      71281529     --R-S----IA-LT--- S VV-E------S--
                Lutiella nitroferrum           224824483    --R-S-E--RV-LTN-- D VV-E-----F---
                Chromobacterium violaceum      34496628     --R-S-E--RV-LNN-- D VV-E-----F---
                Ralstonia pickettii            241662802    -QR-S----RI-L---- E VV-E------A--
                Lautropia mirabilis            319943106    --R-S----RA-LHN-- D VV-E---------
                Methylibium petroleiphilum     124267663    --R-S-E--RL-LAN-- D LV-E---------
                Hydra magnipapillata           221124432    ----S-E--RI-LTN-- D VV-E---------
```

FIGURE A2 | A 1 aa deletion in a conserved region in the Cytochrome c oxidase protein is shown in this partial sequence alignment with the deletion specific for *Opitutaceae bacterium Tav1, Opitutaceae bacterium Tav5,* and *Diplosphaera colitermitum* species. The three species harboring the indel also branch together in the concatenated protein tree.

```
               SpoVG family protein
                all except phyci (which does not have a matching sequence)

                                                         34                                                                      106
               Pirellula staleyi              283777933  DLKIIEGSSGPFVAMPSRK LTSHCHQCGSKNHLKAGYCNHCGARQREDRLVRDQD GRAKLYADIAHPINSACR
               Blastopirellula marina         87310105   -------A----------- --A--P---G----R-----N--F-L-LPPAE-TA- ---------------E--
               Isosphaera pallida             320105243  -------AK-F-------- --DR--H-------RSRF--N----LD-N-AA--P- -----H------------
               Singulisphaera acidiphila      373481189  -------AK-F-------- --DR--H--T----RSRF--Q--S-LD-N-AI--A- -----H---------M--
               Planctomyces brasiliensis      325108722  -----H-AK-A-------- --DR-PK-H-----R-TF--Q--V-LHSE-ASK-D- ---------------E--
Planctomycetes Gemmata obscuriglobus          168703400  -------TK-I-------- --DR-GR--G----RSRF--Q--T-LDDQ-AM-AV- -----H-------H-GA-
               Planctomyces maris             149176657  -----Q-AK-A-------- -MDR-PK-HT----R-SF--Q--I-LD-N-ADK-DA ---R-----------E--
               Planctomyces limnophilus       296123014  -----Q-TR-S-------- -MDR-PR-SC----R-RF--D--CELH-E-ANKAD- ---------------E--
               Can. Kuenenia stuttgartiensis  91202798   ---V---HK-A-------- --DR-PG--G----MSQ---D--T-LD-K-ASKGA  --L--H--T------K--
               Planctomycete KSU-1            386814238  ---V---HK-A-------- --DR-PK--G----M-QH--D--SKLD-K-ASKGA  --L--H--T------K--
               Rhodopirellula baltica         32476479   -----D-T----------- --G--GR-S-----R-T-------KLSGQNAN     SPQ-----V------E--
               Acetivibrio cellulolyticus     366163466  -I----SQN-L-I------                                      APDGEFR-------AET-
               Alkaliphilus oremlandii        158321667  -I-----QN-L-I------                                      MGEGDFR--------ST-
               Blautia hansenii               260589035  -F-V---EK-L-I------                                      ATDGE-R---------T-
               Clostridium botulinum          253681291  -I-V---QN-L-I------                                      TPTGEFK-------TTT-
               Coprococcus eutactus           163815681  -I-----EK-M-I------                                      ASDGE-R-------T-T-
               Desulfitobacterium hafniense   89892897   -V-VV--TN-L--------                                      TPEGEFR------S--A-
               Desulfotomaculum acetoxidans   258513558  -V-VV--QT-L--------                                      TPNGEFR--------SA-
               Dethiobacter alkaliphilus      225181555  -VRV---NN-L------KR                                      TPDGEFK------T-ET-
               Dorea longicatena              153854759  -I-V---EK-L-I----K-                                      ALDGE-R--------GT-
               Eubacterium rectale            238922865  -I-V---EK-L-I------                                      ANDGE-R---------T-
               Heliobacterium modesticaldum   167629337  -V-VV--QK-L-------R                                      TPEGE-R------SAKA-
               Oribacterium sinus             227872713  -I-V---EK-L-I------                                      TTDGE-R------R-TT-
               Peptoniphilus duerdenii        304440542  -I-V-Q-D-SL-I------                                      LSNGEFR-------QEA-
               Ruminococcus gnavus            154503757  -I-V---EK-L-I----K-                                      ALDGE-R--------GT-
               Syntrophobotulus glycolicus    325288384  -V-VV--TN-L--------                                      TPEGDFR------S--A-
Other species  Thermoanaerobacter italicus    289579387  -I-V---QD-L-I------                                      TPGGEFK--------DT-
               Abiotrophia defectiva          229825977  -I---D-DK-L-I------                                      TNDGE-H--------ET-
               Anoxybacillus flavithermus     212637891  -IRV---NN-L------KR                                      TPDGEFR--------TT-
               Bacillus coahuilensis          205372000  -IRV-D-NN-L------KR                                      TPDGEFR--------TT-
               Geobacillus kaustophilus       56418577   -IRV-D-NN-L------KR                                      TPDGEFR---------T-
               Staphylococcus aureus          377747269  --RV---N--L------KR                                      TPDGEFR--------DM-
               Bdellovibrio bacteriovorus     42524211   ---V-Q-T--L------K-                                      RKDGQFR-----L-QET-
               Desulfarculus baarsii          302342163  -I-V-H-NK-L------K-                                      RKDGS-Q-----L--ET-
               Hippea maritima                327399269  -----S-QK-L--------                                      MKDGSFK-V---L-NEM-
               Corallococcus coralloides      383454413  ---V-H--T-L-I---AK-                                      RKDGT-K-----L-ADT-
               Myxococcus xanthus             108757875  ---V-H--T-L-I---AK-                                      RKDGT-K-----L-ADT-
               Stigmatella aurantiaca         310823074  ---V-H-A--L-I---AK-                                      RKDGT-K-----L-ADT-
               Spirochaeta smaragdinae        302338327  NV---D-KN-A-I------                                      T-SGE-K-V------DF-
               Sphaerochaeta pleomorpha       374317349  NI-----KE-D-I-----Q                                      LANGEFK-V----S-EF-
               Treponema succinifaciens       328948676  NV--------L-I------                                      TANGE-K-V----SPDF-
               Fusobacterium gonidiaformans   315917773  G--L---E--K-I------                                      MPDGEFK--V---SPEL-
               Onion yellows phytoplasma      39939218   -IR----ER-I-I------                                      TSKGNFR--------ET-
```

FIGURE A3 | A large, 32–36 aa insert present in all detected species of the Planctomycetes species is presented. The conserved region is present within a conserved region of the SpoVG family protein and is not found in any organism outside of the Planctomycetes phylum.

CHAPTER 5

Protein based molecular markers provide reliable means to understand prokaryotic phylogeny and support Darwinian mode of evolution[1]

The following chapter is a review of comparative genomic analysis 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.

[1] The citation for the manuscript is:

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 ± 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.

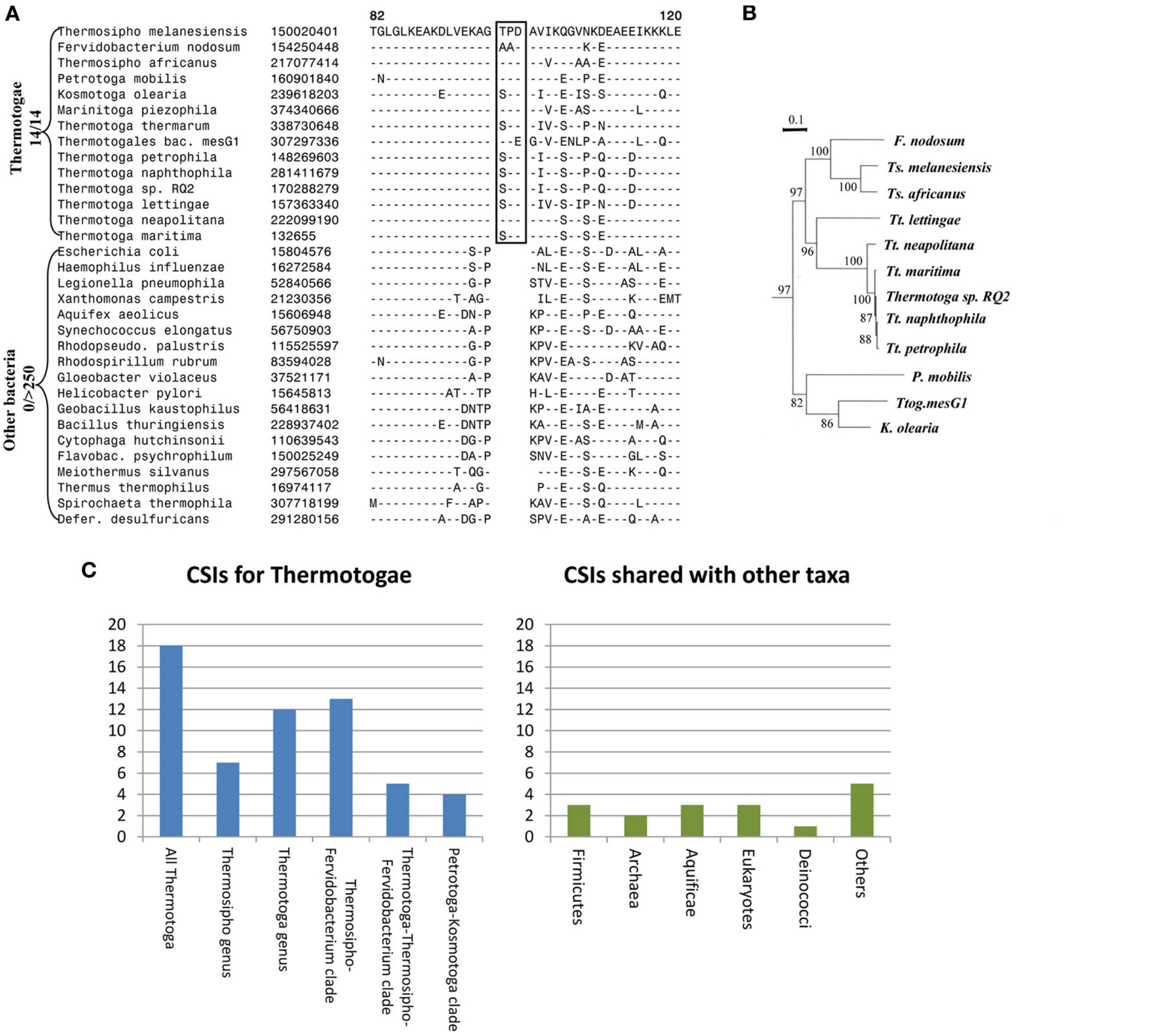


FIGURE 1 | Evolutionary relationships among Thermotogae species based upon CSIs and a Phylogenetic Tree. (A) Partial sequence alignment for the ribosomal protein L7 showing a 3 aa CSI (boxed) that is specific for all detected species of the Thermotogae phylum. The dashes in the alignment (–) indicate 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

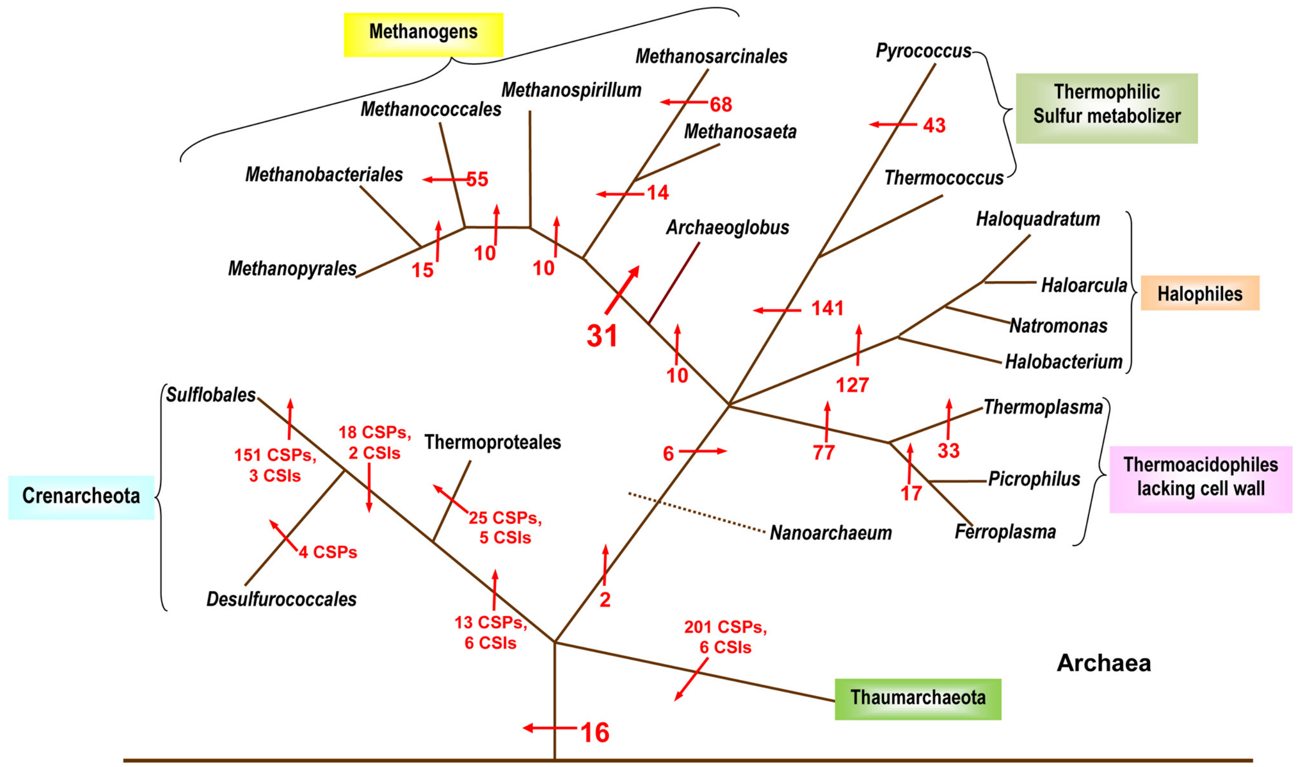


FIGURE 2 | A summary diagram showing the various molecular markers that have been identified for the Archaeal kingdom and its subgroups. The arrows indicate the suggested evolutionary stages where the proteins unique for a particular taxa are proposed to have been introduced. The numbers beside the arrows 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 # **Protein length**	**NP_147640** **262 aa**	**NP_147284** **143 aa**	**BAA81469** **98 aa**	**NP_147588** **228 aa**	**YP_001041009** **127 aa**	**YP_254810** **228 aa**	**YP_254922** **270 aa**	**NP_559041** **626 aa**	**NP_559897** **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 caldifontis	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

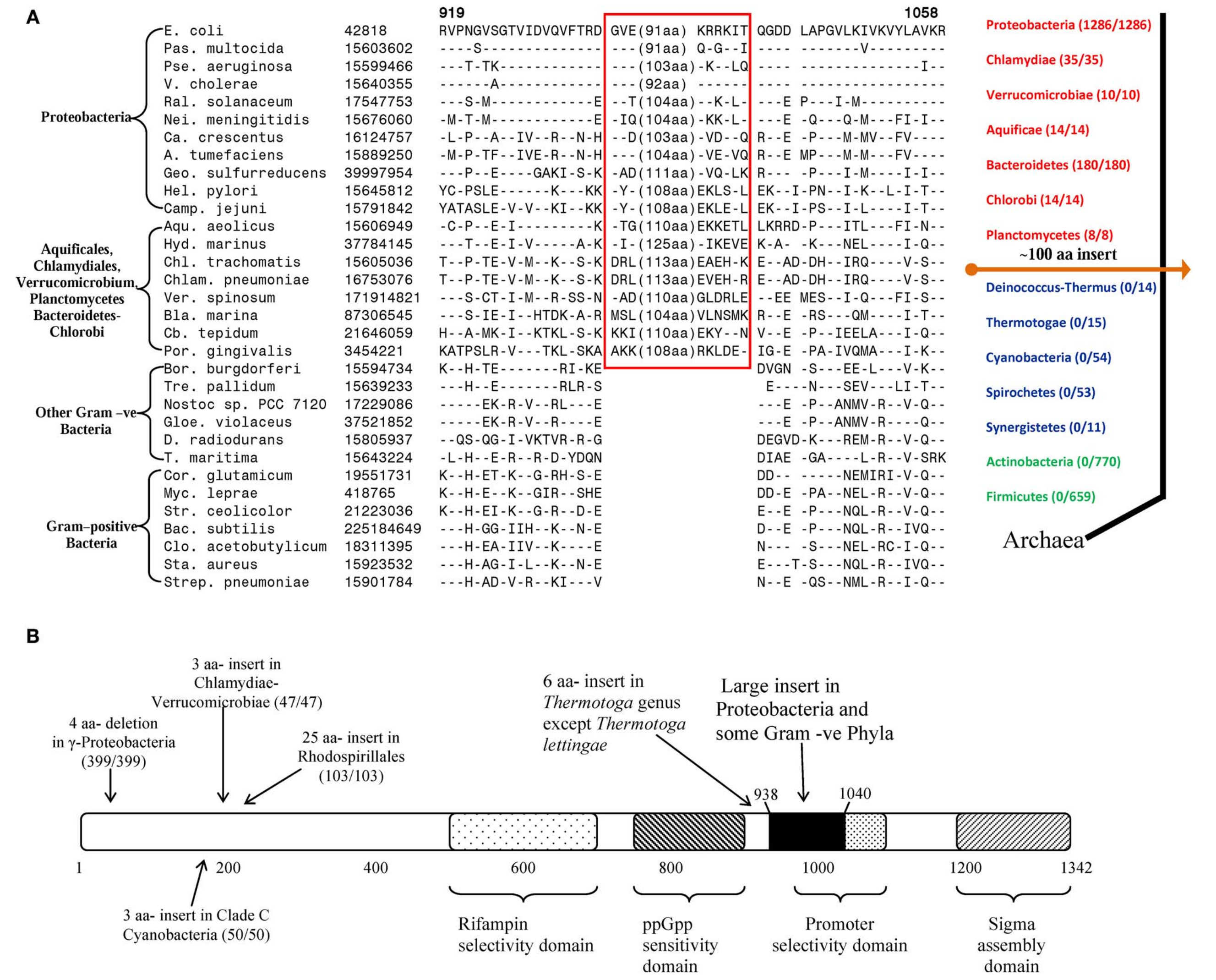

A
919
1058
Proteobacteria
E. coli 42818 RVPNGVSGTVIDVQVFTRD GVE(91aa) KRRKIT QGDD LAPGVLKIVKVYLAVKR
Pas. multocida 15603602 ----S-------------- ---(91aa) Q-G--I ---- -------V---------
Pse. aeruginosa 15599466 ---T-TK------------ ---(103aa)-K--LQ ---- ---------------I--
V. cholerae 15640355 ------A------------ ---(92aa) ------ ---- -----------------
Ral. solanaceum 17547753 ---S-M------------E --T(104aa)--K-L- ---E P---I-M---------
Nei. meningitidis 15676060 -M-T-M------------E -IQ(104aa)-KK-L- ---E -Q---Q-M---FI-I--
Ca. crescentus 16124757 -L-P--A--IV--R--N-H --D(103aa)-VD--Q R--E -P---M-MV--FV----
A. tumefaciens 15889250 -M-P-TF--IVE-R--N-H ---(104aa)-VE-VQ R--E MP---M-M---FV----
Geo. sulfurreducens 39997954 ---P--E----GAKI-S-K -AD(111aa)-VQ-LK R--- -P---I-M----I-I--
Hel. pylori 15645812 YC-PSLE------K---KK -Y-(108aa)EKLS-L EK--I-PN--I-K--L-I-T--
Camp. jejuni 15791842 YATASLE-V-V--KI--KK -Y-(108aa)EKLE-L EK--I-PS--I-L----I-T--
Aquificales, Chlamydiales, Verrucomicrobium, Planctomycetes Bacteroidetes-Chlorobi
Aqu. aeolicus 15606949 -C-P--E-I---------K -TG(110aa)EKKETL LKRRD-P---ITL---FI-N--
Hyd. marinus 37784145 ---T--E-I-V-----A-K -I-(125aa)-IKEVE K-A- -K---NEL----I-Q--
Chl. trachomatis 15605036 T--P-TE-V-M--K--S-K DRL(113aa)EAEH-K E--AD-DH--IRQ----V-S--
Chlam. pneumoniae 16753076 T--P-TE-V-M--K--S-K DRL(113aa)EVEH-R E--AD-DH--IRQ----V-S--
Ver. spinosum 171914821 ---S-CT-I-M--R-SS-N -AD(110aa)GLDRLE --EE MES--I-Q---FI-S--
Bla. marina 87306545 ---S-IE-I--HTDK-A-R MSL(104aa)VLNSMK R--E -RS---QM----I-T--
Cb. tepidum 21646059 H--A-MK-I--KTKL-S-K KKI(110aa)EKY--N V--E -P--IEELA---I-Q--
Por. gingivalis 3454221 KATPSLR-V---TKL-SKA AKK(108aa)RKLDE- IG-E -PA-IVQMA---I-K--
Other Gram –ve Bacteria
Bor. burgdorferi 15594734 K--H-TE-------RI-KE DVGN -S---EE-L---V-K--
Tre. pallidum 15639233 ---H--E-------RLR-S E----N---SEV---LI-T--
Nostoc sp. PCC 7120 17229086 -----EK-R-V--RL---E ---E -P--ANMV-R--V-Q--
Gloe. violaceus 37521852 -----EK-R-V--R----E ---E -P--ANMV-R----Q--
D. radiodurans 15805937 --QS-QG-I-VKTVR-R-G DEGVD-K---REM-R--V-Q--
T. maritima 15643224 -L-H--E-R--R-D-YDQN DIAE -GA----L-R--V-SRK
Gram–positive Bacteria
Cor. glutamicum 19551731 K--H-ET-K--G-RH-S-E DD-- -----NEMIRI-V-Q--
Myc. leprae 418765 K--H-E--K--GIR--SHE DD-E -PA--NEL-R--V-Q--
Str. ceolicolor 21223036 K--H-EI-K--G-R--D-E E--E -P---NQL-R--V-Q--
Bac. subtilis 225184649 ---H-GG-IIH--K--N-E D--E -P---NQL-R--IVQ--
Clo. acetobutylicum 18311395 ---H-EA-IIV--K----E N--- -S---NEL-RC-I-Q--
Sta. aureus 15923532 ---H-AG-I-L--K--N-E E---T-S---NQL-R--IVQ--
Strep. pneumoniae 15901784 ---H-AD-V-R--KI---V N--E -QS--NML-R--I-Q--
Proteobacteria (1286/1286)
Chlamydiae (35/35)
Verrucomicrobiae (10/10)
Aquificae (14/14)
Bacteroidetes (180/180)
Chlorobi (14/14)
Planctomycetes (8/8)
~100 aa insert
Deinococcus-Thermus (0/14)
Thermotogae (0/15)
Cyanobacteria (0/54)
Spirochetes (0/53)
Synergistetes (0/11)
Actinobacteria (0/770)
Firmicutes (0/659)
Archaea
B
3 aa- insert in Chlamydiae-Verrucomicrobiae (47/47)
4 aa- deletion in γ-Proteobacteria (399/399)
25 aa- insert in Rhodospirillales (103/103)
6 aa- insert in Thermotoga genus except Thermotoga lettingae
Large insert in Proteobacteria and some Gram -ve Phyla
938
1040
1
200
400
600
800
1000
1200
1342
3 aa- insert in Clade C Cyanobacteria (50/50)
Rifampin selectivity domain
ppGpp sensitivity domain
Promoter selectivity domain
Sigma assembly domain

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 photosynthesis 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 6

Conclusion

Shortcomings of current methods used in prokaryotic phylogeny and taxonomy

The goal of evolutionary taxonomy has been to provide a classification system for defining bacterial and archaeal relationships reflective of their evolution (Wayne et al., 1987; Woese et al., 1990). Though intended for this purpose, criteria currently used for defining the different taxonomic groupings have so far been quite subjective and sometimes unsatisfactory (Green and Bohannan, 2006). For example, in species classification, 16S rRNA similarity values of 97% have been utilized to suggest that two organisms belong to the same species group (Gevers et al., 2005; Green and Bohannan, 2006; Wayne et al., 1987). Though 16S rRNA is useful for quick classification of species, due to the highly conserved nature of the molecule these values have been unsuitable in multiple occasions (Fox et al., 1992; Martinez-Murcia et al., 1992; Rossello-Mora and Amann, 2001). Similarly, definition of a species group sharing 70% DNA hybridization is an arbitrary value not necessarily reflective of species relationships. Unlike lower taxonomic divisions of *species* and *genus*, definitions of higher ranks have not even these minimal definitive criteria (Oren, 2010b). Thus, the hierarchical divisions are sometimes based on Ad Hoc factors. Consequently, the taxonomic classification has become a system where different groups are defined by different criteria and taxonomic levels do not signify a consistent evolutionary relationship. Due to a lack of consistent definitions for taxonomic divisions, it is said that systematics is utilitarian and does not consistently provide a classification system based on evolutionary concepts (Schleifer, 2009).

A further problem in taxonomy is inaccurate placement of organisms. Most prokaryotic phyla were described years ago with the use of phylogenetic criteria using limited datasets (Ludwig, 2010). Monophyletic clusters separated by relatively long branches were afforded the distinction of being classified as phylum level groupings (Ludwig, 2010). Though this was satisfactory at the time, identification of more organisms has crowded the previously sparse phylogenetic trees and shown phylogenetically made divisions to be inaccurate in some cases (Ludwig et al., 2009). Additionally, inferences derived through use of phylogenetic analyses are largely dependent on the methodology used (Zhi et al., 2012). Many variables can influence the outcome of the trees, including: the use of species for the construction of a tree, the use of outgroups, the use of tree-making algorithms and sequence alignment procedures (Zhi et al., 2012). Thus, groupings of organisms are sometimes unstable and inconsistent due to the various variables involved in tree-making. Perhaps more importantly, even when successfully utilized to its utmost capabilities, phylogenetic and sequence similarity methods fail to provide detectable characteristics to differentiate organisms from each other.

Without universally recognized standards for higher-level classifications and a lack of overtly characteristic markers, grouping together of species into taxonomic divisions higher than genus remains difficult. This is especially the case when species often share the same morphological features with few known physiological or biochemical differences for hierarchical differentiation. An option is to define a high-level taxon based on phylogeny while the other, more conservative option, is to place the species in the lower ranking divisions till characteristics are discovered in the future to reclassify or confirm the placement of the species. Thus, with few reliable characteristics able to differentiate prokaryotes, it is likely that the taxonomic relationships of many species do not reflect their evolutionary relationships.

Use of comparative genomics and CSIs for description of prokaryotic groups

Though phylogenetic analyses depicted some genera to have stronger phylogenetic links than others, no physical markers had been known that could easily differentiate the various groups of the Synergistetes and the Thermotogae phyla. In both phyla, due to lack of such physical evidence, the conglomeration of closely related genera into higher level-groups was avoided. Rather, based on phylogenetic evidence, vastly variant species were all placed into single family-level grouping. The use of synapomorphies had been proposed to improve the level of prokaryotic phylogeny and taxonomic qualifications. In the preceding chapters, utilization of comparative genomics for the identification of taxa defining molecular markers has been depicted.

With the description of CSIs shared by various organisms from the Thermotogae and Synergistetes, molecular markers have for the first time been identified for each of these phyla. Specifically, 18 CSIs were identified to be characteristic of the Thermotogae while 13 CSIs specific for the Synergistetes were presented. These CSIs specific for the phyla link the species of the group together and help differentiate them from other bacterial species. Additionally, many CSIs were identified for previously defined and undefined subdivisions for the two phyla.

Among the Thermotogae, the genera *Thermosipho* and the *Thermotoga* were identified by 7 and 12 CSIs, respectively. Molecular evidence was provided for the reclassification of the species *Thermotoga lettingae* into a different genus. Also, CSIs depicting a relationship between the genera *Fervidobacterium* and *Thermosipho* and CSIs supporting the clade comprised of both of these genera with the genus *Thermotoga* were identified. Though unofficial groupings of several groups have been indicated in chapter 2, official proposals based on these results are in progress. These include a proposal for a new family Fervidobacteriaceae comprising the genera *Fervidobacterium* and *Thermosipho*; a redefinition of the family Thermotogaceae to be comprised of the current *Thermotoga* genus (which is itself to be split into two genera); and a redefining of the order Thermotogales to consist of the Fervidobacteriaceae and Thermotogaceae (Bhandari and Gupta, 2013).

Similarly, CSIs specific for the Synergistetes phyla were found with inter-phylum relationships at different tiers observed among the seven genera for which representative species had their genomic sequences available. The following cladal relationships were supported by numerous CSIs: *Pyramidobacter-Jonquetella*, *Pyramidobacter-Jonquetella-Dethiosulfovibrio*, and *Aminomonas-Thermanaerovibrio*. Consistent with phylogenetic trees, these relationships were for the first time independently (of phylogenetic analyses) verified by multiple molecular markers. As the results in chapter 3 demonstrate, multiple layers of relationships were shared among these species. Using these CSIs may perhaps lead to a taxonomic re-arrangement of its species to more accurately represent the evolutionary relationships among the Synergistetes.

In addition to the Thermotogae and Synergistetes, comparative genomics was also utilized for confirmation of the shared relationships among species of the PVC group of phyla. Among the Verrucomicrobiae, 3 CSIs were identified to support the positioning of the unclassified species *Opitutaceae bacterium* Tav1 and *Opitutaceae bacterium* Tav5 into the genus *Diplosphaerae*. Among the Planctomycetaceae, 2 CSIs were identified to support the reclassification of the anammox bacterium *Candidatus Kuenenia Stuttgartiensis* into a distinct class from its current placement in the class Planctomycetia. More importantly, molecular markers were identified for the first time to tie the phyla Planctomycetes, Verrucomicrobia, Lentisphaerae, and Chlamydiae together as a group with shared ancestry. A conserved signature protein (protein CT421.2) shared by the Planctomycetes, Lentisphaerae, Verrucomicrobia and Chlamydiae was the first molecular marker to link the multiple phyla of the PVC group. A 3 aa insert in the RpoB protein was also discovered to be shared by the Lentisphaerae, Verrucomicrobia and the Chlamydiae. This was consistent with the phylogenetic observation indicating the Planctomycetes to be a deep-branching member of the group. Therefore, the identified CSI and CSP support a cladal relationship among multiple phyla by independent means from phylogenetic analyses, suggesting a shared common ancestor for the so called superphylum. Also, these molecular markers were not found in the Poribacteria species, a group previously included as a member of the superphylum. Genomic sequences for candidate divisions WWE2 and OP3 were unavailable. Due to the specificity of the CSIs and CSPs identified in the analyses, the inclusion of the two candidate divisions into the PVC superphylum may be molecularly determined once the corresponding sequences from species of the two groups become available.

Lateral gene transfer

Some genomic studies have raised alarm over the incidence of LGT events in prokaryotes and their effects in masking of an accurate phylogenetic signal. Though popular, the view on the prevalence of LGT and its influence on prokaryotic evolution is far from the consensus opinion. Throughout chapters 2, 3, 4 and 5, evidence for presence of hierarchical relationships among bacteria and archaea has been provided. Specifically, in chapter 2, it was shown that among 75 CSIs identified, only 16 were shared by Thermotogae and other groups. As CSIs are rare genetic changes, the shared CSIs are indicative of either convergent characteristics or LGT. Considering all such cases to be CSIs, they would represent only ~20% of the total molecular markers with all others being specific for the Thermotogae phylum or its sub-groups. Despite the presence of these shared CSIs, various groups among the Thermotogae were identified based on almost 60 CSIs. If the theory of rampant LGT were true, such clarity and distinction would not have been observed. Additionally, among the 16 shared CSIs, no group of species outside the Thermotogae shared a majority of these CSIs and even when sharing the CSIs, not all members of the particular taxa were observed to contain the indel. Further, recent analyses on the Xanthomonadales genome suggests that even the shared indels among unrelated groups may likely be of independent origin rather than LGT (Naushad and Gupta, 2013).

Further, as highlighted in chapter 5, over the past decade, numerous analyses have also identified molecular markers in the form of CSIs and CSPs for a variety of different prokaryotic taxa. These molecular markers have assisted in depicting prokaryotic relationships at various taxonomic and phylogenetic depths independently from morphological, physiological or phylogenetic means. Specifically, the Archaea along with the Thermotogae are used as sample cases depicting the success of identification of species relationships for these groups based on CSIs and CSPs. A hierarchical relationship structure observed through molecular markers is consistent with the common tree-like pattern observed in phylogenetic trees. The consistencies of these observations are used to support the continuation of the current species-classification system. Though LGT is an ongoing mechanism which has profound effects on prokaryotic life, we suggest that there remains the ability for us to detect and discern species relationship.

Applicability of CSIs

In the preceding chapters, CSIs have been successfully identified that help to define the Thermotogae, Synergistetes, the PVC superphylum and their sub-groups. It is surmised that amino acid insertions and deletions in well-conserved regions of the protein are rare genetic changes unlikely to occur independently in multiple organisms (Gupta, 1998). The parsimonious assumption follows that such genetic changes, due to their rare occurrence, are molecular markers passed down to the progeny of the organism that first harboured the indel (Gupta, 1998). The conservation of amino acid residues surrounding the CSI have been inferred to be retained due to possible functional importance within the proteins that they may be present in (Akiva et al., 2008; Hormozdiari et al., 2009; Itzhaki et al., 2006; Singh and Gupta, 2009). Following this, the presented CSIs provide exciting candidates for work into a variety of diagnostic and functional analyses of prokaryotes.

Primarily, conserved signature indels provide a means to molecularly define groups of species (Gupta, 1998). Thus, along with physiochemical, biochemical and 16S rRNA sequence data, these markers should be incorporated into analyses on prokaryotic relationships and taxonomy rather than the basing prokaryotic relationships on phylogeny or limited morphological/physiological data alone. Among the advantages of CSIs over 16S rRNA and DNA-DNA hybridization is that, once known, CSIs are discernible molecular markers that can be easily identified in an organism without having to make direct comparisons. Consequently, another usage for these characters would be for molecular diagnosis of species in metagenomic data or in clinical and environmental samples. Using CSIs, members of the relevant taxonomic group could be detected in sequence data produced by metagenomic analyses. Similarly, degenerate PCR primers could be designed, based on the conserved region surrounding a group specific CSI, for the identification of the presence of a particular organism in an environmental or clinical sample.

Another major difference between gene comparison analyses and CSIs is that the CSIs present in conserved regions are suggested to also be functionally important elements of the protein (Singh and Gupta, 2009). With little analyses focused on the area, the functional roles of CSIs are relatively unknown. Structural analyses indicate indels, conserved or otherwise, to be mostly present in solute accessible, unstructured regions of the proteins such as loops (Akiva et al., 2008). Within these loops, the CSIs have been reasoned to assist in protein-protein interactions (Itzhaki et al., 2006; Akiva et al., 2008; Hormozdiari et al., 2009). Though the exact roles of CSIs within the protein and the organism remain unknown, functional analyses on CSIs may provide clues to novel biochemical or physiological characteristics of the organisms that they are present in.

Conclusion

Immense genomic data, comprising over 5100 bacterial and almost 200 archaeal species, is now publicly available (Markowitz et al., 2012). Among the many methodologies it has begotten for phylogenetics, CSIs (and CSPs) allow for the identification of rare genomic changes specific for various prokaryotic taxa. Gene based phylogenetic analyses, gene similarity or genome similarity based analyses that define a species and other taxa on degree relatedness. Compared to such analyses, CSIs can suggest upon relationships among organisms with less ambiguity as they divide species into two groups: those with the CSI and those without. Thus CSIs, along with the similar CSPs, act as evolutionary milestones marking prokaryotic branch points during the course of evolution (Gupta, 1998). Basing taxonomic divisions on molecular markers, CSIs provide a means to organize taxonomy on shared derived characters (Hennig, 1966; Williams et al., 2010). Further, they provide tools for diagnostic use and for possible biochemical insight into prokaryotic groups.

Reference List

Akiba,T., Koyama,K., Ishiki,Y., Kimura,S., and Fukushima,T. (1960). On the mechanism of the development of multiple-drug-resistant clones of *Shigella*. Japanese journal of microbiology *4*, 219-227.

Akiva,E., Itzhaki,Z., and Margalit,H. (2008). Built-in loops allow versatility in domain-domain interactions: lessons from self-interacting domains. Proc. Natl. Acad. Sci. U. S. A *105*, 13292-13297.

Allison M.J, Maynerry W.R, McSweeney C.S, and Stahl D.A (1992). *Synergistes jonesii*, gen.nov., sp.nov. : a rumen bacterium that degrades toxic pyridinediols. Syst Appl Microbiol *15*, 522-529.

Amann,R.I., Lin,C., Key,R., Montgomery,L., and Stahl,D.A. (1992). Diversity Among *Fibrobacter* Isolates: Towards a Phylogenetic Classification. Syst Appl Microbiol *15*, 23-31.

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.

Aravind,L., Tatusov,R.L., Wolf,Y.I., Walker,D.R., and Koonin,E.V. (1998). Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. Trends Genet. *14*, 442-444.

Azam,F. (1998). Microbial Control of Oceanic Carbon Flux: The Plot Thickens. Science *280*, 694-696.

Baena,S., Fardeau,M.L., Labat,M., Ollivier,B., Thomas,P., Garcia,J.L., and Patel,B.K. (1998). *Aminobacterium colombiense* gen. nov. sp. nov., an amino acid-degrading anaerobe isolated from anaerobic sludge. Anaerobe. *4*, 241-250.

Baena,S., Fardeau,M.L., Woo,T.H., Ollivier,B., Labat,M., and Patel,B.K. (1999). Phylogenetic relationships of three amino-acid-utilizing anaerobes, *Selenomonas acidaminovorans*, '*Selenomonas acidaminophila*' and *Eubacterium acidaminophilum*, as inferred from partial 16S rDNA nucleotide sequences and proposal of *Thermanaerovibrio acidaminovorans* gen. nov., comb. nov. and *Anaeromusa acidaminophila* gen. nov., comb. nov. Int. J. Syst. Bacteriol. *49 Pt 3*, 969-974.

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

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.

Bhandari, V. and Gupta, Radhey S. The phylum Thermotogae. Prokaryotes . 2013.
Ref Type: In Press

Boussau,B. and Daubin,V. (2010). Genomes as documents of evolutionary history. Trends Ecol Evol *25*, 224-232.

Brenner,D.J., Fanning,G.R., Rake,A.V., and Johnson,K.E. (1969). Batch procedure for thermal elution of DNA from hydroxyapatite. Anal. Biochem. *28*, 447-459.

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.

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.

Coenye,T., Gevers,D., De Peer,Y.V., Vandamme,P., and Swings,J. (2005). Towards a prokaryotic genomic taxonomy. FEMS microbiology reviews *29*, 147-167.

Cohan,F.M. (1994). Genetic exchange and evolutionary divergence in prokaryotes. Trends Ecol. Evol. *9*, 175-180.

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.

Curtis,T.P., Sloan,W.T., and Scannell,J.W. (2002). Estimating prokaryotic diversity and its limits. Proceedings of the National Academy of Sciences *99*, 10494-10499.

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.

Davies,J. (1995). Vicious circles: looking back on resistance plasmids. Genetics *139*, 1465-1468.

Davison,J. (1999). Genetic exchange between bacteria in the environment. Plasmid *42*, 73-91.

de Lillo,A., Ashley,F.P., Palmer,R.M., Munson,M.A., Kyriacou,L., Weightman,A.J., and Wade,W.G. (2006). Novel subgingival bacterial phylotypes detected using multiple universal polymerase chain reaction primer sets. Oral Microbiol. Immunol. *21*, 61-68.

Devol,A.H. (2003). Nitrogen cycle: Solution to a marine mystery. Nature *422*, 575-576.

Diaz,C., Baena,S., Fardeau,M.L., and Patel,B.K. (2007). *Aminiphilus circumscriptus* gen. nov., sp. nov., an anaerobic amino-acid-degrading bacterium from an upflow anaerobic sludge reactor. Int. J. Syst. Evol. Microbiol. *57*, 1914-1918.

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.

Eisen,J.A. (2000). Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. Curr. Opin. Genet. Dev. *10*, 606-611.

Federhen,S. (2012). The NCBI taxonomy database. Nucleic Acids Research *40*, D136-D143.

Fieseler,L., Horn,M., Wagner,M., and Hentschel,U. (2004). Discovery of the novel candidate phylum "Poribacteria" in marine sponges. Appl. Environ. Microbiol *70*, 3724-3732.

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 . (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science *269*, 496-512.

Forterre,P. and Gribaldo,S. (2007). The origin of modern terrestrial life. HFSP. J. *1*, 156-168.

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*. Journal of bacteriology *61*, 675.

Fuerst,J.A. and Sagulenko,E. (2011). Beyond the bacterium: planctomycetes challenge our concepts of microbial structure and function. Nat. Rev. Microbiol *9*, 403-413.

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

Gao, Beile. Comparative and functional genomics of Actinobacteria and Archaea. 2010. McMaster University.
Ref Type: Thesis/Dissertation

Garrity,G.M., Bell,J.A., and Lilburn,T.G. (2004). Taxonomic outline of the Prokaryotes. *Bergey's Manual of Systematic Bacteriology*. 2nd edition, Release 5.0. (New York: Springer-Verlag).

Gevers,D., Cohan,F.M., Lawrence,J.G., Spratt,B.G., Coenye,T., Feil,E.J., Stackebrandt,E., Van de Peer,Y., Vandamme,P., and Thompson,F.L. (2005). Re-evaluating prokaryotic species. Nature Reviews Microbiology *3*, 733-739.

Godon,J.J., Moriniere,J., Moletta,M., Gaillac,M., Bru,V., and Delgenes,J.P. (2005). Rarity associated with specific ecological niches in the bacterial world: the 'Synergistes' example. Environ. Microbiol. *7*, 213-224.

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.

Goodfellow,M., Manfio,G.P., and Chun,J. (1997). Towards a practical species concept for cultivable bacteria. SYSTEMATICS ASSOCIATION SPECIAL VOLUME *54*, 25-60.

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. International Journal of Systematic and Evolutionary Microbiology *57*, 81-91.

Gray,M.W. (2012). Mitochondrial evolution. Cold Spring Harbor Perspectives in Biology *4*.

Green,J. and Bohannan,B.J. (2006). Spatial scaling of microbial biodiversity. Trends in ecology & evolution *21*, 501-507.

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., 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.

Grimont,P.A., Popoff,M.Y., Grimont,F., Coynault,C., and Lemelin,M. (1980). Reproducibility and correlation study of three deoxyribonucleic acid hybridization procedures. Current Microbiology *4*, 325-330.

Guangsheng,C., Plugge,C.M., Roelofsen,W., Houwen,F.P., and Stams,A.J.M. (1992). *Selenomonas acidaminovorans* sp. nov., a versatile thermophilic proton-reducing anaerobe able to grow by decarboxylation of succinate to propionate. Arch Microblol *157*, 169-175.

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 natural evolutionary relationships among prokaryotes. Crit Rev. Microbiol. *26*, 111-131.

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. 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.

Hall,J.B. (1971). Evolution of the prokaryotes. J Theor. Biol *30*, 429-454.

Hennig,W. (1966). Phylogenetic systematics. Urbana, University of Illinois Press.

Hormozdiari,F., Salari,R., Hsing,M., Schonhuth,A., Chan,S.K., Sahinalp,S.C., and Cherkasov,A. (2009). The effect of insertions and deletions on wirings in protein-protein interaction networks: a large-scale study. J Comput. Biol. *16*, 159-167.

Horvath,R.S. (1974). Evolution of anaerobic-energy-yielding metabolic pathways of the procaryotes. J Theor. Biol *48*, 361-371.

Horz,H.P., Citron,D.M., Warren,Y.A., Goldstein,E.J., and Conrads,G. (2006). Synergistes group organisms of human origin. J. Clin. Microbiol. *44*, 2914-2920.

Huber,R. and Hannig,M. (2006). Thermotogales. In The Prokaryotes, M.Dworkin, S.Falkow, E.Rosenberg, K.H.Schleifer, and E.Stackebrandt, eds. (New York: Springer), pp. 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.

Hugenholtz,P., Hooper,S.D., and Kyrpides,N.C. (2009). Focus: Synergistetes. Environ. Microbiol. *11*, 1327-1329.

Itzhaki,Z., Akiva,E., Altuvia,Y., and Margalit,H. (2006). Evolutionary conservation of domain-domain interactions. Genome Biol. *7*, R125.

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. Journal of clinical microbiology *45*, 2761-2764.

Jetten,M.S. (2008). The microbial nitrogen cycle. Environmental Microbiology *10*, 2903-2909.

Jumas-Bilak,E., Carlier,J.P., Jean-Pierre,H., Citron,D., Bernard,K., Damay,A., Gay,B., Teyssier,C., Campos,J., and Marchandin,H. (2007). *Jonquetella anthropi* gen. nov., sp. nov., the first member of the candidate phylum 'Synergistetes' isolated from man. Int. J. Syst. Evol. Microbiol. *57*, 2743-2748.

Jumas-Bilak,E., Roudiere,L., and Marchandin,H. (2009). Description of 'Synergistetes' phyl. nov. and emended description of the phylum 'Deferribacteres' and of the family Syntrophomonadaceae, phylum 'Firmicutes'. Int. J. Syst. Evol. Microbiol. *59*, 1028-1035.

Kallnik,V., Schulz,C., Schweiger,P., and Deppenmeier,U. (2011). Properties of recombinant *Strep*-tagged and untagged hyperthermophilic D-arabitol dehydrogenase from *Thermotoga maritima*. Appl. Microbiol. Biotechnol. *90*, 1285-1293.

Kartal,B., Kuenen,J.G., and van Loosdrecht,M.C. (2010). Engineering. Sewage treatment with anammox. Science *328*, 702-703.

Konstantinidis,K.T. and Tiedje,J.M. (2005). Towards a genome-based taxonomy for prokaryotes. J Bacteriol. *187*, 6258-6264.

Kumar,P.S., Griffen,A.L., Moeschberger,M.L., and Leys,E.J. (2005). Identification of candidate periodontal pathogens and beneficial species by quantitative 16S clonal analysis. J. Clin. Microbiol. *43*, 3944-3955.

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

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.

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

Lawrence,J.G. and Hendrickson,H. (2003). Lateral gene transfer: when will adolescence end? Mol Microbiol *50*, 739-749.

Lee,K.C., Webb,R.I., Janssen,P.H., Sangwan,P., Romeo,T., Staley,J.T., and Fuerst,J.A. (2009). Phylum Verrucomicrobia representatives share a compartmentalized cell plan with members of bacterial phylum Planctomycetes. BMC. Microbiol *9*, 5.

Ludwig,W., Whitman,W.B., and Schleifer,K.H. (2009). Revised road map to the phylum Firmicutes. In Bergey's Manual of Systematic Bacteriology, (New York: Springer-verlag).

Ludwig,W. (2010). Molecular Phylogeny of Microorganisms: Is rRNA Still a Useful. Molecular Phylogeny of Microorganisms 65.

Magot,M., Ravot,G., Campaignolle,X., Ollivier,B., Patel,B.K., Fardeau,M.L., Thomas,P., Crolet,J.L., and Garcia,J.L. (1997). *Dethiosulfovibrio peptidovorans* gen. nov., sp. nov., a new anaerobic, slightly halophilic, thiosulfate-reducing bacterium from corroding offshore oil wells. Int. J. Syst. Bacteriol. *47*, 818-824.

Markowitz,V.M., Chen,I.M., Palaniappan,K., Chu,K., Szeto,E., Grechkin,Y., Ratner,A., Jacob,B., Huang,J., and Williams,P. (2012). IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Research *40*, D115-D122.

Martinez-Murcia,A.J., Benlloch,S., and Collins,M.D. (1992). Phylogenetic interrelationships of members of the genera *Aeromonas* and *Plesiomonas* as determined by 16S ribosomal DNA sequencing: lack of congruence with results of DNA-DNA hybridizations. International Journal of Systematic Bacteriology *42*, 412-421.

McCarren,J., Becker,J.W., Repeta,D.J., Shi,Y., Young,C.R., Malmstrom,R.R., Chisholm,S.W., and DeLong,E.F. (2010). Microbial community transcriptomes reveal microbes and metabolic pathways associated with dissolved organic matter turnover in the sea. Proceedings of the National Academy of Sciences *107*, 16420-16427.

McInerney,J.O., Martin,W.F., Koonin,E.V., Allen,J.F., Galperin,M.Y., Lane,N., Archibald,J.M., and Embley,T.M. (2011). Planctomycetes and eukaryotes: a case of analogy not homology. Bioessays *33*, 810-817.

Medini,D., Serruto,D., Parkhill,J., Relman,D.A., Donati,C., Moxon,R., Falkow,S., and Rappuoli,R. (2008). Microbiology in the post-genomic era. Nature Reviews Microbiology *6*, 419-430.

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.

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.

Nisbet,E.G. and Sleep,N.H. (2001). The habitat and nature of early life. Nature *409*, 1083-1091.

Ochiai,K., Yamanaka,T., Kimura,K., and Sawada,O. (1959). Studies on inheritance of drug resistance between *Shigella* strains and *Escherichia coli* strains. Nippon Iji Shimpo *1861*, 34-46.

Oren,A. (2010a). Concepts about phylogeny of microorganisms - an historical overview. Molecular Phylogeny of Microorganisms 1.

Oren,A. (2010b). The Phyla of Prokaryotes, Cultured and Uncultured. Molecular Phylogeny of Microorganisms 85.

Park,C.S., Yeom,S.J., Lim,Y.R., Kim,Y.S., and Oh,D.K. (2010). Characterization of a recombinant thermostable L: -rhamnose isomerase from *Thermotoga maritima* ATCC 43589 and its application in the production of L-lyxose and L-mannose. Biotechnol. Lett. *32*, 1947-1953.

Peeling,R.W. and Brunham,R.C. (1996). Chlamydiae as pathogens: new species and new issues. Emerg. Infect. Dis. *2*, 307-319.

Pikuta,E.V., Hoover,R.B., and Tang,J. (2007). Microbial extremophiles at the limits of life. Crit Rev. Microbiol *33*, 183-209.

Reysenbach,A.-L. (2001). Phylum BII. Thermotogae phy. nov. In Bergey's manual of systematic bacteriology, G.M.Garrity, D.R.Boone, and R.W.Castenholz, eds. (Berlin: Springer), pp. 369-387.

Richter,M. and Rossello-Mora,R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. Proceedings of the National Academy of Sciences *106*, 19126-19131.

Rivera,M.C., Jain,R., Moore,J.E., and Lake,J.A. (1998). Genomic evidence for two functionally distinct gene classes. Proc. Natl. Acad. Sci. U. S. A *95*, 6239-6244.

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

Rossello-Mora,R. and Amann,R. (2001). The species concept for prokaryotes. FEMS Microbiol. Rev. *25*, 39-67.

Sachse,K., Vretou,E., Livingstone,M., Borel,N., Pospischil,A., and Longbottom,D. (2009). Recent developments in the laboratory diagnosis of chlamydial infections. Vet. Microbiol *135*, 2-21.

Sangwan,P., Kovac,S., Davis,K.E., Sait,M., and Janssen,P.H. (2005). Detection and cultivation of soil verrucomicrobia. Appl. Environ. Microbiol *71*, 8402-8410.

Sapp,J. (2005). The prokaryote-eukaryote dichotomy: meanings and mythology. Microbiology and molecular biology reviews *69*, 292-305.

Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., and Federhen,S. (2011). Database resources of the national center for biotechnology information. Nucleic Acids Research *39*, D38-D51.

Schleifer,K.H. (2009). Classification of Bacteria and Archaea: past, present and future. Syst Appl. Microbiol. *32*, 533-542.

Schloss,P.D. and Handelsman,J. (2004). Status of the microbial census. Microbiol Mol Biol Rev. *68*, 686-691.

Schopf,J.W. (2006). Fossil evidence of Archaean life. Philos. Trans. R. Soc. Lond B Biol Sci. *361*, 869-885.

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.

Snel,B., Bork,P., and Huynen,M.A. (1999). Genome phylogeny based on gene content. Nature genetics *21*, 108-110.

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. (2006). Defining Taxonomic Ranks. The Prokaryotes: Vol. 1: Symbiotic Associations, Biotechnology, Applied Microbiology *1*, 29-57.

Stanier, Roger Y. Some aspects of the biology of cells and their possible evolutionary significance. 20, 1-38. 1970.
Ref Type: Conference Proceeding

Stanier,R.Y. and Niel,C.v. (1962). The concept of a bacterium. Archives of Microbiology *42*, 17-35.

Staudt,L.M. (2003). Molecular diagnosis of the hematologic cancers. The New England journal of medicine *348*, 1777.

Strous,M., Fuerst,J.A., Kramer,E.H., Logemann,S., Muyzer,G., van de Pas-Schoonen KT, Webb,R., Kuenen,J.G., and Jetten,M.S. (1999). Missing lithotroph identified as new planctomycete. Nature *400*, 446-449.

Sutcliffe,I.C. (2010). A phylum level perspective on bacterial cell envelope architecture. Trends Microbiol *18*, 464-470.

Swithers,K.S., Gogarten,J.P., and Fournier,G.P. (2009). Trees in the web of life. J Biol *8*, 54.

Thomas,C.M. and Nielsen,K.M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nature Reviews Microbiology *3*, 711-721.

Tindall,B.J., Rossello-Mora,R., Busse,H.J., Ludwig,W., and K+ñmpfer,P. (2010). Notes on the characterization of prokaryote strains for taxonomic purposes. International Journal of Systematic and Evolutionary Microbiology *60*, 249-266.

Uzzell,T. and Spolsky,C. (1974). Mitochondria and plastids as endosymbionts: a revival of special creation? Am. Sci. *62*, 334-343.

Vartoukian,S.R., Palmer,R.M., and Wade,W.G. (2007). The division "Synergistes". Anaerobe. *13*, 99-106.

Wagner,M. and Horn,M. (2006). The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. Curr. Opin. Biotechnol. *17*, 241-249.

Watanabe,T. and Fukasawa,T. (1961). Episome-mediated transfer of drug resistance in Enterobacteriaceae I. Transfer of Resistance Factors by Conjugation. Journal of bacteriology *81*, 669-678.

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 Turper,H.G. (1987). Report of the Ad Hoc committee on reconciliation of approaches to bacterial systematics. Int J Syst Bacteriol. *37*, 463-464.

West,M., Ginsburg,G.S., Huang,A.T., and Nevins,J.R. (2006). Embracing the complexity of genomic data for personalized medicine. Genome research *16*, 559-566.

Williams,D., Andam,C.P., and Gogarten,J.P. (2010). Horizontal gene transfer and the formation of groups of microorganisms. Molecular Phylogeny of Microorganisms.

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. Biology direct *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.

Xu,J. (2010). Microbial population genetics. Horizon Scientific Press).

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.

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.

Zijnge,V., van Leeuwen,M.B., Degener,J.E., Abbas,F., Thurnheer,T., Gmur,R., and Harmsen,H.J. (2010). Oral biofilm architecture on natural teeth. PLoS. One. *5*, e9321.

Zuckerkandl,E. and Pauling,L. (1965). Molecules as documents of evolutionary history. J. Theor. Biol. *8*, 357-366.